Knowledge-based Semantic Measures: From Theory to Applications

Sébastien Harispe

Collège Doctoral du Languedoc-Roussillon

THÈSE

Pour obtenir le grade de

Docteur

Délivrée par

UNIVERSITE MONTPELLIER 2 (France)

Préparée au sein de l'école doctorale

I2S - Information, Structures, Systèmes

Et de l'unité de recherche

LGI2P - Laboratoire de Génie Informatique et d'Ingénierie de Production de l'**école des mines d'Alès**

Spécialité : **Informatique**

Présentée par

Sébastien Harispe

Knowledge-based Semantic Measures: From Theory to Applications

Mesures sémantiques à base de connaissance : de la théorie aux applicatifs

Soutenue le 25/04/2014 devant le jury composé de

Mme Pascale Kuntz, Professeur *École Polytechnique de l'Université de Nantes*	Rapporteur
M. Jérôme Euzenat, Directeur de Recherche *INRIA Grenoble Rhône-Alpes*	Rapporteur
M. Amedeo Napoli, Directeur de Recherche *CNRS, LORIA - Nancy-Lorraine*	Examinateur Président du jury
Mme Isabelle Mougenot, Maître de conférences *Université Montpellier II*	Examinateur
M. David Sánchez, Enseignant-Chercheur *Universitat Rovira i Virgili, Tarragona, Spain*	Examinateur
M. Jacky Montmain, Professeur (*directeur de thèse*) *École des mines d'Alès*	Examinateur
Mme Sylvie Ranwez, HDR (*encadrement de thèse*) *École des mines d'Alès*	Examinateur
M. Stefan Janaqi, Maître-assistant (*encadrement de thèse*) *École des mines d'Alès*	Examinateur

À la Science,

Acknowledgements

Cette thèse est le fruit de trois années de travail intense et de longues heures de solitude nécessaires à l'exercice de recherche. Cependant une thèse ne peut être réduite à un exercice solitaire. C'est aussi l'occasion de côtoyer des individualités d'une extrême richesse, qui changeront à n'en pas douter votre conception du travail, de l'investissement personnel, et votre regard critique sur votre propre production. Ces rencontrent nourissent votre esprit critique et vous arment pour affronter l'inconnu, parfois déstabilisant, qui rythme le quotidien du chercheur producteur de connaissances nouvelles. Je ne dois cette thèse à personne – bien que je ne l'aurais probablement jamais menée sans les personnes et institutions que je remercie ci-dessous.

Je remercie en premier lieu l'École des mines d'Alès qui a financé mes travaux et qui m'a donné la possibilité d'évoluer au sein du LGI2P pendant ces trois années épanouissantes. J'ai une pensée particulière envers Yannick Vimont, directeur de la recherche de l'école et directeur du laboratoire LGI2P. Toujours enthousiaste et à l'écoute des chercheurs qu'il encadre, il nous a plusieurs fois donné l'opportunité de mobiliser une force de frappe permettant de galvaniser nos projets de recherche ; je lui en suis reconnaissant. J'associe à ces remerciements l'Université de Montpellier 2 (UM2) et l'école doctorale Information Structures Systèmes (I2S) qui m'ont accompagné dans la réalisation de cette thèse.

Je remercie aussi tout particulièrement les deux rapporteurs de ce mémoire : Pascale Kuntz et Jérôme Euzenat. Ils ont su effectuer une analyse critique fine, pertinente et détaillée de mes travaux ; ils ont fourni un travail considérable et ainsi contribuer par leurs nombreuses remarques et suggestions à améliorer la qualité du mémoire que vous allez parcourir, je leur en suis très reconnaissant. Mes remerciements vont également aux chercheurs extérieurs qui ont accepté d'examiner mes travaux : Amedeo Napoli, Isabelle Mougenot et David Sánchez. Les nombreuses discussions proposées lors de la soutenance ont ouvert des perspectives de recherche qui, je n'en doute pas, nourriront mes recherches futures.

Mes remerciements les plus appuyés reviennent tout naturellement aux membres de l'équipe encadrante de ma thèse. Toujours disponibles et à l'écoute, je suis conscient de la chance que j'ai eu d'avoir pu évoluer à leurs côtés et ainsi bénéficier de leurs précieux conseils : mon directeur Jacky Montmain, d'une grande culture, toujours lucide et pertinent ; Sylvie Ranwez, encadrante de proximité joyeuse, investie, toujours de bon conseil, à qui je dois beaucoup ; Stefan Janaqi, encadrant de proximité, toujours de bonne humeur, qui a joué un rôle central dans de nombreuses contributions amenées dans cette thèse. Je ne m'attarde pas, ils savent tout le bien que je pense d'eux – c'est l'essentiel.

Au travers de collaborations scientifiques, cette thèse m'a aussi permis de travailler aux côtés de nombreux chercheurs. Ils ont sans aucun doute donné une autre dimension à mes travaux. Ainsi, dans le cadre d'une collaboration avec l'université de Tarragone (Espagne), j'ai notamment eu la chance de partager de nombreuses réflexions avec Montserrat Batet et David Sánchez, deux chercheurs de grande qualité et de reconnaissance internationale dans le domaine d'étude de cette thèse. J'ai beaucoup appris d'eux, ils ont notamment apporté un regard critique sur mes travaux. Celui-ci a largement contribué à la maturation de nombreuses de mes contributions. J'ai aussi eu l'opportunité de collaborer avec Vincent Ranwez, qui m'a initié au monde de la recherche lors de mon stage de Master ; un chercheur infatigable, d'une grande créativité, perspicace, toujours précis et motivé. Un grand merci aussi à Clément Jonquet qui m'a récemment donné la possibilité de collaborer avec lui. J'ai aussi une pensée pour Isabelle Mougenot, Céline Scornavacca et Manuel Ruiz qui, avec Clément, ont accepté de participer à mes comités de suivi de thèse, et ont ainsi apporté un regard extérieur constructif pour un meilleur positionnement de mes travaux de recherche.

Je remercie tous les collègues du LGI2P, aussi bien les permanents (chercheurs, assistantes administratives, équipe enseignante) que les thésards, élèves et autres contractuels. J'ai partagé avec eux de nombreuses discussions qui m'ont enrichi tant au niveau personnel que professionnel. Ils ont eux aussi largement contribué au plaisir que j'ai eu à travailler sur cette thèse.

Je souhaite aussi remercier du fond du cœur ma famille et en particulier mes parents envers qui j'ai une reconnaissance qui dépasse les quelques mots que je pourrais leur consacrer ici. Je remercie aussi mes amis pour leur soutien constant. Pour finir, j'ai une pensée toute particulière pour la personne exceptionnelle qui m'a accompagné, encouragé, supporté, aidé et donné le sourire pendant ces trois années de thèse, merci Jane, cette thèse est aussi la tienne.

Sébastien Harispe

Contents

List of Figures

List of Tables

Synopsis de la thèse

Mesures sémantiques à base de connaissance : de la théorie aux applicatifs

par Sébastien Harispe

Directeur : Jacky Montmain – Encadrement: Sylvie Ranwez et Stefan Janaqi

Institut : École des mines d'Alès

(The english manuscript begins page 21)

Cet avant-propos introduit les travaux de thèse détaillés dans la suite du manuscrit intitulé "*Knowledge-based semantic measures: from theory to applications*". Il présente dans un premier temps le contexte et le positionnement scientifique de nos travaux ainsi que les objectifs fixés. Dans un second temps, nous discutons les différentes contributions proposées, sans pour autant traiter des aspects techniques qui leurs sont associés. Le lecteur désireux de s'attarder sur ces derniers pourra se référer à la partie du manuscrit correspondante rédigée en anglais scientifique. Ce synopsis se termine par une conclusion générale qui souligne, entre autres, les verrous scientifiques associés à la thématique traitée et les pistes de recherche que nos contributions pourront nourrir.

I Contexte général et objectifs de la thèse

I.I Simuler une *intelligence* : une quête déjà ancienne

L'Intelligence Artificielle (IA) est une branche de l'informatique qui s'attache à développer des approches permettant d'amener la résolution de problèmes complexes par l'utilisation d'ordinateurs. Un de ses objectifs est tout naturellement **de substituer l'homme**[1] **par la machine dans la résolution de tâches complexes nécessitant de fortes capacités cognitives**, i.e. une forme d'intelligence – ici entendue comme la capacité à acquérir et tirer parti de connaissances dans la résolution de problèmes [Oxford Dict., 2012]. Ainsi, depuis 1956, date depuis laquelle l'intelligence artificielle est considérée comme un champ de recherche à part entière, cette discipline fédère un grand nombre de communautés scientifiques dans le but de permettre à l'outil informatique de raisonner, de manipuler la connaissance, d'apprendre, de planifier, de communiquer, ou encore de percevoir le monde qui nous entoure [Russell and Norvig, 2009].

[1] ou tout du moins l'accompagner dans son processus cognitif ; cette démarche appelée "automatisation cognitive" sera discutée plus loin.

Parmi les différentes stratégies explorées en vue de faire émerger une forme d'intelligence artificielle, **nos travaux s'intéressent plus particulièrement à celles basées sur l'utilisation de représentations de connaissance** (e.g. thésaurus, *ontologies*). Ces stratégies reposent sur l'hypothèse, très souvent admise, que la connaissance est l'un des ingrédients requis au développement d'une forme d'intelligence. Elles se concentrent notamment sur la définition de méthodes permettant d'automatiser la résolution de problèmes complexes, et s'intéressent plus spécifiquement aux problèmes qui ont la particularité de nécessiter le recours à d'importantes sources de connaissance pour être résolus, par exemple le diagnostic médical.

Dans ce contexte, de nombreuses communautés et générations de chercheurs se sont intéressées à la résolution d'**un des problèmes fondamentaux de l'IA : exprimer la connaissance de façon à la rendre intelligible et appréciable par l'outil informatique** [Baader et al., 2010; Davis et al., 1993]. Ce défi, toujours d'actualité, a amené la définition de nombreux langages de représentation de connaissance. Basés entre autres sur des formalismes de graphe ou sur des logiques descriptives, ils permettent d'exprimer de façon formelle une connaissance qui pourra être manipulée par ordinateur. Ils offrent, en quelque sorte, la possibilité d'établir une connexion entre la connaissance experte et l'outil informatique, et donnent ainsi la possibilité d'initier un transfert partiel des compétences de l'expert de domaine vers les systèmes informatiques : une condition nécessaire à la mise en place des systèmes informatiques dits *intelligents*.

Ainsi, de façon imagée, **une représentation de connaissance peut être considérée comme le terreau nécessaire à l'émergence d'une forme d'intelligence simulée au travers d'instructions machine**. Des logiciels particuliers, appelés raisonneurs, peuvent notamment les utiliser pour inférer de la connaissance exacte. Ces inférences sont assurées par des procédures de déduction en accord avec la sémantique du langage de représentation de connaissance. Autrement dit, la sémantique du langage définit la façon dont l'outil informatique doit *comprendre* la connaissance exprimée. Les règles d'inférence associées à cette sémantique définissent les interprétations de la connaissance qui permettront l'élaboration de formes de *raisonnement déductif*, i.e. exact. Cette capacité d'inférence est aujourd'hui largement utilisée dans de nombreux domaines des sciences et de l'industrie. Elle permet notamment l'émergence d'une intelligence artificielle sous la forme de programmes informatiques capables d'effectuer des raisonnements déductifs complexes.

Néanmoins, les capacités d'inférence offertes par les représentations de connaissance ne se résument pas aux interprétations strictes et rigides permises par le langage utilisé. En effet, les représentations de connaissance peuvent aussi servir à simuler des formes d'intelligence *débridées*, i.e. non contraintes aux seules règles d'inférence permises par le

language utilisé pour exprimer la connaissance. Celles-ci, basées sur un certain nombre d'hypothèses, permettent l'élaboration de formes de *raisonnement approché*. Ainsi, en permettant de tirer parti des représentations de connaissances sans pour autant être contraintes par la sémantique formelle qui les sous-tend, ces techniques de raisonnement approché offrent d'intéressantes perspectives, notamment pour l'élaboration de stratégies de découverte de connaissance. Pour cela, il est nécessaire de définir des modèles qui permettent de comparer les éléments caractérisés au travers de représentations de connaissance, e.g. pour les regrouper, les analyser et les comprendre plus en détail. Ces modèles reposent sur l'analyse sémantique des éléments comparés et essentiellement sur la notion de *mesure sémantique*. Tout comme **l'homme est capable d'apprécier la similarité d'objets concrets et/ou abstraits** – par exemple, pour la plupart d'entre nous, les concepts `Paix` et `Colombe` seront proches a contrario des concepts `Paix` et `Pigeon` –, **les mesures sémantiques permettent de doter l'outil informatique de cette capacité essentielle à l'élaboration de nombreuses fonctions cognitives**. Pour cela, ces mesures se basent sur la définition de modèles permettant l'analyse de la sémantique exprimée dans des représentations de connaissance et dans des corpus de textes, i.e. du sens porté par ces ressources. **Le rôle capital que jouent ces mesures sémantiques nous a amené à les étudier de façon approfondie et c'est aux résultats de ces recherches que cette thèse est consacrée**. Mais avant de rentrer dans le détail de ce vaste domaine de recherche, il convient de préciser la genèse de ces mesures et les différents cadres applicatifs dans lesquels elles interviennent.

I.II Une accélération portée par les évolutions technologiques

Dans les dernières décennies, nous avons observé **une large adoption des systèmes informatiques à base de connaissance**, i.e. reposant sur l'utilisation de représentations de connaissance. A titre d'exemple, Bioportal, une plateforme dédiée aux représentations de connaissance ayant trait à la biologie et au domaine biomédical, en propose aujourd'hui pas moins d'une centaine [Whetzel et al., 2011]. Elles sont utilisées dans de nombreux applicatifs aussi divers que l'assistance au diagnostic médical, la classification de maladies, l'analyse de gènes, la confection de médicaments [Guzzi et al., 2012; Köhler et al., 2009; Pesquita et al., 2009a].

Les grands acteurs du Web ont, eux aussi, récemment franchi le pas. Ainsi depuis 2011, Google, pour ne citer que lui, tire parti d'un graphe de connaissance composé de milliards de faits lui **permettant de structurer et de désambiguïser un grand nombre d'entités** (e.g., personnes, villes, films). C'est sur ce graphe de connaissance, ou **graphe sémantique**, que se base, par exemple, son système de recherche d'information pour désambiguïser les intentions de ses utilisateurs et améliorer ses résultats [Singhal, 2012].

En effet, grâce à ce modèle de connaissance, une simple recherche portant sur "*Alfred Hitchcock*" permet, au sein même de la page de résultats, de consulter de nombreuses informations associées au réalisateur (date de naissance, films associés). Ainsi, le système informatique *comprend*, en quelque sorte, que le centre d'intérêt de l'utilisateur porte sur un réalisateur particulier et non pas sur une chaîne de caractères jusque-là dénuée de sens ; inutile d'insister sur les larges perspectives offertes par cette désambiguïsation, e.g. recommandations, analyse marketing, informatique décisionnelle.

De nombreuses perspectives ont été amenées par la définition de langages de représentation des connaissances. Parmi elles, l'une des plus ambitieuses et captivantes fait référence à la volonté de tirer parti de l'infrastructure Internet pour **créer un Web de Connaissance**, aussi appelé Web Sémantique. L'objectif est de former un réseau de connaissance mondialement distribué, à la fois exploitable et intelligible par nous autres humains, mais aussi par des agents logiciels [Berners-Lee et al., 2001; Gandon et al., 2012]. Il permet ainsi de pallier les limitations du Web dit de documents, dont le contenu généralement non-structuré et ambigu au regard d'un agent logiciel, n'est que difficilement exploitable par des méthodes automatisées. Dans ce contexte, de nombreuses initiatives font promotion des paradigmes du Web Sémantique et des Données Liées [Heath and Bizer, 2011; Hitzler et al., 2011]. Ces derniers proposent d'amener le développement d'une extension du Web qui permettra une meilleure caractérisation des informations qui y sont exprimées, et donc le **développement d'une synergie entre agents logiciels et humains**. De même que le Web, à sa création, nous a offert la possibilité d'exposer et de relier des documents (multimédias), le Web de Connaissance permet désormais d'exprimer et d'échanger de la connaissance, e.g., "*Alfred Hitchcock est né le 13 août 1899 à Leytonstone*", "*Leytonstone se situe en Angleterre*". Ainsi, en désambiguïsant le contenu exprimé, et en interconnectant différentes bribes d'information existantes, chacun peut aujourd'hui contribuer à l'émergence d'un réseau mondial de connaissance exploitable par tous. Tous les éléments d'une nouvelle révolution numérique semblent réunis.

De nombreux standards permettent d'exprimer des données structurées et désambiguïsées sur le Web. Ainsi, à partir de ces standards, de nombreuses initiatives collaboratives ont permis de créer et d'interconnecter un grand nombre de silos de données, parfois spécialisées, e.g. DBpedia[1] [Auer et al., 2007], Freebase [Bollacker et al., 2008], UniProtKB [UniProt Consortium, 2013]. Ces **données, accessibles publiquement et gratuitement sur le Web**, peuvent être interrogées à l'instar des bases de données classiques. De plus, grâce à la sémantique formelle des langages utilisés, ces données peuvent servir à inférer une connaissance nouvelle, implicite, déductible.

[1] Pendant sémantique de Wikipédia.

Les travaux détaillés dans ce manuscrit ont été menés dans le souci de rester compatibles avec les standards du Web Sémantique, vecteurs d'une richesse encore aujourd'hui largement sous-exploitée.

I.III Les mesures sémantiques au cœur de la démarche

La plupart du temps, les systèmes à base de connaissance sont définis pour inférer de la connaissance exacte sur un domaine, i.e. déduire des faits, à partir d'un ensemble de faits établis. Cependant, **l'utilisation d'une approche déductive n'est pas adaptée à tout type d'application. En particulier lorsque l'objectif est par définition inexact**. C'est souvent le cas en recherche d'information. Il est fréquent de devoir répondre à une question qui ne peut être traitée par de simples opérateurs booléens, e.g. quels sont les groupes similaires aux "*Rolling Stones*" ? **L'ambigüité intrinsèque à la notion de similarité empêche l'utilisation (seule) de techniques de raisonnement déductif**. Cependant, la connaissance définie dans une représentation de connaissance peut fournir des éléments de réponse utiles au système de recherche d'information. Par exemple, l'étude des interconnexions entre les différentes entités définies dans une représentation de connaissance relative au domaine de la musique (e.g. groupes, genres musicaux) permettra sûrement d'établir, à juste titre, que les "*Rolling Stones*" semblent davantage similaires au groupe "*The Who*" qu'à celui des "*Spice Girls*". Dans le domaine biomédical, c'est sur ce même principe que des modèles permettant d'évaluer la pertinence à réutiliser des molécules thérapeutiques ont été proposés [Eronen and Toivonen, 2012]. Cette pratique, appelée extension de médicament, considère qu'une molécule avérée effective dans le traitement d'une condition particulière peut potentiellement être réutilisée pour traiter une condition similaire ; cette similarité est évaluée au regard des informations définies dans des bases de connaissances biomédicales – on retrouve ici une fois de plus la notion centrale de similarité.

De manière plus large, la plupart des approches de raisonnement approximatif ou d'apprentissage automatique reposent sur une mesure permettant de **comparer les entités manipulées**. Cette mesure permet notamment de regrouper des objets au regard de leurs propriétés et de définir, parfois à partir de jeux d'apprentissage, des fonctions discriminantes à même de les classifier. Dans un contexte biomédical, ces fonctions permettront, par exemple, de distinguer des individus malades de ceux qui sont sains. Ainsi, les fonctions de similarité revêtent une importance majeure pour la mise en place de raisonnements approximatifs, pour le développement de techniques d'apprentissage automatique ou encore, pour la mise en place de systèmes de recherche d'information.

Les mécanismes cognitifs de l'homme reposent eux aussi fortement sur la notion de similarité. En effet, la capacité qu'a l'homme à comparer les choses (objets, stimuli) et à identifier des similarités et différences entre celles-ci, a depuis longtemps été caractérisée par les sciences cognitives et la psychologie, comme un élément au coeur de nombreux processus cognitifs [Rissland, 2006]. La similarité joue ainsi un rôle central dans l'apprentissage, dans la prise de décision, dans l'élaboration de certains types de raisonnement, dans la reconnaissance de formes, ou encore dans la définition de plans de résolution [Gentner and Markman, 1997; Ross, 1987; Vosniadou and Ortony, 1989]. En effet, la capacité à reconnaître des situations similaires permet, par exemple, de stimuler notre expérience en activant des traces mentales qui nous permettront de résoudre des problèmes nouveaux, en y appliquant des éléments de résolution appliqués avec succès à des problèmes similaires. Il est donc clairement admis que **cette notion de similarité, ou de façon plus générale, cette capacité à comparer les choses est centrale dans la mise en place de *formes d'intelligence*** ; elle joue donc un rôle essentiel pour les communautés intéressées à l'élaboration d'*intelligences artificielles*.

Ainsi, le développement d'agents intelligents basés sur des représentations de connaissance repose en grande partie sur la définition de fonctions permettant de comparer les éléments qu'elles définissent. Cette comparaison doit être gouvernée par la connaissance définie dans la représentation de connaissance et doit donc tout naturellement reposer sur une mesure à même de tirer parti de la sémantique qui la caractérise. Pour cela, des mesures sémantiques à base de connaissance sont utilisées[1]. De façon plus générale, ces mesures s'inscrivent dans la classe des mesures sémantiques, qui permettent de comparer des entités (unités lexicales, concepts, instances) par l'analyse de *proxies sémantiques* (corpus de textes ou représentations de connaissance). Du fait de leur importance pour de nombreuses communautés, **une vaste littérature est dédiée à ces mesures et de nombreuses approches ont été proposées** pour différents types de traitements. En effet, de la recommandation musicale à l'analyse de données biomédicales (dossiers patients, gènes), en passant par l'étude de données géographiques, **de nombreuses communautés tirent aujourd'hui parti de ces mesures et contribuent à leur étude**.

Les travaux décrits dans ce mémoire ont été effectués au sein de **l'équipe KID[2] du laboratoire LGI2P[3] de l'École des mines d'Alès**. Fédérés autour de l'*automatisation cognitive*, les chercheurs du LGI2P s'intéressent au développement de concepts innovants, méthodologies, et outils pour la conception, la réalisation et l'optimisation de systèmes techniques, de processus collaboratifs ou encore d'organisation sociotechniques. Dans

[1] *Knowledge-based semantic measures* en anglais.
[2] *Knowkedge and Image analysis for Decision making* en anglais.
[3] Laboratoire de Génie Informatique et d'Ingénierie de Production.

ce contexte, l'équipe KID tire parti, entre autres, des représentations de connaissance pour la définition de techniques optimisées de découverte, interrogation et analyse de connaissance ; techniques dans lesquelles les mesures sémantiques jouent très souvent un rôle central [Ranwez, 2013].

Ces travaux sont ancrés dans le domaine de l'*Intelligence Artificielle*, et exploitent plus particulièrement les techniques de représentation de connaissance à l'ère du *Web de Connaissance*. A partir d'une analyse détaillée des *mesures sémantiques*, et plus particulièrement de celles dédiées à la comparaison de concepts ou d'instances définies dans des *représentations de connaissances* structurées sous la forme de *graphes sémantiques*, nous proposons :

1. Un état de l'art étendu sur la notion de mesure sémantique. L'analyse de la littérature nous permet notamment de catégoriser les différentes approches proposées, de caractériser la terminologie d'usage dans le domaine, et de répertorier une large collection de mesures.

2. Un cadre unificateur dédié aux mesures sémantiques à base de connaissance. Celui-ci dote la communauté d'un outil théorique offrant un nouveau regard sur ces mesures. Grâce à lui, nous montrons par exemple que la plupart des mesures publiées de façon indépendante correspondent pour la plupart à des expressions spécifiques de mesures paramétriques génériques. Nous soulignons aussi les perspectives que ce cadre théorique offre pour l'analyse détaillée des mesures.

3. Une librairie logicielle et un ensemble d'outils dédiés au calcul et à l'analyse de ces mesures.

4. Des contributions algorithmiques et théoriques associées à ces mesures.

Ces différentes contributions sont détaillées dans la section qui suit.

II Contributions scientifiques, théoriques et logicielles de la thèse

II.I État de l'art étendu des mesures sémantiques : définitions, analyse détaillée et catégorisation des mesures basées sur une représentation de connaissance

Référence : **Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Base Analysis**. Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain (2013). ArXiv. Computation and Language. `http://arxiv.org/abs/1310.1285v2`

La première partie de ce manuscrit propose une vision détaillée de la notion de mesure sémantique. Nous y présentons une version condensée d'une littérature vaste, interdisciplinaire, et parfois éparpillée relative au domaine. Nous introduisons de nombreuses définitions et nous distinguons un certain nombre de propriétés (mathématiques) permettant de les caractériser, notamment au regard de la sémantique qui leur est associée. Ce travail nous a permis de proposer **une classification générale des différents types de mesures sémantiques évoqués dans la littérature**.

Par la suite, du fait de la diversité du domaine, nous nous sommes concentrés sur **les mesures sémantiques à base de connaissance, et plus particulièrement sur celles reposant sur une structuration de la connaissance sous forme de graphe sémantique**. De nombreux détails techniques relatifs à ce type de mesures sont discutés, et une large collection de mesures proposées dans la littérature est identifiée, classifiée et analysée.

La première contribution majeure de cette thèse est de mutualiser les contributions proposées par des communautés distinctes et de les analyser au travers d'un même prisme. En effet, nous montrons que de nombreuses contributions, pour la plupart relatives aux mesures sémantiques, initialement proposées dans des domaines spécifiques, et parfois exprimées dans des formalismes particuliers, ont souvent une portée plus large que celle initialement escomptée. Ainsi, bien que souvent conçues dans un cadre applicatif bien délimité et dédiées à une problématique très pointue, par exemple l'analyse fonctionnelle de gènes, nous montrons que de nombreuses définitions de mesures sémantiques, pour la plupart *ad hoc*, peuvent souvent profiter à un grand nombre de communautés et ainsi amener la résolution de problèmes divers. Nous soulignons ainsi que **les contributions relatives aux mesures sémantiques s'inscrivent dans un domaine de recherche**

interdisciplinaire d'une large richesse, jusque-là mal identifié, et à l'interface de nombreuses thématiques de recherche.

Cette synthèse ne traite pas de certains sujets importants comme la sélection des mesures. Cependant, nous sommes convaincus qu'elle donne accès, aussi bien au néophyte qu'à l'initié, à une meilleure compréhension des différentes approches proposées et donne ainsi une vision globale, revisitée et organisée du domaine.

Nourris de cette analyse de l'état de l'art, nous avons ensuite distingué un certain nombre de défis que la notion de mesure sémantique offre à nos communautés. Parmi ces défis, nos travaux se sont essentiellement concentrés sur la proposition d'**outils théoriques et pratiques dédiés aux mesures sémantiques**. En effet, un des constats de notre étude préliminaire concerne le cloisonnement de ces mesures en partie dû au caractère *ad hoc* de nombreuses formulations, et à la nature *domaine-spécifique* de la plupart des contributions logicielles associées au domaine.

En réponse à ces limites, notre stratégie de recherche a notamment consisté à abstraire, autant que possible, les mesures sémantiques de leur cadre applicatif et de leur contexte d'utilisation. Cette approche nous a permis d'identifier les éléments constitutifs des mesures sémantiques, parmi lesquels : une représentation des entités manipulées, des estimateurs des parties communes et différentes de ces représentations, une fonction permettant l'aggrégation de ces opérateurs. Ainsi, cette décomposition des mesures nous permet de comprendre plus en détail le mode de fonctionnement et les spécificités de ces mesures. Ce travail constitue la deuxième contribution majeure de la thèse et sera présenté dans la section suivante. Concernant l'aspect applicatif, en réponse à la multiplication de solutions logicielles *domaine-spécifiques*, nous avons développé un outil logiciel générique, performant, représentatif de la diversité de l'état de l'art, et indépendant d'un applicatif particulier. L'objectif visé était de proposer aux communautés utilisatrices et impliquées dans l'étude des mesures sémantiques, une plateforme de développement, d'analyse et de calcul dédiée. Ce sera la troisième contribution, détaillée plus loin.

II.II Un cadre unificateur pour les mesures sémantiques à base de connaissance

Référence :

A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. Sébastien Harispe*, David Sánchez, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. Journal of Biomedical Informatics 2013. `http://dx.doi.org/10.1016/j.jbi.2013.11.006` – publication en collaboration avec David Sánchez de l'Université Rovira i Virgili de Tarragone (Espagne).

En se concentrant sur les mesures sémantiques à base de représentation de connaissance, nous montrons qu'un grand nombre de mesures définies dans la littérature sont des expressions dérivées de **fonctions paramétriques reposant sur un ensemble limité de paramètres abstraits**. Ainsi, dans la continuité de plusieurs travaux portant sur l'étude des similitudes entre mesures [Blanchard, 2008; Blanchard et al., 2008; Pirró and Euzenat, 2010a; Sánchez and Batet, 2011], nous mettons en évidence que la plupart des mesures, jusque-là trop souvent considérées comme indépendantes, sont étroitement liées et reposent essentiellement sur la définition d'opérateurs simples. Cette observation, illustrée tout au long de la thèse par de multiples exemples, propose un nouveau regard sur la large diversité de mesures sémantiques.

A partir de ces travaux, nous avons défini **un cadre unificateur pour les mesures sémantiques à base de connaissance**. En distinguant (i) les composants constitutifs de la plupart des mesures (e.g. points communs et différences), (ii) des expressions particulières de ces composants, et (iii) des formes génériques de mesures permettant l'agrégation de ces composants pour l'expression de mesures concrètes, nous mettons en évidence que le cadre théorique proposé permet à la fois d'**exprimer des mesures et de les analyser en détail**.

Différentes applications pratiques de ce cadre sont illustrées dans le manuscrit. Nous montrons en particulier qu'il permet d'exprimer de nouvelles mesures, d'étudier leurs performances, d'orienter leur sélection au travers d'optimisations paramétriques, et de distinguer les éléments constitutifs des mesures qui semblent jouer un rôle critique dans leur performance. Ces différents applicatifs soulignent la large portée de notre contribution pour l'étude des mesures sémantiques. Par exemple, la caractérisation des éléments centraux des mesures offre des perspectives intéressantes pour la définition, le paramétrage et/ou la sélection des mesures. Nous montrons notamment que le degré de granularité des analyses permises par notre approche laisse envisager l'étude des

mesures à un niveau de détail extrêmement fin. A titre d'exemple, nous montrons comment l'unification des mesures effectuée au niveau théorique, indépendamment de tout contexte applicatif, permet la définition de mesures optimisées pour un contexte applicatif particulier, e.g. pour comparer des gènes annotés par des concepts relatifs au domaine biomédical.

II.III La Semantic Measures Library (SML), une librairie logicielle libre et générique dédiée aux mesures sémantiques

Références :

The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. Oxford Bioinformatics 2013.

From Theoretical Framework to Generic Semantic Measures Library. Sébastien Harispe*, Stefan Janaqi, Sylvie Ranwez, Jacky Montmain. On the Move to Meaningful Internet Systems: OTM 2013 Workshops Lecture Notes in Computer Science Volume 8186, 2013, pp 739-742;

`http://dx.doi.org/10.1007/978-3-642-41033-8_98`

Site internet : `http://www.semantic-measures-library.org`

Tout au long de ce manuscrit, nous soulignons l'importance des évaluations empiriques pour l'analyse des mesures sémantiques, en particulier pour évaluer leur performance. Pourtant la plupart des solutions logicielles existantes ont été développées dans l'objectif de répondre aux attentes d'un domaine applicatif particulier. Ainsi, bien que quelques initiatives aient tenté de proposer des solutions génériques, celles-ci se limitent à l'analyse de représentations de connaissance de tailles réduites et ne sont plus maintenues aujourd'hui – l'exemple le plus abouti reste selon nous *SimPack* [Bernstein et al., 2005]. Ainsi, ces solutions ne permettent pas de répondre aux besoins de nombreux applicatifs reposant sur l'utilisation de représentations de connaissance qui définissent des milliers d'entités – ce qui est de plus en plus fréquent, par exemple dans le domaine biomédical.

C'est dans ce contexte que nous avons initié le développement de la **Semantic Measures Library (SML)** avec comme objectif de rendre accessibles au plus grand nombre des **solutions logicielles robustes dédiées au calcul et à l'analyse des mesures sémantiques**. Dans le cadre de ce projet, nous avons développé une librairie logicielle dédiée aux mesures sémantiques à base de connaissance (développée en Java). Elle met à disposition des chercheurs du domaine un cadre de développement permettant à la fois d'utiliser un grand nombre de mesures, de les analyser, et de facilement développer et

tester de nouvelles approches. Cette librairie implante de nombreux algorithmes relatifs aux mesures sémantiques à base de connaissance. Elle permet aux chercheurs de se concentrer sur leur thématique de recherche, en faisant abstraction par exemple des nombreuses difficultés associées à la manipulation de représentation de connaissance. Ainsi, contrairement aux solutions proposées et utilisées jusque-là, **le caractère générique de la librairie ne contraint pas son utilisation à un contexte applicatif particulier**. Cela est rendu possible grâce à sa compatibilité avec un certain nombre de standards conçus pour la définition de représentation de connaissance, e.g. RDF, OBO. Cette librairie a d'ores et déjà été utilisée dans de nombreux projets pour comparer des entités (concepts et instances) définies dans de nombreuses représentations de connaissance[1] - ces projets sont détaillés dans le chapitre dédié à cette contribution.

Grâce à son aspect générique, la librairie propose de fédérer différentes communautés autour d'un cadre de développement commun. En effet, l'ajout de fonctionnalités à la librairie, e.g. nouvelles implantations de mesures ou optimisations d'algorithmes classiquement utilisés par celles-ci, bénéficiera à toutes les communautés intéressées par les mesures sémantiques. De plus, nous l'avons démontré au travers d'évaluations empiriques, la généricité de la librairie ne se fait pas au détriment de sa performance. En effet, en comparant la librairie à des solutions dédiées à l'analyse de gènes au travers de leurs annotations sémantiques, nous montrons qu'elle offre des performances équivalentes, voire supérieures à celles obtenues par les solutions spécifiques à un domaine, en particulier pour le traitement de gros volumes de données.

Le développement de cette librairie repose sur l'analyse de l'état de l'art et sur l'unification des mesures proposées par le cadre théorique introduit dans nos travaux. En effet, bien que théorique, le cadre unificateur des mesures que nous proposons est parfaitement implantable. De nombreuses évaluations de mesures ont été effectuées à partir d'une implantation (partielle) de ce dernier au sein de la librairie. On voit ici clairement le lien étroit entre les deux contributions théorique et appliquée qui, au final s'enrichissent mutuellement.

En se basant sur la librairie de code, nous avons aussi développé un outil logiciel utilisable en ligne de commande. Il permet aux non-développeurs de tirer parti de certaines fonctionnalités de la librairie, par exemple, pour exploiter les capacités de calcul qu'elle offre. Tout comme la librairie, cet outil générique ne se restreint pas à un domaine particulier et supporte l'utilisation de nombreuses représentations de connaissance. De plus, afin de répondre au plus près aux attentes des utilisateurs, nous avons proposé la mise en place d'interfaces (en ligne de commande) dédiées à des contextes applicatifs particuliers. Elles permettent aux utilisateurs d'interagir avec l'outil générique sans

[1] E.g., la Gene Ontology, le MeSH, SNOMED-CT, Yago, l'ontologie de DBpedia, Schema.org.

pour autant utiliser une terminologie différente de celle communément utilisée dans leurs communautés.

Le développement de la librairie et de l'outil logiciel associé s'accompagne d'une large documentation et d'un support technique assuré au travers d'un groupe de discussions et d'une liste de diffusion. Le lecteur intéressé consultera le site internet : `http://www.semantic-measures-library.org`.

Néanmoins, **le projet SML ne s'arrête pas aux développements logiciels**. De façon plus générale, nous l'avons souligné, ce projet se propose de **fédérer autour de la notion de mesure sémantique**. Ainsi, le travail d'état de l'art sur lequel reposent nos travaux est partagé au travers de ce projet. Pour cela, un document technique relatif aux mesures sémantiques (d'une centaine de pages) a été rendu public et une grande partie de la bibliographie associée est elle aussi partagée [Harispe et al., 2013c] ; ces travaux ont suscité un grand intérêt dans la communauté et de nombreux retours qui devraient conduire à plusieurs collaborations internationales.

L'état de l'art détaillé des mesures et les contributions associées au cadre théorique unificateur et au projet SML correspondent aux trois piliers théoriques et logiciels amenés dans cette thèse. Ils répondent à notre volonté d'initier un état des lieux de la connaissance relative à ce large domaine d'étude, qui a, selon nous, trop longtemps été cloisonné au sein de communautés diverses. **Ces contributions répondent à l'objectif initialement fixé : proposer des outils théoriques et pratiques dédiés aux mesures sémantiques**. En parallèle de ces travaux, qui constituent en quelque sorte le fil rouge de cette thèse, nous proposons différentes contributions algorithmiques et théoriques relatives aux mesures sémantiques à base de connaissance.

II.IV Contributions algorithmiques et autres contributions théoriques associées aux mesures sémantiques

Références :

An information theoretic approach to improve the semantic similarity assessment across multiple ontologies. Batet Montserrat*, Harispe Sébastien, Ranwez Sylvie, Sánchez David, Ranwez Vincent. Information Sciences (Elsevier) 2014 – cette contribution a été réalisée en collaboration avec Montserrat Batet et David Sánchez de l'Université Rovira i Virgili de Tarragone (Espagne) et Vincent Ranwez, Professeur à Montpellier SupAgro.

Robust Selection of Domain-specific Semantic Similarity Measures from Uncertain Expertise. Stefan Janaqi*, Sébastien Harispe, Sylvie Ranwez, Jacky Montmain. IPMU 2014 – Information Processing and Management of Uncertainty in Knowledge-Based Systems

Ces travaux portent sur des aspects spécifiques des mesures sémantiques et se détachent parfois de la vision abstraite qui a été adoptée jusque-là, notamment lors de la définition du cadre unificateur. **Ils proposent d'étudier des aspects particuliers des mesures et reposent sur des contributions algorithmiques ou théoriques ciblant un type de mesure spécifique**, parfois au regard d'un applicatif particulier.

Parmi ces contributions, ce manuscrit présente :

Une technique d'apprentissage semi-supervisée permettant de caractériser les mesures sémantiques adaptées à un contexte applicatif particulier, en tenant compte de l'incertitude intrinsèque aux jeux de tests communément utilisés pour l'évaluation de leur performance.

Les mesures sémantiques sont, la plupart de temps, évaluées par l'étude de leur corrélation avec des scores de similarité définis par des individus, généralement des experts de domaine, e.g. des médecins dans le domaine biomédical [Pakhomov et al., 2011; Pedersen et al., 2007]. C'est, dans certains cas, cette appréciation humaine de la similarité, parfois nourrie d'expertise, que l'on souhaite simuler par l'utilisation de mesures sémantiques. La performance des mesures est donc souvent évaluée à l'aide de jeux de tests composés de scores de similarité *attendus* pour un ensemble de paires d'entités : une mesure sera alors d'autant plus performante que ses résultats seront fortement corrélés avec cet attendu. Ce protocole d'évaluation est très largement utilisé. Cependant, jusque-là, l'incertitude associée aux jeux de test, qui découle de l'incertitude relative aux similarités associées aux paires d'entités qui les composent, n'était pas prise en compte lors de l'évaluation. De nombreuses études le soulignent, l'appréciation de la similarité est

subjective et imprécise, même au sein d'un groupe d'experts. La prise en compte de l'incertitude associée à un jeu de test est donc centrale pour l'évaluation des mesures sémantiques. Ainsi, pour répondre aux limites des protocoles d'évaluation classiques, nous proposons d'adopter un regard nouveau sur les mesures en évaluant leur robustesse : leur capacité de résilience au regard de l'incertitude associée aux jeux de tests classiquement utilisés. Une mesure sera ainsi d'intérêt si elle est fortement corrélée avec l'attendu et si elle le reste lorsque des perturbations (simulant l'incertitude) sont appliquées sur ce dernier. Cette proposition est illustrée par une évaluation empirique dans le domaine biomédical. Nous montrons notamment que, couplée à cette notion de robustesse, la décomposition des mesures permise par le cadre théorique permet d'étudier de nouvelles propriétés des mesures sémantiques.

Une nouvelle approche pour la comparaison d'instances caractérisées au travers d'un graphe sémantique.

Dans cette contribution nous proposons une nouvelle approche pour caractériser une instance au travers de la notion de *projection*. Une projection est utilisée pour représenter une propriété particulière d'une instance en exploitant différentes informations présentes dans un graphe sémantique, e.g. relations directes, indirectes, ou encore, chose nouvelle, en prenant en compte différentes propriétés caractérisées par les types de relation précités. Ainsi, une instance sera analysée au travers de l'ensemble des projections qui la caractérisent. Pour comparer une paire d'instances, ce sont ces projections qui seront examinées à l'aide de mesures (sémantiques) adaptées. Nous proposons ensuite d'estimer la proximité des instances comparées par agrégation des scores associés à la comparaison de leurs projections. L'intérêt de l'approche repose sur la caractérisation détaillée des instances au travers de la notion de projection, et sur les perspectives intéressantes qu'elle offre concernant la traçabilité de la sémantique du résultat produit. En effet, cette approche explicite la sémantique d'un score au regard (i) des différentes projections qui gouvernent la comparaison, (ii) des pondérations qui leurs sont associées, et (iii) des mesures utilisées pour les comparer. Nous soulignons l'importance de cet aspect, en particulier pour la mise en place de systèmes informatiques pour lesquels la justification des résultats obtenus peut représenter une plus-value non-négligeable (par exemple dans le domaine de la décision), notamment pour la mise en place d'interactions homme-machine. A titre d'illustration, nous montrons l'utilité de l'approche proposée pour la définition d'un système de recommandation (semi-supervisé) de groupes de musique. Celui-ci repose sur le paradigme des données liées et tire parti de l'analyse de données issues de DBpedia[1]. L'utilisateur du système a la possibilité de préciser l'importance des

[1]Le prototype développé est accessible à l'adresse `http://www.lgi2p.ema.fr/kid/tools/bandrec` (maintenu au minimum jusqu'en 2015).

propriétés retenues (i.e. les projections) lors de comparaison des groupes de musique. Cela lui permet d'avoir un contrôle fin sur la sémantique des résultats. L'avantage est double : l'utilisateur comprend le pourquoi de la recommandation et a la possibilité d'exprimer plus finement ses attentes – cela permet d'envisager une meilleure interaction avec le système. Cette contribution est détaillée dans un papier écrit en français [Harispe et al., 2013a], cependant, le lecteur pourra aussi se référer au chapitre de la thèse dédié à cette contribution et à [Harispe et al., 2013b].

Une approche pour la définition de mesures sémantiques permettant de comparer deux concepts définis dans des taxonomies différentes.

Cette étude porte sur les mesures sémantiques dédiées à la comparaison de concepts définis dans des représentations de connaissance différentes. Nous avons notamment proposé la redéfinition d'un opérateur communément utilisé dans la définition de ces mesures. Pour cela, en se basant sur des contributions relatives à la théorie de l'information, nous proposons une nouvelle approche pour caractériser la partie commune de deux concepts exprimés dans des taxonomies différentes (non-disjointes). Une fois de plus, ces travaux ne se concentrent pas sur l'étude d'une mesure particulière. En effet, de nombreuses mesures initialement définies pour la comparaison d'une paire de concepts d'une même taxonomie peuvent être utilisées, dans un contexte impliquant l'utilisation de plusieurs représentations de connaissance. En se basant sur différents jeux d'évaluation relatifs au domaine biomédical, nous montrons que l'approche proposée permet d'améliorer la précision des mesures sémantiques (dans le contexte d'évaluation testé[1]).

Une optimisation algorithmique pour calculer certaines mesures sémantiques.

Dans cette contribution, en tirant parti du cadre théorique proposé, nous distinguons une propriété permettant de caractériser une classe particulière de mesures sémantiques. Nous soulignons l'intérêt et les implications algorithmiques de cette propriété pour le calcul des mesures. Nous l'utilisons par la suite pour la définition de solutions algorithmiques dédiées au calcul de la similarité sémantique de l'ensemble des paires de concepts d'une taxonomie. Cette étude ne se limite pas à une analyse de la complexité théorique des solutions proposées, les aspects pratiques des algorithmes sont eux aussi discutés.

[1] De manière générale, trop peu d'analyses sur les mesures sémantiques discutent le caractère potentiellement généralisable des résultats produits par des évaluations empiriques *domaine-spécifiques*. Bien que dans cette étude nous ayons utilisé deux jeux de tests différents, rien ne nous permet de généraliser ce résultat.

III Synthèse et élargissement

L'analyse détaillée d'un grand nombre de contributions associées aux mesures sémantiques nous a permis d'identifier un certain nombre de défis relatifs à ce domaine d'étude. Six d'entre eux, particulièrement importants à nos yeux, sont détaillés :

1. *Proposer une meilleure caractérisation des mesures sémantiques.*
   La plupart des mesures sémantiques méritent d'être mieux caractérisées, notamment en ce qui concerne leur sémantique. En effet, un score de mesure sémantique est encore aujourd'hui trop souvent considéré comme dénué de sens et ramené à une simple valeur numérique. Cependant, dans certains contextes applicatifs, les implications associées à l'utilisation d'une mesure particulière peuvent être lourdes de conséquences et le choix d'une mesure peut, dans certains, cas remettre en cause la cohérence d'un système informatique. Pour relever ce défi, l'analyse des propriétés mathématiques des mesures sémantiques nous semble primordiale.

2. *Proposer des outils théoriques et logicielles pour l'étude des mesures sémantiques.*
   Nous pensons que des efforts soutenus doivent être effectués dans l'objectif de proposer des outils théoriques et logiciels dédiés aux mesures sémantiques. Nous avons notamment souligné l'importance des outils théoriques pour (formellement) caractériser la diversité des mesures proposées dans la littérature. Nous avons aussi attiré l'attention sur le fait qu'un plus grand nombre de jeux de test et d'outils dédiés à l'évaluation empirique des mesures doivent être proposés. De plus, nous avons insisté sur la nécessité de développer des solutions logicielles génériques, en particulier afin de répondre aux limites rencontrées par les solutions *domaine-spécifiques* majoritairement utilisées aujourd'hui.

3. *Standardiser la prise en compte de représentations de connaissance.*
   Nous avons mis en évidence les limitations pratiques induites par le manque de standardisation des traitements effectués sur les représentations de connaissance (lors ou au préalable du calcul de mesures sémantiques). Nous avons notamment souligné le trop grand degré de liberté laissé aux développeurs lors de l'implémentation d'une mesure, ce qui créé souvent un fossé entre une définition théorique d'une mesure et son implémentation. Nous avons par exemple observé que les scores produits par différentes solutions logicielles (en utilisant des mesures déterministes) varient largement dans certains cas. Notre analyse souligne que cette variation peut s'expliquer du fait de la non-standardisation de certains traitements appliqués sur les représentations de connaissance, e.g. la prise en compte des redondances taxonomiques.

4. *Faire la promotion de l'interdisciplinarité dans le domaine.*
   L'état de l'art relatif aux mesures sémantiques le montre bien, de nombreuses communautés contribuent à leur étude. Nous pensons cependant que pour favoriser un enrichissement mutuel de ces communautés, plus d'interactions méritent d'être entretenues. Dans cet objectif, nous avons identifié un certain nombre de communautés aujourd'hui directement impliquées dans l'étude des mesures ou bien qui mériteraient d'être sollicitées pour appuyer ces travaux.

5. *Étudier la complexité algorithmique des mesures sémantiques.*
   Trop peu d'études s'attachent à analyser la complexité algorithmique des mesures sémantiques. Cependant, du fait que la complexité d'une mesure impacte clairement son utilisation pratique, celle-ci constitue souvent un critère de choix important pour l'utilisateur final, soucieux de sélectionner un mesure adaptée à son contexte applicatif.

6. *Proposer des approches permettant d'orienter la sélection de mesures sémantiques au regard d'un contexte d'utilisation.*
   La sélection d'une mesure sémantique est un problème complexe et peu de solutions permettent de faciliter la tâche. En effet, une mesure doit être sélectionnée en fonction du contexte applicatif dans lequel elle sera utilisée. Néanmoins, à l'heure actuelle, la plupart des utilisateurs sélectionnent une mesure "*à l'aveugle*", en justifiant par exemple le choix d'une mesure par sa popularité. La caractérisation des mesures proposée dans nos travaux nous a permis de souligner l'importance à considérer à la fois les propriétés des mesures et la sémantique qui leur est associée. Ainsi, un plus grand nombre d'études méritent d'être effectuées dans ce domaine. Cela permettrait notamment de mieux caractériser la notion de contexte d'utilisation d'une mesure ainsi que les caractéristiques des mesures qui lui sont associées. De plus, des analyses comparatives empiriques doivent être effectuées dans différents domaines applicatifs afin de comparer les performances d'un nombre représentatif de mesures sémantiques. Ces analyses sont essentielles pour déterminer si une classe de mesures tend à obtenir de meilleures performances qu'une autre dans certains contextes, et pour évaluer si ce résultat est, en soit, généralisable.

Les contributions théoriques et logiciels dédiées aux mesures sémantiques qui ont été proposées dans cette thèse fournissent des éléments de réponse à la plupart des défis distingués ci-dessus. Nous avons souligné et illustré leurs apports pour amener une meilleure caractérisation des mesures, au regard à la fois de leurs propriétés mathématiques et de leur sémantique. Nous avons insisté sur l'intérêt d'utiliser les propriétés des mesures afin de les classifier et de les manipuler au travers de familles de mesures. Cela permet

notamment de facilement dériver des propriétés intéressantes pour un grand nombre de mesures. Ainsi, les contributions qui reposent sur ces propriétés (e.g. algorithmes d'optimisation) bénéficient à de nombreuses mesures et trouvent tout naturellement un public plus large. Une stratégie similaire, basée sur l'analyse de familles de mesures, peut être envisagée pour initier l'étude de la complexité algorithmique des mesures.

Le cadre théorique et la solution logicielle proposés, tous deux détachés de contextes applicatifs particuliers, s'inscrivent dans la volonté de créer et d'alimenter des liens étroits entre les communautés impliquées dans l'étude des mesures sémantiques. Cette interaction entre les différents acteurs du domaine est importante, notamment pour tenter d'amener une réponse collective aux défis aujourd'hui offerts à ce domaine de recherche. Nous avons par exemple mis en évidence la nécessité de détailler et de standardiser tant que possible les traitements effectués par les différents logiciels de calcul de mesures sémantiques. Ce travail ne peut être envisagé que si des collaborations larges et interdisciplinaires sont engagées.

Pour finir, nos contributions dotent les communautés de solutions pour orienter la sélection de mesures sémantiques au regard d'un contexte d'utilisation particulier. En effet, la littérature regorge de mesures sémantiques réputées toutes plus performantes les unes que les autres et pourtant trop peu d'études empiriques et théoriques se sont jusque-là intéressées à voir plus clair dans cette diversité, en particulier afin d'identifier les mesures les plus adaptées à un contexte d'utilisation particulier. Nous sommes convaincus que le cadre théorique proposé et le projet Semantic Measures Library ont leur rôle à jouer pour relever ce défi qui s'offre aux communautés investies dans l'étude des mesures sémantiques.

1

Introduction

Contents

Abstract

This first chapter introduces the general context of the thesis and presents several notions and paradigms on which are based our contributions. (i) We present the notion of ontology – how to formally express knowledge to make it *understandable* by software; usages of these ontologies for knowledge inference through exact and approximate reasoning techniques are further discussed. (ii) We introduce the Semantic Web and Linked Data paradigms – how to take advantage of the Internet infrastructure to build a Web of Data/Knowledge: a linked data cloud corresponding to a worldwide network of pieces of data and knowledge interlinked together. (iii) We discuss several contributions related to human appreciation of similarity focusing on the insight provided by cognitive sciences. This section finally defines both objectives and outlines of the thesis as well as chapter summaries.

1.1 General context

1.1.1 Knowledge in the quest to design Artificial Intelligence

One of the main challenges of Artificial Intelligence (AI) is to design intelligent agents which are able to resolve complex problems and to perform elaborated tasks. To this end, AI federates numerous scientific communities to tackle a large diversity of problems in the aim of giving machines the ability to reason, to understand knowledge, to learn, to plan, to manoeuvre, to communicate, and to perceive [Russell and Norvig, 2009].

Back in the 60s, the quest for AI had originally been motivated by the assumption "[...] *that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it* [...]" [McCarthy et al., 2006]; an assumption which has today proved to be pretentious and perhaps even unattainable. Among the various strategies explored to provide machines with intelligence, i.e., the "*ability to acquire and apply knowledge and skills*" [Oxford Dict., 2012], this thesis focuses on those which take advantage of formal expression of knowledge, also denoted as ontologies. Considering that most complex problems have proved to require the analysis of large sources of knowledge in order to be resolved (e.g., medical diagnosis), such strategies are based on the rational assumption that knowledge is one of the central ingredients required for the emergence of intelligence.

In this context, several communities have been involved in working to resolve one of the major problems challenging AI: how to formally express knowledge in order to make it *understandable* by software. These (on-going) efforts have led to the definition of several languages which can be used today to express formal, computer-readable and processable forms of knowledge. The general notion of ontology encompasses a large range of proposals which are commonly defined as formal, explicit and shared conceptualisations [Gruber, 1993]. Nevertheless, more generally, ontologies should be seen as a device used to bridge the gap between domain-specific expertise and computer resources by enabling a partial transfer of expert skills to computer systems. Ontologies will therefore often be considered as the soil from which intelligence can further be simulated using computer instructions. As an example, software denoted *reasoners* can be developed to apply inference procedures defined w.r.t the semantic interpretations associated to ontology languages, i.e., the definition of how the knowledge must be *understood* and processed by computers. This approach is used to simulate intelligence by enabling programmes to automatically perform complex deductive reasoning over a domain of interest. Nevertheless, the use of such ontologies is not only restricted to the rigid and strict interpretations enabled by knowledge representation languages. Indeed,

ontologies are also used to simulate more advanced forms of intelligence which, based on assumptions, will be used to design inexact reasoning or imprecise search techniques. Such techniques open interesting perspectives for AI by enabling the design of systems which think *out-of-the-box* and will, for instance, be used to infer probable knowledge implicitly defined in ontologies.

1.1.2 The growing adoption of knowledge-based systems

A knowledge-based system is characterised by the association between ontologies and software which enables them to be exploited. They are essential for solving complex problems which require the study of domain-specific knowledge to be taken into account. They are therefore largely used in the design of expert systems which support decision making. They are extensively used for the task of classification or, more generally, to answer exact queries w.r.t the knowledge modelled in ontologies. Therefore, from gene analysis to recommendation systems, knowledge-based systems are the backbones of numerous business and research projects today.

In recent decades, we have observed, both in numerous scientific communities and industrial fields, the growing adoption of knowledge-enhanced approaches. As an example, BioPortal and the Open Biological and Biomedical Ontology foundry give access to hundreds of ontologies related to biology and biomedicine [Smith et al., 2007; Whetzel et al., 2011]. These ontologies are used to develop a large range of applications for diagnosis, disease classification, drug design and gene analysis, to mention a few. Even large corporations adopt ontologies to support their large-scale worldwide systems. The most significant example of the recent years is surely the adoption of the Knowledge Graph by Google, a graph built from a large collection of billions of non-ambiguous statements used to formally describe general or domain-specific pieces of knowledge [Singhal, 2012]. This ontology is used to enhance their search engine capabilities and millions of users benefit from it daily.

Therefore, thanks to the large efforts made to standardise the technology stack which can be used to define and take advantage of ontologies (e.g., standard exchange formats, languages, development environment, storage systems, reasoners), a large number of initiatives give access to ontologies and knowledge-based systems in numerous domains (e.g., biology, geography, cooking, sports).

1.1.3 Towards a Web of Data/Knowledge

Numerous exciting perspectives have been opened by early knowledge modellers and specifications of languages enabling formal machine-understandable expressions of knowledge. Among them, one of the most exciting initiatives is the desire to build a Web of Data/Knowledge and services based on shared expressions of unambiguous data/-knowledge exposed through the Internet infrastructure. In short, the Web of hyperlinks between documents, only understood by humans, will be augmented by the definition of interlinked pieces of knowledge in order to make the Web a worldwide knowledge-based system which can be automatically processed by computer agents.

Several initiatives promote the Semantic Web and Linked Data paradigms to provide "*an extension of the current* [Web]*, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" [Berners-Lee et al., 2001]. This Semantic Web enables content publishers to add meaning to their webpages in order to make them more valuable for automatic analyses. In addition, the unambiguous characterisation of resources, a central element of these paradigms, gives collaborative initiatives the opportunity to build large networks of knowledge according to the Umberto Eco quote principle: "*Any fact becomes important when it is connected to another*" [Eco, 1989]. To this end, international consortiums composed of both scientists and organisations, such as the World Wide Web Consortium (W3C), led to the definition of several standards for the publication of structured data and knowledge associated to formal semantics. This knowledge can further be interrogated using a specific standardised query language.

The Semantic Web is now emerging from its cocoon. Thanks to the efforts made to design scalable technological solutions to store and query semantic data, a growing number of companies consider knowledge-based systems as well as Semantic Web technologies to support their business. Billions of pieces of unambiguous machine-understandable knowledge are already exposed on the Internet and several large ontologies are now available, some of them for free: DBpedia [Auer et al., 2007], Freebase [Bollacker et al., 2008], Wikidata [Vrandečić, 2012], Yago [Hoffart et al., 2013].

Another significant example of the increasing adoption of ontologies is the joint effort made by the major search engine companies and web organisations (e.g., Microsoft [Bing], Google, Yahoo!, W3C) to design Schema.org[1]. This set of structured schemas defines a vocabulary which can be used by publishers to define metadata with the aim of characterising the content of their webpages in an unambiguous manner.

[1] `http://schema.org`

All these initiatives converge to the same goal: to express and publish knowledge in a formal machine-understandable fashion in order to enable intelligent agents to take advantage of it.

1.1.4 Thinking outside the box: the importance of inexact searches

Knowledge-based systems and ontology definitions are generally motivated by the desire to reason accurately over domain-specific knowledge. Putting aside the fact that knowledge is not always both accurate and precise, and that numerous efforts are made to formalise such imprecise knowledge, ontologies are also extensively used to support knowledge discovery. Contrary to deductive knowledge inferences which are commonly used to classify and infer new facts based on exact inference procedures, knowledge discovery relies on approximate reasoning.

Approximate reasoning or inexact search techniques are essential for numerous systems and treatments which cannot rely only on asserted knowledge defined in an ontology, e.g., information retrieval or recommendation. They are, for instance, required for query answering based on imprecise goal definitions, e.g., *which bands are similar to the Rolling Stones?*

Given the importance of inexact searches to solve complex problems, numerous contributions have focused on designing algorithmic techniques based on ontologies to support inexact searches, approximate reasoning and knowledge discovery. Here, the aim is not to assert exact facts about a domain, or to search for an exact answer to a query, but rather to evaluate the interconnections between pieces of knowledge w.r.t the ontology in which they are defined. In other words, the aim is to design algorithms which will *think outside the box* by considering specific assumptions. Such algorithms will therefore be used to break the boundary of the formal semantics on which ontologies rely, in order to derive pieces of knowledge neither implicitly nor explicitly defined in an ontology. These approaches rely extensively on the capacity to distinguish features characterising similar cases, and on the capacity to evaluate the similarity of cases represented through specific canonical forms.

Human capacity to evaluate the similarity of *things* (e.g., objects, stimuli) has long been studied by cognitive sciences and psychology. It has been characterised as a central element of the human cognitive system, and is therefore understood nowadays as a pivotal notion to simulate intelligence [Rissland, 2006]. Similarity is indeed a key element in initiating the learning process in which the capacity to recognise similar situations helps us to build our experience, to activate mental traces, to make decisions, to innovate by resolving problems by applying experience which have been gained by resolving similar

problems, etc. Similarity is therefore a central component in memory retrieval, categorisation, pattern recognition, problem solving, reasoning, as well as social judgement. It is therefore clear that intelligent agents must also be endowed with the ability to assess the similarity of *things*. To this end, new approaches need to be defined in order to take advantage of the knowledge defined in ontologies to estimate the similarity of *things*; this is done by means of knowledge-based semantic measures.

1.1.5 Knowledge-based semantic measures

Cornerstones of inexact-search algorithms on ontologies are semantic measures: functions used to estimate the degree of likeness (similarity/relatedness) of semantically characterised entities, e.g., concepts or instances formally characterised in ontologies. These measures are, for example, used to estimate the proximity of resources (e.g., diseases) indexed by concepts structured in an ontology (e.g., syndromes), or more generally, to compare entities w.r.t the knowledge defined in an ontology.

For the sake of clarity, let us specify the notions of entities, concepts, classes and instances which will be used in this manuscript. We considered the notion of *concept* in a broad sense: *an idea or notion; a unit of thought* [W3C, 2009], class of instances which can be of any kind (abstract/concrete, elementary/composite, real/fictive) [Smith, 2004]. Notice that we also consider that a concept can be represented through a *synset*, i.e., a set of synonyms, or more generally, any group of data elements considered as semantically equivalent. A concept can therefore be represented as any set of words or terms referring to the same notion, e.g., the terms *dog* and *Canis lupus familiaris* refer to the concept `Dog`. Note that we use both notions of concept and class interchangeably. However, we will, as much as possible, favour the use of the term concept as specifications used to express ontologies generally refer to it. The notion of a semantically characterised *instance* encompasses several situations in which an object is described through information; information from which semantic analyses can be performed. Semantic characterisations cover a wide range of canonical forms which can be used to characterise an instance, e.g. any description using a specific ontology language or a set of conceptual annotations. The notion of entity encompasses both the notion of concept and instance.

Semantic measures are extensively used to mimic human appreciation of similarity. In this case, the ontology used by the measures can be associated to the human mental representation of knowledge, and a semantic measure can be seen as our capacity to process our knowledge to assess the similarity of things. Semantic measures are therefore

originally framed in cognitive sciences which have, for a long time, studied human appreciation of similarity and which have proposed numerous models of mental representation of knowledge. As we will see, given their importance to fully benefit from ontologies without being restricted to their exact inference procedures, semantic measures are central elements of numerous treatments. This is proved by the extensive literature dedicated to semantic measures which has been published over the last decades - several references are provided in [Harispe et al., 2013c].

1.1.6 General context of this thesis

Anchored in the field of AI, and more particularly interested in techniques based on formal ontologies during the emergence of the Web of Knowledge, this PhD thesis proposes several contributions related to the study of knowledge-based semantic measures. This work has been supported by the KID team (Knowledge and Image analysis for Decision making[1]) of the LGI2P research centre[2] – a laboratory of the engineering school École des mines d'Alès (EMA). Federated around the study of cognitive automation, the LGI2P focuses on the development of innovative concepts, methodologies and tools for the conception, realisation and optimisation of technical systems, collaborative processes and socio-technical organisations. In this effort, the KID team takes advantage of ontologies to define optimised techniques for knowledge discovery, retrieval and analysis [Ranwez, 2013][3]: techniques in which semantic measures are central elements.

1.2 Ontologies from a graph perspective

This section introduces the reader to ontologies which can be processed as graphs. It doesn't aim at: (i) presenting the vast field of Knowledge Representation, (ii) discussing the broad diversity of ontologies which have been proposed in the literature, and (iii) introducing the language and specifications which can be used to express ontologies, e.g., RDF(S), OWL. Here, we assume that the reader is already familiar with knowledge modelling and the associated terminology. However, if required, an introduction to this field of study is provided in Appendix A.

Numerous ontologies can be expressed as graphs. In addition, more complex ontologies can be reduced or used to generate knowledge represented as a graph. This section

[1] `http://kidknowledge.wp.mines-telecom.fr`
[2] Laboratoire de Génie Informatique et d'Ingénierie de Production.
[3] In french.

discusses specific aspects of ontologies related to graph representations. We first introduce simple ontologies which can be represented as graphs (e.g., taxonomies) to further discuss the case of more complex ontologies.

1.2.1 Taxonomies and partially ordered sets

Taxonomies are used to structure elements which have similar characteristics into ordered classes. They were originally used in biology to define *taxa* (classes), by categorising organisms sharing common properties. A taxonomy is a function of a taxonomic scheme which defines the properties considered to distinguish classes. Depending on this scheme, the number of classes and their ordering may vary.

The semantics carried by a taxonomy is non-ambiguous as the interpretation of the taxonomic relationship is formally expressed through particular properties/axioms. Indeed, considering a set of elements C (e.g. concepts), a taxonomy is a non-strict partial order (*poset*) of C. It can be defined by $\preceq_C$, a binary relation $\preceq$ over C which is[1]:

- Reflexive $\forall c \in C : c \preceq c$.
- Antisymmetric $\forall u, v \in C : (u \preceq v \wedge v \preceq u) \Rightarrow u = v$.
- Transitive $\forall u, v, w \in C : (u \preceq v \wedge v \preceq w) \Rightarrow u \preceq w$.

Note that in some rare cases taxonomies are totally ordered, but generally, they are only partially ordered, i.e., $\exists (u, v) \in C : u \npreceq v \wedge v \npreceq u$. Given that they generally contain a root element denoted $\top$ which subsumes all other elements, i.e., $\forall c \in C, c \preceq \top$, they can be represented as a connected, Rooted and Directed Acyclic Graph (RDAG).

A taxonomy of concepts $\preceq_C$ can therefore be formally defined as a semantic graph $O :< C, R, E, A^O >$ with C the set of concepts, R a singleton defining the unique predicate which can be used to order the concepts, i.e. $R = \{\texttt{subClassOf}\}$ and $E \subseteq C \times R \times C$ the set of oriented relationships (edges) which defines the ordering of C.

Only considering $O :< C, R, E >$ leads to a labelled graph structuring elements of C through labelled oriented edges. Nevertheless, by defining the sets of axioms associated to the taxonomic predicate defined in R, e.g., associated relationships are considered reflexive, antisymmetric and transitive, A^O explicitly and formally states that O is a taxonomy *per se* and not a simple graph data structure. These axioms can be used to define inference techniques and more generally to ensure the coherence of specific algorithms w.r.t the knowledge defined in the representations. As an example, Figure 1.1 denotes an example of taxonomy represented by a graph structure.

[1]Note that we adopt the notation used in the literature related to poset instead of the notation commonly used in description logics ($\sqsubseteq$).

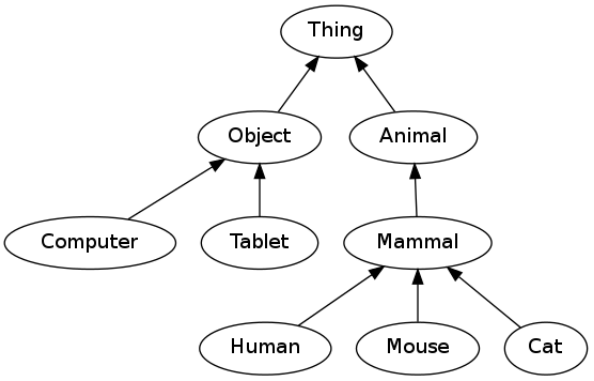

FIGURE 1.1: Taxonomy of concepts represented as a graph

Although simple, taxonomies are ontologies which are used in numerous processes; they are also the backbones of more refined ontologies and are therefore considered as essential components of knowledge modelling. Numerous contributions presented in this manuscript rely on these simple ontologies. The notion of taxonomy has been detailed here through a taxonomy of concepts, we also consider $\preceq_R$ the taxonomic of predicates in which `subPredicateOf` refers to the taxonomic relationship defining that one predicate inherits from another[1].

1.2.2 General discussion on ontologies as graphs

As we have seen, any taxonomy of concepts can be represented as a graph, including those expressed using a logic-based ontology – in some cases reasoners will be used to infer the taxonomy of concepts defined through complex definitions (*subsumption relationships* in description logics [Nardi and Brachman, 2003]). Nevertheless, the taxonomic knowledge encompassed in the ontology can be (partially[2]) manipulated through this taxonomy.

Although some ontologies cannot be reduced to simple graphs, a large part of the knowledge they model can generally be expressed as a graph. Therefore, an important aspect to understand is that ontologies, even if they are not explicitly defined as graphs, can be reduced into graphs. Indeed, in all cases, a partial representation of the knowledge defined in expressive ontologies can be manipulated as a graph. The example of the taxonomy has been underlined but this is also the case for the knowledge which links instances to classes (also obtained by a common reasoning procedure). In this case, the ontology can be reduced as a graph in which instances are represented as nodes and

[1] Generally named `subPropertyOf`, e.g., in RDFS.

[2] Note that we do not directly compare concept descriptions but rather their implicit organisation. The reduction of a set of concept descriptions to a poset implies knowledge loss, i.e., concepts are now considered regarding their ordering.

linked to their class(es) by simple subject–predicate–object (`spo`) statements. Therefore, any complex ontology, in which sets of concepts and instances have been defined, can be represented as a connected graph in which nodes denote concepts or instances.

In this manuscript, we consider a graph-based formalism frequently used to manipulate ontologies. It can be used to express numerous network-based ontologies and, sometimes through reductions, ontologies which rely on complex logic constructs. It corresponds to an extension of the structure $O :< C, R, E, A^O >$ which has been presented to introduce taxonomies as graphs. Extensions have been made to take instances, data values and multiple predicates into consideration. The next subsection presents this formalism in detail, a more detailed discussion regarding the *mapping* between complex ontologies and the specific network-based ontology adopted is further discussed in Appendix A.2.

1.2.3 Types of ontologies considered in this thesis

Regardless of the particularities of some domain-specific ontologies and regardless of the language considered for the modelling, all approaches used to represent knowledge share common components:

- *Concepts* (Classes), set of things sharing common properties, e.g., `Human`.
- *Instances*, i.e., members of classes, e.g., `alan` (an instance of the class `Human`).
- *Predicates*, the types of relationships defining the semantic relationships which can be established between instances or classes, e.g., `subClassOf`.
- *Relationships*, concrete links between classes and instances which carry a specific semantics, e.g., `alan isA Human` – `alan worksAt BletchleyPark`. Relationships form `spo` statements.
- *Attributes*, properties of instances, e.g., `Alan hasName` *Turing*.
- *Axioms*, for instance defined through properties of the predicates, e.g. *taxonomic relationships are transitive*, the definition of the *domain* and the *range* (co-domain) of predicates, or constraints on predicate and attributes, e.g., *Any Human has exactly 2 legs*.

In practice, numerous ontologies do not rely on complex logical constructs or complex concept/predicate definitions but rather correspond to a formal semantic network, here

denoted *semantic graphs*. In addition, we have stressed the fact that complex ontologies can also be regarded as semantic graphs (sometimes considering partial reductions).

A semantic graph, in which instances of classes and data values of specific datatypes are considered, can formally be defined by $O :< C, R, I, V, E, A^O >$, with:

- C the set of concepts.
- R the set of predicates.
- I the set of instances.
- V the set of data values.
- E the set of oriented relationships of a specific predicate $r \in R$:
  $E \subseteq E_{CC} \cup E_{RR} \cup E_{II} \cup E_{IC} \cup E_{CI} \cup E_{CV} \cup E_{RV} \cup E_{IV}$ with:
  - $E_{CC} \subseteq C \times R \times C$
  - $E_{RR} \subseteq R \times R \times R$
  - $E_{II} \subseteq I \times R \times I$
  - $E_{IC} \subseteq I \times R \times C$
  - $E_{CI} \subseteq C \times R \times I$
  - $E_{CV} \subseteq C \times R \times V$
  - $E_{RV} \subseteq R \times R \times V$
  - $E_{IV} \subseteq I \times R \times V$
- A^O the set of axioms defining the interpretations of classes and predicates.

The sets of concepts (C), predicates (R), instances (I), values (V) are expected to be mutually disjoint[1]. We consider that each instance is a member of at least one concept and that the taxonomies of concepts $\preceq_C$, and predicates $\preceq_R$ (if any), correspond to connected RDAGs. In this manuscript we will mainly manipulate such an ontology without considering predicate taxonomies.

In the following, we consider that a lexical reference (*didactical device* [Guarino and Giaretta, 1995]) is used to refer, in an unambiguous manner, to any node which refers to a concept/predicate/instance. Although we will use a literal in this manuscript, in practice, this unique identifier is a URI – except data values (V) which are (typed) literals.

Figure 1.2 presents an example of a semantic graph related to the music domain which involves related concepts, predicates, instances and data values[2].

[1] Note that a set of data types (D) can easily be added.
[2] Representation of a subgraph extracted from DBpedia [Auer et al., 2007].

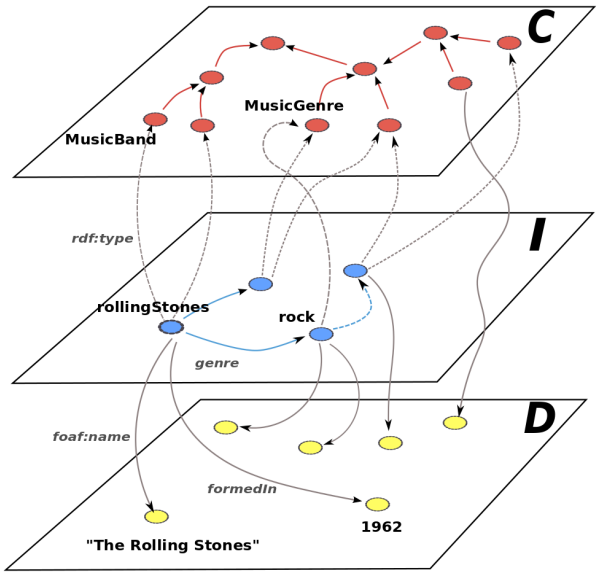

FIGURE 1.2: Example of a semantic graph related to the music domain. Concepts, instances and data values are represented [Harispe et al., 2013b]

In this example, concepts are taxonomically structured in the layer C, e.g. `MusicBand`, `MusicGenre`. Several types of instances are also defined in layer I, e.g. `rollingStones`, `rock`. These instances can be characterised according to specific concepts, e.g. the statement `rollingStones isA MusicBand` defines that `rollingStones` is a member of the class `MusicBand`. In addition, instances can be interconnected through specific predicates, e.g., `rollingStones hasGenre rock`. Specific data values (layer D) can also be used to specify information relative to both concepts and instances, e.g., `rollingStones haveBeenFormedIn "1962-01-01"^^xsd:date`. All relationships which link the various nodes of the graph are directed and semantically characterised, i.e., they carry an unambiguous and controlled semantics. Notice that extra information are not represented in this figure, e.g., the taxonomy of predicates, axiomatic definitions of predicate properties.

Appendix A.2 discusses the treatments which are required to obtain a semantic graph from an ontology. Only the notations introduced in the appendix are presented below.

Reduction of an ontology into a graph: Formally, we denote $G(O)$, shortened as G if there is no ambiguity, the reduction of the ontology O to a semantic graph $G = O \setminus A^O$. This process may involve inference techniques, reduction of some inferred statements, etc.

We denote $G_{R'}(O)$, also shorten $G_{R'}$ if there is no ambiguity, the reduction of O as a semantic graph only considering the relationships having as predicate $r \in R' \subseteq R$. A common reduction of an ontology as a graph is $G_{\texttt{subClassOf}}$, shortened by G_T and named the taxonomic reduction (layer C in Figure 1.2). G_T corresponds to the taxonomy $\preceq_C$, and therefore only contains concepts. As we will see, this reduction is widely used to compute the semantic similarity between concepts; it will be extensively used in this manuscript.

Graph reductions can naturally be more complex. The graph $G_{R'}(O)$, with $R' = \{\texttt{subClassOf},\texttt{isA}\}$, refers to the reduction which is composed of the relationships having as predicate `subClassOf` or `isA`. We denote such a graph G_{TI} (T stands for Taxonomic and I for `isA` relationship). It corresponds to the graph composed of the layers C and I in Figure 1.2 (only considering edges in E_{CC}, E_{II} and E_{IC}[1]).

Knowledge modelling is a vast domain and a large diversity of ontologies have been proposed to express knowledge in a machine understandable form. This section has briefly introduced several ontologies which can be processed as graphs. We have also introduced the formalism adopted in this manuscript to represent such ontologies.

1.2.4 Similarity: a cornerstone of approximate reasoning

Two broad types of reasoning techniques can be used over ontologies[2]: *exact* and *approximate* (inexact) reasoning [Gabbay et al., 1998; Russell and Norvig, 2009].

Exact reasoning is performed by means of *deductive* reasoning: an exact top-down reasoning approach in which general rules defined in a specific domain of discourse are used to infer exact statements. It is commonly used to infer exact facts implicitly defined in ontologies; the validity of the inferences only relies on the validity of the premises taken into account for their derivation – the inferences can only be correct w.r.t the ontologies and the interpretation associated to the language used for their definition. Deductive reasoning is central to knowledge-based systems and ontology modelling, they are, for example, used for the tasks of classification (instance typing, subsumption relationships inference, i.e., taxonomy inference) and for consistency checking. Deductive reasoning on ontologies are extensively discussed in Baader et al. [2010]; Hitzler et al. [2011].

Two types of inexact reasoning techniques are distinguished:

- *Inductive reasoning* is based on generalisation (bottom-up approach); specific observations are used to infer general rules about a domain. In other words, the

[1] Triplets are rarely defined in E_{CI}.

[2] Remember that we do not consider imprecise ontologies, e.g., fuzzy logics. We therefore do not consider inexact ontologies in which uncertain or contradictory knowledge is modelled.

conclusions cannot be guaranteed by the evidence considered [Gabbay et al., 1998; Holland et al., 1989].

- *Abductive reasoning* is also an inexact procedure. It can be used to consider observations and rules to derive possible conclusions. Generally, abductive reasoning considers a set of observations to derive conclusions which better explain them.

Contrary to deductive reasoning for which new knowledge (premise) cannot contradict prior conclusions, in inductive reasoning and abductive reasoning, the hypotheses can be supported or neglected by new observations (i.e., they are non-monotonic). The confidence associated to specific conclusions derived from approximate reasoning techniques is therefore a function of the confidence associated to the evidence considered.

Approximate reasoning may be performed using *supervised* or *unsupervised* learning approaches [MacKay, 2003; Mohri et al., 2012; Witten et al., 2011]. These approaches are used to design automatic classifiers able to correctly label objects (cases) and are generally extensively based on measures assessing the similarity of objects. Learning techniques applied to ontologies, and more particularly lazy learning techniques[1], have been extensively covered in D'Amato [2007].

Despite the fact that exact conclusions cannot be obtained using approximate reasoning techniques, they can be used to better understand complex phenomena, to formulate hypotheses (probable inferences not logically derivable), and to highlight limits of ontologies on which they are based. They are therefore commonly used for automatic construction and enrichment of ontologies, to align ontologies, i.e., to find links between ontologies, or even to evaluate ontologies; D'Amato [2007] provides several references. Approximate reasoning is also central to the design of information retrieval techniques, data analysis or query answering based on ontologies (without being constrained to exact query via SPARQL). Approximate reasoning is therefore a key element of knowledge discovery based on ontologies [Corby et al., 2006; Phillips and Buchanan, 2001].

Both supervised and unsupervised learning techniques rely on functions assessing the similarity/dissimilarity of objects. As an example, lazy learning techniques use a similarity function to distinguish a subset of cases which are relevant to define the discriminative function. Clustering algorithms group cases based on their similarity to further distinguish hidden structures in collection of cases. (Dis-)Similarity measures are therefore central to take advantage of classical learning approaches using ontologies; in this case,

[1] A kind of learning techniques in which the discriminative function (e.g. used for classifying) can be adapted for considering the information carried by new cases.

the measures must not be based on descriptions of objects represented using unstructured features and values, but rather take into account rich semantic characterisation of objects (e.g., concepts, instances).

1.3 Semantic Web and Linked Data paradigms

1.3.1 A natural paradigm shift

Most people understand the Web as a Web of Documents, a graph of interlinked webpages exposed through the Internet. Such a web has been designed to be used by humans; webpages contain information or data which is distilled through texts or multimedia contents that people can read, visualise, listen. People naturally surf the Web, jumping from one webpage to another by following the hyperlinks which structure its massive network of documents. Thanks to the evolution of the Web and the emergence of Web (2.0) communities, it is not only hyperlinks but also *friends* who can be followed. Indeed, changes have allowed for the increasing social commitment of users by enabling them to get connected and to become not only consumers and critics of web content, but also publishers, sometimes of their own lives...

Webpages and more generally speaking web content have long been human-centric and poorly *understood* by computers. Indeed, extraction of knowledge or interesting information from webpages requires the use of complex Natural Language Processing techniques; techniques which are often time consuming, imprecise, and perform poorly with ambiguity (refers to the discussion related to the ambiguity of human language Section A.1.2). Therefore, to overcome these limitations, initiatives have been proposed to semantically characterise webpages through unambiguous metadata. More importantly, propositions have been made to take advantage of the infrastructure offered by the Internet to build a Web of Data/Knowledge, a linked data cloud corresponding to a worldwide network of interlinked pieces of data and knowledge.

Over the last fifteen years, this new Web has technically been made possible through the definition of various specifications and implementations enabling the formal description of ontologies and resources on the Web, e.g., using RDF(S), OWL. These initiatives have led to the Semantic Web and the Linked Data paradigms which envision "*an extension of the current* [Web]*, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*" [Berners-Lee et al., 2001].

Today, a growing amount of semi-structured and structured data sources make up the linked data cloud. This is complementary to open data initiatives which encounter a

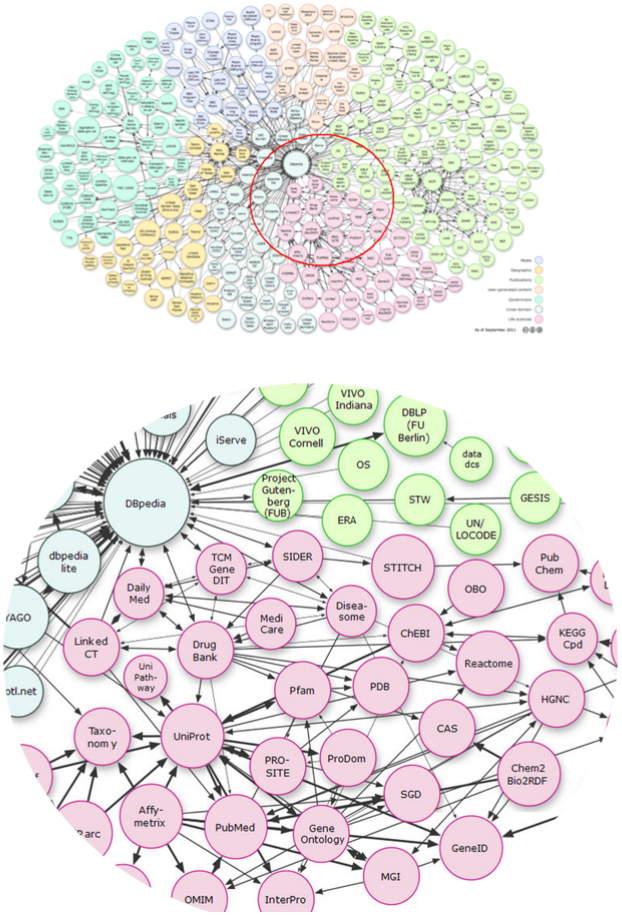

FIGURE 1.3: Linked data cloud showing interlinked data silos available on the Web. Original picture by Cyganiak and Jentzsch [2011]. Reduction from Zieliński [2014]

.

lot of success in governments and industries and which ease data accessibility, sharing and reuse. The data, information and knowledge is thus being freed, structured and semantically characterised, and a new Web conceived paving the way for a potential new automatic process of data and knowledge exposed on this worldwide network. Figure 1.3 presents a famous picture in the Semantic Web community which shows that numerous data silos expressed in RDF are linked together to form a worldwide cloud of semantically characterised data, information and knowledge. Note that, according to the definition presented in Section A.1.1, RDF can be used to express data, information and knowledge. Nevertheless, the distinction between these notions is not generally made when talking about Linked Data and the Semantic Web. Indeed, in some cases the notion of Web of Data is best suited given that only data are represented; in other cases, the notion of Web of Knowledge is more appropriate since RDF graphs are used to express Knowledge, for instance using URI disambiguation and the semantics provided by RDFS and OWL. In this manuscript we prefer the denomination Web of Knowledge.

1.3.2 Technologies and architecture of the Semantic Web

We have already mentioned the central elements of the Semantic Web: URIs, RDF/RDFS and OWL specifications. These specifications are part of the Semantic Web technology stack presented in Figure 1.4.

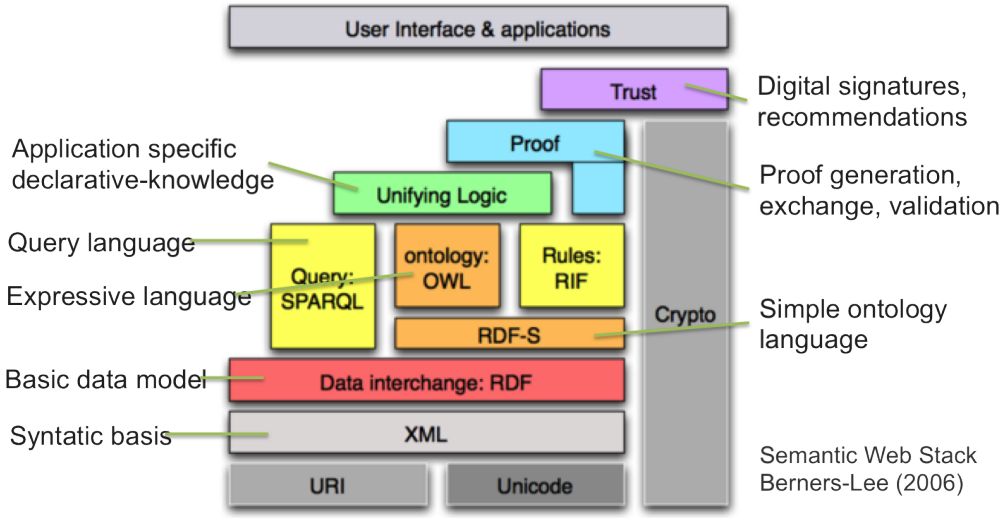

FIGURE 1.4: Technology stack of the Semantic Web.
From `http://projects.kmi.open.ac.uk/euclid`
.

The stack distinguishes several layers of specifications and protocols which must be developed in order to envision a Semantic Web useful for both human and computer agents. Among the different layers, we distinguish the disambiguation layer which enables non-ambiguous characterisation of resources through URIs. These URIs can then be exposed, exchanged and exploited through the HTTP protocol. They can also be linked and further described through various RDF graphs serialisation format.

RDF graphs can be queried using SPARQL, they can also be associated to specific semantics using vocabularies or language constructs provided by RDFS or OWL. At this stage, RDF graphs are not simply oriented and labelled graphs; they correspond to ontologies which can be used to formally characterise (domain-specific) knowledge. Through procedures automatising the inferences which can be made from the semantics of the languages in use, these ontologies can be processed to infer new pieces of knowledge from existing ones.

The layers associated to Proof, Trust and Security refers to both, technologies which have not yet been standardised, and ideas envisioned in the original formulations of the

Semantic Web. They are associated to a problem related to inference quality, confidence associated to pieces of knowledge, privacy...

Finally, the layer associated to user interfaces and applications has recently been subject to numerous contributions. Search engines have for instance been developed to query the Semantic Web, i.e., to find URIs or ontologies, e.g., Watson [D'Aquin and Motta, 2011], Sindice [Tummarello et al., 2007], Falcons [Cheng et al., 2008]. Large companies are also using Semantic Web technologies to disambiguate and enrich the content of their webpages. As an example, the BBC (British Broadcasting Corporation), Times Inc., Elsevier and Boeing are among the numerous companies which have production systems which now benefit from Semantic Web technologies [D'Aquin et al., 2008; Kobilarov et al., 2009].

1.3.3 Inexact searches: a key challenge for the Semantic Web

As we have seen, the Semantic Web and Linked Data paradigms offer the chance to expose structured data and knowledge on the Internet in such a way as to enable their automatic processing. The challenges associated to this research area are numerous.

One of the major challenges is to provide inexact search capabilities. There is indeed a need to develop search engines to distinguish relevant data and sources of knowledge defined on the Web of Knowledge. To this end, techniques must be developed to compare resource descriptions by taking the semantics of their characterisation into account. Resources described through RDF or represented in RDFS and OWL ontologies therefore have to be compared [Corby et al., 2006]. To this end, once again, measures able to assess the similarity or dissimilarity of resources w.r.t their formal descriptions have to be defined.

1.4 Human cognition, similarity and existing models

The human cognitive system is sensitive to similarity, which explains why the capacity to estimate the similarity of things is essential in numerous treatments. It is indeed a key element to initiating the process of learning in which the capacity to recognise similar situations helps us to build our experience[1], to activate mental traces, to make decisions, to innovate by applying experience gained in solving similar problems[2] [Gentner and Markman, 1997; Holyoak and Koh, 1987; Novick, 1988; Ross, 1987, 1989; Vosniadou and Ortony, 1989]. According to the theories of transfer, the process of learning is also subject to similarity since new skills are expected to be easier to learn if similar to skills already learned [Markman and Gentner, 1993]. Similarity is therefore a central component of memory retrieval, categorisation, pattern recognition, problem solving, reasoning, as well as social judgement, e.g., refer to [Goldstone and Son, 2004; Hahn et al., 2003; Markman and Gentner, 1993] for associated references.

As we have seen, the notion of similarity is central in numerous fields and is particularly important for human cognition and *intelligent* system design. In this subsection, we provide a brief overview of the psychological theories of similarity by introducing the main models proposed by cognitive sciences to study and explain (human) appreciation of similarity. Here, the process of similarity assessment should be understood in a broad sense, i.e., as a way to compare objects, stimuli.

Cognitive models of similarity generally aim to study the way humans evaluate the similarity of two mental representations according to some kind of psychological space [Tversky, 2004]. They are therefore based on assumptions regarding the mental representation of the compared objects from which the similarity will be estimated. Indeed, as stated by several authors, the notion of similarity, *per se*, can be criticised as a purely artificial notion. In Goodman [1972], the notion of similarity is defined as "*an imposture, a quack*" because objectively, everything is equally similar to everything else. The authors emphasise that, conceptually, two random objects have an infinitive number of properties in common and infinite different properties[3], e.g. a flower and a computer are both smaller than 10m, 9.99m, 9.98m.... An important notion to understand, which has

[1] Cognitive models based on categorisation consider that human classify things, e.g., experience of life, according to their similarity to some prototype, abstraction or previous examples [Markman and Gentner, 1993].

[2] Here the similarity is associated to the notion of generalisation and is measured in terms of probability of inter-stimulus-confusion errors [Nosofsky, 1992].

[3] This statement also stands if we restrict the comparison of objects to a finite set of properties. The reader may refer to Andersen's famous story of the Ugly Duckling. Proved by Watanabe and Donovan [1969], the *Ugly Duckling* theorem highlights the intrinsic bias associated to classification, showing that all things are equal and therefore that an ugly duckling is as similar to a swan as two swans are to each other. The important teaching is that biases are required to make a judgement and to classify, i.e., to prefer certain categories over others.

been underlined by cognitive sciences, is that differential degrees of similarities emerge only when some predicates are selected or weighted more than others. As stated in Hahn [2011], this important observation doesn't mean that similarity is not an explanatory notion but rather that the notion of similarity is heavily framed in psychology. Similarity assessment must therefore not be understood as an attempt to compare object realisations through the evaluation of their properties, but rather as a process aiming to compare objects as they are understood by the agent which estimates the similarity (e.g., a person). The notion of similarity therefore only makes sense according to the consideration of a partial mental representation on which the estimation of object similarity is based.

Contrary to real objects, representations of objects do not contain infinitesimal properties. As an example, our mental representations of things only capture a limited number of dimensions of the object which is represented. Therefore, the philosophical worries regarding the soundness of similarity vanish given that similarity aim at comparing partial representations of objects and not objects themselves, e.g., human mental representation of objects [Hahn, 2011]. The similarity is thus estimated between mental representations. Considering that these representations are the ones of a human agent, the notion of similarity may thus be understood as how similar objects appear to us. Considering the existential requirement of representations to compare things much of the history of research on similarity in cognitive sciences focuses on the definition of models of the mental representation of objects.

The central role of cognitive sciences regarding the study of similarity relies on the design of cognitive models of both, mental representations and similarity. These models are further used to study how humans store their knowledge, and to interact with, in order to compare objects sometimes represented as pieces of knowledge. Cognitive scientists then test these models according to our understanding of human appreciation of similarity. Indeed, evaluations of human appreciation of similarity help us to distinguish constraints/expectations on the properties an accurate model should have. This approach is essential to reject hypotheses and improve the models. As an example, studies have demonstrated that appreciation of similarity is sometimes asymmetric: the similarity between a person and his portrait is commonly expected to be lower than the inverse.[1] Therefore, the expectation of asymmetric estimation of similarity is incompatible with the mathematical properties of a distance, which is symmetric by definition. Models based on distance axioms therefore appeared inadequate and have therefore to be revised or to be used with moderation. In this context, the introduction of cognitive

[1]Indeed, Tversky [1977] stresses that *We say "the portrait resembles the person" rather than "the person resembles the portrait".*

models of similarity will be particularly useful to understand the foundations of some approaches adopted for the definition of semantic measures.

Cognitive models of similarity are commonly organised into four different approaches: (i) Spatial models, (ii) Feature models, (iii) Structural Models and (iv) Transformational models. We briefly introduce these four models though a more detailed introduction can be found in Goldstone and Son [2004] and Schwering [2008]. A captivating talk introducing cognition and similarity, on which this introduction is based, can also be found in Hahn [2011].

1.4.1 Spatial models

The spatial models, also named geometric models, rely on one of the most influencal theories of similarity in cognitive sciences. They are based on the notion of psychological distance and consider objects (here perceptual effects of stimuli or concepts) as points in a multi-dimensional metric space.

Spatial models consider similarity as a function of the distance between the mental representations of the compared objects. These models derive from Shepard's spatial model of similarity. Objects are represented in a multi-dimensional space and their locations are defined by their dimensional differences [Shepard, 1962].

In his seminal work on generalisation, Shepard [1987] provides a statistical technique in the form of Multi-Dimensional Scaling (MDS) to derive locations of objects represented in a multi-dimensional space. MDS can be used to derive some potential spatial representations of objects from proximity data (similarity between pairs of objects). Based on these spatial representations of objects, Shepard derived the *universal law of generalisation* which demonstrates that various kinds of stimuli (e.g., Morse code signals, shapes, sounds) have the same lawful relationship between distance (in an underlined MDS) and perceive similarity measures (in terms of confusability) – the similarity between two stimuli was defined as an exponentially decaying function of their distance[1].

By demonstrating a negative exponential relationship between similarity and generalisation, Shepard established the first sound model of mental representation on which cognitive sciences will base their studies on similarity. The similarity is in this case

[1] The similarity between two stimuli is here understood as the probability that a response to one stimulus will be generalised to the other [Shepard, 1987]. With $sim(A, B)$ the similarity between two stimuli A, B and $dist(A, B)$ their distance, we obtain the relation $sim(A, B) = e^{-dist(A,B)}$, that is $dist(A, B) = -log\ sim(A, B)$, a form of entropy.

assumed to be inversely proportional to the distance separating the perceptual representations of the compared stimuli [Ashby and Perrin, 1988]. Similarity defined as a function of distance is therefore constrained to the axiomatic properties of distance[1].

A large number of geometric models have been proposed. They have long been among the most popular in cognitive sciences. However, despite their intuitive nature and large popularity, geometric models have been subject to intense criticism due to the constraints defined by the distance axioms. Indeed, several empirical analyses have questioned and challenged the validity of the geometric framework (i.e., both the model and the notion of psychological distance), by underlying inconsistencies with human appreciation of similarity, e.g., violation of the symmetry, triangle inequality and identity of the indiscernibles, e.g., [Tversky, 1977; Tversky and Gati, 1982; Tversky and Itamar, 1978][2].

1.4.2 Feature models

To respond to the limitation of the geometric models, Tversky [1977] proposes the feature model in which evaluated objects are manipulated through sets of features. A feature "*describes any property, characteristic, or aspect of objects that are relevant to the task under study*" [Tversky and Gati, 1982]. Therefore, feature models evaluate the similarity of two stimuli according to a feature-matching function F which makes use of their common and distinct features:

$$sim_F(u, v) = F(U \cap V, U \setminus V, V \setminus U) \tag{1.1}$$

The function F is expected to be non-decreasing, i.e., the similarity increases when common (distinct) features are added (removed). Feature models are therefore based on the assumption that F is monotone and that common and distinct features of compared objects are enough for their comparison. In addition, an important aspect is that the feature-matching process is expressed in terms of a matching function as defined in set theory (i.e., binary evaluation).

The similarity of two objects is further derived as a parametrised function of their common and distinct features. Two models, the *contrast model* (sim_{CM}) and the *ratio model* (sim_{RM}) were initially proposed by Tversky [1977]. They can be used to compare

[1] Properties which will be detailed in the following chapter, Section 2.1.3.1

[2] Note that recent contributions propose to answer these inconsistencies by generalising the classical geometric framework through quantum probability [Pothos et al., 2013]. Compared objects are represented in a quantum model in which they are not seen as points or distributions of points, but entire subspaces of potentially very high dimensionality, or probability distributions of these spaces.

two objects u and v represented through sets of features U and V:

$$sim_{CM}(u,v) = \gamma f(U \cap V) - \alpha f(U \setminus V) - \beta f(V \setminus U) \tag{1.2}$$

$$sim_{RM}(u,v) = \frac{f(U \cap V)}{\alpha f(U \setminus V) + \beta f(V \setminus U) + f(U \cap V)} \tag{1.3}$$

The symmetry of the measures produced by the two models can be tuned according to the parameters α and β. This enables the design of asymmetric measures. In addition, one of the major constructs of the feature model is the function f which is used to capture the salience of a (set of) feature(s). The salience of a feature is defined as a notion of specificity: "*the salience of a stimulus includes intensity, frequency, familiarity, good form, and informational content*" [Tversky, 1977]. Therefore, the operators $\cap, \cup$ and $\setminus$ are based on feature matching (F) and the function f evaluates the contribution of the common or distinct features (distinguished by previous operators) to estimate the similarity[1].

1.4.3 Structural alignment models

Structural models are based on the assumption that objects are represented by structured representations. Indeed, a strong criticism of the feature model was that (features of) compared objects are considered to be unstructured, contrary to evidence suggesting that perceptual representations are well characterised by hierarchical systems of relationships, e.g., [Gentner and Markman, 1994; Markman and Gentner, 1993].

Structural alignment models are structure mapping models in which the similarity is estimated using matching functions which will evaluate the correspondence between the compared elements [Gentner and Markman, 1994; Markman and Gentner, 1993]. Here, the process of similarity assessment is expected to involve a structural alignment between two mental representations in order to distinguish correspondences. Therefore, the greater the number of correspondences, the more similar the objects will be considered. In some cases, the similarity is estimated in an equivalent manner to analogical mapping [Markman and Gentner, 1990] and similarity is expected to involve mapping between both features and relationships.

[1] As an example, the notion of the salience associated to a feature implicitly defines the possibility of designing measures which do not respect the identity of the indiscernibles, i.e. which enable non-maximal self-similarity.

Another example of a structural model was proposed by Goldstone [1994a, 1996]. The authors proposed to model similarity as an interactive activation and mapping model using connectionism activation networks based on mappings between representations.

1.4.4 Tranformational models

Transformational models assume that similarity is defined by the transformational distance between mental representations [Hahn et al., 2003]. The similarity is framed in *representational distortion* [Chater and Hahn, 1997] and is expected to be assessed based on the analysis of the modifications required to transform one representation to another. The similarity, which can be explained in terms of the Kolmogorov complexity theory [Li and Vitányi, 1993], is therefore regarded as a decreasing function of transformational complexity [Hahn et al., 2003].

1.4.5 Unification of cognitive models of similarity

Several studies highlighted links and deep parallels between the various cognitive models. Tenenbaum and Griffiths [2001] propose a unification of spatial, feature-based and structure-based models through a framework relying on the generalisation of Bayesian inference (see Gentner [2001] for criticisms). Alternatively, Hahn [2011] proposes an interpretation of the models in which the transformational model is presented as a generalisation of the spatial, feature and structure-based models.

In this section, we have briefly presented several cognitive models which have been proposed to explain and study (human) appreciation of similarity. These models are characterised by particular interpretations and assumptions on the way knowledge is mentally represented and processed. The fundamental differences between the models also rely on their conceptual approach used to explain similarity assessment and their mathematical properties, e.g., symmetry, triangle inequality... Nevertheless, despite these strong differences, several meaningful initiatives have been initiated in order to unify the cognitive models. To this end, researchers have proposed to develop frameworks which generalise existing models of similarity.

1.5 Objectives and outlines of the thesis

Taking into consideration the critical importance of similarity and dissimilarity measures for information retrieval, approximate reasoning, learning techniques, or more generally for any treatments in which imprecise search is required, this thesis proposes an in-depth study of knowledge-based semantic measures. These semantic measures can be used to compare concepts or instances semantically characterised in ontologies. As we have seen, they are central for the design of most knowledge-based systems or expert systems which take advantage of ontologies. They also play an important role for the communities involved in the definition and practical adoption of the Semantic Web and Linked Data paradigms. In this context, we propose to answer specific needs related to both the study and practical use of knowledge-based semantic measures.

Our first aim is to provide a wide overview of the interdisciplinary field related to semantic measures. By studying and analysing numerous contributions made by different communities, we propose to extract lessons in the aim of highlighting important research perspectives for this domain. This analysis will help us to define **the main objective of this thesis: to develop theoretical and software tools dedicated to knowledge-based semantic measures**. It also warns us about the breadth of this field of study; due to which we decided to mainly restrict the technical discussions presented in this manuscript to semantic measures which rely on semantic graphs, i.e., a specific type of commonly used network-based ontologies introduced in Section 1.2. Therefore, although this work covers the use of any ontologies expressed as a semantic graph[1], we will not cover, as such, the comparison of complex logic-based entity descriptions.

In this manuscript, we propose **an in-depth study on the theoretical basis of semantic measures**. This will help us to better understand the large diversity of measures proposed in the literature over recent decades. Our strategy has been to focus on the unification of numerous semantic measures. To this end, we propose **a general theoretical framework** which enables the breakdown of measures through parametric abstract formulas. As we will see, this theoretical tool opens interesting perspectives to express and study knowledge-based semantic measures. We will, for instance, highlight how the framework has been used to: (i) better understand the relationships between numerous existing proposals, (ii) characterise central elements of semantic measures, (iii) identify potential room for improvement of measures, and (iv) distinguish best suited context-specific configurations of semantic measures.

We will also examine **a general approach to define semantic measures in the aim of comparing instances described in a semantic graph**. In addition, we will

[1] A discussion related to the mapping of ontologies to semantic graphs is proposed in Appendix A.2.

also explore an approach dedicated to the **comparison of concepts characterised by several semantic graphs**. We are convinced that strong links can be established between these two problems and the fields of ontology alignment/matching and instance matching. However, this manuscript does not cover or propose any approach related to these fields – our contributions are instead anchored on the study of inexact search techniques for the comparison of entities defined in similar or different ontologies.

Nowadays, both the growing number of ontologies available and the development of large semantic graphs composed of millions of facts[1] challenge semantic measure designers. Based on the in-depth characterisation of semantic measures carried through the proposed framework, **several optimisation techniques** are proposed in the aim of improving measure accuracy and reducing their computational complexity. Nevertheless, this thesis doesn't aim to provide an extensive study of the algorithmic complexity of semantic measures.

Due to their interdisciplinary nature, semantic measures are generally defined considering ontologies which are not expressed using Semantic Web standardised technologies. Since RDF(S) and OWL are cornerstones of the Semantic Web, this work will, as much as possible, consider the definition of semantic measures w.r.t ontologies defined using those standards. To this aim, existing limits and considerations associated to the use of semantic measures on RDF(S) ontologies will be underlined. This effort is essential to make concrete implementations of semantic measures possible.

The practical use of semantic measures also played a particularly important role in our work. This is justified by two important aspects of this field of study.

First, the communities studying the Semantic Web and Linked Data, as well as knowledge-based system designers are extremely committed to demonstrate the feasibility of their proposals through concrete implementations. This particular aspect is critical for investors and companies who (mostly) agreed on the theoretical soundness of the paradigms but are now waiting for their full capabilities to be unearthed. Practical applications are also of major importance for the technological transfer from academic institutions to businesses. This aspect has been central to our work given that this thesis has been carried out in the LGI2P laboratory of the Engineering school École des mines d'Alès, a laboratory focusing mainly on applied sciences[2].

Secondly, the importance given to practical applications of semantic measures is also motivated by the fact that most knowledge-based systems and evaluations of semantic

[1] E.g., DBpedia, Freebase.

[2] The school is associated to Innovup (`http://www.innovup.com` – french website), an incubator supporting IT entrepreneurs.

measures are governed by empirical analyses. Indeed, as we will see, several benchmarks have been developed to assess performance of semantic measures according to context-specific expectations. Therefore, large analysis of semantic measures can only be supported through **efficient, generic and open-source software solutions** – solutions that were not available when this thesis was initiated.

1.6 Chapter summaries

This manuscript is structured as follows:

- Chapter 2 defines the notion of semantic measures and proposes an overview of the different types of measures which have been defined in the literature. This chapter is therefore dedicated to an overview of the broad diversity of semantic measures. It will help us to clearly classify the measures, to characterise important notions on which our work relies, and to clearly define the scope of our contributions.
- Chapter 3 is dedicated to the specific state-of-the-art related to knowledge-based semantic measures. This chapter first presents a broad overview of the different measures which have been proposed to compare entities defined in a semantic graph. Next, it presents a technical and detailed state-of-the-art dedicated to knowledge-based semantic similarity measures defined for the comparison of concepts defined in a semantic graph.
- Chapter 4 focuses on technical aspects relative to knowledge-based semantic similarity measures. Based on an in-depth analysis of the core elements of similarity measures, and on related contributions on abstract expression of measures, we unify a large diversity of measures through a theoretical framework.
- Chapter 5 presents several use cases highlighting the theoretical and practical perspectives opened up by the aforementioned theoretical framework. At the light of the framework we propose: (i) a theoretical analysis of semantic measures, (ii) an empirical analysis of a particular family of measures, (iii) a study of robustness of semantic measures. The biomedical domain is considered for practical use case scenario.
- Chapter 6 introduces semantic relatedness measures for comparing instances which are semantically characterised in semantic graphs (RDF graphs). This proposal corresponds to the definition of a new canonical form which can be used for highly expressive characterisation of instances described in semantic graphs. Based on this contribution, we further define a semi-supervised content-based recommendation system.

- Chapter 7 presents two algorithmic contributions related to semantic measures. We present algorithms to speed-up computation of a specific type of semantic measures. In addition, we define a semantic similarity measure which can be used to compare concepts through the use of several (non-disjoint) taxonomies. Experimental studies and validation are performed in a use case relative to the biomedical domain.
- Chapter 8 introduces the Semantic Measures Library and toolkit: robust open-source software solutions dedicated to the computation, development and analysis of semantic measures. They have been developed, distributed and maintained during this thesis.

2

The notion of semantic measures

Contents

Abstract

This chapter introduces the notion of semantic measures. It starts by presenting their practical usages in different application contexts. Next, we expose general definitions associated to the notion, positioning with regard to related contributions in Mathematics, and we propose a detailed study of the semantics associated to semantic measures. This latter point will help us to better capture the meaning of semantic measures (results). This is done by defining the terminology classically found in the literature, i.e., semantic similarity/proximity/relatedness/distance/dissimilarity/etc. and by proposing an organisation of the notions commonly used. In a second step, to better understand the characteristics of these measures, we distinguish several central aspects of measures which can be used to categorising the large diversity of measure proposals. As a result, a general classification of the variety of semantic measures defined in the literature is presented.

2.1 From usages towards formalisation

Semantic measures are widely used to compare units of language (e.g., terms, sentences, documents), concepts or semantically characterised instances, according to information supporting their meaning[1] [Harispe et al., 2013c]. Otherwise stated, semantic measures can be used to estimate their *semantic likeness*, i.e., the strength of the semantic relationship which links pairs of the aforementioned *semantic elements*. As we will see, the broad notion of semantic likeness of a pair of semantic elements can in some cases be understood intuitively as the probability of a mental activation of one element when the other is discussed (e.g., `Sand` often brings to mind `Beach`). For clarity, let us note that the notion of semantic measure is not framed in the rigorous mathematical definition of measure. It encompasses any theoretical tool or function which can be used to summarise, using a (numerical) value, the result of the comparison of two semantic elements[2]. In this case, the comparison is assumed to be supported by the analysis of *semantic evidence*.

A large diversity of measures exists to estimate the similarity or the difference/distance between specific mathematical objects (e.g., vectors, matrices, graphs, [fuzzy] sets), data structures (e.g., lists, objects) and data types (e.g., numbers, strings, dates). The main particularity of semantic measures compared to traditional similarity or distance functions relies on two aspects: (i) they are dedicated to the comparison of semantic elements, and (ii) they are based on the analysis of *semantic proxies* from which semantic evidence can be extracted. This semantic evidence is expected to directly or indirectly characterise the meaning of compared elements. As an example, measures used to compare two words according to their sequence of characters cannot be considered as semantic measures, as only word characters and their ordering is taken into account, not their meaning. Indeed, according to such lexical measures, the two words *car* and *vehicle* would be regarded as *distant* despite their closely related semantics. Semantic measures rely on two broad types of semantic proxies: corpora of texts and ontologies.

The first type of semantic proxy corresponds to unstructured or semi-structured texts (e.g., plain texts, dictionaries). They have been proved to contain informal evidences of the semantic relationship(s) between units of language. Let us consider a simple example. Since it is common to drink *coffee* with *sugar* and nothing particular links *coffee* to *cats*, most will agree that the pair of words *coffee* – *sugar* is more semantically *coherent* than the pair of words *coffee* – *cat*. Interestingly, corpus of texts can be used to

[1] Note that the notion of semantic measure is used in cognitive sciences since the sixties, e.g. [Moss, 1960], its use for referring to mathematical tools used to compare objects based on their meaning go back, to our knowledge, to the end of the eighties, e.g. [Su et al., 1990].

[2] For convenience they will simply be denoted *elements* in the following.

derive the same conclusion. To this end, a semantic measure will take advantage of the fact that the word *coffee* is more likely to co-occur with the word *sugar* than with the word *cat*. Simply stated, it is possible to use observations regarding the distribution of words in a corpus in order to estimate the strength of their semantic relationship, e.g., based on the assumption that semantically related words tend to co-occur.

The second type of semantic proxy from which semantic evidence can be extracted encompasses the large range of existing ontologies (refer to Appendix A). From structured vocabularies to highly formal ontologies, these proxies are structured and model, in an explicit manner, knowledge about the elements they define. As an example, in an ontology defining the concepts `Coffee` and `Sugar`, a specific relationship will probably explicitly define the link between the two concepts, e.g., that `Coffee canBeDrunkWith Sugar`. Semantic measures based on knowledge analysis rely on techniques which take advantage of network-based (e.g., thesaurus, taxonomies), or logic-based ontologies to extract semantic evidence on which the comparison will be based.

From gene analysis to recommendation systems, semantic measures have recently been found to cover a broad field of applications and are now essential metrics for leverage data mining, data analysis, classification, knowledge extraction, textual processing or even information retrieval based on text corpora or ontologies. They play an essential role in numerous treatments requiring the analysis of the meaning of compared elements (i.e., semantics). In this context, the study of semantic measures has always been an interdisciplinary effort. Communities of Psychology, Cognitive Sciences, Linguistics, Natural Language Processing, Semantic Web, and Biomedical informatics being among the most active contributors. Due to the interdisciplinary nature of semantic measures, recent decades have been highly prolific in contributions related to the notion of semantic relatedness, semantic similarity and semantic distance, etc. Yet before introducing the technical aspects required to further introduce semantic measures, we will briefly discuss their large diversity of applications.

2.1.1 Semantic measures in action

Semantic measures are used to solve problems in a broad range of applications and domains. They are therefore essential tools for the design of numerous algorithms and treatments in which semantics matters. In this section, we present diverse practical applications involving semantic measures. Three domains of application are considered in particular: (i) Natural Language Processing, (ii) Knowledge Engineering/Semantic Web and Linked Data, and (iii) Biomedical informatics and Bioinformatics. Since they are transversal, additional applications related to information retrieval and clustering

are also briefly considered. The list of the applications of semantic measures presented in this section is far from being exhaustive. An extensive classification of contributions related to semantic measures is proposed in the extended version of the state-of-the-art presented in this manuscript [Harispe et al., 2013c]. This classification underlines the broad range of applications of semantic measures and highlights the large number of communities involved.

2.1.1.1 Natural Language Processing

Linguists have, quite naturally, been among the first to study semantic measures in the aim of comparing units of language (e.g., words, sentences, paragraphs, documents). The estimation of word/concept relatedness plays an important role in detecting paraphrasing, e.g., duplicate content and plagiarism [Fernando and Stevenson, 2008], in generating thesauri or texts [Iordanskaja et al., 1991], in summarising texts [Kozima, 1993], in identifying discourse structure, and in designing question answering systems [Bulskov et al., 2002; Freitas et al., 2011; Wang et al., 2012a] to mention a few. The effectiveness of semantic measures to resolve both syntactic and semantic ambiguities has also been demonstrated on several occasions, e.g., [Patwardhan, 2003; Resnik, 1999; Sussna, 1993].

Several surveys related to the usage of semantic measures and to the techniques used for their design for natural language processing can be found in [Curran, 2004; Dinu, 2011; Mohammad and Hirst, 2012a; Panchenko, 2013; Sahlgren, 2008; Weeds, 2003].

2.1.1.2 Knowledge engineering, Semantic Web and Linked Data

In this field, semantic measures can be used as part of processes aiming to integrate heterogeneous ontologies (refer to ontology alignment and instance matching) [Euzenat and Shvaiko, 2013]; they are used to find similar/duplicate entities defined in different ontologies. Applications to provide inexact search capabilities over ontologies or to improve classical information retrieval techniques have also been proposed, e.g., [Hliaoutakis, 2005; Hliaoutakis et al., 2006; Kiefer et al., 2007; Pirró, 2012; Sy et al., 2012; Varelas et al., 2005]. In this context, semantic measures have also been successfully applied to learning tasks using Semantic Web technologies [D'Amato, 2007]. Their benefits for designing recommendation systems based on the Linked Data paradigm have also been stressed, e.g., [Passant, 2010].

2.1.1.3 Biomedical Informatics & Bioinformatics

A large number of semantic measures have been defined in biomedical informatics and bioinformatics. In these domains, semantic measures are commonly used to take advantage of biomedical ontologies in order to study various types of instances which have been semantically characterised (genes, proteins, drugs, diseases, phenotypes)[1]. Several surveys related to the usage of semantic measures underline the diversity of their applications in the biomedical domain [Guzzi et al., 2012; Pedersen et al., 2007; Pesquita et al., 2009a]. As an illustration, here we focus on applications related to studies on the Gene Ontology (GO) [Ashburner et al., 2000].

The GO is the preferred example with which to highlight the large adoption of ontologies in biology[2]; it is extensively used to conceptually annotate gene (products) on the basis of experimental observations or automatic inferences. A gene is classically annotated by a set of concepts structured in the GO. These annotations formally characterise genes regarding their molecular functions, their cellular location and the biological processes in which they are involved. Thanks to semantic measures, these annotations make the automatic comparison of genes possible, not on the basis of particular gene properties (e.g. nucleotidic/proteic sequence, structural similarity, gene expression), but rather on the analysis of biological aspects which are formalised by the GO. Genes can be further analysed by considering their representation in a semantic space expressing our current understanding of particular aspects of biology. In such cases, conceptual annotations bridge the gap between global knowledge of biology defined in the GO (e.g., organisation of molecular functions) and fine-grained understanding of specific instances (e.g., the specific role of a gene at molecular level). In this context, semantic measures enable computers to take advantage of this knowledge to analyse genes and therefore open interesting perspectives for inferencing new knowledge.

As an example, various studies have highlighted the relevance of semantic measures to assess the functional similarity of genes [Du et al., 2009; Wang et al., 2007], to build gene clusters [Sheehan et al., 2008], to validate and to study protein-protein interactions [Xu et al., 2008], to analyse gene expression [Xu et al., 2009], to evaluate gene set coherence [Diaz-Diaz and Aguilar-Ruiz, 2011] or to recommend gene annotations [Couto et al., 2006], among others. A survey dedicated to semantic measures applied to the GO can be found in [Guzzi et al., 2012].

[1] Biology and biomedicine are heavy users of ontologies, e.g., BioPortal, a portal dedicated to ontologies related to biology and the biomedical domain, references hundreds of ontologies [Whetzel et al., 2011].

[2] More than 11k citations between 2000 and 2013!

2.1.1.4 Other applications

Information Retrieval (IR) uses semantic measures to overcome the limitations of techniques based on plain lexicographic term matching, i.e., simple IR models consider that a document is relevant according to a query only if the terms specified in the query occur in the document. Semantic measures enable the meaning of words to be taken into account by going over syntactic search. They can therefore be used to improve classic models, e.g., synonyms will no longer be considered as totally different words. As an example, semantic measures have been successfully used in the design of ontology-based information retrieval systems and for query expansion, e.g., [Baziz et al., 2007; Hliaoutakis, 2005; Hliaoutakis et al., 2006; Saruladha et al., 2010b; Sy et al., 2012; Varelas et al., 2005].

An important aspect is that semantic measures based on ontologies allow for the analysis and querying of non-textual resources and therefore do not restrict IR techniques in text analysis, e.g., genes annotated by concepts can be queried [Sy et al., 2012].

GeoInformatics actively contributes to the study of semantic measures. In this domain, measures have for instance been used to compute the similarity between locations according to semantic characterisations of their geographic features [Janowicz et al., 2011], e.g. estimating the semantic similarity of *tags* defined in the OpenStreetMap Semantic Network[1] [Ballatore et al., 2012]. Readers interested in the application of semantic measures in this field may also refer to the various references proposed in [Harispe et al., 2013c], e.g.[Akoka et al., 2005; Andrea Rodríguez and Egenhofer, 2004; Anna, 2008; Janowicz, 2006; Janowicz et al., 2008; Rodríguez et al., 2005].

2.1.2 Semantic measures: definitions

2.1.2.1 Generalities

The goal of semantic measures is easy to understand – they aim to capture the strength of the semantic interaction between semantic elements (e.g., words, concepts) based on their meaning. Are the words *car* and *auto* more semantically related than the words *car* and *mountain*? Most people would agree that they are. This has been proved in multiple experiments, inter-human agreement on semantic similarity ratings is high, e.g. [Miller and Charles, 1991; Pakhomov et al., 2010; Rubenstein and Goodenough, 1965][2].

[1] `http://wiki.openstreetmap.org/wiki/OSM_Semantic_Network`

[2] As an example, considering three benchmarks, [Schwartz and Gomez, 2011] observed 73% to 89% human inter-agreement between scores of semantic similarity associated to pairs of words.

Appreciation of similarity is obviously subject to multiple factors. Our personal background is an example of such a factor, e.g., elderly people and teenagers will probably not associate the same score of semantic similarity between the two concepts `Phone` and `Computer`[1]. However, most of the time, a consensus regarding the estimation of the strength of the semantic link between elements can be reached [Miller and Charles, 1991] – this is what makes the notion of semantic measures intuitive and fascinating[2].

The majority of semantic measures try to mimic the human capacity to assess the degree of relatedness between things according to semantic evidence. However, strictly speaking, semantic measures evaluate the strength of the semantic interactions between things according to the analysis of semantic proxies (texts, ontologies), nothing further. Therefore, not all measures aim at mimicking human appreciation of similarity. In some cases, designers of semantic measures only aim to compare elements according to the information defined in a semantic proxy, no matter if the results produced by the measure correlate with human appreciation of semantic similarity/relatedness. This is, for instance, often the case in the design of semantic measures based on domain-specific ontologies. In these cases, the ontology can be associated to our brain and the semantic measure can be regarded as our capacity to take advantage of our knowledge to compare things. The aim, therefore, is to be coherent with the knowledge expressed in the considered semantic proxy, without regards to the coherence of the modelled knowledge. As an example, a semantic measure based on an ontology built by animal experts would not consider the two concepts `Sloth` and `Monkey` to be similar, even if most people think sloths are monkeys. Given that semantic measures aim at comparing things according to their meaning captured from semantic evidence, it is difficult to further define the notion of semantic measures without defining the concepts of *Meaning* and *Semantics*.

Though risking the disappointment of the reader, this section will not face the challenge of demystifying the notion of Meaning. As stressed by Sahlgren [2006] "*Some 2000 years of philosophical controversy should warn us to steer well clear of such pursuits*". The reader can refer to the various theories proposed by linguists and philosophers. In this contribution, we only consider that we are dealing with the notion of semantic meaning proposed by linguists: how meaning is conveyed through signs or language. Regarding the notion of semantics, it can be defined as the meaning or interpretation of any lexical

[1] Given that nowadays smartphones are kinds of computers and very different to the first communication devices.

[2] Despite some hesitations and interrogations regarding the notion of (semantic) similarity, it is commonly admitted that the notions related to similarity make sense. However, there are numerous examples of authors who question their relevance, e.g. "*Similarity, ever ready to solve philosophical problems and overcome obstacles, is a pretender, an impostor, a quack.*"[Goodman, 1972] or "*More studies need to performed with human subjects in order to discover whether semantic distance actually has any meaning independent of a particular person, and how to use semantic distance in a meaningful way*" [Delugach, 1993], see also [Goldstone, 1994b; Hahn and Ramscar, 2001; Murphy and Medin, 1985].

units, linguistic expressions or instances which are semantically characterised according to a specific context.

> **Definition**: *Semantic Measures* are mathematical tools used to estimate the strength of the semantic relationship between units of language, concepts or instances, through a (numerical) description obtained according to the comparison of information supporting their meaning.

It is important to stress the diversity of the domain (in a mathematical sense) in which semantic measures can be used. They can be used to drive word-to-word, concept-to-concept, text-to-text or even instance-to-instance comparisons. In this manuscript, when we do not focus on a specific type of measure, we will refer, as much as possible, to any element of the domain of measures through the generic term element. An element can therefore be any unit of language (e.g. word, text), a concept/class, an instance which is semantically characterised in an ontology (e.g., gene products, ideas, locations, persons).

We formally define a semantic measure as a function:

$$\sigma_k : E_k \times E_k \rightarrow \mathbb{R} \tag{2.1}$$

with E_k the set of elements of type $k \in \mathbb{K}$ and $\mathbb{K}$, the various types of elements which can be compared regarding their semantics, e.g., $\mathbb{K}$ =$\{$words, concepts, sentences, texts, webpages, instances annotated by concepts... $\}$.

This expression can be generalised so as to take into account the comparison of different types of elements. This could be interesting to evaluate entailment of texts or to compare words and concepts, among others. However, in this thesis, we restrict our study to the comparison of pairs of elements of the same nature (already a vast subject of research). We stress that semantic measures must implicitly or explicitly take advantage of semantic evidence. As an example, as we have said, measures comparing words through their syntactical similarity cannot be considered as semantic measures; recall that semantics refers to evidence regarding the meaning of compared elements.

The distinction between approaches that can and cannot be assimilated to semantic measures is sometimes narrow; there is no clear boundary distinguishing non-semantics to semantic-augmented approaches, but rather a range of approaches. Some explanations can be found in the difficulty of clearly characterising the notion of meaning. For instance, someone can say that measures used to evaluate lexical distance between words

capture semantic evidence related to the meaning of the words. Indeed, the sequence of characters associated to a word derives from its etymology which is sometimes related to its meaning, e.g., words created through morphology derivation such as *subset* from *set*.

Therefore, the notion of semantic measure is sometimes difficult to distinguish from measures used to compare specific data structures. This fine line can also be explained by the fact that some semantic measures compare elements which are represented through canonical forms corresponding to specific data structures for which specific (non-semantic) similarity measures have been defined. As an example, pure graph similarity measures can be used to compare instances which are semantically characterised through semantic graphs.

In some cases, the semantics of the measure is therefore not captured by the measure used to compare the canonical forms of the compared elements. It is rather the process of mapping an element (e.g., word, concept) from a semantic space (text, ontology) to a specific data structure (e.g., vector, set), which semantically enhances the comparison. This, however, is an interesting paradox, the definition of the rigorous semantics of the notion of semantic measure is hard to define – this is one of the challenges this contribution faces.

2.1.2.2 Semantic relatedness and semantic similarity

Among the various notions associated to semantic measures, this section defines the two central notions of semantic relatedness and semantic similarity, which are among the most commonly referred to in the literature. Several authors have already distinguished them in different communities, e.g., [Pedersen et al., 2007; Resnik, 1999]. Based on these works, we propose the following definitions.

Definition *Semantic relatedness*: the strength of the semantic interactions between two elements with no restrictions on the types of semantic links considered.

Note that compared to the general definition of semantic measure, the notion of interaction used to define semantic relatedness refers to a positive value, i.e. the more two elements interact the more related they will be considered. As an example, compared to semantic relatedness, semantic distance refers to the degree of repulsion between the two compared elements.

Definition *Semantic similarity*: subset of the notion of semantic relatedness only considering taxonomic relationships in the evaluation of the semantic interaction between two elements.

In other words, semantic similarity measures compare elements regarding the constitutive properties they share and those which are specific to them. The two concepts `Tea` and `Cup` are therefore highly related despite the fact that they are not similar: the concept `Tea` refers to a `Drink` and the concept `Cup` refers to a `Vessel`. Thus, the two concepts share few of their constitutive properties. This highlights a potential interpretation of the notion of similarity, which can be understood in term of substitution, i.e., evaluating the implication to substitute the compared elements: `Tea` by `Coffee` or `Tea` by `Cup`.

In some specific cases, communities or linguists will consider a more complex definition of the notion of semantic similarity for words. Indeed, word-to-word semantic similarity is sometimes evaluated not only considering (near-)synonymy, or the lexical relations which can be considered as equivalent to the taxonomic relationships for words, e.g., hyponymy and hypernymy or even troponymy for verbs. Indeed, in some contributions, linguists also consider that the estimation of the semantic similarity of two words must also take into account other lexical relationships, such as antonymy [Mohammad and Hirst, 2012a].

In other cases, the notion of semantic similarity refers to the approach used to compare the elements, not the semantics associated to the results of the measure. As an example, designers of semantic measures relying on ontologies sometimes use the term semantic similarity to denote measures based on a specific type of semantic relatedness which only considers meronymy, e.g., partial ordering of concepts defined by `partWhole` relationships. The semantics associated to the scores of relatedness computed from such restrictions differs from semantic similarity. Nevertheless, as we will see, technically speaking, most approaches defined to compute semantic similarities based on ontologies can be used on any restriction of semantic relatedness considering a type of relationship which is transitive, reflexive and antisymmetric. In this manuscript, for the sake of clarity, we consider that only taxonomic relationships are used to estimate the semantic similarity of compared elements.

Older contributions relative to semantic measures do not stress the difference between the notions of similarity and relatedness. The reader must understand that in the literature, authors sometimes introduce semantic similarity measures which estimate semantic relatedness and *vice versa*. In addition, despite the fact that the distinction

between the two notions is now commonly admitted by most communities, it is still common to observe improper use of both notions.

Extensive terminology refers to the notion of semantic measures and contributions related to the domain often refer to the notions of semantic distance, closeness, nearness or taxonomic distance, etc. The following subsection proposes to clarify the semantics associated to the terminology which is commonly used in the literature.

2.1.2.3 The diversity of types of semantic measures

We have so far introduced the broad notion of semantic measures and have also distinguished the two central notions of semantic relatedness and semantic similarity. Extensive terminology has been used in the literature to refer to the notion of semantic measure. Thus, we here define the meaning of the terms commonly used (the list may not be exhaustive):

- *Semantic relatedness*, sometimes called *proximity*, *closeness* or *nearness*, refers to the notion introduced above.
- *Semantic similarity* has also already been defined. In some cases, the term *taxonomic semantic similarity* is used to stress the fact that only taxonomic relationships are used to estimate the similarity.
- *Semantic distance* is generally considered as the inverse of semantic relatedness, and all semantic interactions between the compared elements are considered. These measures respect (for the most part) the mathematical properties of distances which will be introduced later. Semantic distance is also sometimes denoted as *farness*.
- *Semantic dissimilarity* is understood as the inverse of semantic similarity.
- *Taxonomic distance* also corresponds to the semantics associated to the notion of dissimilarity. However, these measures are expected to respect the properties of distances.

Figure 2.1 presents a graph in which the various notions related to semantic measures are (informally) structured through semantic relationships. Most of the time, the notion considered to be the inverse of semantic relatedness is denoted as semantic distance, whether or not the measure respects the mathematical properties characterising a distance. Therefore, for the purpose of organising the different notions, we introduce the term *semantic unrelatedness* to denote the set of measures whose semantics is the inverse to the one carried by semantic relatedness measures, without necessarily respecting

the properties of a distance. To our knowledge, this notion has never been used in the literature.

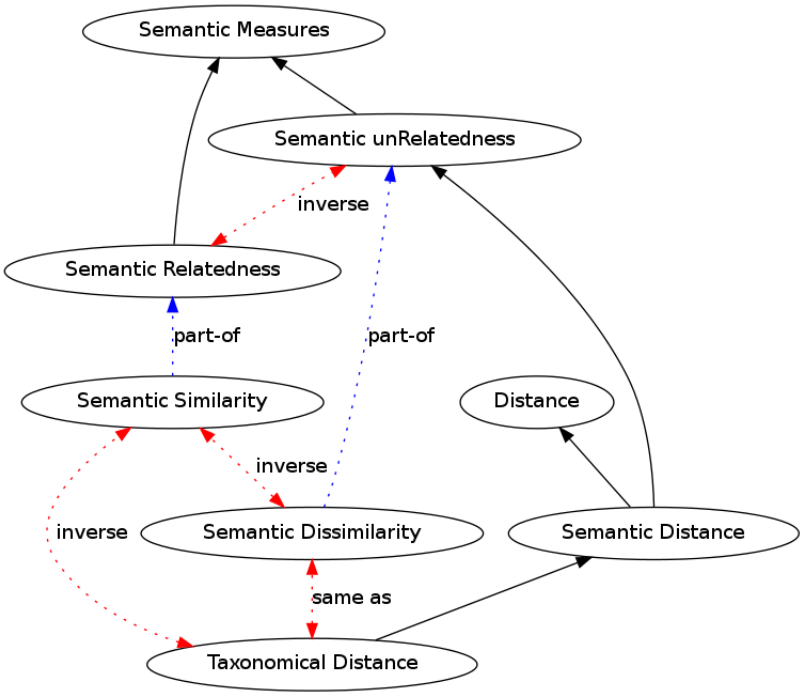

FIGURE 2.1: Informal semantic graph of the terminology related to semantic measures. It structures various types of semantics which have been associated to semantic measures in the literature. Black (plain) relationships correspond to taxonomic relationships, inverse relationships refer to the semantic interpretation associated to the score of the measure, e.g., semantic similarity and dissimilarity measures have inverse semantic interpretations

2.1.3 From distance and similarities to semantic measures

Are semantic measures mathematical measures? What are the specific properties of a distance or a similarity measure? Do semantic similarity measures correspond to similarity measures in the way mathematicians understand them? As we have seen in Section 2.1.2, contributions related to semantic measures do not for the most part rely on formal definitions of the notion of measure or distance. Indeed, generally, contributions related to semantic measures rely on the commonly admitted and intuitive expectations regarding these notions, i.e. similarity (resp. distance) must be higher (resp. lower) the more (resp. less) the two compared elements share commonness[1]. However, the notions of measure and distance have been rigorously defined in mathematics through specific axioms from which particular properties derive. These notions have been expressed for

[1] The works of [Blanchard et al., 2008; D'Amato, 2007] are among the exceptions.

well-defined objects (element domain). Several contributions rely on these axiomatic definitions and interesting results have been demonstrated according to them. This section briefly introduces the mathematical background relative to the notions of distance and similarity. It will help us to rigorously define and better characterise semantic measures in mathematical terms; it is a prerequisite to clarify the fuzzy terminology commonly used in studies related to semantic measures.

For more information on the definition of measures, distance and similarity, the reader can refer to: (i) the seminal work of Deza and Deza [2013] – *Encyclopedia of Distances*, (ii) the work of [Hagedoorn, 2000, Chapter 2] – *A theory of similarity measures*, and (iii) the definitions proposed by D'Amato [2007]. Most of the definitions proposed in this section have been formulated based on these contributions. Therefore, for convenience, we will not systematically refer to them. In addition, contrary to most of the definitions presented in these works, here we focus on highlighting the semantics of the various definitions according to the terminology introduced in Section 2.1.2.

2.1.3.1 Distance and similarity in Mathematics

For the definitions presented hereafter, based on D'Amato [2007], we consider a set D which defines the elements of the domain we want to compare and a totally ordered set $(V, \preceq)$. We also consider the element min_V as the element of V such as $\forall v \in V : min_V \preceq v$, $max_V \in V$ such as $\forall v \in V : v \preceq max_V$; and $0_V \in V$ such as $min_V \preceq 0_V \preceq max_V$.[1]

Definition *Distance*: a function $dist : D \times D \rightarrow V$ is a distance on D if, $\forall x, y \in D$, the function is:

- Non-negative, $dist(x, y) \succeq 0_V$.
- Symmetric, $dist(x, y) = dist(y, x)$.
- Reflexive $dist(x, x) = 0_V$ and $\forall y \in D \wedge y \neq x : dist(x, x) \prec dist(x, y)$.

To be considered as a distance in a metric space, the distance must additionally respect two properties:

- The *identity of indiscernibles* also known as *strictness* property, *minimality* or *self-identity*, that is $dist(x, y) = 0_V$ iff $x = y$.
- The *triangle inequality*, when $V \subseteq \mathbb{R}$, the distance between two *points* must be the shortest distance along any path: $dist(x, y) \leq dist(x, z) + dist(z, y)$.

[1] E.g. different definitions of V could be $V = \mathbb{R}, V = [0, 1], V = \{$very low, low, medium, high, very high$\}$.

Despite the fact that some formal definitions of similarity have been proposed, e.g., [Deza and Deza, 2013; Hagedoorn, 2000], contrary to the notion of distance, there is no axiomatic definition of similarity that sets the standard; the notion appears in different fields of mathematics, e.g., figures with the same shape are denoted similar (in geometry), similar matrices are expected to have the same eigenvalues, etc. In this manuscript, we consider the following definition.

Definition *Similarity*: a function $sim : D \times D \rightarrow V$ is a similarity on D if, for all $x, y \in D$, the function sim is non-negative ($sim(x, y) \succeq 0_V$), symmetric ($sim(x, y) = sim(y, x)$) and reflexive, i.e., $sim(x, x) = max_V$ and $\forall x, y \in D \wedge y \neq x : sim(x, x) \succ sim(x, y)$.

Definition *Normalised function*: any function f on D (e.g. similarity, distance) with values in $[0, 1]$.

Notice that a normalised similarity *sim* can be transformed into a distance *dist* considering multiple approaches; inversely, a normalised distance can also be converted into a similarity. Some of the approaches used for the transformations are presented in [Deza and Deza, 2013, Chapter 1].

As we have seen, distance and similarity measures are formally defined in mathematics as functions with specific properties. These properties are extensively used to demonstrate results and to develop proofs. However, the benefits of fulfilling some of these properties, e.g., triangle inequality for distance metric, have been subject to debate among researchers. As an example, Jain et al. [1999] stress that the mutual neighbour distance used in clustering tasks does not satisfy the triangle inequality but perform well in practice – to conclude by "*This observation supports the viewpoint that the dissimilarity does not need to be a metric*".

A large number of properties which are not presented in this section have been distinguished to further characterise distance or similarity functions, e.g., see [Deza and Deza, 2013]. These properties are important as specific theoretical proofs require studied functions to fulfil particular properties. However, as we have seen, the definition of semantic measures proposed in the literature is not framed in the mathematical axiomatic definitions of distance or similarity. In some cases, such a distortion among the terminology creates difficulties in bridging the gap between the various communities involved in the study of semantic measures and similarity/distance. As an example, in the *Encyclopedia of distances*, Deza and Deza [2013] do not distinguish the notions of distance and

dissimilarity, which is the case in the literature related to semantic measures (refer to Section 2.1.2). In this context, the following section defines the terminology commonly adopted in the study of semantic measures w.r.t the mathematical properties already introduced.

2.1.3.2 Flexibility of semantic measures

Notice that we haven't introduced the precise and technical mathematical definition of a measure proposed by *measure* theory. This can be disturbing considering that this manuscript extensively refers to the notion of semantic measure. The notion of measure we use is indeed not framed in the rigorous mathematical definition of *measure*. It refers to any "*measuring instruments*" which can be used to "*assess the importance, effect, or value of (something)*" [Oxford Dict., 2012] – in our case, any functions answering the definitions of semantic distance/relatedness/similarity/etc. proposed in Section 2.1.2.

Various communities have used the concepts of similarity or distance without considering the rigorous axiomatic definitions proposed in mathematics but rather using their broad intuitive meanings[1]. To be in accordance with most contributions related to semantic measures, and to facilitate the reading of this manuscript, we will not limit ourselves to the mathematical definitions of distance and similarity.

The literature related to semantic measures generally refers to a semantic distance as any (non-negative) function, designed to capture the inverse of the strength of the semantic interactions linking two elements. Such functions must respect that: the higher the strength of the semantic interactions between two elements, the lower their distance. The axiomatic definition of a distance (metric) may not be respected. A semantic distance is, most of the time, what we define as a function estimating semantic unrelatedness. However, to be in accordance with the literature, we will use the term semantic distance to refer to *any* function designed to capture semantic unrelatedness. We will explicitly specify that the function respects (or does not respect) the axiomatic definition of a distance (metric) when required.

Semantic relatedness measures are functions which are associated to an inverse semantics of the one associated to semantic unrelatedness: the higher the strength of the semantic interactions between two elements, the higher the function will estimate their semantic relatedness.

[1] As we have seen, researchers in cognitive science have demonstrated that human expectations regarding (semantic) distance challenges the mathematical axiomatic definition of distance. Thus, the communities involved in the definition of semantic measures mainly consider a common vision of these notions without always clearly defining their mathematical properties.

Properties	**Definitions**
Non-negative	$dist(x, y) \succeq 0_V$
Symmetric	$dist(x, y) = dist(y, x)$
Reflexive	$dist(x, x) = 0_V$
Normalised	$V = [0, 1]$
Identity of indiscernibles	$dist(x, y) = 0_V$ *iff* $x = y$
Triangle inequality ($V \subseteq \mathbb{R}$)	$dist(x, y) \leq dist(x, z) + dist(z, y)$

TABLE 2.1: Properties which can be used to characterise any function which aims to estimate the notion of distance between two elements. Refer to the notations introduced page 64.

Properties	**Definitions**
Non-negative	$sim(x, y) \succeq 0_V$
Symmetric	$sim(x, y) = sim(y, x)$
Reflexive	$sim(x, x) = max_V$
Normalised	$V = [0, 1]$
Identity of indiscernibles	$sim(x, y) = max_V$ *iff* $x = y$
Integrity	$sim(x, y) \preceq sim(x, x)$

TABLE 2.2: Properties which can be used to characterise any function which aims to estimate the notion of similarity/relatedness between two elements. Refer to the notations introduced page 64.

The terminology we use (distance, relatedness, similarity) refers to the definitions presented in Section 2.1.2. To be clear, the terminology refers to the semantics of the functions, not their mathematical properties. However, we further consider that semantic measures must be characterised through mathematical properties. Table 2.1 and Table 2.2 summarise some of the properties which can be used to formally characterise any function designed in order to capture the intuitive notions of semantic distance and relatedness/similarity. These properties will be used in the manuscript to characterise some of the measures that we will consider. They are essential to further understand the semantics associated to the measures and to distinguish semantic measures which are adapted to specific contexts and usage.

2.2 Classification of semantic measures

We have seen that various mathematical properties can be used to characterise technical aspects of semantic measures. This section distinguishes other general aspects which may be interesting to classify semantic measures. They will be used to introduce the large diversity of approaches proposed in the literature. First we present some of the general aspects of semantic measures which can be relevant for their classification, and subsequently introduce two general classes of measures.

2.2.1 How to classify semantic measures

The classification of semantic measures can be made according to several aspects; we propose to discuss four of them:

- The type of elements that the measure aims to compare.
- The semantic proxies used to extract the semantics required by the measure.
- The semantic evidence and assumptions considered during the comparison.
- The canonical form adopted to represent an element and how to handle it.

2.2.1.1 Types of elements compared: words, concepts, instances...

Semantic measures can be used to compare various types of elements:

- Units of language: words, sentences, paragraphs, documents.
- Concepts/Classes, groups of concepts.
- Semantically characterised instances.

Semantic measures can therefore be classified according to the type of elements they aim to compare.

2.2.1.2 Semantic proxies from which semantics is distilled

Semantic measures require a source of information to compare two semantic elements. It will be used to characterise compared elements and to extract the semantics required by the measure.

Definition *Semantic proxy*: any source of information from which indication of the semantics of the compared elements, which will be used by a semantic measure, can be extracted.

Two broad types of semantic proxies can be distinguished:

- Unstructured or semi-structured texts: text corpora, controlled vocabularies, dictionaries.
- Structured: ontologies, e.g., thesaurus, structured vocabularies, taxonomies.

2.2.1.3 Semantic evidence and considered assumptions

Depending on the semantic proxy used to support the comparison of elements, various types of semantic evidence can be considered. The nature of this evidence conditions the assumptions associated to the measure.

> **Definition** *Semantic evidence*: any clue or indication based on semantic proxy analysis from which, often based on assumptions, a semantic measure will be based.

As an example, considering the measures which rely on text analysis, we have already mentioned that the proximity or relatedness of terms can be assessed considering that pairs of terms which co-occur frequently are more related. In this case, the co-occurrence of words is considered as semantic evidence; its interpretation is governed by the assumption that relatedness of terms is a function of their degree of co-occurrence.

2.2.1.4 Canonical forms used to represent compared elements

The canonical form (representation) chosen to process a specific element can also be used to distinguish the measures defined for comparing a specific type of element. Since a canonical form corresponds to a specific reduction of the element, the degree of granularity with which the element is represented may highly impact the analysis. The selected canonical form is of major importance since it influences the semantics associated to the score produced by a measure, that is to say, how a score must/can be understood/interpreted. This particular aspect is essential when inference must be driven from scores produced by semantic measures.

> A semantic measure is defined to process a given type of element represented through a specific canonical form.

Figure 2.2 presents a partial overview of the landscape of semantic measures which can be used to compare various types of elements (words, concepts, instances...). It

summarizes one of the classifications of semantic measures which can be proposed. As we have seen, measures can first be classified based on the elements they can compare. Based on this aspect, we distinguish two main types of measures:

- Distributional measures used to compare units of language, concepts or instances from text analysis, i.e. unstructured semantic proxies. Distributional measures are generally used to compare words or more generally units of language. However, they can also be used to compare concepts or instances by considering that disambiguation techniques have been used to identify concepts or instance denotation in texts.
- Knowledge-based measures which are designed for comparing entities defined in ontologies, i.e. structured semantic proxies. Knowledge-based measures can also be used to compare units of language, e.g., sentences or texts, for instance by considering that disambiguation techniques have been used for establishing bridges between texts and ontologies.

Hybrid strategies can also be defined mixing both distributional and knowledge-based measures. Nevertheless, in the literature, measures are generally defined for comparing a specific type of elements: units of language or entities defined in an ontology. Therefore, classifying measures based on the elements they compare and the semantic proxy which is used in the analysis, i.e. texts or ontologies (knowledge representation system), helps to distinguish the general types of measures which have been proposed. These measures are based on different semantic evidence and assumptions which are used to capture the semantics of compared elements, e.g. the distributional hypothesis, intentional or extensional evidence expressed into ontologies. Based on these evidence and assumptions, a model is defined for comparing two elements – such a model is generally denoted a semantic measure. Various specific types of approaches have been proposed for distributional and knowledge-based measures, the figure structures several broad categories. Depending on the strategy which is used for defining the measure and the evidence and assumptions which are considered, the semantics of the measure, i.e. the meaning which can be associated to the scores it produces, may vary. Therefore, the measure can be used to estimate, among others, the semantic relatedness or the semantic similarity between the compared elements.

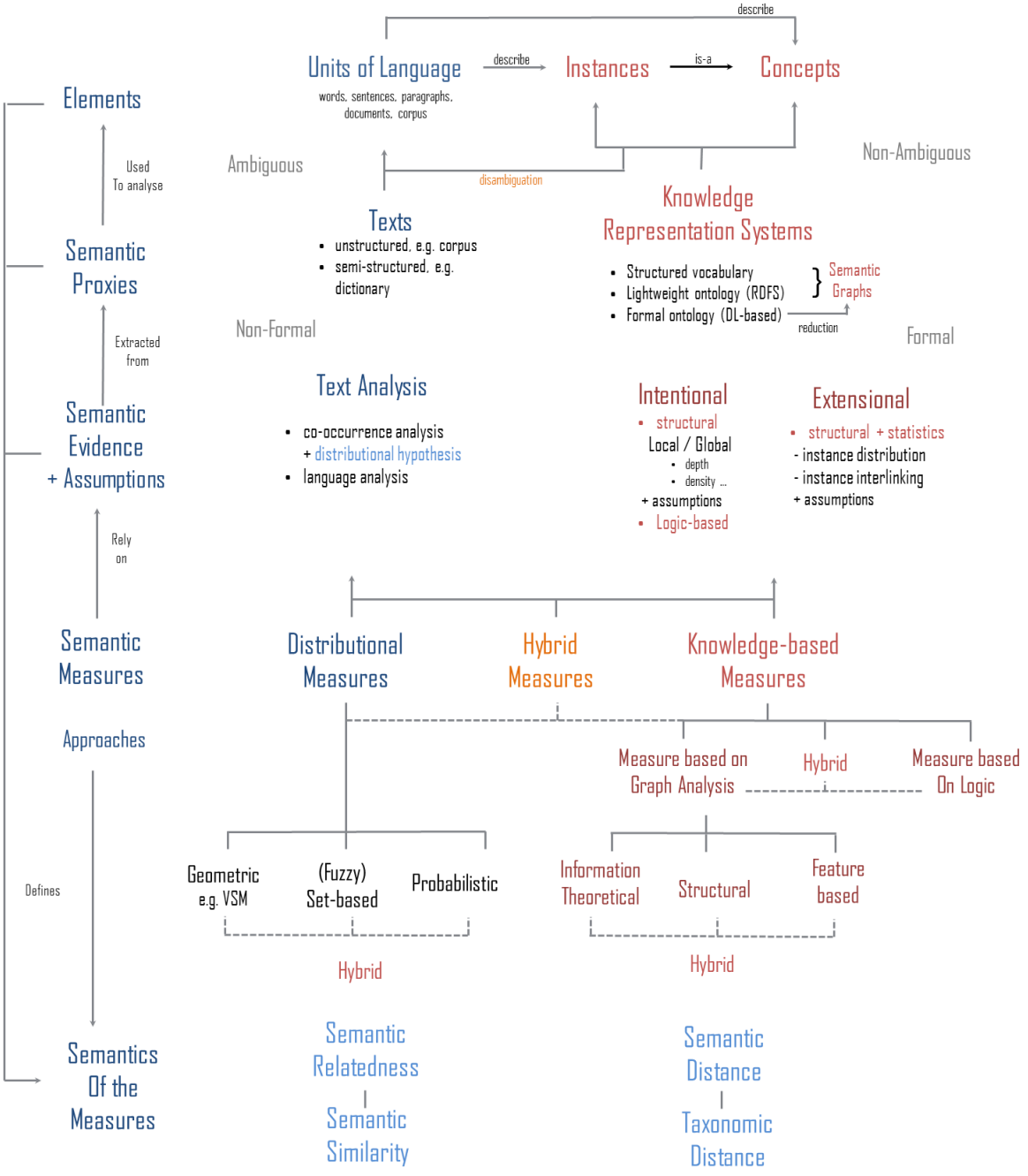

FIGURE 2.2: Partial overview of the landscape of the types of semantic measures which can be used to compare various types of elements (words, concepts, instances...) [Harispe et al., 2013c]

2.2.2 Distributional measures

2.2.2.1 Generalities

- Semantic proxy: unstructured/semi-structured texts.
- Type of elements compared: units of language, i.e., words, sentences, paragraphs, documents.

Distributional measures enable the comparison of units of language through the analysis of unstructured texts. They are mainly used to compare words, sentences or even documents studying the repartition of words in texts (number of occurrences, location in texts)[1]. An introduction to this type of measures for the comparison of a pair of words can be found in [Curran, 2004; Mohammad and Hirst, 2012b].

Several contributions have been proposed to tackle the comparison of pairs of sentences or documents (text-to-text measures) [Mihalcea et al., 2006]. Some of these measures derive from word-to-word semantic measures; others rely on specific strategies based on lexical/ngram overlap analysis, Latent Semantic Analysis extensions [Lintean et al., 2010], or even topic model using Latent Dirichlet Allocation [Blei et al., 2003].

Distributional measures rely on the *distributional hypothesis* which considers that words occurring in similar contexts tend to be semantically close [Harris, 1981]. This hypothesis is one of the main tenets of statistical semantics. It was made popular through the idea belonging to Firth [1957]: "*a word is characterised by the company it keeps*"[2]. Considering that the context associated to a word can be characterised by the words surrounding it, the distributional hypothesis states that words occurring in similar contexts, i.e., often surrounded by the same words, are likely to be semantically similar as "*similar things are being said about both of them*" [Mohammad and Hirst, 2012b]. It is therefore possible to build a distributional profile of a word according to the contexts in which it occurs.

A word is classically represented through the vector space model: a geometric metaphor of meaning in which a word is represented as a point in a multidimensional space modelling the diversity of the vocabulary in use [Sahlgren, 2006]. This model is used to characterise words through their distributional properties in a specific corpus of texts. To this end, words are generally represented through a matrix of co-occurrence – it

[1] In the literature, distributional measures are sometimes defined as a specific type of a more general type of measures, denoted as corpus-based measures [Panchenko and Morozova, 2012]. In this manuscript we consider the most common classification by considering distributional measures as any measure which relies on location and number of occurrences of words in text. There is therefore no need to distinguish them from corpus-based measures.

[2] Also implicitly discussed in [Weaver, 1955] originally written in 1949 [ACL, 2013].

can either be a word-word matrix or more generally a word-context matrix in which the context is any lexical unit (surrounding words, sentences, paragraphs or even documents). Such a characterisation of a word regarding a specific corpus, sometimes denoted as word-space model [Sahlgren, 2006], is analogue to the vector space model which is widely known in Information Retrieval [Salton and McGill, 1983].

Generally, the design of a semantic measure for the comparison of words corresponds to the definition of a function which will assess the similarity of two context vectors. The various distributional measures are therefore mainly distinguished by the:

- Type of context used to build the co-occurrence matrix.
- Frequency weighting (optional). The function used to transform the raw counts associated to each context in order to incorporate frequency and informativeness of the context [Curran, 2004].
- Dimension reduction technique (optional) used to reduce the co-occurrence matrix. This aspect defines the type of co-occurrences which is taken into account (e.g. first order, second order, etc.).
- Vector measure used to assess the similarity/distance of the vectors which represent the words in the co-occurrence matrix. In some cases, vectors will be regarded as (fuzzy) sets.

Several distributional measures have been proposed. Due to a lack of space these measures are not presented in this manuscript but an introduction and references to related surveys can be found in [Harispe et al., 2013c].

2.2.2.2 Advantages and limits of distributional measures

Advantages

- Unsupervised, they can be used to compare the relatedness of words expressed in corpora without prior knowledge regarding their meaning or usage.

Limits

- The words to compare must occur at least a few times on the considered corpus.
- They highly depend on the corpus used. This specific point can also be considered as an advantage as the measure is context-dependent.
- Sense-tagged corpora are generally unavailable [Resnik, 1999; Sánchez and Batet, 2011]. The construction of a representative corpus of texts can be challenging in some usage context, e.g., biomedical studies.

- Difficulties are found when attempting to estimate the relatedness between concepts or instances due to the disambiguation process required prior to the comparison. Distributional measures are mainly designed for the comparison of words. However, some pre-processing and disambiguation techniques can be used to enable concept or instance comparison from text analysis. Nevertheless, their computational complexity is a drawback the majority of the time, making such approaches impracticable to be used with large corpora analysis.
- Difficulty arises on estimating the semantic similarity. Though different observations are nevertheless provided in the literature, it is commonly said that distributional measures can only be used to compare words regarding their semantic relatedness, i.e., co-occurrence can only be seen as evidence of relatedness, e.g., [Batet, 2011a]. However, Mohammad and Hirst [2012b] specifies that similarity can be captured performing specific pre-processing. In any case, capturing the similarity between words from text analysis requires elaborate techniques which are not tractable for large corpora analysis.
- There are difficulties explaining and tracing the semantics of the relatedness. The interpretation of the score is almost only driven by the distributional hypothesis; it is difficult, however, to deeply understand the semantics associated to co-occurrences.

2.2.3 Knowledge-based measures

This section is more detailed than the previous one since our works in this thesis mainly focused on this kind of measures.

2.2.3.1 Generalities

- Semantic proxy: network-based ontologies (e.g., thesaurus, taxonomy, semantic graph), logic-based ontologies.
- Type of elements compared: words/terms, concepts, groups of concepts, semantically characterised instances.

Knowledge-based measures rely on any form of ontologies from which the semantics associated to the compared elements will be extracted. A large diversity of measures have been defined to compare both concepts[1] and instances. Two main types of measures can be distinguished considering the type of ontology which is taken into account:

[1]Note that predicates can also be compared using some measures defined for the comparison of concepts.

- *Measures based on graph analysis*, also denoted as semantic measures framed in the relational setting in [D'Amato, 2007]. They consider ontologies as semantic graphs. They rely on the analysis of the structural properties of the semantic graph and elements are compared studying their interconnections.
- *Measures relying on logic-based semantics* such as description logics. These measures use a higher degree of semantic expressivity; they can take logical constructors into account, and can be used to compare rich descriptions of knowledge, mainly concepts.

Most semantic measures have been defined to compare elements defined in a single ontology. Nevertheless, some semantic measures have also been proposed to compare elements defined in different ontologies. In this section, we mainly consider the measures defined for a single ontology. Semantic measures which have been defined to take advantage of multiple ontologies are briefly presented next.

2.2.3.2 Semantic measures based on graph analysis

Semantic measures based on graph analysis do not take into account logical constructors which can sometimes be used to define the semantics of an ontology. These measures only consider the semantics carried by the semantic relationships (relational setting), e.g., specific treatments can be performed regarding the type of relationship processed. Some properties associated to the relationships defined in the graph can be considered by the measures. The transitivity of the taxonomic relationship will for instance be implicitly or explicitly used in the design of these measures. In other cases, the taxonomy of predicates (i.e., the types of semantic relationships) can also be taken into account.

A large number of approaches have been proposed to express semantic measures using this strategy. Chapter 3 is dedicated to them. Here, we only present a non-technical overview of these measures focusing on those used to compare a pair of concepts. The idea is to give a first insight into this type of semantic measures by presenting, through simple and intuitive examples, the main approaches which have been proposed in the literature.

Semantic measures based on graph analysis are commonly classified into four approaches: (i) Structure-based, (ii) Feature-based, (iii) Information-Theory and (iv) Hybrid.

The structural approach: Semantic measures based on the structural approach compare elements defined in the semantic graph by studying the structure of the graph induced by its relationships. The measures are generally expressed as a function of the strength of the interconnections of the compared elements in the graph. Conceptually, the structural approach corresponds, in some sense, to the design of semantic measures according to the structural model defined in cognitive sciences (refer to Section 1.4). The graph corresponds to a structured space in which compared elements are described.

The first measures based on this approach proposed to compare two concepts according to the length of shortest path linking them in the graph (in terms of number of edges); the shorter the path, higher their semantic relatedness [Rada et al., 1989]. The types of relationships considered in order to distinguish the shortest path define the semantics of the measures, e.g., only the taxonomic relationships will be considered to estimate the semantic similarity. As an example, considering Figure 2.3, the length of the shortest path between the concepts `Computer` and `Tablet` is two. Considering only the taxonomic relationship, the length of the shortest path between the concepts `Computer` and `Rodent` is five. As expected, the concept `Computer` will therefore be considered to be more similar to the concept `Tablet` than to the concept `Rodent`.

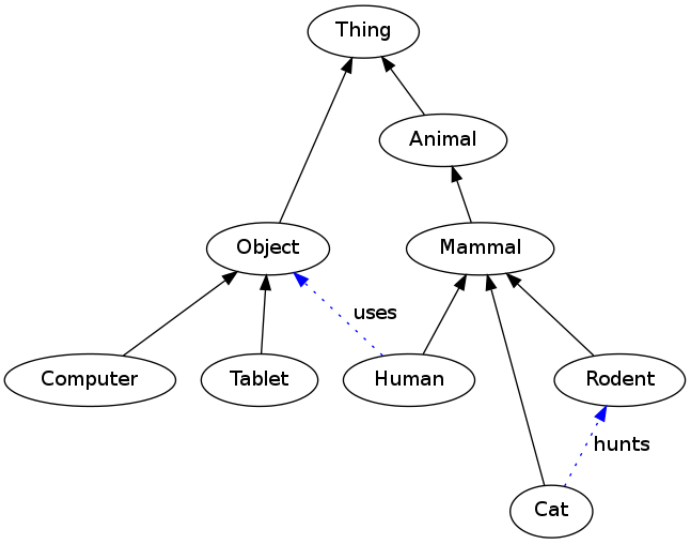

FIGURE 2.3: Semantic graph representing a taxonomy of a set of concepts and their relationships. Dotted edges refer to non-taxonomic relationships between concepts, others are taxonomic relationships

A large diversity of structural measures have been proposed to compare elements structured in a graph as a function of the strength of their interconnections (e.g., random-walk approaches). More refined measures take advantage of the analysis of intrinsic factors to better estimate the similarity/relatedness, e.g., by considering non-uniform weights of relationships.

The feature-based approach: semantic measures associated to the feature model defined by Tversky which was introduced in Section 1.4.2. Semantic measures are based on the evaluation of compared elements represented through sets of properties derived from graph analysis.

A central element of measures based on this approach is the function which characterises the features of the elements on which their comparison will be based. Among the various strategies proposed, the features characterising a concept can be considered as the *senses* it encompasses, which corresponds to its ancestors in the graph, i.e., all concepts which subsumes the concept according to the partial ordering defined by the taxonomy of concepts. By adopting this strategy, the following feature-based representation of concepts can be considered:

- `Computer` = { `Computer, Object, Thing` }
- `Tablet` = { `Tablet, Object, Thing` }
- `Rodent` = { `Rodent, Mammal, Animal, Thing` }

The comparison of two concepts represented as sets of elements, here sets of concepts, can therefore be made by evaluating the number of features they share according to a feature-matching function. This approach is therefore framed in set theory; relaxing the degree of membership of elements defined in the set, semantic measures based on this approach are also sometimes defined in terms of fuzzy set theory. Considering classical set-based feature matching functions, i.e., the boolean function, the pair `Computer` - `Tablet` will also be estimated as more similar than the pair `Computer` - `Rodent` as the former pair share more *senses*[1] than the latter.

The information theoretical approach: it is based on Shannon's Information Theory [Shannon, 1948] and relies more particularly on the notion of information content of concepts introduced by Resnik [1995]. Compared elements are regarded in terms of the information they convey. Therefore, the elements, generally concepts, are compared according to the amount of information they share and the one amongst them which is distinct.

This approach relies on the central notion of Information Content (IC) which will be covered in detail in the following chapter. In short, the IC of a concept was initially defined as a function of its probability of occurrence in a corpus considering the ordering defined by the taxonomy, i.e., in Figure 2.3 the concept `Mammal` is also considered to occur when the concept `Cat` is encountered in the corpus. Therefore, the IC is defined as inversely proportional to the probability of occurrence of the concepts; informally, the

[1] Note that we talk about about senses when we adopt a synset vision. If we consider concepts we can refer to properties.

more a concept is used (the more general it is), the less informative it will be considered. Thus, intuitively, the IC of the concept `Thing` will be smaller than the IC of the concept `Cat`; consider for instance the informativeness of the following sentences: *Yesterday Lucie bought a(n)* `Tablet` (`Object`). Technically speaking, the original definition of the notion of IC is based on an hybrid approach which involves both an ontology and a corpus of texts. Nevertheless, as we will see in the next chapter, numerous approaches have also been proposed to estimate the IC of concepts based solely on the analysis of the structural properties of semantic graphs.

Using a measure to capture the informativeness of a concept, the similarity of two concepts can easily be defined as a function of the informativeness (IC) of their Most Informative Common Ancestor (MICA), i.e., the more informative (specific) their MICA, the more similar the two compared concepts will be considered. The MICAs of the pairs of concepts `Computer - Rodent` and `Computer - Tablet` are respectively the concepts `Thing` and `Object` (Figure 2.3). By definition, we know that the informativeness of the concept `Object` can only be higher than the informativeness of the concept `Thing`. Therefore, a measure which defines the similarity of a pair of concepts as directly proportional to the IC of the MICA of the compared concepts will estimate the pair of concepts `Computer - Tablet` more semantically similar than the pair of concepts `Computer - Rodent`.

The hybrid approach: semantic measures are defined mixing some of the specificities of the approaches briefly introduced above.

2.2.3.3 Semantic measures based on logic-based semantics

Semantic measures based on the relational setting cannot be used to directly compare complex descriptions of classes or instances which rely on logic-based semantics, e.g. description logics (DLs). To this end, semantic measures have been proposed which are capable of taking into account logic-based semantics. They are for example used to compare complex concept definitions expressed in OWL.

Among the diversity of proposals, measures based on simple DLs, e.g., only allowing concept conjunction (logic $\mathcal{A}$), were initially proposed through extensions of semantic measures based on graph analysis [Borgida et al., 2005]. More refined semantic measures have since been designed to exploit high expressiveness of DLs, e.g. $\mathcal{ALC}$, $\mathcal{ALN}$, $\mathcal{SHI}$, $\mathcal{ELH}$ description logics [Araújo and Pinto, 2007; D'Amato et al., 2005a,b, 2008; Fanizzi and D'Amato, 2006; Hall, 2006; Janowicz, 2006; Janowicz and Wilkes, 2009; Lehmann and Turhan, 2012; Stuckenschmidt, 2009].

As an example D'Amato et al. [2005a] proposed to compare complex concept descriptions by aggregating functions which consider various components of their $\mathcal{ALC}$ normal forms[1]. These measures rely mostly on extensions of the feature model proposed by Tversky. They have been extensively covered in the thesis of D'Amato [2007]. The contributions presented in this manuscript do not focus on this type of approaches.

2.2.3.4 Semantic measures for multiple ontologies

Several approaches have been designed to estimate the relatedness of concepts or instances using multiple ontologies. These approaches are sometimes named cross-ontology semantic similarity/relatedness measures in the literature, e.g., [Petrakis et al., 2006]. Their aim is twofold:

- To enable the comparison of elements which have not been defined in the same ontology (the ontologies must model a subset of equivalent elements).
- To refine the comparison of elements by incorporating a larger amount of information during the process.

These measures are in some senses related to those commonly used for the task of ontology alignment and instance matching [Euzenat and Shvaiko, 2013]. Therefore, prior to their introduction we will first highlight the relationship between these measures and those designed for the aforementioned processes.

Comparison with ontology alignment and instance matching

The task of ontology mapping aims at finding links between the classes and predicates defined in a collection of ontologies. These mappings are further used to build an alignment between ontologies. Instance matching focuses on finding similar instances defined in a collection of ontologies. These approaches generally rely on multiple matchers which will be aggregated for evaluating the similarity of the compared elements [Euzenat and Shvaiko, 2013; Shvaiko and Euzenat, 2013]. The commonly distinguished matchers are:

- *Terminological* – based on string comparison of the labels or definitions.
- *Structural* – mainly based on the structuration of classes and predicates.
- *Extensional* – based on instance analysis.
- *Logic-based* – rely on logical constructs used to define the elements of the ontologies.

The score produced by these matchers is generally aggregated; a threshold is used to estimate if two (groups of) elements are similar enough to define a mapping between

[1] *Primitives* and restrictions (both existential and universal) are considered.

them. In some cases, the mapping will be defined between an element and a set of elements, e.g., depending on the difference of granularity of the compared ontologies, a concept can be mapped to a set of concepts. The problem of ontology alignment and instance matching is a field of study in itself. The techniques used for this purpose involve semantic similarity measures for the design of structural, extensional and logic-based matchers (terminological matchers are not semantic). However, the measures used in this context aim to find exact matches and are therefore generally not suited for the comparison of non-equivalent elements defined in different ontologies. Indeed, techniques used for ontology alignment are for instance not suited to answering questions such as: to which degree are the two concepts `Coffee` and `Cup` related?

In every instance, technically speaking, nothing prevents the use of matching techniques to estimate the similarity between elements defined in different ontologies. Indeed, the problem of knowing if two elements must be considered as equivalent can be reformulated as a function of their degree of semantic similarity. Nevertheless, a clear distinction of the problem of ontology alignment and semantic measure design exists in the literature. This can be partially explained by the fact that, in practice, compared to approaches used for ontology alignment and instance matching, semantic measures based on multiple ontologies:

- Can be used to estimate the semantic relatedness and not only the semantic similarity of compared elements.
- Sometimes rely on strong assumptions and approximations which cannot be considered to derive alignments, e.g., measures based on shortest path techniques.
- Focus on the design of techniques for the comparison of elements defined in different ontologies which generally consider a set of existing mappings between ontologies.

In short, ontology alignment and instance matching are complex processes which use specific types of (semantic) similarity measures and which can be used to support the design of semantic measures involving multiple ontologies. We briefly present the main approaches which have been proposed for the definition of semantic measures based on multiple ontologies.

Main approaches for the definition of semantic measures using multiple ontologies

The design of semantic measures for the comparison of elements defined in different ontologies have attracted less attention than classical semantic measures designed for single ontologies. They have been successfully used to support data integration [M.C.

Lange, D.G. Lemay, 2007; Rodríguez and Egenhofer, 2003], clustering [Batet et al., 2010b], or information retrieval tasks [Xiao and Cruz, 2005], to cite a few. In this context, several contributions have focused on the design of semantic measures based on multiple ontologies without focusing on specific application contexts.

The measures proposed in the literature can be distinguished according to the approach they adopt – we consider the same classification used for semantic measures defined for a single ontology (the list of references may not be exhaustive):

- Structural approach: [Al-mubaid and Nguyen, 2009].
- Feature-based approach: [Batet et al., 2010b, 2013; Petrakis et al., 2006; Sánchez and Batet, 2013; Solé-Ribalta et al., 2014].
- Information Theoretical approach: [Sánchez and Batet, 2013; Saruladha, 2011; Saruladha and Aghila, 2011; Saruladha et al., 2010a].
- Hybrid approach: [Rodríguez and Egenhofer, 2003].

2.2.3.5 Advantages and limits of knowledge-based measures

Advantages

- They can be used to compare all types of elements defined in an ontology, i.e., terms, concepts, instances. These measures can therefore be used to compare elements which cannot be compared using text analysis, e.g., comparison of gene products according to conceptual annotations corresponding to their molecular functions.
- Give access to fine control on the semantic relationships taken into account to compare the elements. This aspect is important to understand the semantics associated to a score of semantic measures, e.g., semantic similarity/relatedness, and can therefore be essential in a decision process.
- Generally easier and less complex to compute than distributional measures.

Limits

- Require an ontology describing the elements to compare, which can be a strong limitation if no ontology is available for the domain to consider.
- The use of logic-based measures can be challenging to compare elements defined in large ontologies (high computational complexity).
- Measures based on graph analysis generally require the knowledge to be modelled in a specific manner in the graph and are not designed to take non-binary relationships into account. Such relationships are used in specific ontologies and play

an important role in defining specific properties to relationships/statements. In Section A.1.4.1, we show that reification techniques can be used to express such knowledge by defining a ternary relationship, i.e., the (binary) relationship is expressed by a node of the graph. Most measures based on graph analysis are not adapted to this case. This aspect is relative to the mapping of an ontology to a semantic graph; a more detailed discussion of this specific aspect is proposed in Appendix A.

2.2.4 Mixing knowledge-based and distributional approaches

Hybrid measures have been proposed to take advantage of both corpora and ontology analyses to estimate the semantic similarity or relatedness of units of language, concepts and instances. We distinguish two broad types of measures, *Pure-hybrid measures* and *Aggregated measures*:

- *Pure-hybrid measures* correspond to measures which are not based on the aggregation of several measures; they are designed by defining a strategy which takes advantage of both corpus and ontology analysis. First and most common examples of pure-hybrid measures are semantic measures based on the information theoretical approach. As an example, Resnik [1995] proposed to estimate the amount of information carried by a concept as the inverse of the probability of the concept occurring in texts. The information content is the cornerstone of information theoretical measures, it can therefore be used to take advantage of several knowledge-based measures by considering corpus-based information. Other authors have also proposed to mix text analysis and structure-based measures. The extended gloss overlap measure introduced by Banerjee and Pedersen [2002], and the two measures based on context vectors proposed by Patwardhan [2003] are good examples. Interested readers may also consider [Banerjee and Pedersen, 2003; Patwardhan et al., 2003; Patwardhan and Pedersen, 2006].

- *Aggregated measures* derive from the aggregation combining distributional and knowledge-based semantic measures[1]. Scores of selected measures are aggregated according to the average, min, max, median or any aggregation function which can be designed to aggregate matrix of scores[2].

Several studies have demonstrated the benefits of performance mixing knowledge-based and distributional approaches in specific usage contexts [Panchenko and Morozova, 2012; Petrakis et al., 2006].

[1] Pure-hybrid measures can also be part of the aggregation.

[2] Several aggregations will be discussed in the introduction of semantic similarity measures which can be used to compare groups of concepts – Section 3.6.2.2.

This chapter has introduced the notion of semantic measures. We have presented their practical usages in different application contexts, we have proposed general definitions associated to the notion, and we have distinguished different semantics which can be associated to them. This latter point helped us to better capture the meaning of semantic measures (results). To this end, we define the terminology classically found in the literature, e.g., semantic similarity/proximity/relatedness/distance, and we proposed an organisation of the notions commonly used, e.g. semantic similarity is a component of semantic relatedness. In a second step, to better understand the characteristics of semantic measures, we distinguished several central aspects of measures which can be used to categorising the large diversity of measure proposals. As a result, a general classification of the variety of semantic measures defined in the literature has been presented. Such a classification highlights the similarities and differences of the numerous measures and approaches which have been proposed in the literature. It can therefore be used to better understand the large diversity of measures and to characterise areas of research which have not been explored for designing measures. Importantly, this overview of semantic measures and the proposed classification also stresses the breadth of this field of study and the difficulty to define the notions on which are based semantic measures, e.g., semantic relatedness and semantic similarity.

3

Semantic measures based on semantic graph analysis

Contents

Abstract

This chapter focuses on semantic measures based on semantic graph analysis, the measures on which the rest of thesis is dedicated to. First we underline the particular role played by these measures in order to bring to light their preponderant role in the literature. Notations used to manipulate semantic graphs are also introduced. Particular attention is given to the presentation of semantic evidence which can be derived from a semantic graph, and its role in the definition of semantic measures. This will help us to introduce the diversity of proposals which have been introduced in the literature, in particular to compare a pair of (groups of) concepts. Based on the in-depth analysis of this particular type of measures, and more generally on the survey which supports this study, we finally distinguish several perspectives and challenges offered to the communities involved in the study of semantic measures.

3.1 Importance of semantic measures based on semantic graph analysis

As we have seen, two main families of semantic measures can be distinguished: distributional measures, which take advantage of unstructured or semi-structured texts, and knowledge-based measures which rely on ontologies.

Distributional measures are essential for comparing units of languages such as words, or even concepts, when there is no formal expression of knowledge available to drive the comparison. As we have stressed, these measures rely on algorithms governed by assumptions to capture the semantics of the elements they compare (i.e., mainly the *distributional hypothesis*). On the contrary, knowledge-based semantic measures rely on formal expressions of knowledge explicitly defining how the compared elements must be understood. Thus, they are not constrained to the comparison of units of language and can be used to drive the comparison of any formally described pieces of knowledge, which encompasses a large diversity of elements, e.g., concepts, genes, person, music bands, etc.

The rest of this thesis focuses on knowledge-based measures and more particularly on those which rely on ontologies processed as semantic graphs; this positioning is motivated below.

We have underlined the main limitation of knowledge-based measures: their strong dependence on the availability of an ontology – an expression of knowledge which can be difficult to obtain and may therefore not be available for all fields of studies. However, in recent decades, we have observed, both in numerous scientific communities and industrial fields, the growing adoption of knowledge-enhanced approaches based on ontologies. As an example the Open Biological and Biomedical Ontology (OBO) foundry gives access to hundreds of ontologies related to biology and biomedicine. Moreover, thanks to the large efforts made to standardise the technology stack which can be used to define and take advantage of ontologies (e.g., RDF(S), OWL, SPARQL – triple stores implementations) and thanks to the increasing adoption of the Linked Data and Semantic Web paradigms, a large number of initiatives give access to ontologies in numerous domains (e.g., biology, geography, cooking, sports).

In the introduction, we also point out that several large corporations adopt ontologies to support large-scale worldwide systems. The most significant example over recent years is the adoption of the Knowledge Graph by Google, a graph built from a large collection of billions of non-ambiguous subject-predicate-object statements used to formally describe general or domain-specific pieces of knowledge. This ontology is used to enhance their

search engine capabilities and millions of users benefit from it on a daily basis. Several examples of such large ontologies are now available: DBpedia, Freebase, Wikidata, Yago.

Another significant fact about the increasing adoption of ontologies is the joint effort made by the major search engines companies, e.g., Bing (Microsoft), Google, Yahoo! and Yandex, to design Schema.org, a set of structured schemas defining a vocabulary which can be used to characterise the content of webpages in an unambiguous manner.

Another interesting aspect of the last few years is the growing adoption of graph databases (e.g., Neo4J[1], OrientDB[2], Titan[3]). These databases rely on a graph model to describe information in a NoSQL fashion. They actively contribute to the growing adoption of the graph property model – thinking in terms of connected entities [Robinson et al., 2013].

In this context, a lot of attention has been given to ontologies, which in numerous cases merely correspond to semantic graphs – characterised elements (concepts, instances and relationships) are defined in an unambiguous manner without using complex logical constructs. Such semantic graphs have the interesting properties of being easily expressed and maintained while ensuring a good ratio between semantic expressivity and effectiveness (in term of computational complexity). This justifies the large number of contributions related to the design of semantic measures dedicated to semantic graphs – a diversity of measures to which this chapter is dedicated.

3.2 Formal notations used to manipulate semantic graphs

We further introduce the notations used to refer to particular constitutive elements of a semantic graph. Please refer to Section 1.2 for the notations which have already been introduced.

3.2.1 Relationships – statements – triplets

The relationships of a semantic graph are distinguished according to their predicate and to the pair of elements they link. The triplet (u, t, v) corresponds to the unique relationship of type $t \in R$ which links the elements u, v: u is named the subject, t the predicate and v the object. Relationships are central elements of semantic graphs and will be used to define algorithms and to characterise paths in the graph.

[1] `http://www.neo4j.org/`
[2] `http://www.orientechnologies.com/orientdb/`
[3] `http://thinkaurelius.github.io/titan/`

Since the relationships are oriented, we denote t^- the type of relationship carrying the inverse semantic of t. We therefore consider that any relationship (u, t, v) implicitly implies (v, t^-, u), even if the type of relationship t^- and the relationship (v, t^-, u) are not explicitly defined in the graph. As an example, the relationship `Human subClassOf Mammal` implies the inverse relationship `Mammal subClassOf`$^-$ `Human` (even if the ontology defines `subClassOf`$^-$ $\equiv$ `superClassOf`). The notion of inverse predicate will be considered to discuss detailed paths. In some ontology languages, inverse relationships between predicates are explicitly defined by specific construct, e.g., `owl:inverseOf` in OWL.

3.2.2 Graph traversals

Graph traversals are often represented through paths in a graph, i.e., sequence of relationships linking two nodes. To express such graph paths, we adopt the following notations[1].

Path: Sequence of relationships $[(c_{i-1}, t_i, c_i), (c_i, t_{i+1}, c_{i+1}), \ldots]$. To lighten the formalism, if a single predicate is used the path is denoted $[c_{i-1}, c_i, c_{i+1}, \ldots]^t$.

Path Pattern: We denote $\pi =< t_1, \ldots, t_n >$ with $t_n \in R$, a path pattern which corresponds to a list of predicates[2]. Therefore, any path which is a sequence of relationships is an instance of a specific path pattern π.

We extend the use of the path pattern notation to express concise expressions of paths:

- $< t_* >$ corresponds to the set of paths of any length composed only of relationships having for predicate t.
- $< t_*^* >$ corresponds to the set of paths of any length composed of relationships associated to the predicate t or t^-.

As an example, $\{$`Human`$, <$ `subClassOf`$_* >$, `Animal`$\}$ refers to all paths which link concepts `Human` and `Animal` and which are only composed of relationships `subClassOf` (they do not contain relationships of type `subClassOf`$^-$).

We also mix the notations to characterise set of paths between specific elements. As an example, $\{u, <$ `t, subClassOf`$_* >, v\}$ represents the set of paths which (i) link the elements u and v, (ii) start by a relationship of predicate t, and (iii) end by a (possibly empty) path of `subClassOf` relationships. As an example the concept membership function $\mathcal{I}$ which characterises instances of a specific concept can formally be redefined

[1] These notations are based on an adaptation of the notations used by Lao [2012].
[2] In SPARQL 1.1, such paths are denoted using path properties $t_1/t_2/t_3$.

by:

$$\mathcal{I}(\texttt{X}) = \{i | \{i, < \texttt{isA}, \texttt{subClassOf}_* >, \texttt{X}\} \neq \emptyset \} \tag{3.1}$$

To lighten the formalism, we consider that the set of paths $\{u, < p_* >, v\}$ can be shortened by $\{u, p, v\}$, e.g. $\{\texttt{Human}, < \texttt{subClassOf}_* >, \texttt{Animal}\} = \{\texttt{Human}, \texttt{subClassOf}, \texttt{Animal}\}$ and $\{\texttt{Human}, < \texttt{subClassOf}^*_* >, \texttt{Animal}\} = \{\texttt{Human}, \texttt{subClassOf}^*, \texttt{Animal}\}$

3.2.3 Notations for taxonomies

The taxonomy G_T is the semantic graph associated to the non-strict partial order defined over the set of concepts C. We introduce the notations used to characterise G_T as well as its concepts; some of them have already been introduced and are repeated for clarity:

- $C(G_T)$ shortened by C refers to the set of concepts defined in G_T.
- $E(G_T)$ shortened by E_T refers to the set of relationships defined in G_T with:

$$E_T \subseteq C \times \{\texttt{subClassOf}\} \times C \subseteq E_T \subseteq E_{CC}$$[1]

- A concept v subsumes another concept u if $u \preceq v$, i.e., $\{u, \texttt{subClassOf}, v\} \neq \emptyset$. Several additional denominations will be used; it is commonly said that v is an ancestor of u, that u is subsumed by v and that u is a descendant of v.
- $C^+(u) \subseteq C$, with $u \in C$, the set of concepts such as:

$$C^+(u) = \{c | (u, \texttt{subClassOf}, c) \in E_T\}$$

- $C^-(u) \subseteq C$, with $u \in C$, the set of concepts such as:

$$C^-(u) = \{c | (c, \texttt{subClassOf}, u) \in E_T\}$$

- $C(u) \subseteq C$, with $u \in C$, the set of neighbours of concepts such as:

$$C(u) = C^+(u) \cup C^-(u)$$

- $A(u)$ the set of concepts which subsumes u, also named the ancestors of u, i.e., $A(u) = \{c | \{u, \texttt{subClassOf}, c\} \neq \emptyset\} \cup \{u\}$. We also denote $A^-(u) = A(u) \setminus \{u\}$ the exclusive set of ancestors of u.
- $parents(u)$ the minimal subset of $A^-(u)$ from which $A^-(u)$ can be inferred according to the taxonomy G_T, i.e., if G_T doesn't contain taxonomic redundancies[2] we obtain: $parents(u) = C^+(u)$.

[1] E_{CC} were used to introduce semantic graphs

[2] Taxonomic redundancies are introduced in Section A.2.

- $D(u)$ the set of concepts which are subsumed by u, also named the descendants of u, i.e., $D(u) = \{c | \{c, \texttt{subClassOf}, u\} \neq \emptyset\} \cup \{u\}$. We also denote $D^-(u) = D(u) \setminus \{u\}$ the exclusive set of descendants of u.

- $children(u)$ the minimal subset of $D^-(u)$ from which $D^-(u)$ can be inferred according to the taxonomy G_T, i.e., if G_T doesn't contain taxonomic redundancies we obtain: $children(u) = C^-(u)$.

- $roots(G_T)$, shortened by *roots*, the set of concepts $\{c | A(c) = \{c\}\}$. We call the *root*, denoted as $\top$, the unique concept (if any) which subsumes all concepts, i.e., $\forall c \in C, c \preceq \top$.

- $leaves(G_T)$, shortened by *leaves*, the set of concepts without descendants, i.e. $leaves = \{c | D(c) = \{c\}\}$. We also note $leaves(u)$ the set of leaves subsumed by a concept (inclusive if u is a leaf), i.e., $leaves(u) = D(u) \cap leaves$.

- $depth(u)$, the length of the longest path in $\{u, \texttt{subClassOf}, \top\}$, for convenience we also consider $depth(G_T) = \underset{c \in C}{\text{argmax}}\ depth(c)$.

- $G_T^+(u)$ the graph composed of $A(u)$ and the set of relationships which link two concepts in $A(u)$.

- $G_T^-(u)$ the graph composed of $D(u)$ and the set of relationships which link two concepts in $D(u)$.

- $G_T(u) = G_T^+(u) \cup G_T^-(u)$ the graph induced by $A(u) \cup D(u)$.

- $\Omega(u, v)$, the set of Non Comparable Common Ancestors (NCCAs) of the concepts u, v. $\Omega(u, v)$ is formally defined by: $\forall (x, y) \in \Omega(u, v), (x, y) \in \{A(u) \cap A(v)\} \times \{A(u) \cap A(v)\} \wedge x \notin A(y) \wedge y \notin A(x)$. NCCAs are also called the Disjoint Common Ancestors (DCAs) in some contributions, e.g. [Couto et al., 2005][1].

- A *taxonomic tree* is defined as a special case of taxonomy in which: $\forall c \in C : |parents(c)| < 2$.

Despite the fact that these notations are used to characterise the taxonomy of concepts G_T and that specific semantics is associated to the notations (e.g., parents, children), they can be used to characterise any poset.

[1] The modification of the terminology has been made in agreement with the reviewers of [Harispe et al., 2013d] which stressed that the term DCAs was misleading.

3.3 Semantic evidence in semantic graphs and their interpretations

A semantic graph carries explicit semantics, e.g. through the taxonomy defining concepts partial ordering. It also contains implicit semantic evidence. According to Section 2.2.1.3, we consider semantic evidence as any information on which interpretations can be based according to the meaning carried by the ontology or the elements it defines (concepts, instances, relationships).

Semantic evidence derives from the study of specific factors (e.g., number of concepts, depth of a concepts, average relationships associated to a concept) which can be used to discuss particular properties of the semantic graph (e.g., coverage, expressiveness) or particular properties of its elements (e.g. specificity of concepts). Figure 3.1 illustrates the acquisition of semantic evidence. Based on the analysis of specific factors using particular metrics, some properties of both the semantic graph and the elements it defines can be obtained. Based on these properties, and either based on high assumptions or theoretically justified by the core semantics on which relies the ontology, semantic evidence can be obtained. As an example of semantic evidence, the number of concepts described in a taxonomy can be interpreted as a clue on the degree of coverage of the ontology. One can also consider that the deeper a concept w.r.t the depth of G_T, the more specific the concept.

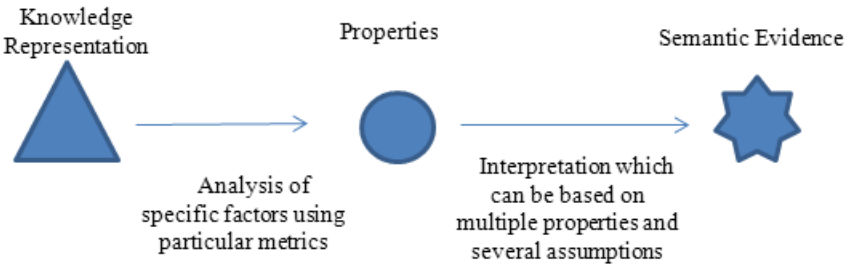

FIGURE 3.1: General process showing how semantic evidence can derive from an ontology analysis

As we will see, several properties are used to consider extra semantics from semantic graphs; they are especially important for the design of semantic measures. Indeed, semantic evidence is core elements of measures; it has been used for instance to: (i) normalise measures, (ii) estimate the specificity of concepts and to (iii) weigh the relationships defined in the graph, i.e., to estimate the *strength of connotation* between

concept/instances. It is therefore central for both designers and users of semantic measures to know: (i) the properties which can be used to derive semantic evidence, (ii) how it is computed, and (iii) the assumptions on which its interpretation relies.

Most of the properties used to derive semantic evidence are well-known graph properties defined by graph theory. In this section, we only introduce the main properties which are based on the study of a taxonomy of concepts (G_T). We go on to introduce two applications of these properties: the estimation of the specificity of concepts and the estimation of the strength of connotation between concepts.

3.3.1 Semantic evidence in taxonomies

In this section we mainly focus on semantic evidence commonly exploited in taxonomies. Two kinds of semantic evidence can be distinguished:

- *Intentional evidence* which can also be called *intrinsic evidence*, which is based on the analysis of properties associated to the topology of G_T.
- *Extensional evidence* which is based on the analysis of both the topology of G_T and the distribution of concepts' usage, i.e., the number of instances associated to concepts.

Notice that we don't consider semantic evidence *purely* extensional, i.e., only based on concepts' usage, without taking the taxonomy into account. Indeed, in most cases, the distribution of concepts' usage must be evaluated considering the transitivity of the taxonomic relationship. If this is not the case, incoherent results could be obtained, As an example, if the transitivity of the taxonomic relationship is not considered to propagate the usage of concepts (instance membership), the distribution of instances can be incoherent w.r.t the partial order defined by the taxonomy, i.e., a concept can have more instances than one of its ancestors.

We further distinguish the evidence which is based on *global properties* (i.e., derived from the full taxonomy), from that based on *local properties* of concepts.

3.3.1.1 Intentional evidence

Global properties

- Depth of the taxonomy – maximal number of ancestors of a concept

The depth of the taxonomy corresponds to the maximal depth of a concept in G_T. It informs on the degree of expressiveness/granularity of the taxonomy. As an example, the deeper G_T, the more detailed the taxonomy is expected to be.

The maximal number of ancestors of a concept is also used as an estimator of the upper bound of the degree of expressivity of a concept. Inversely, the number of concepts defined in G_T, i.e., $|D(\top)|$ if $\top$ exists, can also be used as an upper bound of the degree of generality of a concept.

- Diameter – width of the taxonomy

The width of the taxonomy corresponds to the length of the longest shortest path[1] which links two concepts in G_T. It also informs on the degree of coverage of the taxonomy. G_T is generally assumed to better cover a domain the bigger its diameter.

Local properties

- Local density

It can be considered that relationships in dense parts of a taxonomy represent smaller taxonomic distances. Metrics such as compactness can be used to characterise local density [Botafogo et al., 1992][2]. Other metrics such as the (in/out)-branching factor of a concept ($|C^+(u)|$, $|C^-(u)|$), the number of neighbours of a given concept ($|C(u)|$), can also be used [Sussna, 1993]. It is generally assumed that the higher the number of neighbours of a concept, the more general it is.

- Number of ancestors – depth – number of descendants – number of subsumed leaves – distance to leaves.

The number of ancestors of a concept is often considered to be directly proportional to its degree of expressiveness. The more a concept is subsumed, the more detailed/restrictive the concept is expected to be. The number of ancestors can also be interpreted w.r.t the maximal number of ancestors a concept of the taxonomy can have. The depth of a concept is also expected to be directly proportional to its degree of expressiveness. The

[1] Backtracks, loops or detours excluded, ref: `http://mathworld.wolfram.com/GraphDiameter.html`.
[2] Author also introduces interesting factors for graph-based analysis; the depth of a node is also introduced.

deeper the concept (according to the maximal depth), the more detailed/restrictive the concept is regarded[1]. A local depth of a concept can also be evaluated according to the depth of the branch in which it is defined.

In a similar fashion, in some cases the distance of a concept to the leaves it subsumes, or the number of leaves it subsumes, will be considered as an estimator of expressiveness: the greater the distance/number the less expressive the concept is considered.

3.3.1.2 Extensional evidence

Global Properties

- Distribution of instances among the concepts.

The distribution of instances among concepts, i.e., concept usage, can be used to design local correction factors, e.g., to correct estimations of the expressiveness of a concept. This is generally made by evaluating the balance of the distribution.

Local Properties

- Number of instances associated to a concept

The number of instances of a concept is expected to be inversely proportional to its expressiveness, the less instances a concept has, the more specific it is expected to be.

This semantic evidence and its interpretations have been used to characterise notions extensively used by semantic measures. They are indeed used to estimate the specificity of concepts as well as the strength of connotations between concepts.

3.3.2 Estimation of concept specificity

Not all concepts have the same degree of specificity. Indeed, most people will agree that `Dog` is a more specific description of a `LivingBeing` than `Animal`. The notion of specificity can be associated to the concept of *salience* which has been defined by Tversky [1977] to characterise a stimulus according to its "*intensity, frequency, familiarity, good form, and informational content*". In Bell et al. [1988], it is also specified that "*salience*

[1]Note that the depth of a concept as an estimator of its degree of expressiveness can be seen as an inverse function of the notion of *status* introduced by Harary et al. [1965] for organisation study.

is a joint function of intensity and what Tversky calls diagnosticity, which is related to the variability of a feature in a particular set [i.e., universe, collection of instances]". The idea is to capture the amount of information carried by a concept – this amount is expected to be directly proportional to its degree of specificity and generality.

The notion of specificity of a concept is not totally artificial and can be explained by the roots of taxonomies. Indeed, the transitivity of the taxonomic relationship specifies that not all concepts have the same degree of specificity or detail. In knowledge modelling, the ordering of two concepts $u \prec v$ defines that u must be considered as more abstract (less specific) than v. In fact, the taxonomy explicitly defines that all instances of u are also instances of v. This expression is illustrated by Figure 3.2; we can see that the more a concept is subsumed by numerous concepts: (A) the number of properties which characterise the concept increases (intentional interpretation), and (B) its number of instances decreases (extensional interpretation).

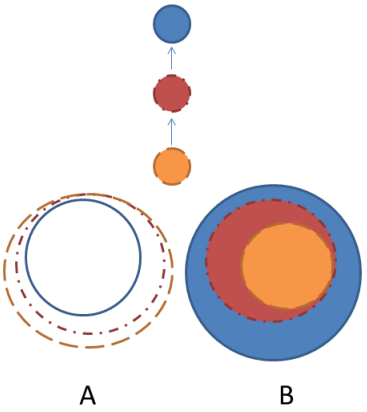

FIGURE 3.2: Set-based representations of ordered concepts according to (A) their intentional expressions in term of properties characterising the concepts, and (B), in term of their extensional expressions, i.e., the set of instances which compose the concept. Figure based on Blanchard [2008]

Therefore, another way of comparing the specificity of concepts defined in a total order[1] is to study their usage, analysing their respective number of instances. The concept which contains the highest number of instances will be the least specific (its universe of interpretation is larger). In this case, it is therefore possible to assess the specificity of ordered concepts either studying the topology of the graph, or the set of instances associated to them.

Nevertheless, in taxonomies, concepts are generally only partially ordered. This implies that presented evidence used to compare the specificity of two ordered concepts cannot be used without assumptions, i.e., concepts which are not ordered are in some sense

[1] For any pair of concepts u, v either $u \preceq v$ or $v \preceq u$.

not comparable. This aspect is underlined in Figure 3.3. It is impossible to compare, in an exact manner, the specificity of two non-ordered concepts. This is due to the fact that the amount of shared and distinct properties between these concepts can only be estimated w.r.t the properties which characterise the common concepts they derive from, i.e., their NCCAs. However, this estimation can only be a lower bound of their commonality since extra properties shared by the two concepts may not be carried by such NCCAs.

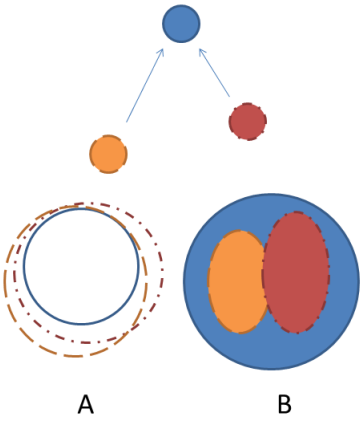

FIGURE 3.3: Potential set-based representations of non-ordered concepts according to (A) their intentional expressions in term of properties characterising the concepts, and (B), in term of their extensional expressions, i.e., the set of instances associated to the concepts. Figure based on Blanchard [2008]

As we will see, the estimation of the degree of specificity of concepts is of major importance in the design of semantic measures. Therefore, given that discrete levels of concept specificity are not explicitly expressed in a taxonomy, various approaches and functions have been explored to evaluate concept specificity. We denote such a function as θ:

$$\theta : C \rightarrow \mathbb{R}_+ \tag{3.2}$$

The function θ may rely on the intrinsic and/or extrinsic properties presented above. It must be in agreement with the taxonomic representation which defines that concepts are always semantically broader than their specialisations[1]. Thus, θ must monotonically decrease from the leaves (concepts without descendants) to the root(s) of the taxonomy:

$$x \preceq y \Rightarrow \theta(x) \geq \theta(y) \tag{3.3}$$

We present examples of θ functions which have been defined in the literature.

[1]This explains that the specificity of concepts cannot be estimated only considering extrinsic information.

3.3.2.1 Basic intrinsic estimators of concept specificity

The specificity of concepts can be estimated considering the location of its corresponding node in the graph. A naive approach will define the specificity of the concept c, $\theta(c)$, as a function of some simple properties related to c, e.g., $\theta(c) = f(depth(c))$, $\theta(c) = f(A(c))$ or $\theta(c) = f(D(c))$ with $A(c)$ and $D(c)$ the ancestors and descendants of c.

The main drawback of simple specificity estimators is that concepts with a similar depth or an equal number of ancestors/descendants will have similar specificities, which is a heavy assumption. In fact, two concepts can be described with various degrees of detail, independently of their depth, e.g., [Yu et al., 2007a]. More refined θ functions have been proposed to address this limitation.

3.3.2.2 Extrinsic information content

Another strategy explored by designers of semantic measures has been to characterise the specificity of concepts according to Shannon's Information Theory. The specificity of a concept will further be regarded as the amount of information the concept conveys, its Information Content (IC). The IC of a concept can for example be estimated as a function of the size of the universe of interpretations associated to it. The IC is a common expression of θ and was originally defined by Resnik [1995] to assess the informativeness of concepts from a corpus of texts.

The IC of the concept c is defined as inversely proportional to $p(c)$, the probability to encounter an instance of c in a collection of instances (negative entropy). The original IC definition was based on the number of occurrences of a concept in a corpus of texts.

We denote *eIC* any IC which relies on extrinsic information, i.e., information not provided by the ontology[1] and generally provided by the analysis of concept usage in a corpus of texts or by analysing a collection of instances for which associated concepts are known[2]. We consider the formulation of *eIC* originally defined by Resnik [1995]:

$$p(c) = \frac{|\mathcal{I}(c)|}{|I|}$$

[1] Note that if the instances are represented in the graph, some *eIC* are indeed *iIC*.

[2] As an example, usage of concepts defined in the Gene Ontology can be known studying gene annotations which provide genes and associated Gene Ontology concepts, e.g. refer to UniprotKB.

with $\mathcal{I}(c)$ the set of instances of c, e.g., occurrences of c in a corpus, instances in an ontology $\{i|\{i, < \texttt{isA}, \texttt{subClassOf}_* >, c\} \neq \emptyset\}$.

$$\begin{aligned} eIC_{Resnik}(c) &= -log(p(c)) \\ &= log(|I|) - log(|\mathcal{I}(c)|) \end{aligned} \tag{3.4}$$

The suitability of the log function can be supported by the work of Shepard [1987][1]. Notice also the link with Inverse Document Frequency (IDF) which is commonly used in information retrieval [Jones, 1972]:

$$\begin{aligned} IDF(c) &= log(\frac{|I|}{|\mathcal{I}(c)|}) \\ &= log(|I|) - log(|\mathcal{I}(c)|) \\ &= IC(c) \end{aligned} \tag{3.5}$$

The main drawback of θ functions based on extrinsic information lies in their tight dependence on concepts usage: they will automatically reflect its bias[2]. Nevertheless, in some cases, the consideration of such bias is desired as all concepts which are highly represented will be considered less informative, even the concepts which seem specific w.r.t intrinsic factors (e.g., depth of concepts). However, in some cases, bias in concept usage can badly affect IC estimation and may not be adapted. In addition, IC computation based on text analysis can be both time consuming and challenging given that, in order to be accurate, complex disambiguation techniques have to be used to detect which concept refers an occurrence of a word.

3.3.2.3 Intrinsic information content

In order to avoid the dependency of eIC calculus to concept usage, various intrinsic IC formulations (iIC) have been proposed. They can be used to define θ functions by only considering structural information extracted from the ontology, e.g., the intrinsic factors presented in Section 3.3.1. iIC formulations extend basic specificity estimators presented above.

Multiple topological characteristics can be used to express iIC, e.g., number of descendants, ancestors, depths, etc. [Sánchez et al., 2011; Schickel-Zuber and Faltings, 2007;

[1] Shepard derived his universal law of stimulus generalisation based on the consideration that logarithm functions are suited to approximate semantic distance [Al-Mubaid and Nguyen, 2006].

[2] As an example, this can be problematic for GO-based studies as some genes are studied and annotated more than others (e.g., drug related genes) and annotation distribution patterns among species reflect abnormal distortions, e.g. human – mouse [Thomas et al., 2012].

Seco et al., 2004; Zhou et al., 2008]. As an example, the formulation proposed by Zhou et al. [2008] enables to consider the contribution of both the depth and the number of descendants of a concept to compute its specificity:

$$iIC_{Zhou}(c) = k\left(1 - \frac{log(|D(c)|)}{log(|C|)}\right) + (1-k)\frac{log(depth(c))}{log(depth(G_T))} \tag{3.6}$$

with $|C|$ the number of concepts defined in the taxonomy, $depth(c)$ the depth of c, $depth(G_T)$ the maximal depth of a concept in G_T and $k \in [0,1]$ a parameter used to set the contribution of both components (originally set to 0.6).

In [Sánchez et al., 2011], the iIC incorporates additional semantic evidence in the aim of better distinguishing the concepts with the same numbers of descendants but different degrees of *concreteness* – here captured as a function of the number of ancestors a concept has.

$$iIC_{Sanchez}(c) = -log\left(\frac{\frac{|leaves^-(c)|}{|A(c)|} + 1}{|leaves| + 1}\right) \tag{3.7}$$

We denote $leaves^-(c)$ the exclusive set of leaves of the concept c, i.e., if c is a leaf $leaves^-(c) = \emptyset$. Note that $iIC_{Sanchez}$ will set the same iIC for each leaf. To avoid this, we propose the following modification:

$$iIC_{Sanchez'}(c) = -log\left(\frac{\frac{|leaves(c)|}{|A(c)|}}{|leaves| + 1}\right) \tag{3.8}$$

$iICs$ are of particular interest as only the topology of the taxonomy is considered. They prevent errors related to bias on concept usage. However, the relevance of iIC relies on the assumption that G_T expresses enough knowledge to rigorously evaluate the specificities of concepts. Therefore, as a counterpart, $iICs$ are sensitive to structural bias in the taxonomy and are therefore sensitive to *unbalanced taxonomy, degrees of completeness, homogeneity and coverage* of the taxonomy [Batet et al., 2010a].

3.3.2.4 Non-taxonomic information content

Both introduced iIC and eIC only take taxonomic relationships into account. In order to take advantage of all predicates and semantic relationships, Pirró and Euzenat [2010a] proposed the *extended IC* ($extIC$).

$$extIC(c) = \alpha EIC(c) + \beta IC(c) \tag{3.9}$$

$$EIC(c) = \sum_{r \in R} \frac{\sum_{u \in C_r(u)} iIC(u)}{|C_r(u)|}$$

With $C_r(u)$ the set of concepts linked to the concept c by any relationship of type $r \in R$ (i.e., generalisation of $C(u)$). In this formula, the contribution of the various relationships of the same predicate is averaged. However, the more a concept establishes relationships of different predicates, the higher its EIC will be. We thus propose to average EIC by $|R|$, or to weigh the contribution of the different predicates.

3.3.2.5 List of functions defined to estimate concept specificity

We have presented various strategies which can be used to estimate the specificities of concepts defined in a partially ordered set (θ functions). It is important to understand that these estimators are based on assumptions regarding ontologies. Table 3.1 lists some of the properties of some of the θ functions proposed in the literature – proposals are ordered by date.

Names – References	**Use Extensional inf.**	**Co-domain**	**Comments**
Depth	No	[0, 1]	Normalised depth or max depth can be used. In a graph, considering the minimal depth of a concept doesn't ensure that the specificity increases according to the partial ordering (due to multi-inheritance).
IC Resnik [Resnik, 1995]	Yes	$[0, +\infty[$, [0, 1]	Depend on concept usage. $IC(c) = -log(\|\mathcal{I}(c)\|/\|I\|)$. Normalised versions have also been proposed, e.g., [Sheehan et al., 2008].
IC Resnik intrinsic [Resnik, 1995]	No	[0, 1]	Resnik's IC with $\forall c \in C$ the number of instances associated to c (without taxonomic inferences) set to one.
IC Seco [Seco, 2005]	No	[0, 1]	IC estimated from the number of descendants.
Depth non-linear [Seco, 2005]	No	[0, 1]	Use log to introduce non-linear estimation of depth.
TAM [Yu et al., 2007a]	Yes	$[0, +\infty[$	The probability $p(c)$ associated to a concept is computed as the number of pairs of instances which are members of c divided by total the number of pairs.
IDF [Chabalier et al., 2007]	Yes	$[0, +\infty[$	Inverse Document Frequency (IDF) obtained by dividing the number of instances by the number of instances of the concept [Jones, 1972], i.e. $IDF(c) = log(\|I\|/\|\mathcal{I}(c)\|)$. As we saw, in Section 3.3.2.2, this formulation is similar to IC proposed by Resnik [1995].
APS [Schickel-Zuber and Faltings, 2007]	No	[0,1/2]	iIC based on the number of descendants of a concept.
IC Zhou [Zhou et al., 2008]	No	[0,1]	Parametric hybrid iIC mixing Seco's IC and nonlinear depth.
$extIC$ [Pirró and Euzenat, 2010a]	No	[0, 1]	iIC based on all predicates.
IC Sanchez et al (A) [Sánchez et al., 2011]	No	$[0, +\infty[$	Consider the number of leaves contained in $D(c)$, the higher it is, the less specific c is considered.
IC Sanchez et al. (B) [Sánchez et al., 2011]	No	$[0, +\infty[$	Refined version A (see above) exploiting $D(c)$.

TABLE 3.1: Selection of θ functions which can be used to estimate the specificity of a concept defined in a taxonomy. The estimation can be based on an intrinsic strategy, i.e., only evaluating the topology of the taxonomy, or taking advantage of the extensional information associated to the concepts

3.3.3 Estimation of the strength of connotations between concepts

A notion strongly linked to concept specificity is the strength of connotation between a pair of concepts/instances, i.e., the strength of the relationship(s) which links two concepts/instances. Otherwise stated, this notion can be used to assess the strength of interaction associated to a specific relationship.

Considering taxonomic relationships, it is generally considered that the strength of connotation between concepts is stronger the deeper two concepts are in the taxonomy. As an example, the taxonomic relationship linking `SiberianTiger` to `Tiger` will generally be considered to be stronger than the one linking `Animal` to `LivingBeing`. Such a notion is quite intuitive and has for instance been studied by Quillian and Collins in the early studies of semantic networks [Collins and Quillian, 1969] – hierarchical network models were built according to response time to questions, i.e., mental activations evaluated w.r.t the time people took to correctly answer questions related to two concepts, e.g., a `Canary` is an `Animal` – a `Canary` is a `Bird` – a `Canary` is a `Canary`. Based on the variation of times taken to correctly answer questions involving two ordered concepts (e.g., `Canary` – `Animal`), the authors highlighted human sensibility to non-uniform strength of connotation and its link to concept specificity.

It is worth noting that the estimation of the strength of connotation of two linked concepts is in some sort a measure of the semantic similarity or taxonomic distance between the two directly ordered concepts. The models used to estimate the strength of connotation between two concepts are generally based on the assumption that the taxonomic distance associated to a taxonomic relationship *shrinks* with the depth of the two concepts it links [Richardson et al., 1994]. Given that the strength of connotation between concepts is not explicitly expressed in a taxonomy, it has been suggested that several intrinsic factors need considering in order to refine its estimation, e.g., [Richardson et al., 1994; Sussna, 1993; Young Whan and Kim, 1990].

A taxonomy only explicitly defines the partial ordering of its concepts, which means that if a concept v subsumes another concept u, all the instances of u are also instances of v, i.e., $u \preceq v \Rightarrow \mathcal{I}(u) \subseteq \mathcal{I}(v)$. Nevertheless, non-uniform strength of connotation aim to consider that all taxonomic relationships do not convey the same semantics.

Strictly speaking, taxonomic relationships only define concept ordering and concept inclusion. Therefore, according to the extensional interpretation which can be made of a taxonomy, the size of the universe of interpretation of a concept, i.e., the size of the set of its possible instances w.r.t the whole set of instances, must reduce the more a concept is specialised[1]. This reduction of the universe of eligible interpretations associated to

[1] We here consider a finite universe.

a concept (i.e. instances), corresponds to a specific understanding of the semantics of non-uniform strengths of connotation. Alternative explanations which convey the same semantics can also be expressed according to the insights of the various cognitive models which have been introduced in Section 1.4:

- Spatial/Geometric model: it states that the distance between concepts is a non-linear function which must take salience of concept into account.
- Feature model (which represents a concept as a set of properties): It can be seen as the difficulty to further distinguish a concept which is relevant to characterise the set of instances of a domain.
- Alignment and Transformational models: the effort of specialisation which must be done to extend a concept increases the more a concept has been specialised.

All these interpretations state the same central notion – the strength of connotation which links two concepts is a function of two factors: (i) the specificities of the linked concepts, and (ii) the variation of these specificities. The semantic evidence introduced in the previous section, as well as the notion of IC, can be used to assess the strength of connotation of two concepts.

As an example, the strength of connotation w which characterises a taxonomic relationship linking two concepts u, v, with $u \preceq v$, can be defined as a function of the ICs of u and v [Jiang and Conrath, 1997]: $w(u, v) = IC(u) - IC(v)$.

It is important to stress that estimations of the strength of connotations based on the density of concepts, the branching factor, the maximal depth or the width of the taxonomy, are based on assumptions regarding the definition of the ontology.

We have presented various semantic evidence which can be used to extract knowledge from an ontology represented as a semantic graph. We have also presented two applications of such semantic evidence for assessing the specificity of a concept defined in a taxonomy and the strength of interaction between two elements defined in a semantic graph. As we will see, semantic evidence are central for the definition of semantic measures. We will now introduce the various semantic measures which can be considered depending on particular of the semantic graph in use.

3.4 Types of semantic measures w.r.t graph properties

Two main groups of measures can be distinguished w.r.t the properties of semantic graphs:

- Measures adapted to semantic graphs composed of (multiple) predicate(s) which potentially induce cycles.
- Measures adapted to taxonomies, i.e., acyclic semantic graphs composed of a unique predicate inducing transitivity.

The two types are presented in this section.

3.4.1 Semantic measures on cyclic semantic graphs

Considering all predicates defined in a semantic graph potentially leads to a cyclic graph. Nevertheless, only few semantic measures framed in the relational setting have been designed to deal with cycles. Since these measures take advantage of all predicates, they are generally used to evaluate semantic relatedness (and not semantic similarity). Notice that they can be used to compare concepts and instances. Two types of measures can be further distinguished:

- *Measures based on graph traversal*, i.e., pure graph-based measures. These measures have initially been proposed to study node interactions in a graph and essentially derive from graph theory contributions. They can be used to estimate the relatedness of nodes considering that greater the (direct or indirect) interconnection between two nodes, the more related they are. These measures are not semantic measures *per se* but rather graph measures used to compare nodes. However, they can be used on semantic graphs and can also be adapted in order to take into account evidence of semantics defined in the graph (e.g. strength of connotation).

- *Measures based on the graph property model*. These measures consider concepts or instances as sets of properties distinguished from the graph.

The two types of measures are presented.

3.4.1.1 Semantic measures based on graph traversals

Measures based on graph traversals can be used to compare any pair of concepts or instances represented as nodes. These measures rely on algorithms designed for graph

analysis which are generally used in a straightforward manner. Nevertheless, some adaptations have been proposed in order to take into account the semantics defined in the graph. Among the large diversity of measures and metrics which can be used to estimate the relatedness (distance, interconnection, etc.) of two nodes in a graph, we distinguish:

- Shortest path approaches.
- Random-walk approaches.
- Other interconnection measures.

The main advantage of these measures is their unsupervised nature. Their main drawback is the absence of extensive control over the semantics which are taken into account; this generates difficulties in justifying, explaining, and therefore analysing the resulting scores. However, in some cases, these drawbacks are reduced by enabling fine-grain control over the predicates considered during the comparison. This is done by tuning the contribution of each relationship or predicate.

Shortest path approaches

The shortest path problem is one of oldest problems of graph theory. It can be applied to compare both pairs of instances and concepts considering their relatedness as a function of the distance between their respective nodes. More generally, the relatedness is estimated as a function of the weight of the shortest path linking them. Classical algorithms proposed by graph theory can be used. The algorithm to use depends on specific properties of the graph, e.g., Do the constraints applied to the shortest path (really) induce cycles? Are there non-negative weights associated to relationships? Is the graph considered to be oriented?

Rada et al. [1989] were among the first to use the shortest path technique to compare two concepts defined in a semantic graph (initially a taxonomy). This approach is sometimes denoted as the *edge-counting strategy* in the literature (edge refers to relationship). As the shortest path may contain relationships of any predicate we call it unconstrained shortest path (*usp*).

One of the drawbacks of the *usp* in the design of semantic measures lies in the fact that the meaning of the relationships from where the relatedness derives is not taken into account. In fact, complex semantic paths which involve multiple predicates and only those composed of taxonomic relationships are considered equally. Therefore, propositions to penalise any *usp* reflecting complex semantic relationships have been proposed [Bulskov et al., 2002; Hirst and St-Onge, 1998]. Approaches for considering particular predicates

in a specific manner have also been described. To this end, a *weighting scheme* can be considered in order to tune the contribution of each relationship or predicate in the computation of the final score – this weighting scheme can be derived from the notion of strength of connotation (Section 3.3.3).

Random walk approaches

These approaches are based on a Markov chain model of random walks [Spitzer, 1964]. The random walk is defined through a transition probability associated to each relationship. The walker can therefore walk from node to node – each node represents a state of the Markov chain. Several measures can be used to compare two nodes u and v based on this technique; a selection of measures introduced in [Fouss et al., 2007] is listed:

- The average first-passage time, hitting time, i.e., the average number of steps needed by the walker to go from u to v.
- The average commute time, Euclidean commute time distance.
- The average first passage cost.
- The pseudo inverse of the Laplacian matrix.

These approaches are closely related to spectral clustering and spectral embedding techniques [Saerens et al., 2004]. Examples of measures based on random walk techniques are defined and discussed in [Alvarez and Yan, 2011; Fouss et al., 2007; Garla and Brandt, 2012; Hughes and Ramage, 2007; Ramage et al., 2009].

As an example, the hitting time $H(u, v)$ of two nodes u, v is defined as the expected number of steps needed by a random walker to go from u to v. The hitting time can recursively be defined by:

$$H(u, v) = 1 + \sum_{k \in N^+(u)} p(u, k) H(k, v) \tag{3.10}$$

With $N^+(u)$ the set of nodes which are linked to u by an outgoing relationship starting from u and $p(u, k)$ the transition probability of the Markov Chain:

$$p(u, k) = \frac{w(u, k)}{\sum_{i \in N^+(u)} w(u, i)}$$

With $w(u, k)$ the weight of the relationship between u and k.

The commute time $C(u, v) = H(u, v) + H(v, u)$ corresponds to the expected time needed for a random walker to travel from u to v and back to u. Intuitively, the more paths

that connect u and v, the smaller their commute distance becomes. Several technical criticisms of classical approaches used to evaluate hitting and commute times, as well as associated extensions, have been formulated in the literature, e.g., [Sarkar et al., 2008; von Luxburg et al., 2011].

In a similar vein, approaches based on graph-kernel can also be used to estimate the relatedness of two nodes in a graph [Kondor and Lafferty, 2002]; they have already been applied to the design of semantic measures in [Guo et al., 2006].

Note that these measures take advantage of second-order information which is generally hard to interpret (in terms of semantics).

Other measures based on interaction analysis

Several approaches exploiting graph structure analysis can be used to estimate the relatedness of two nodes through their interconnections. Chebotarev and Shamis [2006a,b] proposed the use of indirect paths linking two nodes by means of the matrix-forest theorem. sim_{Rank}, proposed by Jeh and Widom [2002], is an example of such a measure. Considering N as the set of nodes of the graph, $N^-(n)$ as the nodes linked to the node n by a single relationship ending with n (i.e., in-neighbours), sim_{Rank} similarity is defined by:

$$sim_{Rank}(u,v) = \frac{|N|}{|N^-(u)||N^-(v)|} \sum_{x \in N^-(u)} \sum_{y \in N^-(v)} sim_{Rank}(x,y) \tag{3.11}$$

Note that sim_{Rank} is a normalised function. Olsson et al. [2011] propose an adaptation of the measure for semantic graphs built from Linked Data.

3.4.1.2 Semantic measures for the graph property model

The second type of measures which can be used to compare a pair of instances/concepts defined in a (potentially) cyclic semantic graph relies on the graph property model. Here the graph is not only considered as a data structure which highlights the interactions between the different elements it defines. It is considered as a data model in which concepts and instances are describes through sets of properties. The properties may sometimes refer to specific data types. Therefore, the nodes of the graph may refer to data values, concepts, instances or even predicates – the semantic graphs generally correspond to RDF graphs or labelled graphs.

In this case the measures take advantage of semantic graphs by encompassing expressive definitions of concepts/instances through properties. The measures rely on the comparison of the different properties which characterise the concepts or instances being compared. Therefore the study of these measures inherits from early work related to both the comparison of objects defined into knowledge base and the comparison of entities defined in a subset of the first order logic [Bisson, 1992, 1995]. As an example, these measures have been extensively studied for comparing objects analysing their different properties. They are based on the aggregation of specific measures enabling the comparison of each of the properties characterising compared objects [Valtchev, 1999a,b; Valtchev and Euzenat, 1997]. Considering the domain of knowledge representation, these contributions have formed the basis of several frameworks which are used for comparing instances or concepts in the field of ontology alignment or instance matching, e.g., OWL Lite Alignment (OLA) method has been proposed to compare ontologies based on aggregations of several measures [Euzenat et al., 2004; Euzenat and Valtchev, 2004].

In this presentation, we do not introduce the expressive formalisms which have been introduced in earlier contributions [Bisson, 1992, 1995; Euzenat et al., 2004; Euzenat and Valtchev, 2004], e.g. for comparing objects defined in a knowledge base [Valtchev, 1999a,b; Valtchev and Euzenat, 1997]. We rather distinguish two general approaches which have been proposed and which are commonly used to compare concepts or instances.

Elements represented as a list of direct property

An element can be evaluated by studying its direct properties, i.e., the set of values associated to the element according to a specific predicate. As an example, focusing on relationships related to instances, two types of relationships can be distinguished:

- *Taxonomic relationships* (`isA`) – relationships which link instances to concepts.
- *Non-taxonomic relationships*:
  - Which link two instances (*object properties* in OWL).
  - Which link instances to data values (*datatype properties* in OWL).

Two elements will be compared w.r.t values associated to each property considered. To this end, for each property considered, a specific measure will be used to compare associated values (concepts, data values, instances).

Properties which link two instances associate a set of instances to the instance which is characterised. Considering Figure 3.4, the property `genre` can be used to characterise the instance `rollingStones` through a set of instances $\{i | \exists(\texttt{rollingStones}, \texttt{genre}, i)\}$,

i.e., {`rock,...`}. Such properties therefore refer to sets, they are often compared using simple set-based measures – they will for example evaluate the cardinality of the intersection (e.g., the number of music genres that two bands have in common).

Taxonomic properties are evaluated using semantic measures adapted to concept comparison. These measures will be presented in Section 3.5.

Properties associated to data values can be compared using measures adapted to the type of data considered, e.g., a measure for comparing dates if the corresponding property refers to a date.

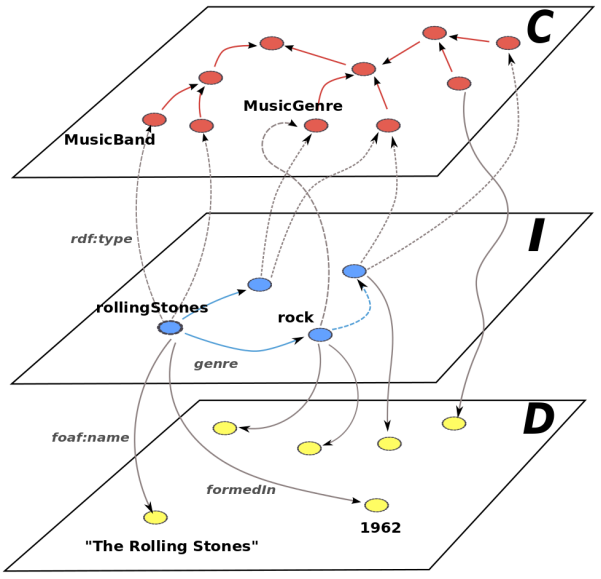

FIGURE 3.4: Example of a semantic graph related to the music domain. Concepts (C), instances (I), and data values (D) are represented [Harispe et al., 2013b]

Finally, the scores produced by the various measures (associated to the various properties) are aggregated in order to obtain a global score of relatedness of the two elements [Euzenat and Shvaiko, 2013]. Such a representation has been formalised in the framework proposed by Ehrig et al. [2004]. This is a strategy which is commonly adopted in ontology alignment, instance matching or link discovery between instances; SemMF [Oldakowski and Bizer, 2005], SERIMI [Araujo et al., 2011] and SILK [Volz et al., 2009] are all based on this approach. The reader can also refers to the extensive survey presented in [Euzenat and Shvaiko, 2013].

Consideration of indirect properties of elements

Several contributions underline the relevance of indirect properties in comparing entities represented through graphs, especially in object models [Bisson, 1995]. Referring to

Figure 3.4, indirect properties might be used to consider properties of music genres (e.g., `rock`, `rockNroll`) to compare two music bands (e.g., `rollingStones` - `doors`).

This approach relies on a representation of the compared elements which is an extension of the canonical form used to represent an element as a list of properties. This approach can be implemented to take into account the indirect properties of compared elements, e.g., properties induced by the elements associated to the element that we want to characterise.

Albertoni and De Martino [2006] extended the formal framework proposed in Ehrig et al. [2004] to allow for the consideration of some indirect properties. This framework is dedicated to instance comparison. It formally defines an indirect property of an instance along a path in the graph. The indirect properties to be taken into account depend on the context of use of the framework, e.g., application context.

From a different perspective, Andrejko and Bieliková [2013] suggested an unsupervised approach to compare two instances by considering their indirect properties. Each direct property which is shared between the compared instances plays a role in computing the global relatedness. When the property links two instances, a recursive process is applied to take into account properties of associated instances with the instances being processed. Lastly, the measure aggregates the scores obtained during the recursive process. The authors have also proposed to weigh the contribution of the various properties in the aggregation so as to define a personalised information retrieval approach.

All the measures which can be used on the whole semantic graph G can also be used for any acyclic reduction $G_R \subseteq G$. Nevertheless, numerous specific semantic measures have been defined to work on a reduction of G. Depending on the topological properties of the reduction, two cases can be distinguished:

1. The reduction G_R leads to a cyclic graph. Measures presented for cyclic graphs can be used.
2. G_R is acyclic – particular techniques and algorithms can be used. Most semantic measures defined for acyclic graphs focus on taxonomic relationships defined in G_R and consider the reduction to be the taxonomy of concepts G_T. However, some measures consider a specific subset of R, e.g., $R = \{\texttt{isA}, \texttt{partOf}\}$, which also produces an acyclic graph [Wang et al., 2007]. The measures which can be used in this case are usually a generalisation of semantic similarity measures designed for G_T.

3.4.2 Semantic measures on acyclic graphs

Semantic measures applied to graph-based ontologies were originally designed for taxonomies. Since most ontologies are usually composed mainly of taxonomic relationships or represent poset structures, substantial literature is dedicated to semantic similarity measures[1]. In particular, a large diversity of semantic measures focus on G_T and have been defined for the comparison of pairs of concepts. These measures are presented in details in the following section.

3.5 Semantic similarity between a pair of concepts

The majority of semantic measures framed in the relational setting have been proposed to assess the semantic similarity or taxonomic distance of a pair of concepts defined in a taxonomy. Given that they are designed to compare two concepts, these measures are denoted as *pairwise measures* in some communities, e.g., bioinformatics [Pesquita et al., 2009a]. As we will see, extensive literature is dedicated to these measures – they can be used to compare any pairs of nodes expressed in a graph which defines a (partial) ordering, that is to say, any graph structured by relationships which are transitive, reflexive and antisymmetric (e.g., `isA`, `partOf`).

In Section 2.2.3.2, we distinguished the main approaches used to compare concepts defined in a taxonomy. Let us remember those measures which can be applied to acyclic graphs:

- **Measures based on graph structure analysis**. They estimate the similarity as a function of the degree of interconnection between concepts. They are generally regarded as measures which are framed in the spatial model – the similarity of two concepts is estimated as a function of their distance in the graph, e.g., based on the analysis of the lengths of the paths which link the concepts. These measures can also be considered as being framed in the transformational model by considering them as functions which estimate the similarity of two concepts regarding the difficulty to transform one concept to another.

- **Measures based on concept features analysis**. This approach extracts features of concepts from the graph. These features will be subsequently analysed to estimate the similarity as a function of shared and distinct features of the compared concepts. This approach is conceptually framed in the feature model. The

[1] According to the literature we consider that semantic measures on G_T are necessarily semantic similarity measure.

diversity of feature-based measures relies on the diversity of strategies which have been proposed to characterise concept features, and to take advantage of them in order to assess the similarity.

- **Measures based on Information Theory**. Based on a function used to estimate the amount of information carried by a concept, i.e., its Information Content (IC), these measures assess the similarity w.r.t the amount of information which is shared and distinct between compared concepts. This approach is framed in information theory; it can however be seen as a derivative of the feature-based approach in which features are not compared using a boolean feature-matching evaluation (shared/not shared), but also incorporate their saliency, i.e. their degree of informativeness.

- **Hybrid measures**. Measures which are based on multiple paradigms.

The broad classification of measures that we propose is interesting as an introduction to basic approaches defined to assess the similarity of two concepts – and to put them in perspective with the models of similarity proposed by cognitive sciences. It is however challenging to constrain the diversity of measures to this broad classification. It is important to understand that these four main approaches are highly interlinked and cannot be seen as disjoint categories. As an example, all measures rely in some sense on the analysis of the structure of the taxonomy, i.e., they all take advantage of the partial ordering defined by the (structure of the) taxonomy. These categories must be seen as devices used by designers of semantic measures to introduce approaches and highlight relationships between several proposals. Indeed, as we will see, numerous approaches can be regarded as hybrid measures which take advantage of techniques and paradigms used to characterise measures of a specific approach. Therefore, the affiliation of a specific measure to a particular category is often subject to debate, e.g., as it is exposed in [Batet, 2011b]. This can be partially explained by the fact that several measures can be redefined or approximated using reformulations, in a way that further challenge the classification. Indeed, the more you analyse semantic measures, the harder it is to restrict them to specific boxes; the analogy can be made with the relationship between cognitive models of similarity[1].

Several classifications of measures have been proposed. The most common one is to distinguish measures according to the elements of the graph that they take into account [Pesquita et al., 2009a]. This classification distinguishes three approaches: (i) *edge-based* – measures focusing on relationship analysis, (ii) *node-based* – measures based on node analysis, and (iii) *hybrid measures* – measures which mix both approaches. In the

[1] Refer to dedicated Section 1.4 and more particularly to efforts made for the unification of the various models.

literature, edge-based measures often refer to structural measures, node-based measures refer to measures framed in the feature-model and those based on information theory. Hybrid measures are those which implicitly or explicitly mix several paradigms.

Another interesting way to classify measures is to study whether they are (i) *intentional*, i.e., based on the explicit definition of the concepts expressed by the taxonomy, (ii) *extensional*, i.e., based on the analysis of the realisations of the concepts (i.e., instances), or (iii) *hybrid*, measures which mix both intentional and extensional information about concepts. Refer to [Aimé, 2011; Gandon et al., 2005][1] for examples of such classifications.

In some cases, authors will mix several types of classifications to present measures. In this section, we will introduce the measures according to the four approaches presented above: (i) structural, (ii) feature-based, (iii) framed in information theory, and (iv) hybrid. We will also specify the extensional, intentional, or hybrid nature of the measures.

Numerous concept-to-concept measures have been defined for trees, i.e. special graphs without multiple inheritances. In the literature, these measures are generally considered to be applied *as it is* on graphs. However, in graphs, some adaptations deserve to be made and several components of measures generally need to be redefined in order to avoid ambiguity, e.g., to be implemented on computer software. For the sake of clarity, we first highlight the diversity of proposals by introducing the most representative measures defined according to the different approaches. In most cases, measures will be presented according to their original definitions. When the measures have been defined for trees, we will not necessarily stress the modifications which must be taken into account for them to be used on DAGs. These modifications will be discussed after the introduction of the diversity of measures. For convenience, `subClassOf` relationships will be denoted *isa* (there is no ambiguity with `isA` since G_T only contains concepts).

3.5.1 Structural approach

Structural measures rely on the graph-traversal approaches presented in Section 3.4.1.1 (e.g., shortest path techniques, random walk approaches). They focus on the analysis of the interconnection between concepts to estimate their similarity. However, most of the time, they consider specific tuning in order to take into account specific properties and interpretations induced by the transitivity of the taxonomic relationships. In this context, some authors, e.g., [Hliaoutakis, 2005], have linked this approach to the spreading activation theory [Collins and Loftus, 1975]. The similarity is in this case seen as a function of propagation between concepts through the graph.

[1] In french.

Back in the eighties, Rada et al. [1989] expressed the taxonomic distance of two concepts defined in a taxonomic tree as a function of the shortest path linking them[1]. We denote $sp(u, isa^*, v)$ the shortest path between two concepts u and v, i.e., the path of minimal length in $\{u, isa^*, v\}$. Remember that the length of a path has been defined as the sum of the weights associated to the edges which compose the path. When the edges are not weighted we refer to the edge-counting strategy – the length of the shortest path is the number of edges it contains. The taxonomic distance is therefore defined by[2]:

$$dist_{Rada}(u, v) = sp(u, isa^*, v) \tag{3.12}$$

Distance-to-similarity conversions can also be applied to express a similarity from a distance. A semantic similarity can therefore be defined in a straightforward manner:

$$sim_{Rada}(u, v) = \frac{1}{dist_{Rada}(u, v) + 1} \tag{3.13}$$

Notice the importance of considering the transitive reduction of the tree/graph to obtain coherent results using measures based on the shortest path. In the following presentation, we consider that the taxonomy G_T doesn't contain redundant relationships (here redundancies refer to relationships which can be inferred due to the transitivity of taxonomic relationships).

In a tree, the shortest path $sp(u, isa^*, v)$ contains a unique common ancestor of u and v. This common ancestor is the Least Common Ancestor (LCA)[3] of the two concepts according to any function θ (since the θ function is monotonically decreasing)[4]. Therefore, in trees, we obtain $dist_{Rada}(u, v) = sp(u, isa, LCA(u, v)) + sp(v, isa, LCA(u, v))$.

Several issues with the shortest path techniques have been formulated. The edge-counting strategy, or more generally any shortest path approach with uniform edge weight, has been criticised for the fact that the distance represented by an edge linking two concepts does not take concept specificities/salience into account[5]. Several modifications have therefore been proposed to break this constraining uniform appreciation of edges. Implicit or explicit models defining non-uniform strength of connotation between

[1] It is worth noting that they didn't invent the notion of shortest path in a graph. In addition, in Foo et al. [1992], the authors refer to a measure proposed by Gardner et al. [1987] to compare concepts defined in a conceptual graph using the shortest path technique.

[2] In this chapter, equations named *dist* refer to taxonomic distances.

[3] The Least Common Ancestor is also denoted as the Last Common Ancestor (LCA), the Most Specific Common Ancestor (MSCA), the Least Common Subsumer/Superconcept (LCS) or Lowest SUPERordinate (LSuper) in the literature.

[4] Here relies the importance of applying the transitive reduction of the taxonomic graph/tree, redundant taxonomic relationships can challenge this statement and therefore heavily impact the semantics of the results.

[5] As an example, Foo et al. [1992] quotes remarks made in Sowa personal communication.

concepts have therefore been introduced e.g., [Richardson et al., 1994; Sussna, 1993; Young Whan and Kim, 1990].

One of the main challenges of designers of semantic measures over the years has therefore been to refine measures by (implicitly or explicitly) taking advantage of semantic evidence related to concept specificity and the strength of connotation between concepts. The different strategies and factors used to appreciate concept specificity as well as strength of connotations have already been introduced in Section 3.3. Another use of the various semantic evidence which can be extracted from G_T has been to normalise the measures. As an example, Resnik [1995] suggested considering the maximal depth of the taxonomy to bound the edge-counting strategy:

$$sim_{Resnik-eb}(u,v) = 2 \cdot depth(G_T) - sp(u, isa, LCA(u,v)) - sp(v, isa, LCA(u,v)) \quad (3.14)$$

To simulate non uniform edge weighing, Leacock and Chodorow [1998][1] introduced a logarithmic transformation of the edge counting strategy:

$$sim_{LC}(u,v) = -log\left(\frac{N}{2 \cdot depth(G_T)}\right) = log(2 \cdot depth(G_T)) - log(N) \quad (3.15)$$

with N the cardinality of the union of the sets of nodes involved in the shortest paths $sp(u, isa, LCA(u,v))$ and $sp(v, isa, LCA(u,v))$.

Authors have also proposed taking into account the specificity of compared concepts, e.g., [Mao and Chu, 2002], sometimes as a function of the depth of their LCA, e.g.,[Pekar and Staab, 2002; Wang et al., 2012b; Wu and Palmer, 1994]. As an example, Wu and Palmer [1994] proposed expressing the similarity of two concepts as a ratio taking into account the shortest path linking the concepts as well as the depth of their LCA.

$$sim_{WP}(u,v) = \frac{2 \cdot depth(LCA(u,v))}{2 \cdot depth(LCA(u,v)) + sp(u, isa, LCA(u,v)) + sp(v, isa, LCA(u,v))} \quad (3.16)$$

This function is of the form:

$$f(x,y,z) = \frac{x}{(x + (y+z)/2)}$$

with x the depth of the LCA of the two concepts u, v and $y+z$ the length of the shortest path linking u, v. It is easy to see that for any given non-null length of the shortest path, this function increases with x; otherwise stated, to a given shortest path length, $sim_{WP}(u,v)$ increases with the depth of $LCA(u,v)$. In addition, as expected, for a given

[1]Note that according to Resnik [1995], this approach was already proposed in an 1994 unpublished paper by the same authors [Leacock and Chodorow, 1994].

depth of the LCA, the longer the shortest path which links u, v, less similar they will be considered.

Based on a specific expression of the notion of depth, a parameterised expression of sim_{WP} has been proposed in Wang and Hirst [2011]. A variation was also proposed by Pekar and Staab [2002]:

$$sim_{PS}(u,v) = \frac{depth(LCA(u,v))}{sp(u,isa,LCA(u,v)) + sp(v,isa,LCA(u,v)) + depth(LCA(u,v))} \tag{3.17}$$

Zhong et al. [2002] also proposed comparing concepts taking into account the notion of depth:

$$dist_{Zhong}(u,v) = 2 \cdot \frac{1}{2k^{depth(LCA(u,v))}} - \frac{1}{2k^{depth(u)}} - \frac{1}{2k^{depth(v)}} \tag{3.18}$$

with $k > 1$ a factor defining the contribution of the depth.

In a similar fashion, Li et al. [2003, 2006] defined a parametric function in which both the length of the shortest path and the depth of the LCA are taken into account:

$$sim_{LB}(u,v) = e^{-\alpha dist_{Rada}(u,v)} \times df(u,v) \tag{3.19}$$

with,

$$df(u,v) = \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}$$

The parameter h corresponds to the depth of the LCA of the compared concepts, i.e. $h = depth(LCA(u,v))$. The parameter $\beta > 0$ is used to tune the depth factor (df) and to set the importance given to the degree of specificity of concepts. The function used to express df corresponds to the hyperbolic tangent which is normalised between 0 and 1. It defines the degree of non-linearity to associate to the depth of the LCA. In addition, $\alpha \geq 0$ controls the importance of the taxonomic distance expressed as a function of the length of the shortest path linking the two concepts.

Approaches have also been proposed to modify existing measures in order to obtain particular properties. As an example, Slimani et al. [2006] proposed an adaptation of the measure proposed by Wu and Palmer [1994] (Equation 3.16) in order to avoid the fact that, in some cases, neighbour concepts can be estimated as more similar than ordered concepts. To this end, the authors introduced sim_{tbk} which is based on a factor used to penalise concepts defined in the neighbourhood:

$$sim_{tbk}(u,v) = sim_{WP}(u,v) \times pf(u,v) \tag{3.20}$$

with,

$$pf(u,v) = (1-\lambda)(min(depth(u), depth(v)) - depth(G_T)) + \lambda(depth(u) + depth(v) + 1)^{-1}$$

In the same vein [Ganesan et al., 2012; Shenoy et al., 2012] recently proposed alternative measures answering the same problem. The approach proposed by Shenoy et al. [2012] is presented[1]:

$$sim_{Shenoy}(u,v) = \frac{2 \cdot depth(G_T) \cdot e^{-\lambda L / depth(G_T)}}{depth(u) + depth(v)} \tag{3.21}$$

with L the weight of the shortest path computed by penalising paths with multiple changes of type of relationships, e.g. a path following the pattern $< isa, isa^-, isa, \ldots >$. Note that the penalisation of paths inducing complex semantics, e.g., which involves multiple types of relationships, was already introduced in [Bulskov et al., 2002; Hirst and St-Onge, 1998].

Several approaches have also been proposed to consider density of concepts, e.g., through analysis of cluster of concepts [Al-Mubaid and Nguyen, 2006]. Other adaptations also proposed taking into account concepts' distance to leaves [Wu et al., 2006], and variable strengths of connotation considering particular strategies [Lee et al., 1993; Zhong et al., 2002], e.g., using IC variability among two linked concepts or multiple topological criteria [Alvarez et al., 2011; Jiang and Conrath, 1997].

In terms of the spreading activation theory, measures have also been defined as a function of transfer between the compared concepts [Schickel-Zuber and Faltings, 2007]. Wang et al. [2007] use a similar approach based on a specific definition of the strength of connotation. Finally, pure graph-based approaches defined for the comparison of nodes can also be used to compare concepts defined in a taxonomy (refer to Section 3.4.1.1). As an example, Garla and Brandt [2012] and Yang et al. [2012] define semantic similarity measures using random walk techniques such as the personalised page rank approach.

As we have seen, most structural semantic similarity measures are extensions or refinements of the intuitive shortest path distance considering intrinsic factors to consider both the specificity of concepts and variable strengths of connotations. Nevertheless, the algorithmic complexity of the shortest path algorithms hampers the suitability of these measures for large semantic graphs[2]. To remedy this problem, we have seen that shortest path computation can be substituted by approximation based on the depth

[1] Note that we assume that the paper contains an error in the equation defining the measure. The formula is considered to be $X/(Y+Z)$, not $X/Y+Z$ as written in the paper.

[2] A linear algorithm in $O(C+E)$ exists for DAGs; nevertheless search for $sp(u, isa^*, v)$ requires the consideration of cyclic graphs for which algorithms, such as Dijkstra's, are in $O(C^2)$ or $O(E + C \cdot logC)$ using sophisticated implementation.

of the LCA of the compared concepts[1], and that several measures proposed by graph theory can be used instead.

Towards other estimators of semantic similarity

Most criticisms related to the initial edge-counting approach were linked to the uniform consideration of edge weights. As we have seen, to remedy this, several authors proposed considering a great deal of semantic evidence to differentiate strengths of connotation between concepts.

One of the central findings conveyed by early developments in structure-based measures is that the similarity function can be broken down into several components, in particular those distinguished by the feature model: commonality and difference. Indeed, the shortest path between two concepts can be seen as the difference between the two concepts (considering that all specialisation add properties to a concept). More particularly, in trees, or under specific constraints in graphs, we have seen that the shortest path linking two concepts contains their LCA. It can therefore be broken down into two parts corresponding to the shortest paths which link compared concepts to their LCA: in most cases, $sp(u, isa^*, v) = sp(u, isa, LCA(u, v)) + sp(v, isa, LCA(u, v))$. Therefore, the LCA can be seen as a *proxy* which partially summarises the commonality of compared concepts[2]. Distances between compared concepts and their LCA can therefore be used to estimate their differences.

The fact that measures can be broken down into specific components evaluating commonalities and differences is central in the design of the approaches which will further be introduced: the feature-based strategy and the information theoretical strategy. As we will see, they mainly define alternative strategies to characterise compared concepts in order to express semantic measures as a function of their commonalities and differences.

[1] The algorithmic complexity of the LCA computation is significantly lower than the computation of the shortest path: constant after linear preprocessing [Harel and Tarjan, 1984].

[2] The LCA only partially summarises commonality. Indeed, it can only be considered as an upper-bound of the commonality since highly similar concepts (`Man`, `Women`) may have a general concept for LCA (`LivingBeing`). This LCA will only encompass a partial amount of their commonalities. Please refer to Section 3.3.2. In addition, notice that in some cases the set of NCCAs contains other concepts than the LCA.

3.5.2 Feature-based approach

The feature-based approach generally refers to measures which rely on a taxonomic interpretation of the feature model proposed by Tversky [1977] (introduced in Section 1.4.2). However, as we will see, contrary to the original definition of the feature model, this approach is not necessarily framed in set theory[1].

The main idea is to represent concepts as collections of features, i.e., characteristics describing the concepts, to further express measures based on the analysis of their common and distinct features. The score of the measures will only be influenced by the strategy adopted to characterise concept features[2], and the strategy adopted for their comparison.

As we will see, the reduction of concepts to collections of features makes it possible to set the semantic similarity estimation back in the context of classical binary similarity or distance measures (e.g., set-based measures).

An approach commonly used to represent the features of a concept is to consider its ancestors as features[3]. We denote $A(u)$ the set of ancestors of the concept u. Since the Jaccard index that was proposed 100 years ago, numerous binary measures have been defined in various fields. A survey of these measures distinguishes 76 of them in Choi et al. [2010]. Considering that the features of a concept u are defined by $A(u)$, an example of a semantic similarity measure expressed from the Jaccard index was proposed in Maedche and Staab [2001][4]:

$$sim_{CMatch}(u,v) = \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|} \tag{3.22}$$

Another example of a set-based expression of the feature-based approach is proposed in Bulskov et al. [2002]:

$$sim_{Bulskov}(u,v) = \alpha \frac{|A(u) \cup A(v)|}{|A(u)|} + (1-\alpha)\frac{|A(u) \cup A(v)|}{|A(v)|} \tag{3.23}$$

with $\alpha \in [0,1]$ a parameter used to tune the symmetry of the measure.

[1] You will recall that the feature matching function on which the feature model is based, relies on binary evaluations of the features "*In the present theory, the assessment of similarity is described as a feature-matching process. It is formulated, therefore, in terms of the set-theoretical notion of a matching function rather than in terms of the geometric concept of distance*" [Tversky and Itamar, 1978].

[2] As stressed in Schickel-Zuber and Faltings [2007], there is a narrow link with the multi-attribute utility theory [Keeney, 1993] in which the utility of an item is a function of the preference on the attributes of the item.

[3] Its implicit senses if the concept refers to a synset.

[4] This is actually a component of a more refined measure.

Rodríguez and Egenhofer [2003] also proposed a formulation derived from the *ratio model* defined by Tversky (introduced in Section 1.4.2):

$$sim_{RE}(u,v) = \frac{|A(u) \cap A(v)|}{\gamma|A(u) \setminus A(v)| + (1-\gamma)|A(v) \setminus A(u)| + |A(u) \cap A(v)|} \tag{3.24}$$

with $\gamma \in [0,1]$, a parameter that enables the tuning of the symmetry of the measure.

Sánchez et al. [2012a] define the taxonomic distance of two concepts as a function of the ratio between their distinct and shared features:

$$dist_{Sanchez}(u,v) = log_2\left(1 + \frac{|A(u) \setminus A(v)| + |A(v) \setminus A(u)|}{|A(u) \setminus A(v)| + |A(v) \setminus A(u)| + |A(u) \cap A(v)|}\right) \tag{3.25}$$

Various refinements of these measures have been proposed, e.g., to enrich concept features by taking their descendants into account [Ranwez et al., 2006].

The feature-based measures may not be intentional, i.e., they are not expected to solely rely on the knowledge defined in the taxonomy. When instances of the concepts are known, the feature of a concept can also be seen by extension and be defined on the basis of instances associated to concepts. As an example, the Jaccard index can be used to compare two ordered concepts according to their shared and distinct features, here characterised by extension:

$$sim_{JacExt}(u,v) = \frac{|\mathcal{I}(u) \cap \mathcal{I}(v)|}{|\mathcal{I}(u) \cup \mathcal{I}(v)|} \tag{3.26}$$

with $\mathcal{I}(u) \subseteq I$ the set of instances of the concept u. Note that this approach makes no sense if the desire is to compare concepts which are not ordered – the set $\mathcal{I}(u) \cap \mathcal{I}(v)$ will tend to be empty.

D'Amato et al. [2008] also define an extensional measures considering:

$$sim_{D'Amato}(u,v) = \frac{min(|\mathcal{I}(u)|, |\mathcal{I}(v)|)}{|\mathcal{I}(LCA(u,v))|}\left(1 - \frac{|\mathcal{I}(LCA(u,v))|}{|I|}\right)\left(1 - \frac{min(|\mathcal{I}(u)|, |\mathcal{I}(v)|)}{|\mathcal{I}(LCA(u,v))|}\right) \tag{3.27}$$

Classical feature-based measures summarise the features of a concept through a set representation which generally corresponds to a set of concepts or instances. However, alternative approaches can also be explored. Therefore, even if, to our knowledge, such approaches have not been defined, the features of a concept could also be represented as a set of relationships, as a subgraph, etc.

In addition, regardless of the strategy adopted to characterise the features of a concept (other concepts, relationships, instances), the comparison of the features is not necessarily driven by a set-based measure. Indeed, the collections of features can also be seen as vectors. As an example, a concept u can be represented by a vector U in a chosen real space of dimension $|C|$, e.g., considering that each dimension associated to an ancestor of u is set to 1. Vector-based measures will evaluate the distance of two concepts by studying the coordinates of their respective projections.

In this vein, Bodenreider et al. [2005] proposed the comparison of two concepts according to their representation through the Vector Space Model. Considering a concept-to-instance matrix, a weight corresponding to the IC[1] of the concept u is associated to the cell (u, i) of the matrix if the instance $i \in \mathcal{I}(u)$. The vectors representing two concepts are then compared using the classical dot product of the vectors, e.g., discussed in [Salton, 1968].

3.5.3 Information theoretical approach

The information theoretical approach relies on Shannon's information theory [Shannon, 1948]. As with the feature-based strategy, these measures rely on the comparison of two concepts according to their commonalities and differences, here defined in terms of information. This approach formally introduces the notion of salience of concepts through the definition of their informativeness – Information Content (IC) – Section 3.3.2 introduces the notion of IC.

Resnik [1995] defines the similarity of a couple of concepts as a function of the IC of their common ancestor which maximises an IC function (originally eIC), i.e., their Most Informative Common Ancestor (MICA).

$$sim_{Resnik}(u, v) = IC(MICA(u, v)) \tag{3.28}$$

Resnik's measure doesn't explicitly capture the specificities of compared concepts. Indeed, pairs of concepts with an equivalent MICA will have the same semantic similarity, whatever their respective ICs. To correct this limitation, several authors refined the measure proposed by Resnik to incorporate specificities of compared concepts. We here present the measures proposed by Lin [1998][2] – sim_{Lin} , [Jiang and Conrath, 1997] – $dist_{JC}$, [Mazandu and Mulder, 2013] – $sim_{Nunivers}$, [Pirró, 2009; Pirró and Seco, 2008]

[1] Originally the authors used the IDF but we saw that both the IC and the IDF are similar (Section 3.3.2.2).

[2] Originally defined as: $sim_{Lin}(u, v) = \frac{2 \times log(MICA(u,v))}{log(u)+log(v)}$

– sim_{PSec} and [Pirró and Euzenat, 2010b] – sim_{Faith}:

$$sim_{Lin}(u,v) = \frac{2 \cdot IC(MICA(u,v))}{IC(u) + IC(v)} \tag{3.29}$$

$$dist_{JC}(u,v) = IC(u) + IC(v) - 2 \cdot IC(MICA(u,v)) \tag{3.30}$$

$$sim_{Nunivers}(u,v) = \frac{IC(MICA(u,v))}{max(IC(u), IC(v))} \tag{3.31}$$

$$sim_{PSec}(u,v) = 3 \cdot IC(MICA(u,v)) - IC(u) - IC(v) \tag{3.32}$$

$$sim_{Faith}(u,v) = \frac{IC(MICA(u,v))}{IC(u) + IC(v) - IC(MICA(u,v))} \tag{3.33}$$

Taking into account specificities of compared concepts can lead to high similarities (low distances) when comparing general concepts. As an example, when comparing general concepts using sim_{Lin}, the maximal similarity will be obtained comparing a (general) concept to itself. In fact, the identity of the indiscernibles is generally ensured (except for the root which generally has an IC equal to 0). However, some treatments require this property not to be respected. Authors have therefore proposed to lower the similarity of two concepts according to the specificity of their MICA, e.g. [Li et al., 2010; Schlicker et al., 2006]. The measure proposed by Schlicker et al. [2006] is presented:

$$sim_{Rel}(u,v) = sim_{Lin}(u,v) \times (1 - p(MICA(u,v))) \tag{3.34}$$

with $p(MICA(u,v))$ the probability of occurrence of the MICA. An alternative approach proposed by Li et al. [2010] relies on the IC of the MICA and can therefore be used without extensional information on concepts, i.e., using an intrinsic expression of the IC.

Authors have also proposed to characterise the information carried by a concept by summing the ICs of its ancestors [Cross and Yu, 2011; Mazandu and Mulder, 2011]:

$$sim_{Mazandu}(u,v) = \frac{2 \cdot \sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u)} IC(c) + \sum_{c \in A(v)} IC(c)} \tag{3.35}$$

$$sim_{JacAnc}(u,v) = \frac{\sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u) \cup A(v)} IC(c)} \tag{3.36}$$

These measures can also be considered as hybrid strategies between the feature-based and information theory approaches. One can consider that these measures rely on a redefinition of the way to characterise the information conveyed by a concept (by summing the IC of the ancestors). Other interpretations can simply consider that features are weighted. Thus, following the set-based representations of features, authors have also studied these measures as fuzzy measures [Cross, 2004, 2006; Cross and Sun, 2007; Cross and Yu, 2010, 2011; Popescu et al., 2006], e.g., defining the membership function of a feature corresponding to a concept as a function of its IC.

Finally, other measures based on information theory have also been proposed, e.g., [Cazzanti and Gupta, 2006; Maguitman and Menczer, 2005; Maguitman et al., 2006]. As an example, in Maguitman and Menczer [2005] the similarity is estimated as a function of prior and posterior probability regarding instances and concept membership.

3.5.4 Hybrid approach

Other techniques take advantage of the various aforementioned paradigms. Among the numerous proposals, [Bin et al., 2009; Jiang and Conrath, 1997] defined measures in which density, depth, strength of connotation and ICs of concepts are taken into account. We present the measure proposed by Jiang and Conrath [1997][1]. The strength of association $w(u, v)$ between two concepts u, v is defined as follows:

$$w(u,v) = (\beta + (1-\beta))\frac{\overline{dens}}{|children(v)|} \times \left(\frac{depth(v)+1}{depth(v)}\right)^{\alpha} \times (IC(u) - IC(v)) \times T(u,v)$$

The factor $\overline{dens}$ refers to the average density of the whole taxonomy, see Jiang and Conrath [1997] for details. The factors $\alpha \geq 0$ and $\beta \in [0,1]$ control the importance of the density factor and the depth respectively. $T(u,v)$ defines weights associated to predicates. Finally, the similarity is defined by the weight of the shortest path which links compared concepts and which contains their LCA:

$$dist_{JC-Hybrid}(u,v) = \sum_{(s,p,o) \in sp(u,isa,LCA(u,v)) \cup sp(v,isa,LCA(u,v))} w(s,o)$$

Defining $\alpha = 0$, $\beta = 1$ and $T(u,v) = 1$, we obtain the information theoretical measure proposed by the same authors, i.e., $dist_{JC}(u,v) = IC(u) + IC(v) - 2 \cdot IC(MICA(u,v))$ (Equation 3.30).

[1] This measure is a parametric distance. Couto et al. [2003] discuss the implementation, Othman et al. [2008] propose a genetic algorithm which can be used to tune the parameters and Wang and Hirst [2011] propose a redefinition of the notion of depth and density initially proposed.

Singh et al. [2013] proposed a mixing strategy based on [Jiang and Conrath, 1997] IC-based measure $dist_{JC}$. They consider transition probabilities between concepts relying on a depth-based estimation of the strength of connotation.

Rodríguez and Egenhofer [2003] also proposed mixing a feature-based approach considering structural properties such as the concepts' depth. Finally, Paul et al. [2012] defined multiple measures based on an aggregation of several existing measures.

3.5.5 Considerations when comparing concepts in semantic graphs

Several measures introduced in the previous sections were initially defined to compare concepts expressed in a tree. However, despite the fact that this subject is almost never discussed in the literature, several considerations must be taken into account in order to estimate the similarity of concepts defined in a semantic graph [Blanchard, 2008][1] – please refer to notations introduced in Section 3.2.

3.5.5.1 Shortest path

A tree is a specific type of graph in which multiple inheritances cannot be encountered, i.e. $\forall c \in C, |parents(c)| < 2$. This implies that two concepts u, v which are not ordered will have no common descendants, i.e., $G_T^-(u) \cap G_T^-(v) = \emptyset$. Therefore, if there is no redundant taxonomic relationship, the shortest path which links u, v always contains a single common ancestor of u, v: $LCA(u, v)$. However, in a graph, since two non-ordered concepts u, v can have common descendants, i.e., $G_T^-(u) \cap G_T^-(v) \neq \emptyset$, the shortest path which links u, v can in some cases not contain one of their common ancestors. Figure 3.5 illustrates the modifications induced by multiple inheritances.

In Figure 3.5, the shortest path linking the two non-ordered concepts $C5$ and $C7$ in the tree (i.e. without considering red dotted edges) is $[C5-C3-C1-Root-C2-C4-C7]$. However, if we consider multiple inheritances (red dotted edges), it is possible to link $C5$ and $C7$ through paths which do not contain one of their common ancestors, e.g., $[C5-C3-C6-C4-C7]$ or even $[C5-C8-C7]$. The shortest path which contains a common ancestor of the compared concepts is defined in the search space corresponding to the graph $G_T^+(u) \cup G_T^+(v)$. In practice, despite the fact that in most graphs $G_T^-(u) \cap G_T^-(v) \neq \emptyset$ (for two non-ordered concepts), it is commonly admitted that the shortest path must contain a single ancestor of the two compared concepts. Given this constraint, the edge-counting taxonomic distance of u and v in $G_T^+(u) \cup G_T^+(u)$ is generally (implicitly[2]) defined by: $dist_{SP}(u, v) = sp(u, isa, LCA(u, v)) + sp(v, isa, LCA(u, v))$.

[1] In french.

[2] Generalisation of measures defined from trees to graphs is poorly documented in the literature.

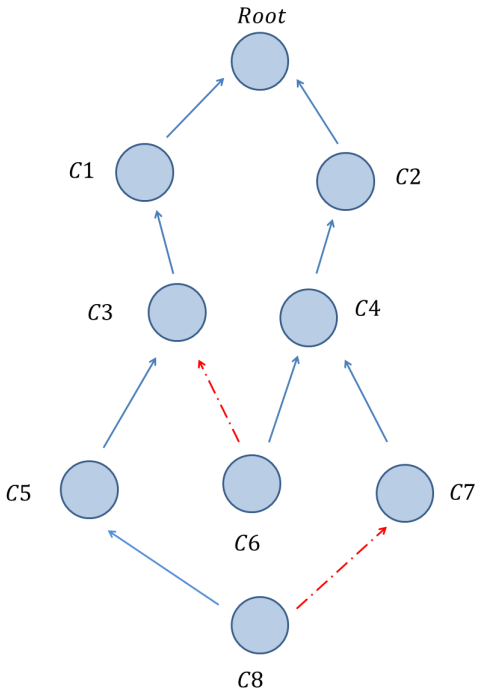

FIGURE 3.5: The graph composed of the plain (blue) edges is a taxonomic tree, i.e., it doesn't contain concepts with multiple parents. If the (red) dotted relationships are also considered, the graph is a directed acyclic graph (e.g., a taxonomic graph)

Note that when non comparable common ancestors (NCCAs) are shared between compared concepts, the ancestor which maximises the similarity is expected to be considered. Depending on the θ function which is used, the shortest path doesn't necessarily involve the concept of the NCCAs which maximise θ, e.g. the deeper in the taxonomy. As an example, in order to distinguish which NCCA to consider, Schickel-Zuber and Faltings [2007] took into account a mix between depth and reinforcement (number of different paths leading from one concept to another).

Nevertheless, the shortest path techniques can also be relaxed to consider paths which do not involve common ancestors or which involve multiple common ancestors:

$$sim_{SP-R}(u,v) = \frac{1}{sp(u, isa^*, v) + 1}$$

3.5.5.2 Notion of depth

The definition of the notion of depth must also be reconsidered when the taxonomy is not a tree. Remember that, in a tree without redundancies, the depth of a concept has been defined as the length of the shortest path linking the concept to the root. The depth of a concept is a simple example of specificity estimator. In a tree, this estimator makes perfect sense since the depth of a concept is directly correlated to its number of ancestors since $depth(c) = |A(c)| - 1$.

In a graph, or in a tree with redundant taxonomic relationships, we must ensure that the depth is monotonically decreasing according to the ordering of concepts. As an example, to apply depth-based measures to graphs, we must ensure that $depth(LCA(u, v))$ is lower or equal to both $depth(u)$ and $depth(v)$. To this end, the maximal depth of a concept must be used, i.e., the length of the longest path in $\{u, isa, \top\}$, denoted $lp(u, isa, \top)$. As an example, the measure proposed by Pekar and Staab [2002] – Equation 3.17 – is therefore implicitly generalised to:

$$sim_{PS-G}(u, v) = \frac{lp(LCA(u, v), isa, \top)}{lp(u, isa, LCA(u, v)) + lp(v, isa, LCA(u, v)) + lp(LCA(u, v), isa, \top)}$$

3.5.5.3 Notion of least common ancestors

Most measures which have been presented take advantage of the notions of LCA or MICA. However, in graphs, these measures do not consider the whole set of NCCAs – denoted $\Omega(u, v)$ for the concepts u and v. To remedy this, several authors have proposed adaptations of existing measures. As an example, Couto and Silva [2011]; Couto et al. [2005] proposed *GraSM* and *DiShIn* strategies.

In [Couto et al., 2005] the authors proposed the modification of information theoretical measures based on the notion of MICA. The authors recommended substituting the IC of the MICA by the average of the ICs of the concepts which compose the set of NCCAs. A redefinition of the measure proposed by [Lin, 1998] – Equation 3.29 – is presented:

$$\begin{aligned} sim_{Lin-GraSM}(u, v) &= \frac{2 \cdot \frac{\sum_{c \in \Omega(u,v)} IC(c)}{|\Omega(u,v)|}}{IC(u) + IC(v)} \\ &= \frac{2 \cdot \sum_{c \in \Omega(u,v)} IC(c)}{|\Omega(u, v)| \times (IC(u) + IC(v))} \end{aligned} \tag{3.37}$$

Wang et al. [2012b] also proposed averaging the similarity between the concepts according to their multiple NCCAs:

$$sim_{Wang}(u,v) = \frac{\sum_{a \in \Omega(u,v)} \frac{2 \cdot depth(a)^2}{d_a(\top,u) \times d_a(\top,v)}}{|\Omega(u,v)|}$$

With $d_a(\top, u)$ the average length of the set of paths which contain the concept a and which link the concept u to the root of the taxonomy ($\top$).

As we have underlined, numerous approaches have been defined to compare pairs of concepts defined in a taxonomy, these measures can be used to compare any pair of nodes defined in a poset. Table 3.2 to Table 3.5 present some properties of a selection of measures defined to compare pairs of concepts.

3.5.6 List of pairwise semantic similarity measures

Several semantic measures which can be used to compare concepts defined in a taxonomy or any pair of elements defined in a poset. Measures are ordered according to their date of publication. Other contributions studying some properties of pairwise measures can be found in Slimani [2013]; Yu [2010]. IOI: Identify of the Indiscernibles. Some of the values associated to specific measures have not been complete yet. This is generally because the reference associated to the measure was not available or because the properties of the measure are still under study.

Structural Measures					
Name	**Type**	**Const.**	**Range**	**IOI**	**Comment**
Shortest Path strategy	Sim / Rel	None	$\mathbb{R}_+$	Yes	Measures are defined as a function of the weight of the shortest path linking compared concepts. Several modifications can be considered depending on the strategy adopted, e.g., weighing of the relationships (according to their predicate), constraints on the inclusion of a common ancestor of the compared concepts, etc.
Rada et al. [1989]	Dist (ISA)	DAG	$\mathbb{R}_+$	Yes	Specific shortest path strategy with uniform edge weight. The shortest path is constrained to containing the LCA of the compared concepts.
Young Whan and Kim [1990]	x	x	x	x	x
Lee et al. [1993]	x	x	x	x	x
Sussna [1993]	Dist	RDAG	$\mathbb{R}_+$	Yes	Originally defined as a parametric semantic relatedness. Under specific constraints, this measure can be used as a semantic similarity. Shortest path technique which take into account non-uniform strengths of connotation. This latter is tuned according to the depth of the compared concepts and to the weights associated to predicates.

Richardson et al. [1994]	x	x	x	x	The authors propose several intrinsic metrics (e.g., depth, density) to weigh the relationships and define hybrid measures by mixing the structural and information theoretical approach. No measure explicitly is defined.
Wu and Palmer [1994]	Sim	RDAG	[0,1]	Yes	The similarity is assessed as a function of the depth of the compared concepts and the depth of their LCA.
Leacock and Chodorow [1994, 1998]	Sim	RDAG	$\mathbb{R}_+$	Yes	Rada et al. [1989] formulation penalising long shortest path between the compared concepts according to the depth of the taxonomy.
Resnik [1995]	Sim	RDAG	$[0, 2D]$	Yes	Similarity based on the shortest path technique which has been bound by (twice) the depth of the taxonomy (D).
Hirst and St-Onge [1998]	Sim / Rel	None	$\mathbb{R}_+$	Yes	Shortest path penalising multiple changes of predicate. Can be used as a similarity or relatedness measure depending on the relationships which are considered.
Zhong et al. [2002]	Dist	RDAG	$[0,max[$	Yes	Taxonomic distance taking into account the depth of the compared concepts. With max defined as $max = 1/k^{depth(LCA(u,v))}$ with k a given constant.

Pekar and Staab [2002]	Sim	RDAG	[0,1]	Yes	Shortest path technique which takes into account the depth of the LCA of the compared concepts.
Mao and Chu [2002]	Sim	DAG	$\mathbb{R}_+$	No	Modification of Rada et al. [1989] formulation taking into account concept specificity as a non-linear function of the number of descendants a concept has.
Li et al. [2003]	x	x	x	x	x
Li et al. [2006]	Sim	RDAG	$\mathbb{R}_+$	No	Measure considering both the length of the shortest path linking the compared concepts and their depth.
Ganesan et al. [2003]	Sim	x	x	x	Refer to the function named *leafsim* in the publication.
Yu et al. [2005]	Sim	RDAG	[0,1]	Yes	Measure allowing non-null similarity only to ordered pairs of concepts.
Wu et al. [2006]	Sim	RDAG	[0,1]	No	Take into account compared concepts (i) commonality (length of the longest shared path from the concepts to the root), (ii) specificity (defined as a function of the shortest path from the concept to the leaves it subsumes) and (iii) local distance (Rada et al. [1989] distance).

Slimani et al. [2006]	Sim	RDAG	[0,1]	Yes	Modification of Wu and Palmer [1994] measure to avoid cases in which neighbours of a concept might have higher similarity values than concepts which are ordered with it.
Blanchard et al. [2006]	Sim	RDAG	[0,1]	No	Measure which compare two concepts w.r.t their depth and the depth of their LCA. Originally defined for trees and extended for DAG in [Blanchard, 2008]
Nagar and Al-Mubaid [2008]	Sim	DAG	$\mathbb{R}_+$	Yes	Use a modification of shortest path constrained by the LCA.
Cho et al. [2003]	Sim	DAG	$\mathbb{R}_+$	No	Multiple factors are considered to take into account the specificity of the compared concepts.
Alvarez and Yan [2011] (SSA)	Sim / Rel	RDAG	[0,1]	No	This measure relies on three components evaluating (i) the shortest path linking the compared concepts (a weighing scheme is applied to the graph), (ii) their LCA, and (iii) their literal definitions.
Wang et al. [2012b]	Sim	RDAG	[0,1]	Yes	Approach taking into account the depth of the compared concepts, as well as the depth of all their DCAs.

Shenoy et al. [2012]	Sim	RDAG	x	x	Modification of Wu and Palmer [1994] measure to avoid cases in which neighbours of a concept might have higher similarity values than concepts which are ordered with it.
Ganesan et al. [2012]	Sim	RDAG	x	x	Modification of Wu and Palmer [1994] measure to avoid cases in which neighbours of a concept might have higher similarity values than concepts which are ordered with it.

TABLE 3.2: Semantic similarity measures or taxonomic distances defined using a structural approach. These measures can be used to compare a pair of concepts defined in a taxonomy or any pair of elements defined in a partially ordered set

Information theoretical Measures					
Name	**Type**	**Const.**	**Range**	**IOI**	**Comment**
Resnik [1995]	Sim	DAG	$[0, 1]$, $[0, +\infty[$	No	The similarity is defined as the IC of MICA. The range of the measure depends on the IC.
Jiang and Conrath [1997]	Dist	DAG	[0,1]	Yes	The taxonomic distance computed as a function of the IC of the compared classes and their MICA.
Lin [1998]	Sim	DAG	[0,1]	Yes	The similarity is computed as a ratio between the IC of the MICA of compared classes and the sum of their respective ICs.
Schlicker et al. [2006]	Sim	DAG	[0,1]	No	Modification of Lin [1998] in order to take into account the specificity of the MICA, i.e., to avoid high score of similarity comparing two general classes (due to the ratio).
Couto et al. [2007]	Sim	DAG	[0,1]	No	Derivative of Lin [1998] measure in which all the DCAs of the compared classes are taken into account.
Yu et al. [2007b]	Sim	DAG	$[0, +\infty[$	No	Total Ancestry Measure (TAM) – measure based on a specific definition of the LCA.

Pirró [2009]; Pirró and Seco [2008]	Sim	DAG	$[0, x]$	No	With x the maximal IC value. Formulation similar to Jiang and Conrath [1997] but which gives more importance to the informativeness of the MICA.
Li et al. [2010]	Sim	DAG	[0,1]	No	Lin [1998] measure modified to take specificity into account, i.e. to avoid high score of similarity comparing two general classes.
Pirró and Euzenat [2010b]	Sim	DAG	[0,1]	Yes	Ratio formulation similar to the measure proposed by Lin [1998] but which gives more importance to the difference between the compared concepts.
Mazandu and Mulder [2011] simDIC	Sim	DAG	[0,1]	Yes	Measure similar to Lin [1998] but which uses a new approach to characterise the IC of a concept.
Mazandu and Mulder [2013] Sim Nuniver	Sim	DAG	[0,1]	Yes	IC of the MICA of the compared classes divided by the maximal IC of the compared classes.

TABLE 3.3: Semantic similarity measures or taxonomic distances defined using an information theoretical approach. These measures can be used to compare a pair of concepts defined in a taxonomy or any pair of elements defined in a partially ordered set

Feature-based Measures					
Name	**Type**	**Const.**	**Range**	**IOI**	**Comment**
Maedche and Staab [2001]	Sim	DAG	[0,1]	Yes	Feature-based expression relying on the Jaccard index.
Bodenreider et al. [2005]	Sim	DAG	[0,1]	x	Cosine similarity on a vector-based representation of the classes. The vector representation is built according to the set of instances of the classes.
Ranwez et al. [2006]	Dist	DAG	$\mathbb{R}_+$	Yes	The distance is defined as a function of the number of descendants of the LCA of the compared concepts. This measure respects distance axioms (i.e., positivity, symmetry, triangle inequality).
Jain and Bader [2010]	Sim	DAG	[0,1]	No	Build a meta-graph reducing the original ontology into cluster of related concepts. Similarity is assessed through a specific function evaluating LCA information content.
Batet et al. [2010b]	Sim	DAG	$\mathbb{R}_+$	Yes	Comparison of the classes according to their ancestors. Formulation expressed as a distance converted to a similarity using negative log.

TABLE 3.4: Semantic similarity measures or taxonomic distances designed using a feature-based approach. These measures can be used to compare a pair of concepts defined in a taxonomy or any pair of elements defined in a partially ordered set

Hybrid Measures					
Name	**Type**	**Const.**	**Range**	**IOI**	**Comment**
Couto et al. [2003]; Jiang and Conrath [1997]; Othman et al. [2008]	Sim / Dist	RDAG	[0,1]	Var.	Strategy based on the shortest path constrained by the LCA of the compared classes. Relationships are weighted according to the difference of IC of the classes they link.
Al-Mubaid and Nguyen [2006]	Sim	DAG	$[0, +\infty[$	Yes	Assigns cluster(s) to classes. The similarity is computed considering multiple metrics.
Wang et al. [2007]	Sim/Rel	RDAG	[0,1]	Yes	This measure as originally been defined as a semantic relatedness. It can also be used to compute semantic similarity. It relies on a non-linear approach to characterise the strength of connotation and a specific approach to characterise the informativeness of a concepts.
Alvarez and Yan [2011]	Sim / Rel	RDAG	[0,1]	No	Exploits three components evaluating concepts, their shortest path (a weighting scheme is applied to the graph), their LCA, and their literal definitions.
Paul et al. [2012]	Sim	x	x	x	Multiple approaches are mixed

TABLE 3.5: Semantic similarity measures or taxonomic distances designed using an hybrid approach. These measures can be used to compare a pair of concepts defined in a taxonomy or any pair of elements defined in a partially ordered set

3.6 Semantic similarity between groups of concepts

Two main approaches are commonly distinguished to introduce semantic similarity measures designed for the comparison of two sets of concepts, i.e., *groupwise measures*:

- *Direct approach*, the measures which can be used to directly compare the sets of concepts according to information characterising the sets w.r.t the information defined in the taxonomy.
- *Indirect approach*, the measures which assess the similarity of two sets of concepts using one or several pairwise measures, i.e. measures designed for the comparison of a pair of concepts. They are generally simple aggregations of the scores of similarities associated to the pairs of concepts defined in the Cartesian product of the two compared sets.

Note that the sets are generally expected to not contain semantically redundant concepts, i.e., they do not contain any pair of ordered concepts – $\forall (u, v) \in X, u \not\preceq v \wedge v \not\preceq u$.

Once again, a large diversity of measures have been proposed, some of which are presented in the next subsections.

3.6.1 Direct approach

The direct approach corresponds to a generalisation of the approaches defined for the comparison of pairs of concepts in order to compare two sets of concepts. It is worth noting that classical set-based approaches can be used. The sets can also be compared through their vector representations, e.g., using the cosine similarity measure. Nevertheless, these measures are in most cases not relevant to be used considering the semantics they convey – they do not take into account the similarity of the elements composing compared sets[1], e.g., $sim(\{\texttt{Man}, \texttt{Girl}\}, \{\texttt{Women}, \texttt{Boy}\}) = 0$.

3.6.1.1 Structural approach

Considering $G_T^+(X)$ as the graph induced by the union of the ancestors of the concepts which compose the set X, Gentleman [2007] defined the similarity of two sets of concepts (U, V) according to the length of the longest $sp(c, isa, \top)$ which links the concept $c \in G_T^+(U) \cap G_T^+(V)$ to the root ($\top$).

[1] These simple approaches are generally used when the compared sets contain semantically redundant concepts.

3.6.1.2 Feature-based approach

The feature-based measures are characterised by the approach adopted to express the features of a set of concepts.

Several measures have been proposed from set-based measures. We introduce sim_{UI} [Gentleman, 2007][1], and the Normalised Term Overlap measure sim_{NTO} [Mistry and Pavlidis, 2008]. For convenience, we consider $C_T^+(X)$ as the set of concepts contained in $G_T^+(X)$:

$$sim_{UI}(U,V) = \frac{|C_T^+(U) \cap C_T^+(V)|}{|C_T^+(U) \cup C_T^+(V)|} \tag{3.38}$$

$$sim_{NTO}(U,V) = \frac{|C_T^+(U) \cap C_T^+(V)|}{min(|C_T^+(U)|, |C_T^+(V)|)} \tag{3.39}$$

3.6.1.3 Information theoretical measures

Among others, Pesquita et al. [2007] proposed considering the information content of the concepts (originally an eIC expression):

$$sim_{GIC}(U,V) = \frac{\sum_{c \in C_T^+(U) \cap C_T^+(V)} IC(c)}{\sum_{c \in C_T^+(U) \cup C_T^+(V)} IC(c)} \tag{3.40}$$

3.6.2 Indirect approach

In Section 3.5, we introduced numerous measures for comparing a pair of concepts (pairwise measures). They can be used to drive the comparison of sets of concepts.

3.6.2.1 Improvements of direct measures using concept similarity

One of the main drawbacks of basic vector-based measures is that they consider dimensions as mutually orthogonal and do not exploit concept relationships. In order to remedy this, vector-based measures have been formulated to:

- Weigh dimensions considering concept specificity evaluations (e.g., IC) [Benabderrahmane et al., 2010b; Chabalier et al., 2007; Huang et al., 2007].
- Exploit an existing pairwise measure to perform vector products [Benabderrahmane et al., 2010b; Ganesan et al., 2003].

[1] Also published through the name *Term Overlap* (TO) in Mistry and Pavlidis [2008].

Therefore, pairwise measures can be used to refine the measures proposed to compare sets of concepts using a direct approach.

3.6.2.2 Aggregation strategies

A two-step indirect strategy can also be adopted in order to take advantage of pairwise measures to compare sets of concepts:

1. The similarity of pairs of concepts obtained from the Cartesian product of the two compared sets has to be computed.
2. Pairwise scores are then summarised using an aggregation strategy, also called mixing strategy in the literature.

Classic aggregation strategies can be applied (e.g. max, min, average); more refined strategies have also been proposed. Among the most commonly used we present: Max average (Max-Avg), Best Match Max – BMM [Schlicker et al., 2006] and Best Match Average – BMA [Pesquita et al., 2008]:

$$sim_{Avg}(U,V) = \frac{\sum_{u \in U} \sum_{v \in V} sim(u,v)}{|U| \times |V|} \tag{3.41}$$

$$sim_{Max-Avg}(U,V) = \frac{1}{|U|} \sum_{u \in U} max_{v \in V} sim(u,v) \tag{3.42}$$

$$sim_{BMM}(U,V) = max(sim_{Max-Avg}(U,V), sim_{Max-Avg}(V,U)) \tag{3.43}$$

$$sim_{BMA}(U,V) = \frac{sim_{Max-Avg}(U,V) + sim_{Max-Avg}(V,U)}{2} \tag{3.44}$$

3.6.3 List of groupwise semantic similarity measures

Direct Groupwise Measures						
Name	**Type**	**Approach**	**Const.**	**Range**	**IOI**	**Comment**
[Ganesan et al., 2003] Optimistic Genealogy Measure	Sim	Hybrid	RDAG	[0,1]	Yes	Feature-based approach taking into consideration structural properties during the comparison.
[Popescu et al., 2006] A	Sim	Feature-based	DAG	[0,1]	yes	Weighted Jaccard
[Popescu et al., 2006] B	Sim	Feature-based	DAG	[0,1]	x	Fuzzy Measure
[Chabalier et al., 2007]	Sim	Feature-based (Vector)	RDAG	[0,1]	Yes	Groups of concepts are represented using the Vector Space Model and compared using the cosine similarity.
[Gentleman, 2007] SimLP	Sim	Structural	RDAG	[0,1]	Yes	Similarity as a function of the longest common path found in the graph induced by the compared groups of concepts.
[Cho et al., 2007]	Sim	Feature-based	RDAG	$\mathbb{R}_+$	No	Feature-based measure taking into account the specificity of the compared concepts.
[Pesquita et al., 2008] SimGIC	Sim	Feature-based	DAG	[0,1]	Yes	Jaccard measure in which a set of concepts is represented by the concepts contained in the graph it induces.
[Sheehan et al., 2008] SSA	Sim	Feature-based	RDAG	[0,1]	Yes	Extends the notion of MICA to pair of groups of concepts then redefine the Dice coefficient.

[Ali and Deane, 2009]	Sim	Feature-based	DAG	[0,1]	No	Commonality is assessed considering shared nodes in the graph induced by the ancestors of the compared sets of concepts.
[Jain and Bader, 2010] TCSS	Sim	Feature-based	RDAG	[0,1]	No	Max Strategy considering a specific pairwise measure
[Diaz-Diaz and Aguilar-Ruiz, 2011]	Sim	Structural	RDAG	[0,1]	Yes	Distance taking into account the shortest path between the concepts and the depths of the compared concepts.
[Alvarez and Yan, 2011]	Sim/Rel	Structural	None	$\mathbb{R}_+$	Yes	Measure based on the analysis of structural properties of the graph.
[Alvarez et al., 2011] SPGK	Sim	Structural		None	Yes	The set of concepts is represented by the graph induced by the concepts it subsumes. A similarity measure is used to compare the two graphs.
[Teng et al., 2013]	Sim	x	x	x	x	x

TABLE 3.6: Semantic similarity measures or taxonomic distances designed using a direct approach. These measures can be used to compare a pair of groups of concepts defined in a taxonomy or any pair of group of elements defined in a partially ordered set

Indirect Groupwise Measures based on a direct approach						
Name	**Type**	**Approach**	**Const.**	**Range**	**IOI**	**Comment**
[Ganesan et al., 2003] GCSM	Sim	Feature-based	RDAG	[0,1]	Yes	GCSM: Generalised Cosine-Similarity Measure. Groups of concepts are represented using the Vector Space Model. Dimensions are not considered independent, i.e. the similarity of two dimensions is computed using an approach similar to the one proposed by Wu and Palmer [1994] measure. The similarity between the vector representations of two groups of concepts is estimated using to the cosine similarity.
[Huang et al., 2007]	x	x	RDAG	[0,1]	Yes	x
[Benabderrahmane et al., 2010b] Intelligo	Sim	Feature-based (Vector)	RDAG	[0,1]	Yes	Groups of classes are represented using the Vector Space Model - Also consider [Benabderrahmane et al., 2010a]. The dimensions are not considered to be independent.

TABLE 3.7: Semantic similarity measures or taxonomic distances designed using an indirect approach. These measures can be used to compare a pair of groups of concepts defined in a taxonomy or any pair of group of elements defined in a partially ordered set

Indirect Groupwise Measures (Mixing strategy)		
Mixing strategies	**Range**	**IOI**
Classic approaches Max/Min/AVG, etc.	depends	depends
Best Match Max (BMM)		
Best Match Average [Azuaje et al., 2005]		

TABLE 3.8: Semantic similarity measures or taxonomic distances designed using an indirect approach (mixing strategy). These measures can be used to compare a pair of groups of concepts defined in a taxonomy or any pair of group of elements defined in a partially ordered set

3.7 Challenges

In the light of the state-of-the-art of the large diversity of semantic measures presented in this chapter, and based on the survey presented in [Harispe et al., 2013c], this section highlights some of the challenges faced by the communities involved in the study of semantic measures.

3.7.1 Better characterise semantic measures and their semantics

Throughout the introduction of semantic measures, we have stressed the importance of controlling their semantics, i.e., the meaning of the scores they produce. This particular aspect is of major importance since the semantics of measures must explicitly be understood by end-users: it conditions the relevance to use a specific measure in a particular context.

Nevertheless, the semantics of semantic measures is generally not discussed in proposals (except some broad distinction between the notion of semantic similarity and relatedness). However, semantic similarity based on taxonomies can have different meanings depending on the assumptions on which they rely. In this introduction, we have underlined that the semantics associated to semantic measures can only be understood w.r.t: (i) the semantic proxy used to support the comparison, (ii) the mathematical properties associated to the measures, and (iii) the semantic evidence and assumptions on which the measures are based.

The semantics of the measures can therefore only be captured if a deep characterisation of semantic measures is provided. In recent decades, researchers have mainly focused on the design of semantic measures, and despite the central role of the semantics of semantic measures, few contributions have focused on this specific aspect. This can be partially explained by the fact that numerous semantic measures have been designed in order to mimic human appreciation of semantic similarity/relatedness. In this case, the

semantics to be carried by the measures is expected to be implicitly constrained by the benchmarks used to evaluate the accuracy of measures. Nevertheless, despite evaluation protocols based on *ad hoc* benchmarks being relevant to compare semantic measures in specific contexts of use, they do not give access to a deep understanding of measures and therefore do not sufficiently provide the information needed to take advantage of semantic measures in other contexts of use.

There are numerous implications involved in a better characterisation of semantic measures. We have already stressed its importance for the selection of semantic measures in specific contexts of use. Such a characterisation could also benefit cognitive sciences. Indeed, as we saw in Section 1.4, cognitive models aiming to explain human appreciation of similarity have been supported by the study of properties expected by the measures. As an example, remember that spatial models have been challenged according to the fact that human appreciation of similarity has proven not to be in accordance with axioms of distance. Therefore, characterising: (i) which semantic measures best performed according to human expectations of semantic similarity/relatedness and (ii) the properties satisfied by these measures could help cognitive scientists to improve existing models of similarity or to derive more accurate ones.

In [Harispe et al., 2013c], we have proposed an overview of the various semantic measures which have been proposed to compare units of language, concepts or instances which are semantically characterised. In Chapter 2, we distinguished various aspects of semantic measures which must be taken into account for their broad classification:

- The types of elements which can be compared.
- The semantic proxies used to extract semantic evidence on which the measures will be based.
- The canonical form adopted to represent the compared elements and therefore enable the design of algorithms for their comparison.

In Section 2.1.3, we recalled some of the mathematical properties which can be used to further characterise semantic measures. In Section 2.1.2, based on the several notions introduced in the literature, we proposed a characterisation of the general semantics which can be associated to semantic measures (e.g., similarity, relatedness, distance, taxonomic distance). Finally, throughout this introduction, and particularly in Section 3.3, we distinguished extensive semantic evidence on which semantic measures can be based, and we underlined the assumptions associated to their consideration.

We encourage designer of semantic measures to provide an in-depth characterisation of measures they propose. To this end, they can use the various aspects and properties of the measures distinguished in our survey. We also encourage the communities involved

in the study of semantic measures to better define a good semantic measure and exactly what makes one measure better than another. Within this goal, the study of the role of contexts seems to be of major importance. Indeed, as discussed in [Harispe et al., 2013c], the accuracy of measures can only be discussed w.r.t specific expectations of measures. Several other properties of measures could also be taken into account and further investigated:

- Algorithmic complexity.
- Degree of control on the semantics of the scores produced by the measures.
- The trust which can be associated to a score.
- The robustness of a measure, i.e., the capacity for a measure to produce robust scores considering the uncertainty associated to expected scores, or disturbances of the semantic proxies on which the measure relies (modification of the ontologies, corpus modifications).
- The discriminative power of the measure, i.e., the distribution of the scores produced by a measure.

3.7.2 Provide tools for the study of semantic measures

The communities studying and using semantic measures require software solutions, benchmarks, and theoretical tools to compute, compare and analyse semantic measures.

3.7.2.1 Develop benchmarks

There are a host of benchmarks for evaluating semantic similarity and relatedness [Harispe et al., 2013c]. Most of them aim at evaluating the accuracy of semantic measures according to human appreciation of similarity/relatedness. For the most part, they are composed of a reduced number of entries, e.g., pairs of words/concepts, and have been computed using a reduced pool of subjects.

Initiatives for the development of benchmarks must be encouraged in order to obtain larger benchmarks in various domains of study. Word-to-word benchmarks must be conceptualised (as much as possible)[1] in order for them to be used to evaluate knowledge-based semantic measures. It is also important to propose benchmarks which are not based on human appreciation of similarity, i.e., benchmarks relying on an indirect evaluation strategy – evaluations based on the analysis of the performance of processes which rely on semantic measures [Harispe et al., 2013c].

[1] E.g. using DBpedia URIs.

3.7.2.2 Develop generic open-source software

In [Harispe et al., 2013c], we proposed an overview of the main software solutions dedicated to semantic measures. They are of major importance to: (i) ease the use of the theoretical contributions related to semantic measures, (ii) support large scale comparisons of measures and therefore (iii) better understand the measures and (iv) develop new proposals.

Software solutions dedicated to distributional measures are generally developed without being restricted to a specific corpus of texts. They can therefore be used in a large diversity of contexts of use, as long as the semantic proxy considered corresponds to a corpus of texts.

Software solutions dedicated to knowledge-based semantic measures are generally developed for a specific domain (e.g., refer to the large number of solutions developed for the Gene Ontology alone [Harispe et al., 2013c]). Such a diversity of software is limiting for designers of semantic measures since implementations made for a specific ontology cannot be reused in applications relying on others ontologies. In addition, it hampers the reproducibility of results since some of our experiments have shown that specific implementations tend to produce different results[1]. In this context, we encourage the development of generic open-source software solutions which are not restricted to specific ontologies. This is challenging since the formalism used to express ontologies is not always the same and specificities of particular ontologies sometimes deserve to be taken into account in order to develop semantic measures. However, there are several cases in which generic software can be developed. As an example, numerous knowledge-based semantic measures rely on data structures corresponding to poset or more generally semantic graphs. Other measures are designed to take advantage of ontologies expressed in standardised languages such as RDF(S), OWL. Generic software solutions can be developed to encompass these cases. Reaching such a goal could open interesting perspectives. Indeed, based on such generic and robust software supported by several communities, domain specific tools and various programming language interfaces can subsequently be developed to support specific use cases and ontologies.

The diversity of software solutions is also beneficial as it generally stimulates the development of robust solutions. Therefore, another interesting initiative, complementary to the former, could be to provide generic and domain specific tests to facilitate both the development and the evaluation of software solutions. Such tests could for instance be expected scores of semantic measures for a reduced example of a corpus/ontology. This

[1] This will be discussed in Chapter 8.

specific aspect is important in order to standardise software solutions dedicated to semantic measures and to ensure the users of specific solutions that the score produced by measure implementations are in accordance with the original definitions of the measures.

As discussed in [Harispe et al., 2013c], the evaluation of semantic measures is mainly governed by empirical studies used to assess their accuracy according to expected scores/behaviours of the measures. Therefore, the lack of open-source software solutions implementing a large diversity of measures hampers the study of semantic measures. It explains, for instance, that evaluations of measures available in the literature only involve the comparison of a subset of measures which is not representative of the diversity of semantic measures available today. Initiatives aiming at developing robust open-source software solutions which give access to a large catalogue of measures must therefore be encouraged. It is worth noting the importance of these solutions being open-source. Our communities also lack open-source software dedicated to the evaluation of semantic measures. Indeed, despite some initiatives in specific domains[1], evaluations are not made through a common framework as is done in most communities, e.g. information retrieval [NIST, 2012; Voorhees and Harman, 2005], ontology alignment [Euzenat and Shvaiko, 2013; Grau et al., 2013].

3.7.2.3 Develop theoretical tools

It is currently difficult to study the overwhelming amount of proposed semantic measures, e.g., deriving the interesting properties of measures requires the analysis of each measure. However, as we will see in the following chapter, several initiatives have proposed theoretical tools to ease the characterisation of measures, e.g., by means of measure unification in some cases. These contributions open interesting perspectives on studying groups of measures. They are also essential to better understand the limitation of existing measures and the benefits of new proposals. Finally, they are central to distinguishing the main components on which measures rely, and to improve families of semantic measures based on this characterisation.

3.7.3 Standardise ontology handling

In Appendix A, we discuss the process required to transform an ontology to a semantic graph, a data structure commonly adopted to compute semantic measures. Such a process is currently overly subject to interpretations and deserves to be carefully discussed

[1] E.g., CESSM to evaluate semantic measures designed for the Gene Ontology [Pesquita et al., 2009b]. Note that this solution is not open-source, it can therefore not be used to support large scale evaluations and it is impossible to reproduce experiments and conclusions derived from them.

and formalised. Indeed, as an example, we stress that numerous measures consider ontologies as semantic graphs despite the fact that the formalism on which some ontologies rely cannot be mapped to semantic graphs without reductions – this is the case for some expressive logic-based ontologies. The impact of such a reduction of ontologies is of major importance since it can highly impact semantic measure results[1]. The treatment performed to map an ontology to a semantic graph is generally not documented, which explains some of the difficulties encountered to reproduce the results of some experiments.

3.7.4 Promote interdisciplinarity

From cognitive sciences to biomedical informatics, the study of semantic measures involves numerous communities. Efforts have to be made to promote interdisciplinary studies and to federate the contributions made in the various fields. We briefly provide a non-exhaustive list of the main communities involved in semantic measure study and the communities/fields of study which must be relevant to solicit to further analyse semantic measures. The list is alphabetically ordered and may not be exhaustive:

- *Biomedical Informatics* and *Bioinformatics*: very active in the definition and study of semantic measures, these communities are also active users of semantic measures.
- *Cognitive Sciences*: propose cognitive models of similarity and mental representations which can be used to (i) improve the design of semantic measures and (ii) better understand human expectations w.r.t similarity/relatedness. These communities can also use empirical evaluation studies of semantic measures to discuss the cognitive models they propose.
- *Complexity Theory*: important field of study which is essential to analyse complexity of semantic measures.
- *Geoinformatics*: defines and studies semantic measures. Members of this community are also active users of semantic measures.
- *Graph Theory*: several major contributions relative to graph processing have been proposed in this domain. Such theoretical works are essential for the optimisation of measures relying on network-based ontologies. This community will probably play an important role on knowledge-based semantic measures in the near future, since large semantic graphs composed of billions of relationships are now available – processing such graphs require the development of optimisation techniques.

[1] Consider, for instance, a taxonomy in which redundant relationships have been defined – redundancies highly impact shortest path computation. Should they be considered?

- *Information Retrieval*: defines and studies semantic measures taking advantage of corpus of texts or ontologies.
- *Information Theory*: it plays an important role in better understanding the notion of information and in defining metrics which can be used to capture the amount of information which is conveyed, shared and distinct between compared elements, e.g., notion of information content.
- *Knowledge Engineering*: this community studies and defines ontologies which will further be used by some semantic measures. It could, for instance, play an important role in characterising the assumptions made by semantic measures.
- *Linguistics and Natural Language Processing*: people from this community are actively involved in the definition of distributional measures. They propose models to characterise corpus-based semantic proxies and to define measures for the comparison of units of language.
- *Logic*: defines formal methods to express and take advantage of knowledge. This community can play an important role in characterising the complexity of knowledge-based semantic measures, for instance.
- *Machine Learning*: plays an important role in the definition of techniques and parameterised functions which can be used for the definition and tuning of semantic measures.
- *Measure Theory*: defines a mathematical framework to study and define the notion of measure. Essential for deriving properties of measures, better characterising semantic measures and taking advantage of theoretical contributions proposed by this community.
- *Metrology*: studies both theoretical and practical aspects of measurements.
- *Optimisation area*: important contributions which can be used to optimise measures, to study their complexity and to improve their tuning.
- *Philosophy*: plays an important role in the definition of essential concepts on which semantic measures rely, e.g., definition of the notions of *Meaning*, *Context*.
- *Semantic Web* and *Linked Data*: define standards (e.g., languages, protocols) and processes to take advantage of ontologies. The problem of ontology alignment and instance matching are actively involved in the definition of (semantic) measures based on ontologies.

- *Statistics* and *Data Mining*: important contributions which can be used to characterise large collection of data. Major contributions in clustering which can, for instance, be used to better understand semantic measures.

3.7.5 Study the algorithmic complexity of semantic measures

Most contributions have focused on the definition of semantic measures. However, their algorithmic complexity is *near inexistent* despite the fact that this aspect is essential for practical applications. Therefore, to date, no comparative studies can be made to discuss the benefits of using computationally expensive measures. These aspects are, however, essential for comparing semantic measures. Indeed, in most application contexts, users will prefer to reduce measure accuracy for a significant reduction of the computational time and resources required to use a measure. To this end, designers of semantic measures must, as much as possible, provide the algorithmic complexity of their proposals. In addition, as the theoretical complexity and the practical efficiency of an implementation may differ, developers of software tools must provide metrics to discuss and compare the performance of the measures' implementation.

3.7.6 Support context-specific selection of semantic measures

Both theoretical and software tools must be proposed to orient end-users of semantic measures in the selection of measures according to the needs defined by their application contexts. Indeed, despite the fact that most people only (*blindly*) consider benchmark results in order to select a measure, efforts have to be made in order to orient end-users in the selection of best suited approaches according to their usage context – understanding the implications (if any) of using one approach compared to another. The numerous properties of the measures presented in this introduction can be used to guide the selection of semantic measures. In addition, numerous large-scale comparative studies have to be performed in order to better understand the benefits of selecting a specific semantic measure in a particular context of use.

4

Unification of knowledge-based semantic similarity measures

Contents

Abstract

A plethora of *ad hoc* and domain-specific semantic similarity measures have been defined over the recent years. In order to shed some light onto the diversity of proposals, this chapter performs an in-depth technical analysis of existing knowledge-based measures to identify the core elements of semantic similarity assessment. Based on existing works related to abstract expression of semantic measures, we present a unifying framework that aims to improve the understanding of measures, to highlight their relationships and to propose bridges linking their theoretical bases. By demonstrating that groups of measures are simply particular instantiations of parameterised functions, we unify a large number of state-of-the-art semantic measures through common expressions. Finally, we underline the application of the proposed framework and its practical usefulness for the design of measures. Other applications of the framework will be presented in the following chapter.

Associated references on which this chapter is based:

- **A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain**. Sébastien Harispe*, David Sánchez, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. Journal of Biomedical Informatics 2013. `http://dx.doi.org/10.1016/j.jbi.2013.11.006`
- **From Theoretical Framework to Generic Semantic Measures Library**. Sébastien Harispe*, Stefan Janaqi, Sylvie Ranwez, Jacky Montmain. On the Move to Meaningful Internet Systems: OTM 2013 Workshops Lecture Notes in Computer Science Volume 8186, 2013, pp 739-742; `http://dx.doi.org/10.1007/978-3-642-41033-8_98`
- **Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Base Analysis**. Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain (2013). ArXiv. Computation and Language. `http://arxiv.org/abs/1310.1285v2`

Special thanks to:

- David Sànchez from the university of Tarragona (URV) who collaborates on this work and to Montserrat Batet (URV) for her relevant advices and recommendations on early versions of the theoretical framework presented in this chapter.

4.1 Introduction

4.1.1 Motivation

As we saw in Chapter 3, a large diversity of knowledge-based semantic measures have been proposed over recent decades. Although some measures are context-independent, most of them were designed in an *ad hoc* manner and were expressed on the basis of domain-specific or application-oriented formalisms. Therefore, most proposals related to these measures target a specific audience and fail to benefit other communities. In this way, a non-specialist can only interpret the plethora of state-of-the-art proposals as an extensive list of measures (refer to Tables presented in Sections 3.5.6 and 3.6.3). As a consequence, the selection of an appropriate measure for a specific usage context is a challenging task. Actually, no extensive studies have characterised the large diversity of proposals, even though a few important contributions focusing on theoretical aspects of knowledge-based semantic similarity measures exist, e.g., [Blanchard, 2008; Blanchard and Harzallah, 2005; Blanchard et al., 2008; Cross, 2006; Cross and Yu, 2010, 2011; D'Amato, 2007; Pirró and Euzenat, 2010a; Sánchez and Batet, 2011].

Despite the large number of contributions related to knowledge-based semantic similarity measures nowadays, the understanding of their foundations is limited. For a designer/practitioner, some fundamental questions remain: Why does one measure work better than another one? How does one choose or design a measure? Is it possible to distinguish families of measures sharing specific properties? How can one identify the most appropriate measure(s) according to particular criteria?

To fill these gaps, this chapter presents an extensive study of knowledge-based semantic similarity measures leading to our proposal of a unifying framework which dissects measures using a set of intuitive core elements. For convenience, knowledge-based semantic measures will be denoted as semantic measures and knowledge-based semantic similarity measures as semantic similarity measures.

4.1.2 Contributions and plan

Based on existing works on the unification of semantic measures, the framework presented in this chapter proposes to model, in a generic and flexible way, the core elements on which most available semantic measures rely. We subsequently demonstrate that particular semantic measures can be properly characterised and directly obtained as instantiations of the framework components. This brings new insights for the measures by:

- *Distinguishing the core elements on which measures rely.* The theoretical characterisation of semantic measures helps to understand the different measure paradigms and the large diversity of expressions proposed in the state-of-the-art.

- *Unifying measures through parameterised measures.* Based on the characterisation of the core elements of semantic measures, our framework enables the identification of commonalities, bridges and equivalences between existing measures. Indeed, even if many semantic measures are: (i) of an *ad hoc* nature, (ii) domain-specific, or (iii) based on different theoretical principles, their design could be unified through abstract expressions. Expressing semantic measures through parameterised functions can therefore facilitate the detection of their common properties and the analysis of their behaviour in specific applications.

- *Selecting appropriate domain-specific measures.* Such a framework provides a systematic, theoretically-coherent and direct way to define or tune the semantic similarity assessment for particular application scenarios. semantic similarity measures expressed through parameterised functions could therefore be used to optimise measure tuning in domain-specific applications.

- *Designing new families of semantic measures.* New measures can be easily defined due to the modularity provided by the framework. Their design can take into account: (i) the elements that affect the semantic assessment the most (e.g. estimation of concept specificity) and (ii) the particularities of ontology/application to which it will be applied (e.g., the presence of multiple inheritances).

- *Identifying the crucial aspects of semantic similarity assessment.* Based on the analysis of specific expressions of measures derived from the framework, empirical studies could be used to highlight core elements best impacting the measures' accuracy. As a result, the framework could be used to guide research efforts towards the aspects that can improve measure performances.

An important aspect is that such an approach will not only benefit a single measure designed for a domain-specific application (which is to date the focus of most related works); it will instead result in improvements on a wide set of measures and applications.

The rest of the chapter is organised as follows. Section 4.2 introduces the reader to previous works regarding the unification of semantic measures. Section 4.3 describes the proposed framework from which state-of-the-art measures are unified. Section 4.4 presents a first application of the framework to design existing and new semantic similarity measures. Section 4.5 concludes the chapter.

4.2 Related work on the unification of semantic measures

This section presents state-of-the-art contributions related to the unification of semantic measures dedicated to the comparison of concepts.

4.2.1 Similitude between semantic similarity measures

Numerous authors have underlined similitudes between semantic measures. As an example, in a tree, the edge-counting strategy defined by Rada and Bicknell [1989] (Equation 3.12) can also be expressed as a function of the depths of compared concepts and their LCA [Blanchard, 2008]:

$$dist_{Rada}(u,v) = depth(u) + depth(v) - 2 \cdot depth(LCA(u,v))$$

Indeed, as we have seen in Section 3.3.2, the depth of a concept can be seen as an estimator of the specificity of a concept. In addition, we have generalised such estimators using the function θ. The edge-counting strategy can thus be defined through an abstract expression of the symmetric difference[1]:

$$dist_{\triangle^*}(u,v) = \theta(u) + \theta(v) - 2 \cdot \theta(LCA(u,v))$$

As stressed by several authors, e.g., [Blanchard et al., 2008; Cross and Yu, 2010; Sánchez and Batet, 2011], we can see that this expression generalises the information theoretical distance proposed by Jiang and Conrath [1997]:

$$dist_{JC}(u,v) = IC(u) + IC(v) - 2 \cdot IC(MICA(u,v))$$

In the same manner, it has also been stressed that, in a tree[2], the measure proposed by Wu and Palmer [1994] (Equation 3.16) can be reformulated by:

$$sim_{WP}(u,v) = \frac{2 \cdot depth(LCA(u,v))}{depth(u) + depth(v)}$$

[1] In set theory the symmetric difference between two sets is $A \triangle B = A \setminus B \cup B \setminus A$. The exponent * is used to denote abstract semantic measures, e.g., the abstract form of sim_x is denoted sim_{x^*}.

[2] In which a transitive reduction has been performed.

Therefore, once again, this expression can be generalised by an abstract similarity measure which corresponds to an abstract formulation of the Dice index:

$$sim_{Dice^*}(u,v) = \frac{2 \cdot \theta(LCA(u,v))}{\theta(u) + \theta(v)}$$

Such an abstract expression highlights the relationship between structural and information theoretical approaches – here exemplified through the relationships between sim_{Dice^*} and sim_{Lin} (Equation 3.29):

$$sim_{Lin}(u,v) = \frac{2 \cdot IC(MICA(u,v))}{IC(u) + IC(v)}$$

A similar approach can be adopted to underline the relationship between some feature-based measures and information theoretical measures [Pirró and Euzenat, 2010a; Sánchez and Batet, 2011]. Indeed, under specific tunings, comparing two concepts using a feature-based measure (i.e., according to their shared and distinct features), can be equivalent to considering a particular expression of an information theoretical measure. As an example, characterising the features of the concept u by $A(u)$, and using a semantic measure based on sim_{Dice^*}, we obtain:

$$sim_{Dice-FB}(u,v) = \frac{2|A(u) \cap A(v)|}{|A(v)| + |A(u)|}$$

Since, in a tree, two concepts have a unique LCA, this feature-based expression can be reformulated as:

$$sim_{Dice-FB}(u,v) = \frac{2|A(LCA(u,v))|}{|A(v)| + |A(u)|}$$

Thus, this expression is a specific instantiation of the abstract measure sim_{Dice^*} defining $\theta(u) = |A(u)|$.

Using a similar reformulation of the measure proposed by Stojanovic et al. [2001], [Blanchard, 2008; Blanchard et al., 2008] underlined that, in trees, several feature-based expressions can be reformulated using the depth of concepts (since $|A(c)| = depth(c) + 1$). Thus, in a tree, we obtain:

$$\begin{aligned} sim_{CMatch}(u,v) &= \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|} \\ &= \frac{depth(LCA(u,v)) + 1}{depth(u) + depth(v) - depth(LCA(u,v)) + 1} \end{aligned}$$

As we have seen, several contributions have stressed that links exist between semantic similarity measures. Such links have also been highlighted for other (non-semantic) measures which have been designed to compare specific mathematical objects (e.g., sets [Choi et al., 2010], probability distribution functions [Cha, 2007], and fuzzy sets [Bouchon-Meunier et al., 1996]). Similarly, correspondences have also been underlined between different types of measures. As an example, Borgida et al. [2005] discusses logic-based semantic measures through derivation of measures proposed for semantic graph analysis. These findings have highlighted that measures can be seen as particular expressions of more abstract measures, i.e., abstract formula expressed using abstract components which are commonly used to compare objects. The components required to design such abstract measures are generally defined in abstract frameworks, some of those proposed for semantic measures are presented hereinafter.

4.2.2 Existing frameworks of semantic measures

The *feature model* proposed by Tversky is probably the best known framework dedicated to similarity [Tversky, 1977]. It distinguished parametric formulations of measures from which several similarity measures can be expressed. We already introduced this framework in Section 1.4.2. Among the assumptions associated to the feature model, compared objects are expected to be represented by sets of features. This framework therefore requires the features of compared elements to be specified in order to obtain a concrete implementation of a measure. This is why the feature model can be considered as an abstract framework. It doesn't define concrete implementations but rather *backbones* (i.e., constrained parametric functions) from which measures can be expressed.

To be used for the comparison of concepts defined in an ontology, the feature model therefore requires the definition of a function characterising the features of a concept. The similarity is then intuitively defined based on the common and distinctive features of the compared concepts. Assessing the similarity of objects based on their common/distinct properties has been used for a long time to compare sets according to the study of their shared and distinct elements (e.g., Jaccard Index, Dice coefficient). As we have seen, Tversky defined the *contrast model* and the *ratio model* as functions which can be used to compare objects represented as sets of features. Below we recall the formulation of the *ratio model*:

$$sim_{RM}(u,v) = \frac{f(U \cap V)}{\alpha \cdot f(U \setminus V) + \beta \cdot f(V \setminus U) + f(U \cap V)} \tag{4.1}$$

Such a general parameterised formulation of a similarity measure can be used to derive a large number of concrete measures. As an example, considering the function f which

estimates the salience of a set of features as the cardinality of the set, and $\alpha = \beta = 1$, the *ratio model* leads to the original definition of the Jaccard index. Setting $\alpha = \beta = 0.5$ leads to the Dice coefficient, e.g. [Bradshaw, 1997]. Indeed, a large diversity of set-based measures can be expressed from specific instances of such parameterised functions. In other words, such general measures are abstract similarity measures which can be used to instantiate concrete similarity measures through the definition of a limited set of parameters. As an example, to be used in order to compute the similarity between two elements (e.g., concepts), the *ratio model* requires the concrete definition of: (i) the function mapping an object to a set of features, (ii) a function f which can assess the salience of a set of features, and, (iii) values of the parameters α and β.

The framework proposed by Tversky constrains compared objects to be represented by sets of features in order to further assess the similarity as a function of the commonalities and differences of the two sets. By definition, the *contrast model* and the *ratio model* are therefore *constrained* to set-based formulations of measures. To be more precise, the feature model is thus constrained to fuzzy set theory, since, originally, Tversky defined the commonalities and differences of two objects as a function of the salience of their shared and distinct features (defined by the aforementioned function f). Nevertheless, in the literature, the *feature model* is generally regarded as a pure set-based framework and the function f is generally understood as the cardinality of the set, i.e. $f(X) = |X|$[1].

Therefore, using a specialisation process, both ratio and contrast models can also be regarded as *pure* parametric set-based functions. The literature relative to set-based similarity/distance function is rich. Nevertheless, several contributions have focused on the unification of the numerous formulations proposed over the years, e.g., [Choi et al., 2010]. As an example, it has been shown that most set-based measures can be expressed using Caillez and Kuntz [1996] (σ_α), and Gower and Legendre [1986] (σ_β) parametric measures [Blanchard et al., 2008]. Therefore, since set-based measures can be used to design semantic measures considering that compared elements are represented as sets of features; σ_α and σ_β can intuitively be generalised in a straightforward manner in order to define new feature-based *models*:

$$\sigma_\alpha^f(u,v) = \frac{f(U \cap V)}{\left(\frac{f(U)^\alpha + f(V)^\alpha}{2}\right)^{1/\alpha}} \tag{4.2}$$

$$\sigma_\beta^f(u,v) = \frac{\beta \cdot f(U \cap V)}{f(U) + f(V) + (\beta - 2) \cdot f(U \cap V)} \tag{4.3}$$

[1]Originally, as stressed in Section 1.4.2, the operators $\cap, \cup$ and $\setminus$ are based on feature matching (F) and the function f evaluates the contribution of the common or distinct features (distinguished by previous operators) to estimate the similarity.

Therefore, defining the function $f(X)$ (e.g., as the cardinality of the set of features X), the abstract formulations σ_α and σ_β can be used to derive a large number of set-based measures. As an example, Simpson and Ochiai coefficients [Choi et al., 2010] can be expressed from σ_α setting α to $-\infty$ and 0 respectively. The σ_β reformulation can also be used to express other numerous measures, e.g. Sokal and Sneath ($\beta = 0.5$), Jaccard index ($\beta = 1$) and Dice coefficient ($\beta = 2$) [Blanchard et al., 2008; Choi et al., 2010]. By extension they can also be seen as primitive abstract semantic measures.

Other frameworks and models of similarity measures have been proposed in the literature. For instance, in Roddick et al. [2003], the authors propose a model of semantic distance relying on a graph-based approach. This model quantifies the distance between data values as a function of graph traversals. It can therefore be generalised in order to compare any elements structured in a graph. Nevertheless, this kind of model has proved not to be easily workable to express and study semantic measures as it has not been extensively studied and used in the literature.

An interesting contribution relative to the study of semantic similarity measures through abstract functions was made by Blanchard and collaborators. They were the first to take advantage (in an explicit manner) of abstract definitions of measures for the comparison of a pair of concepts defined in a taxonomy [Blanchard et al., 2008]. In their studies, the authors focused on an information theoretical expression of semantic similarity measures to highlight relationships between several measures proposed in the literature. Their extensive work was mainly concentrated on the comparison of concepts structured in a tree-based taxonomy; generalisation of their framework for multiple inheritance was then conducted. Based on the intuitive notions of commonalities and differences, and on a particular expression of the notion of specificity, the authors underlined several links between measures. As an example, they underlined that measures proposed by Wu and Palmer [1994] and Lin [1998] can be derived from an abstract expression of the Dice coefficient (see previous subsection). They also stressed that the general expression of the Dice coefficient, here named sim_{Dice^*}, corresponds to the expression of an abstract formulation of the *ratio model* defining $\alpha = \beta = 0.5$, and can also be seen as particular expression of σ_β setting $\beta = 2$[1]. Several other abstract expressions of measures, and links between measures can be found in Blanchard [2008]; Blanchard and Harzallah [2005]; Blanchard et al. [2008]. In their studies, summarised in the PhD thesis [Blanchard, 2008][2], the authors stressed an essential point, which has been poorly understood by the communities studying semantic measures: the relevance of dissecting semantic measures through abstract expressions in order to further characterise their properties and to study groups of measures. Nevertheless, the technical background required to fully capture the

[1] Also highlighted in Bradshaw [1997].
[2] In french.

Description	**Feature-based model**	**Information-theoretic model**
Salience of common features	$f(U \cap V)$	$IC(MICA(u,v))$
Salience of the features of u not shared with the features of v	$f(U \setminus V)$	$IC(u) - IC(MICA(u,v))$
Salience of the features of v not shared with the features of u	$f(V \setminus U)$	$IC(v) - IC(MICA(u,v))$

TABLE 4.1: Mapping proposed by Pirró and Euzenat [2010a] between the feature model and the information theoretic approach (reproduction with some modifications to be in accordance with the notions and notations introduced)

relevance of such an abstract framework hampered its use and only few contributions related to semantic measures took advantage of this important contribution.

Next to the contributions of Blanchard and collaborators, other authors have also demonstrated relationships between different similarity measures and have taken further advantage of abstract frameworks to design new measures or to study existing ones [Cross, 2006; Cross and Yu, 2010; Cross et al., 2013; Mazandu and Mulder, 2013; Pirró and Euzenat, 2010a; Sánchez and Batet, 2011]. These contributions mainly focused on establishing local relationships between set-based measures and measures framed in Information Theory. Note that several contributions have been proposed during the period covered during this thesis; Cross et al. [2013] and Mazandu and Mulder [2013] contributions were for instance published after the design of the proposal introduced hereafter (they will nevertheless be discussed).

Pirró and Euzenat [2010a] present an information theoretical expression of the component distinguished by the *feature model* (commonalities and differences). Based on this contribution, numerous information theoretical measures can be expressed from abstract expressions of the *ratio model* and the *contrast model*. Table 4.1 presents the mapping between feature-based and information theoretical similarity models proposed by the authors. As an example, using the *ratio model* with $\alpha = \beta = 1$, the authors proposed the definition of a new measure which corresponds to a particular expression of an abstract form of the Jaccard coefficient:

$$sim_{Faith}(u,v) = \frac{IC(MICA(u,v))}{IC(u) + IC(v) - IC(MICA(u,v))}$$

Alternatively, Sánchez and Batet [2011] also proposed a framework grounded in information theory. It allows several measures (i.e., edge-counting and set-based coefficients) to be uniformly redefined according to the notion of IC. The authors defined a mapping able to take advantage of set-based measures in order to express measures framed in

Expressions found in set-based similarity coefficients	**Approximation in terms of IC**
$\|U\|$	$IC(u)$
$\|V\|$	$IC(v)$
$\|U \cap V\|$	$IC(MICA(u,v))$
$\|U \setminus V\| = \|U\| - \|U \cap V\|$	$IC(v) - IC(MICA(u,v))$
$\|V \setminus U\| = \|V\| - \|U \cap V\|$	$IC(v) - IC(MICA(u,v))$
$\|U \cup V\| = \|U\| + \|V\| - \|U \cap V\|$	$IC(u) + IC(v) - IC(MICA(u,v))$
$\|U\| + \|V\|$	$IC(u) + IC(v)$

TABLE 4.2: Mapping proposed by Sánchez and Batet [2011] between expressions found in set-based similarity measures and the information theoretic approach (reproduction with some modifications to be in accordance with the notations introduced)

information theory. Based on the links defined in Table 4.2, the authors derived several semantic measures from set-based measures. This contribution extends the work of Pirró and Euzenat [2010a] by enriching the mappings already proposed (Table 4.1). In addition, the authors also proposed several redefinitions of structural measures using the notion of IC. As an example, among other links between measures, they underlined the link between the edge-counting strategy and the information theoretical measure defined by Jiang and Conrath[1] (a link which was already presented Section 4.2.1).

In the same vein, in a series of papers, Cross [2006]; Cross and Yu [2010]; Cross et al. [2013] proposed a similar contribution in which feature-based approaches and measures based on information theory are expressed through the frame of fuzzy set theory. This work has recently led to a unification proposal grounded in fuzzy set theory [Cross et al., 2013] – it *only* targets pairwise similarity measures and is limited to approaches relying on canonical forms of concepts which can be expressed using fuzzy sets.

Recently, Mazandu and Mulder [2013] proposed another general framework and unified description of measures relying on the notion of IC for the comparison of pairs of concepts. Like Blanchard et al. [2008], the authors focused on an information theoretical definition of measures to underline similarities between existing measures.

Despite the suitability of these frameworks for studying some properties of semantic measures, few works rely on them to express measures [Cross et al., 2013; Sánchez and Batet, 2011]. Moreover, current frameworks generally only focus on a specific paradigm to express measures (e.g., feature-based, information theoretical). In fact, most existing frameworks only encompass a limited number of measures and were not defined with the purpose of unifying measures expressed using the different paradigms reviewed in

[1] This link has also been underlined in Blanchard [2008].

Chapter 3. These frameworks derive from the feature model or an information theoretical expression of the feature model, they are therefore limited to these paradigms by definition.

The main limitations associated to existing works were due to the constraints induced by the canonical forms adopted to manipulate compared elements, e.g. a set of features for the feature-based approach, an amount of information for the information theoretical approach. To overcome this limitation, we propose a new unifying framework for semantic measures in which the representation of compared elements is defined as a central parametric component. This framework has its roots in the teaching of cognitive sciences in the central role played by the representation adopted to characterise compared elements. Therefore, contrary to other existing frameworks, this proposal is not limited to specific approaches constrained by a canonical form of the compared elements (feature-based, structural, information theoretical). Indeed, this framework gives the possibility of explicitly defining the strategy adopted to characterise the representation of a concept (set-based representation, information-theoretical, graph-based, etc.). The framework further distinguishes the primitive functions commonly found in measure expressions (e.g., functions used to characterise the commonalities and the differences of the compared representations, the degree of specificity or amount of information carried by a representation).

4.3 A unifying framework for semantic similarity measures

The analysis of the state-of-the-art allowed us to distinguish a few core elements underlying most semantic similarity measures. Their notation and meaning are given in this section. The abstract measures which can be defined as a function of these core elements are then introduced and discussed. Finally, we illustrate the suitability of the proposed framework to express a selection of well-known semantic similarity measures available in the literature. Other applications will be presented in the following chapter.

4.3.1 Reminder of the notations

This subsection recall some of the notations introduced so far which will be used for presenting the framework. These notations have already been presented in Section 3.2.3 and are repeated for clarity.

The taxonomy G_T is the semantic graph associated to the non-strict partial order defined over the set of concepts C. The notations used to characterise G_T as well as its concepts are :

- $C(G_T)$ shortened by C refers to the set of concepts defined in G_T.
- $E(G_T)$ shortened by E_T refers to the set of relationships defined in G_T with:

$$E_T \subseteq C \times \{\texttt{subClassOf}\} \times C \subseteq E_T \subseteq E_{CC}$$[1]

- A concept v subsumes another concept u if $u \preceq v$, i.e., $\{u, \texttt{subClassOf}, v\} \neq \emptyset$. Several additional denominations will be used; it is commonly said that v is a ancestor of u, that u is subsumed by v and that u is a descendant of v.
- $C^+(u) \subseteq C$, with $u \in C$, the set of concepts such as:

$$C^+(u) = \{c | (u, \texttt{subClassOf}, c) \in E_T\}$$

- $C^-(u) \subseteq C$, with $u \in C$, the set of concepts such as:

$$C^-(u) = \{c | (c, \texttt{subClassOf}, u) \in E_T\}$$

- $C(u) \subseteq C$, with $u \in C$, the set of neighbours of concepts such as:

$$C(u) = C^+(u) \cup C^-(u)$$

- $A(u)$ the set of concepts which subsumes u, also named the ancestors of u, i.e., $A(u) = \{c | \{u, \texttt{subClassOf}, c\} \neq \emptyset\} \cup \{u\}$. We also denote $A^-(u) = A(u) \setminus \{u\}$ the exclusive set of ancestors of u.
- $parents(u)$ the minimal subset of $A^-(u)$ from which $A^-(u)$ can be inferred according to the taxonomy G_T, i.e., if G_T doesn't contain taxonomic redundancies[2] we obtain: $parents(u) = C^+(u)$.
- $D(u)$ the set of concepts which are subsumed by u, also named the descendants of u, i.e., $D(u) = \{c | \{c, \texttt{subClassOf}, u\} \neq \emptyset\} \cup \{u\}$. We also denote $D^-(u) = D(u) \setminus \{u\}$ the exclusive set of descendants of u.
- $children(u)$ the minimal subset of $D^-(u)$ from which $D^-(u)$ can be inferred according to the taxonomy G_T, i.e., if G_T doesn't contain taxonomic redundancies we obtain: $children(u) = C^-(u)$.
- $roots(G_T)$, shortened by *roots*, the set of concepts $\{c | A(c) = \{c\}\}$. We call the *root*, denoted as $\top$, the unique concept (if any) which subsumes all concepts, i.e., $\forall c \in C, c \preceq \top$.
- $leaves(G_T)$, shortened by *leaves*, the set of concepts without descendants, i.e. $leaves = \{c | D(c) = \{c\}\}$. We also note $leaves(u)$ the set of leaves subsumed by a concept (inclusive if u is a leaf), i.e., $leaves(u) = D(u) \cap leaves$.

[1] E_{CC} were used to introduce semantic graphs

[2] Taxonomic redundancies are introduced in Section A.2.

- $depth(u)$, the length of the longest path in $\{u, \texttt{subClassOf}, \top\}$, for convenience we also consider $depth(G_T) = \underset{c \in C}{\text{argmax}}\ depth(c)$.
- $G_T^+(u)$ the graph composed of $A(u)$ and the set of relationships which link two concepts in $A(u)$.
- $G_T^-(u)$ the graph composed of $D(u)$ and the set of relationships which link two concepts in $D(u)$.
- $G_T(u) = G_T^+(u) \cup G_T^-(u)$ the graph induced by $A(u) \cup D(u)$.
- $\Omega(u, v)$, the set of Non Comparable Common Ancestors (NCCAs) of the concepts u, v. $\Omega(u, v)$ is formally defined by: $\forall (x, y) \in \Omega(u, v), (x, y) \in \{A(u) \cap A(v)\} \times \{A(u) \cap A(v)\} \wedge x \notin A(y) \wedge y \notin A(x)$. NCCAs are also called the Disjoint Common Ancestors (DCAs) in some contributions, e.g. [Couto et al., 2005].

4.3.2 Core elements of semantic similarity measures

We first present the core elements of semantic similarity measures which are distinguished by the framework. Each of them are then further detailed through concrete examples.

As stated in Chapter 3, semantic similarity measures are designed according to specific paradigms. Therefore, designers of measures first adopt a specific paradigm from which estimators of commonalities and differences will be defined. They then adopt a strategy by which these estimators will be aggregated to express a similarity measure or a taxonomic distance. Indeed, in a broad sense, when comparing two things, their commonalities and differences are the only evidence from which similarity (or dissimilarity) can be evaluated. In the aim of distinguishing the core elements of semantic similarity measures, estimators of commonalities and differences intuitively appear as critical elements of semantic measures. In fact, they are the roots of all existing similarity measures. As we have seen, these two functions are the cornerstone of all existing frameworks, e.g. the feature model.

The definition of the estimators of commonalities and differences depends on the paradigm which has been chosen to formulate semantic similarity measures. For instance, for some structural approaches, the difference of two concepts is assessed as a function of the length of the shortest path linking them, while for feature-based approaches, concept differences are computed as a function of the features characterising one concept (e.g. $A(u)$), which are not shared with the other.

The main differences between existing paradigms depend on the strategy adopted to represent a concept. Such a representation will determine the expressions of the estimators of commonalities/differences, and is therefore critical for the design of semantic measures. We are therefore convinced that abstract frameworks may distinguish such a function. Thus, we formally introduce a function aiming at representing a concept, or more generally, a set of concepts.

Definition *Semantic representation* (ρ): the mapping of a set of concept $C' \subseteq C$ to its semantic representation, denoted $\widetilde{C'}$, is defined by the function $\rho(C')$:

$$\rho : \mathcal{P}(C) \rightarrow \mathbb{K} \tag{4.4}$$

with $\mathbb{K}$ a domain containing any subset or subgraph of G_T, e.g. C, E_T.

For convenience, we note $\rho(u)$ and $\tilde{u}$, the representation of a single concept u, i.e. $\{u\}$. Remember that concrete examples of the core elements will be discussed later.

We also formally define the functions aiming to estimate the commonalities and differences of two concepts (u, v), according to their semantic representations $(\tilde{u}, \tilde{v})$:

Definition *Commonality of two semantic representations* (Ψ): the commonality of two concept representations $(\tilde{u}, \tilde{v})$ is estimated using a function $\Psi(\tilde{u}, \tilde{v})$:

$$\Psi : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}_+ \tag{4.5}$$

Definition *Difference between two semantic representations* (Φ): the difference between $\tilde{u}$ not found in $\tilde{v}$ is estimated using a function $\Phi(\tilde{u}, \tilde{v})$:

$$\Phi : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}_+ \tag{4.6}$$

The three abstract functions ρ, Ψ, Φ are the core elements of most similarity measures. In the context of semantic similarity estimation, they can be used to reformulate, in an abstract manner, *all* semantic similarity measures based on commonalities and differences of compared concepts.

As an example, the shortest path linking two concepts u, v can be abstracted to the sum of their differences, being estimated according to their LCA: $sp(u, isa^*, v) \approx \Phi(\tilde{u}, \tilde{v}) + \Phi(\tilde{v}, \tilde{u})$ where $\Phi(\tilde{u}, \tilde{v}) = sp(u, isa, LCA(u, v))$ and $\Phi(\tilde{v}, \tilde{u}) = sp(v, isa, LCA(u, v))$, with $LCA(u, v) \in sp(u, isa^*, v)$.

Designers occasionally integrate information regarding the universe in which compared elements are defined [Choi et al., 2010]. We therefore introduce a function with the aim of capturing this information.

Definition *Global information on the universe* (ζ): the amount of knowledge defined in G_T (i.e., modelled in the taxonomy), which is neither found in $\tilde{u}$ nor in $\tilde{v}$, can be estimated by a function $\zeta(\tilde{u}, \tilde{v})$:

$$\zeta : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}_+ \tag{4.7}$$

Most measures can be expressed in an abstract manner using the functions ρ, Ψ, Φ and, in some particular cases, ζ. However, there are situations in which functions Ψ and Φ may also be expressed according to the specificity of a (group of) concept(s) or, more generally, according to the amount of information carried by a representation (e.g., information theoretic measures). Thus, we further define two functions capturing these notions.

Definition *Specificity of a concept* (θ): the specificity of a concept u is estimated by a function $\theta(u)$:

$$\theta : C \rightarrow \mathbb{R}_+ \tag{4.8}$$

This function has already been introduced in Section 3.3.2 and is briefly repeated for clarity. The expressions used to compute the IC of a concept are particular expressions of function θ.

Finally, we also generalise the notion of specificity of a concept to a semantic representation.

Definition *Specificity of a semantic representation* (Θ): the degree of specificity of a semantic representation $\tilde{u}$ can be estimated by a function $\Theta(\tilde{u})$:

$$\Theta : \mathbb{K} \rightarrow \mathbb{R}_+ \tag{4.9}$$

The Θ function generalises the function θ defined to estimate concept specificity. This is required to express state-of-the-art semantic measures based on an aggregation of θ [Mazandu and Mulder, 2011; Pesquita et al., 2007]. As an example, considering the representation of concept $\tilde{u} = A(u)$, in Equation 3.35, Mazandu and Mulder [2011] defined $\Theta(\tilde{u})$ as:

$$\Theta(A(u)) = \sum_{c \in A(u)} \theta(u)$$

Figure 4.1 presents an intuitive feature-based representation of the functions introduced by the framework. The representation of the concept u, i.e. $\rho(u)$, is here defined as $A(u)$. The commonalities and differences (Ψ, Φ) of two concept representations are intuitively defined by the set operators ($\cap$ and $\setminus$ respectively). The part of the universe which is not contained in compared representations is denoted by ζ.

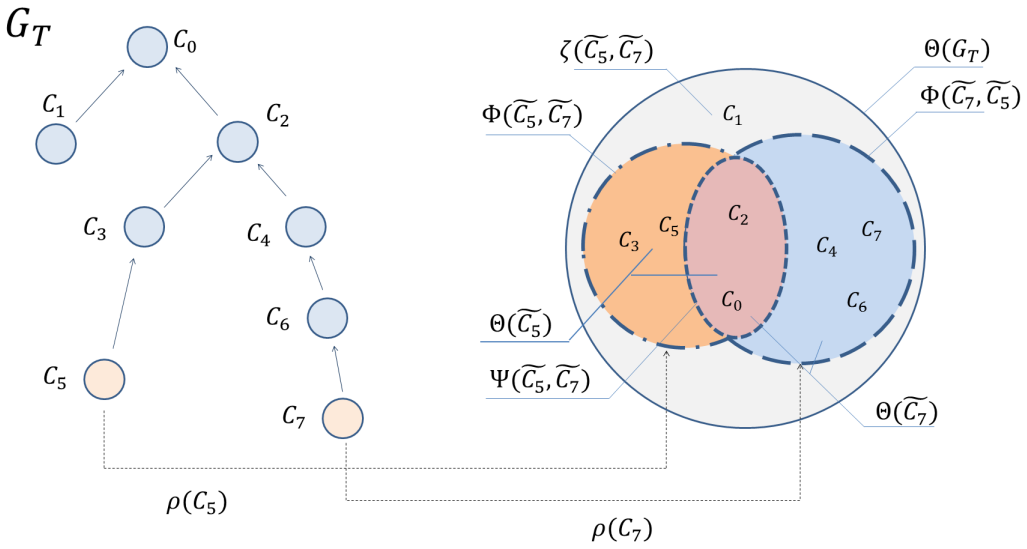

FIGURE 4.1: Example of expressions of the framework's core elements according to the feature-based approach

We further detail the various core elements distinguished by the framework.

4.3.2.1 Mapping a concept to its semantic representation (ρ)

"*For AI Systems, what 'exists' is that which can be represented*" [Guarino et al., 2009].

$$\rho : \mathcal{P}(C) \rightarrow \mathbb{K}$$

The semantic representation of a set of concepts can be viewed as a subset of the knowledge that the taxonomy models. Thus, the function ρ defines the mapping between a set of concepts and its semantic representation in the ontology. We first consider the case in which the set of concepts only contains a single concept. This case is central for the study of pairwise measures. Figure 4.2 shows some semantic representations of a concept that are commonly used to design semantic measures.

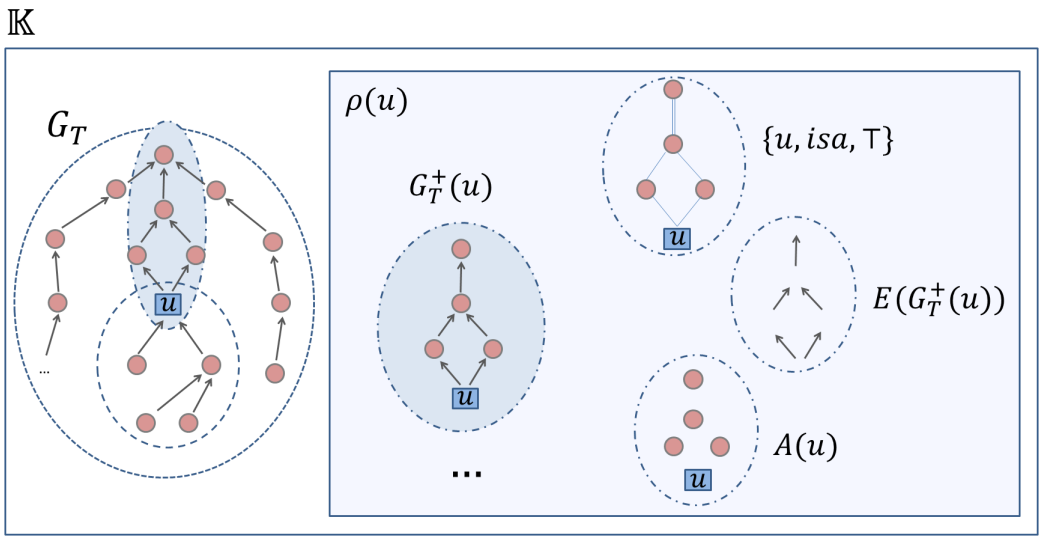

FIGURE 4.2: Representations of a concept commonly used to design semantic measures

One of the most general semantic representations of a concept u is $G_T(u)$, i.e., the graph induced by the ancestors ($A(u)$) and the descendants ($D(u)$) of u. However, in most cases, semantic similarity measures are based on $G_T^+(u)$, the graph induced by $A(u)$. Indeed, as stressed in Figure 4.2, from $G_T^+(u)$, multiple concept representations can be derived, such as the set of ancestors $A(u)$ or the set of paths linking the concept to the root, here named $\{u, isa, \top\}$ or $E_T^+(u)$, i.e. $E(G_T^+(u))$, the set of edges composing $G_T^+(u)$.

As we saw in Section 3.5.2, representing a concept by $A(u)$ is extensively used to express measures based on the feature approach [Rodríguez and Egenhofer, 2003; Sánchez et al., 2012a], or based on Information Theory [Jiang and Conrath, 1997; Maedche and Staab, 2001; Resnik, 1999]. Moreover, the representation of a concept through the paths linking it to the root of the taxonomy is commonly adopted in defining measures based on the edge-counting approach [Pekar and Staab, 2002; Rada and Bicknell, 1989; Wu and Palmer, 1994].

Given that the function ρ is defined for a set of concepts, we consider that union operators are defined for the proposed concept representations. This is indeed the case for all representations based on sets and of those corresponding to graphs. Formally, the representation of a set of concepts $C' \subseteq C$ can be derived from the representation of a single concept, i.e., $\rho(C') = \bigcup_{u \in C'} \rho(u)$, e.g., defining $\rho(C') = \bigcup_{u \in C'} A(u)$.

4.3.2.2 The specificity of concepts and representations (θ and Θ)

$$\theta : C \rightarrow \mathbb{R}_+$$

$$\Theta : \mathbb{K} \rightarrow \mathbb{R}_+$$

Numerous measures rely on the amount of information captured by a concept. Measures based on the notion of IC extensively rely on it. Other strategies, which are not grounded in information theory, have also proposed to evaluate the specificity of a concept according to, for instance, its depth in the taxonomy. In Section 3.3.2 we therefore generalise the notion of IC by introducing a function θ which estimates the specificity of a concept. Since the central element of the framework is the representation of a group of concepts (ρ), we also introduce a function Θ which assesses the specificity of that semantic representation. This function generalises θ and, in coherency with the taxonomic structure, it decreases monotonically from the leaves to the root of the taxonomy when single concept representations are compared, i.e., $u \preceq v \rightarrow \Theta(\tilde{u}) \geq \Theta(\tilde{v})$.

Various strategies can be defined to evaluate $\Theta(\tilde{u})$ depending on the representation defined by ρ. Without loss of generality, we focus here on the case where $\Theta(\tilde{u})$ is assessed for $\tilde{u} \subseteq G_T^+(u)$, e.g. $\tilde{u} = A(u)$.

Two commonly used strategies are briefly discussed:

- A *direct strategy* which will define a way to evaluate $\tilde{u}$. As an example, when $\tilde{u}$ corresponds to a set of elements (concepts, edges, paths...) the cardinality of the set can be evaluated. Considering $\tilde{u} = A(u)$, we obtain $\Theta(\tilde{u}) = |A(u)|$, which can be substituted by $\theta(u)$ so that $\theta(u) = |A(u)|$. In this case, a commonly used strategy is to define $\Theta(\tilde{u}) = max_{c \in A(u)} \theta(c) = \theta(u)$. This strategy was adopted by Lin, Resnik, Wu & Palmer and numerous other designers of semantic measures.

- An *indirect strategy* from which the specificity of elements composing the representations will be taken into account. As an example, the specificity of the concept contained in $A(u)$ will be aggregated by considering a particular θ function. This leads to $\Theta(\tilde{u}) = \sum_{c \in A(u)} \theta(c)$. Mazandu and Mulder [2011] (Equation 3.35) recently implicitly proposed a Θ function using such a strategy to evaluate the specificity of a concept – defining $\Theta(\tilde{u}) = \sum_{c \in A(u)} IC(c)$.

4.3.2.3 Estimating the commonality of two representations (Ψ)

$$\Psi : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{R}_+$$

The commonality between concept representations is evaluated by the function Ψ. According to some paradigms, the commonality can be regarded as the amount of information captured by features shared among the semantic representations of these concepts, i.e., intuitively $\Theta(\tilde{u} \cap \tilde{v})$. For example, when $\tilde{u}$ is associated to a set-based representation, e.g. $\tilde{u} = A(u)$, a commonly used strategy is to define the commonality to $|A(u) \cap A(v)|$ (sim_{RE} Equation 3.24). In other words, the function Ψ assesses the specificity of the part of the semantic representations of the compared concepts which is shared. This stresses that the function $\Psi(\tilde{u}, \tilde{v})$ can, in some case, implicitly be seen as: $\Psi(\tilde{u}, \tilde{v}) = \Theta(\Psi'(\tilde{u}, \tilde{v}))$ with, $\Psi' : \mathbb{K} \times \mathbb{K} \rightarrow \mathbb{K}$[1]. Nevertheless, to lighten the formalism we do not consider this extension.

Numerous similarity measures consider taxonomies as tree structures. In a tree, there is just a single concept ω that subsumes two other concepts u, v such as $A(\omega) = A(u) \cap A(v)$. The notions of LCA and MICA correspond to this concept ω. Thus, in trees, the function Ψ can assess the commonalities of two concepts by just considering ω.

However, because of the presence of multiple inheritances in most widely used taxonomies (e.g., in the biomedical domain for instance), the notion of a single subsuming concept ω characterising the whole commonality of two concepts is not usually fulfilled. Therefore, in order to capture the commonalities of two concepts (u, v), $\Psi(\tilde{u}, \tilde{v})$ must define an aggregation strategy while taking into account the specificity of all concepts which compose $\Omega(\tilde{u}, \tilde{v})$, that is, the set of non-comparable common ancestors of concepts u and v (NCCAs, introduced in Section 3.2.3). In other words, for most ontologies, $\Omega(u, v)$ will (theoretically) be a more accurate estimator of the commonality than ω.

Each concept in $\Omega(u, v)$ represents a particular *semantic facet* of the commonality between the concepts u and v. Some approaches which evaluate the commonality explicitly aggregate the amount of information carried by the semantic facets defined in Ω. However, most measures adopt the maximal strategy as they only exploit ω^*, that is, the concept from Ω which maximises a selected θ function. Measures relying on the MICA (e.g. [Lin, 1998; Resnik, 1995]) or on the LCA (e.g. [Wu and Palmer, 1994]) are examples of this strategy. Nevertheless, other aggregations have been proposed [Couto and Silva, 2011; Couto et al., 2005]. For example, GraSM strategy proposes to average the specificities of concepts in Ω using a specific θ function, it can therefore be generalised

[1] The domain of the function Φ and ζ could also be modified.

by:

$$\Psi_{GraSM}(\tilde{u},\tilde{v}) = \frac{\sum_{c\in G_{\Omega^+}} \theta(c)}{|\Omega|} \tag{4.10}$$

Note that for ontologies incorporating multiple inheritances, the commonality of a pair of concepts can also be estimated by taking into account their common descendants (which can be seen as their shared potential extensions). The problem is symmetrical to the estimation of the commonality based on shared ancestors Ω (which could be renamed Ω^+). Likewise, a set Ω^- representing the non-comparable common descendants of two concepts can also be expressed. Estimation of the commonality of concepts based on the study of their descendants has been recently introduced in [Yang et al., 2012].

As we have seen, evaluating the commonalities of two concepts is, in most cases, equivalent of evaluating the specificity of the semantic representation built from the group of concepts Ω , i.e. $\Theta(\widetilde{\Omega})$. Existing approaches (LCA/MICA, e.g., [Resnik, 1995; Wu and Palmer, 1994], GraSM and DiShIn [Couto and Silva, 2011; Couto et al., 2005]) only define an aggregation strategy over the specificity of elements defined in Ω.

4.3.2.4 Estimating the difference of two representations (Φ)

$$\Phi : \mathbb{K}\times\mathbb{K} \rightarrow \mathbb{R}_+$$

Some measures also rely on the differences between the semantic representations associated to compared concepts, which we refer to as function Φ. Considering two concepts u,v, the amount of knowledge contained in $\tilde{u}$ that is not in $\tilde{v}$ is intuitively expressed by:

$$\Phi(\tilde{u},\tilde{v}) = \Theta(\tilde{u}) - \Psi(\tilde{u},\tilde{v}) \tag{4.11}$$

In practice, Φ is usually computed as $\Phi(\tilde{u},\tilde{v}) = \Theta(\tilde{u}) - \Theta(\tilde{\Omega})$. Moreover, similarly to Ψ, numerous Φ approaches only consider ω^*[1] to estimate the difference of representations associated to singleton. This results in $\Phi(\tilde{u},\tilde{v}) = \Theta(\tilde{u}) - \Theta(\widetilde{\omega^*})$, which is usually expressed by $\Phi(\tilde{u},\tilde{v}) = \theta(u) - \theta(\omega^*)$. We present an example of such a formulation used in the well-known Jiang and Conrath measure:

$$\begin{aligned} dist_{JC}(u,v) &= IC(u) + IC(v) - 2\times IC(MICA(u,v)) \\ &\approx \theta(u) - \theta(\omega^*) + \theta(v) - \theta(\omega^*) \\ &\approx \Phi(\tilde{u},\tilde{v}) + \Phi(\tilde{v},\tilde{u}) \end{aligned}$$

[1]The concept from Ω which maximises a selected expression of the function θ.

We thus obtain the definition of the distance by setting: $\Phi(\tilde{u}, \tilde{v}) = IC(u) - IC(MICA(u, v))$.

For edge-counting approaches, as introduced in Section 3.5.1, the differences of a concept u with respect to v are usually assessed from the length of the shortest path between the concept and their LCA:

$$\Phi(\tilde{u}, \tilde{v}) = sp(u, isa, LCA(u, v)) \tag{4.12}$$

Other strategies can be defined to aggregate the differences between a concept and those contained in Ω. As an example, some information theoretical measures (e.g. sim_{DIC}, Equation 3.35) take into account all the information related to Ω, as follows:

$$\Phi(\tilde{u}, \tilde{v}) = \Theta\left(A(u) \setminus \bigcup_{c \in \Omega} A(c)\right) = \Theta\left(A(u) \setminus (A(u) \cap A(v)\right)$$

Despite some measures use particular instantiations of Φ, the vast majority of measures exploiting semantic differences estimate the difference between two concepts as $\Phi(\tilde{u}, \tilde{v}) = \Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$.

4.3.2.5 Other components

As we will see, most measures can be expressed using the abstract functions introduced so far. Nevertheless, some semantic similarity measures also use supplementary functions. They can be used, for example, to aggregate scores of multiple semantic measures or to impact the final score produced by a measure. Table 4.3 shows some functions which can be used to tune numerous semantic measures. This is done by taking information not originally captured by the original measure definition.

As an example, we present two supplementary functions used for the evaluation of the similarity of two concepts. The first function is used in conjunction with a pairwise measure respecting the property of the identity of the indiscernibles. The idea is to lower the score between concepts which are characterised as broader. For this, the final score is modified according to a function aiming to capture the relevance of the score. A technique proposed to assess such relevance is based on the evaluation of the amount of information carried by the part shared by two representations, i.e., $\Theta(\tilde{u} \cap \tilde{v})$, e.g., the IC of the MICA of the compared concepts. This function was originally proposed in sim_{Rel} (Equation 3.34).

Considering sim' as a semantic similarity measure respecting the identity of indiscernibles and $rfactor \in [0, 1]$ as a function capturing the relevance of a score of similarity

Name	Comment
GraSM [Couto et al., 2007] / Dishin [Couto and Silva, 2011]	Estimation of the commonalities between a pair of concepts averaging the information contained by all NCCAs i.e. Ω. If $\theta(c)$ is set to $IC(c)$, using GraSM we obtain $\Psi(\tilde{u},\tilde{v}) = \frac{\sum_{c\in\Omega} IC(c)}{\lvert\Omega\rvert}$
Relevance factor [Li et al., 2010; Schlicker et al., 2006]	Impact the score of a measure sim which respects the identity of indiscernibles, e.g., $sim(u,v) = sim'(u,v) * rfactor$ with $rfactor$ a metric which captures the relevance of the score ($rfactor \in [0,1]$), for example $rfactor = 1 - p(MICA(u,v))$ [Schlicker et al., 2006] with $p(MICA(u,v))$ the probability of occurrence associated to the MICA.
Descendant + Open World Assumption [Yang et al., 2012]	Composite measure aggregating scores of semantic measures evaluating different aspects of the compared elements, e.g., coupling classical measures based on $\rho(c) = G_T^+(c)$ with a measure taking into account shared descendants and their uncertainty.

TABLE 4.3: Strategies that can be used to tune semantic measures

(e.g. $\theta(\omega^*)$) we obtain:

$$sim(u,v) = sim'(u,v) \times rfactor$$

A second function can be used to aggregate pairwise scores obtained by various semantic similarity measures on a similar pair of concepts [Yang et al., 2012]. The function can be used to weigh the contribution of semantic similarity measures focusing on particular aspects of the compared concepts, e.g., G_T^+ and G_T^-, respectively denoted by $sim_{G_T^+}$ and $sim_{G_T^-}$. An example is proposed below:

$$sim(u,v) = \alpha \cdot sim_{G_T^+}(u,v) + \beta \cdot sim_{G_T^-}(u,v)$$

Note that other functions can also be added to design new measures.

4.3.3 Unification of abstract similarity measures

In this subsection, we demonstrate the relationships between known abstract expressions of measures through the definition of a new parameterised measure.

In previous sections we have identified the core elements of semantic similarity measures. Moreover, we have underlined that set-based measures can be used to express abstract measures. By extension, we also stressed that Caillez & Kuntz σ_α and Gower & Legendre σ_β formulas (presented in Section 4.2.2) may be considered as abstract parameterised measures. By focusing on the unification of measure expressions, we here demonstrate that under some conditions, σ_α and σ_β can be partially unified and extended through a

common expression. Formulas used in the demonstration are abstracted and repeated for convenience:

$$\sigma_{\alpha^*}(\tilde{u},\tilde{v}) = \frac{\Psi(\tilde{u},\tilde{v})}{(\frac{\Theta(\tilde{u})^\alpha+\Theta(\tilde{v})^\alpha}{2})^{1/\alpha}} \tag{4.13}$$

$$\sigma_{\beta^*}(u,v) = \frac{\beta\cdot\Psi(\tilde{u},\tilde{v})}{\Theta(\tilde{u})+\Theta(\tilde{v})+(\beta-2)\cdot\Psi(\tilde{u},\tilde{v})} \tag{4.14}$$

$$sim_{RM^*}(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{x\cdot\Phi(\tilde{u},\tilde{v})+y\cdot\Phi(\tilde{v},\tilde{u})+\Psi(\tilde{u},\tilde{v})}$$

We first demonstrate that σ_{α^*} can be easily extended to the well-known *generalised mean of order* α [Webster, 1994] (Result 1). In addition, we show that σ_{β^*} is a particular case of the *ratio model* proposed by Tversky (Result 2). Finally, based on Results 1 & 2, we demonstrate that a new abstract tunable measure can be used to express a large diversity of abstract measures (Result 3).

Result 1. First, note that Cauchy's mean σ_{α^*} implies a symmetric contribution of $\Theta(\tilde{u})$ and $\Theta(\tilde{v})$. In a straightforward manner, we extend σ_{α^*} to the generalised mean of order α. This is done by introducing two parameters x and y enabling us to tune $\Theta(\tilde{u})$ and $\Theta(\tilde{v})$ contributions.

$$\sigma_{\alpha,x,y^*}(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{(x\cdot\Theta(\tilde{u})^\alpha+y\cdot\Theta(\tilde{v})^\alpha)^{1/\alpha}} \tag{4.15}$$

with $x+y=1$ and $x,y\geq 0$. σ_α is a special case of σ_{α,x,y^*} when $x=y=1/2$.

Result 2. We demonstrate the relationship between σ_{β^*} and the abstract formulation of the *ratio model* (sim_{RM^*}). Recall that $\Theta(\tilde{u})$ (resp. $\Theta(\tilde{v})$) represents the amount of knowledge carried by a concept representation $\tilde{u}$ (resp. $\tilde{v}$). The function $\Theta(\tilde{u})$ is commonly considered as additive, i.e., $\Theta(\tilde{u}\cup\tilde{v}) = \Theta(\tilde{u})+\Theta(\tilde{v})$ for any pair of non-comparable[1] semantic representations $(\tilde{u},\tilde{v})$. With this condition we can demonstrate the following lemma.

Lemma. Considering $\Theta(\tilde{u}) = \Phi(\tilde{u},\tilde{v})+\Psi(\tilde{u},\tilde{v})$ for any $\tilde{v}$, σ_{β^*} is a particular case of the abstract formulation of the *ratio model* (sim_{RM^*})[2].

[1] This notion depends on the consideration of the function ρ.

[2] By extension, this applies to any specific case derived from instantiation of the framework which respects the given properties – such as set-based formulations.

Proof.
Considering the inverse of both σ_{β^*} and the sim_{RM^*}, we obtain:

(a) setting $x = y$ in sim_{RM^*}:

$$\frac{1}{sim_{RM^*}(u,v)} = 1 + x\frac{\Phi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})} + x\frac{\Phi(\tilde{v},\tilde{u})}{\Psi(\tilde{u},\tilde{v})}$$

(b) in addition,

$$\begin{aligned}\frac{1}{\sigma_{\beta^*}(u,v)} &= 1 - \frac{2}{\beta} + \frac{1}{\beta}\frac{\Theta(\tilde{u})}{\Psi(\tilde{u},\tilde{v})} + \frac{1}{\beta}\frac{\Theta(\tilde{v})}{\Psi(\tilde{u},\tilde{v})} \\ &= 1 - 2x + x \cdot \frac{\Theta(\tilde{u})}{\Psi(\tilde{u},\tilde{v})} + x \cdot \frac{\Theta(\tilde{v})}{\Psi(\tilde{u},\tilde{v})} \text{ (with } x = \frac{1}{\beta})\end{aligned}$$

Thus, considering $\Theta(\tilde{u}) = \Phi(\tilde{u},\tilde{v}) + \Psi(\tilde{u},\tilde{v})$, (a) and (b), we obtain:

$$\begin{aligned}\frac{1}{\sigma_{\beta^*}(u,v)} &= 1 - 2x + x \cdot \frac{\Phi(\tilde{u},\tilde{v}) + \Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})} + x \cdot \frac{\Phi(\tilde{v},\tilde{u}) + \Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})} \\ &= 1 + x \cdot \frac{\Phi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})} + x \cdot \frac{\Phi(\tilde{v},\tilde{u})}{\Psi(\tilde{u},\tilde{v})} \\ &= \frac{1}{sim_{RM^*}(u,v)}\end{aligned}$$

□

Therefore, σ_{β^*} is a particular case of the abstract *ratio model* sim_{RM^*} considering an equal contribution of $\Phi(\tilde{u},\tilde{v})$ and $\Phi(\tilde{v},\tilde{u})$ (i.e. $x = y$).

Result 3. σ_{α,x,y^*} and the sim_{RM^*} (which includes σ_{β^*}, see Result 2) may be expressed by the general function Σ_{α,x,y,z^*} (shorten by Σ_*).

$$\Sigma_*(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{(x \cdot \Theta(\tilde{u})^\alpha + y \cdot \Theta(\tilde{v})^\alpha + z \cdot \Psi(\tilde{u},\tilde{v})^\alpha)^{1/\alpha}} \tag{4.16}$$

with $x, y, z \geq 0$ and $x + y + z = 1$. Note that by setting $\alpha = 1$ and $\Theta(\tilde{u}) = \Phi(\tilde{u},\tilde{v}) + \Psi(\tilde{u},\tilde{v})$, the abstract measure Σ_* can also be formulated as:

$$\Sigma_*(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{x \cdot \Phi(\tilde{u},\tilde{v}) + y \cdot \Phi(\tilde{v},\tilde{u}) + (x + y + z) \cdot \Psi(\tilde{u},\tilde{v})} \tag{4.17}$$

In this subsection, we have demonstrated that existing abstract measures can be generalised to the Σ_* abstract measure and that a large diversity of measures can be derived from it. Unifying abstract measures opens interesting perspectives for measure optimisation. Indeed, expressing measures through a common parameterised formula enables better understanding of the relationships between the various proposals. Moreover, as we

will see, a large variety of measures can easily be instantiated by tuning few parameters. Unification of measures is therefore a prerequisite in order to distinguish parameters best impacting measure accuracy.

The proposed framework – abstract components and measures distinguished – can easily be used to define semantic measures to compare a pair of concepts or groups of concepts. In the state-of-the-art, we have seen that groupwise measures (i.e., measures used to compare groups of concepts) can be expressed according to two strategies: *direct* and *indirect* (Section 3.6).

Groupwise measures built using the direct approach have already been taken into account by the proposed framework. Indeed, all the measures rely on the function ρ which has been defined to represent a set of concepts. Therefore, all (abstract formulations of) measures which have been defined to compare a pair of concepts can be used to derive groupwise measures.

Measures built using the indirect approach rely on an aggregation of pairwise measures. Therefore, to be encompassed by the current framework, we only need to consider an extra aggregation function which will aggregate the similarity matrix corresponding to the similarity scores of the Cartesian product of the compared sets.

4.4 Expression of measures using the framework

4.4.1 Guidelines for framework instantiation

We define the guidelines to instantiate/design semantic similarity measures from the proposed framework. Two main steps can be distinguished. To ease the presentation, we focus on the design of measures for the comparison of a pair of concepts:

1. *Selection of an abstract measure*, such as Σ_*, σ_{α^*}, sim_{RM^*} (see Section 4.3.3).
2. *Definition of the expression of the core elements.* This step consists of selecting a specific semantic representation of a concept (ρ function) and the definition of the expression of the abstract operators on which the selected abstract measure relies – for instance to estimate the commonality (Φ) or the difference (Ψ) between two concept representations.

4.4.1.1 Selection of an abstract measure

The first step in designing a semantic measure is to select an abstract measure. This measure is defined through the core elements distinguished by the framework. The multiple parameterised abstract measures discussed in the previous section can be used to express a large diversity of measures. In addition, set-based expressions proposed in the literature can also easily be *abstracted* using the core elements of the framework.

Indeed, the proposed framework enables the full use of studies made for other types of measures. As an example, the proposed core elements can be mapped to existing theoretical tools used by other communities to study binary measures (e.g., measures used to compare vectors or sets). Table 4.4 shows abstract expressions of the Operational Taxonomic Units (OTUs) classically used to represent binary measures. In a similar manner to the approach relying on information theory, the amount of information expressed in a taxonomy G_T can be viewed as $\Theta(G_T)$. The amount of information encompassed in the semantic representation of a concept is expressed by $\Theta(\tilde{c})$, and the amount of information expressed in G_T which is not found in $\tilde{c}$ can be defined by $\overline{\Theta(\tilde{c})}$.

$u \setminus v$	$\Theta(\tilde{v})$	$\overline{\Theta(\tilde{v})}$
$\Theta(\tilde{u})$	$\Psi(\tilde{u},\tilde{v})$	$\Phi(\tilde{u},\tilde{v})$
$\overline{\Theta(\tilde{u})}$	$\Phi(\tilde{v},\tilde{u})$	$\zeta(\tilde{u},\tilde{v})$

TABLE 4.4: Links between Operation Taxonomic Units (OTUs) commonly used for the definition of binary measures and the theoretical framework core elements – see Choi et al. [2010] for numerous expressions of binary measures using OTUs

Name	Expressions
$sim_{\cap^*}$	$\Psi(\tilde{u},\tilde{v})$
sim_{CM^*}	$\lambda\cdot\Psi(\tilde{u},\tilde{v})-\alpha\cdot\Phi(\tilde{u},\tilde{v})-\beta\cdot\Phi(\tilde{v},\tilde{u})$
sim_{RM^*}	$\frac{\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})+\alpha\cdot\Phi(\tilde{u},\tilde{v})+\beta\cdot\Phi(\tilde{v},\tilde{u})}$
$sim_{Simpson^*}$	$\frac{\Psi(\tilde{u},\tilde{v})}{min(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v}),\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u}))}$
$sim_{Braun-Blanquet^*}$	$\frac{\Psi(\tilde{u},\tilde{v})}{max(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v}),\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u}))}$
$sim_{Maryland-Bridge^*}$	$\frac{1}{2}\left(\frac{\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v})}+\frac{\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u})}\right)$
sim_{Bader^*}	$\frac{\Psi(\tilde{u},\tilde{v})^2}{(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v}))(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u}))}$
sim_{Knappe^*}	$k\frac{\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v})}+(1-k)\frac{\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u})}$
sim_{Ochaia^*}	$\frac{\Psi(\tilde{u},\tilde{v})}{\sqrt{(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v}))(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u}))}}$
sim_{Cosine^*}	$\frac{\Psi(\tilde{u},\tilde{v})}{\sqrt{(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{u},\tilde{v}))^2}\sqrt{(\Psi(\tilde{u},\tilde{v})+\Phi(\tilde{v},\tilde{u}))^2}}$

TABLE 4.5: Examples of abstract semantic measures derived from classical binary measures. Most of the binary measures have been abstracted from [Choi et al., 2010] considering Table 4.4

The mapping proposed in Table 4.4 can be used to easily express semantic similarity measures based on binary measure expressions defined through OTUs. As an example, in Choi et al. [2010], a large characterisation of binary measures through OTUs is performed. The authors distinguish more than seventy expressions of binary measures. Using Table 4.4, these expressions can be used to easily express a large diversity of abstract semantic measures. The main idea is to generalise existing binary measures using the proposed core elements of the framework in order to derive semantic similarity measures; examples of abstract measures are presented in Table 4.5.

The abstract measure which will be selected to instantiate a concrete measure partially defines the semantics of the compared concepts which will be taken into account during the comparison, e.g. commonalty (Ψ), difference (Φ), and also their weight in the similarity assessment. As an example, we have seen that both the Jaccard index and the Dice coefficient can be derived from the Tversky's *ratio model* by setting $\alpha,\beta=1$ and $\alpha,\beta=0.5$, respectively. It is therefore explicit that the Dice coefficient gives more importance to commonalities (and less importance to differences) for similarity estimation, compared to the Jaccard Index. The selection of the abstract measure is

therefore important to finely control the meaning of the scores produced by a measure. This aspect may be particularly important for context-specific applications.

4.4.1.2 Definition of the expression of the core elements

The next step consists of defining how to represent a concept according to the ontology. Such representation is defined by the function ρ. It is required to derive the expression of the operators used by the abstract measure. Indeed, expressions for abstract operators (e.g. estimators of commonalities or differences) must be defined in accordance with the selected expression of ρ. Finally, the selection of a specific representation, e.g. set of concepts $\tilde{u} = A(u)$ to represent a concept u, also partially defines which semantics will be considered in the similarity assessment. Examples will be provided in the next subsections.

4.4.1.3 How to select adapted parameters

The users will therefore have to consider (i) specific expressions of the primitive functions distinguished by the framework, (ii) abstract semantic measures and (iii) specific parameter freedom. Two scenarios can therefore be distinguished:

1. The designer has a very clear idea about the more relevant elements that guide the similarity assessment in the concrete scenario and their relative weights. He thus tunes and obtains the measure accordingly. Some of the parameters on which the measures rely can, for example, be restricted due to constraints defined by the context of use (e.g. the measure must be symmetric: the user will therefore only consider setting where $\alpha = \beta$ in the abstract *ratio model*).
2. The designer optimises semantic measure parameters using a benchmark from which the accuracy of measures can be evaluated. As an example, the designer has a training set of similarity scores (human-rated) that would be expected to be produced by a semantic measure[1]. The scores can be used to evaluate the accuracy of measures resulting from the framework instantiation. The set of measures to be evaluated can eventually be restricted according to specific properties induced by specific core element expressions or abstract measures, e.g., algorithmic complexity (cf. scenario 1). The selection of the best suited measures will therefore be performed empirically using the training set from which performances of measures can be estimated. Such a training set or test sample must be composed of expected scores of similarity for a reasonable amount of pairs of concepts. It

[1]This is the most common type of benchmarks in the domain. Refer to [Harispe et al., 2013c] for a review.

must be built alongside the experts of the domain according to the behaviour we want the system to have. The selection of the most appropriate measures can be made studying correlations between expected scores and semantic measures scores of similarity.

With the aforementioned method, our framework can be used to easily instantiate existing or new semantic similarity measures, while finely controlling the semantics considered during the similarity assessment. Such a constructive approach draws interesting perspectives for evaluating semantic measures, such as testing the influence of the various components (i.e., abstract measures, core element expressions) over the accuracy of concrete measures in domain-specific tasks.

4.4.2 Expression of semantic similarity measures

This subsection presents concrete examples of semantic similarity measures derived from the framework.

4.4.2.1 Expression of pairwise measures

To illustrate the generality and potential of the proposed framework, we present some instantiations corresponding to existing measures that can be obtained from the abstract form of the Jaccard index:

$$sim_{Jaccard^*}(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{\Psi(\tilde{u},\tilde{v}) + \Phi(\tilde{u},\tilde{v}) + \Phi(\tilde{v},\tilde{u})}$$

considering $\Phi(\tilde{u},\tilde{v}) = \Theta(\tilde{u}) - \Psi(\tilde{u},\tilde{v})$,

$$sim_{Jaccard^*}(u,v) = \frac{\Psi(\tilde{u},\tilde{v})}{\Theta(\tilde{u}) + \Theta(\tilde{v}) - \Psi(\tilde{u},\tilde{v})}$$

Based on specific expressions of the functions Ψ and Φ presented in Table 4.6, $sim_{Jaccard^*}$ can be used to express sim_{PS}, sim_{Faith} or sim_{CMatch} (see Equations 4.18, 4.19, 4.20):

$$sim_{PS}(u,v)^{1} = \frac{lp(LCA(u,v), isa, \top)}{lp(u, isa, LCA(u,v)) + lp(v, isa, LCA(u,v)) + lp(LCA(u,v), isa, \top)} \tag{4.18}$$

[1]Compared to Equation 3.17 presented in Section 3.5.1, $depth(u,v)$ is here substituted by $sp(LCA(u,v), isa, \top)$. In DAGs which contain redundancies the longest shortest path should be considered instead, i.e., the shortest path in the graph without redundancies.

Elements	sim_{PK}	sim_{Faith}	sim_{CMatch}	sim_{cGIC}
$\rho(u) = \tilde{u}$	$G_T^+(u)$	$A(u)$	$A(u)$	$A(u)$
$\Theta(\tilde{u})$	$lp(u, isa, \top)$	$IC(u)$	$\lvert A(u)\rvert$	$\sum_{c \in A(u)} IC(c)$
$\Psi(\tilde{u}, \tilde{v})$	$lp(LCA(u,v), isa, \top)$	$IC(MICA(u,v))$	$\lvert A(u) \cap A(v)\rvert$	$\sum_{c \in A(u) \cap A(v)} IC(c)$
$\Phi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$

TABLE 4.6: Examples of particular expressions of core elements from which pairwise semantic similarity measures can be obtained as instantiations of an abstract form of the Jaccard index. These can also be used to obtain other measures using different set-based coefficients

$$sim_{Faith}(u,v) = \frac{IC(MICA(u,v))}{IC(u) + IC(v) - IC(MICA(u,v))} \tag{4.19}$$

$$sim_{CMatch}(u,v) = \frac{|A(u) \cap A(v)|}{|A(u)| + |A(v)| - |A(u) \cup A(v)|} = \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|} \tag{4.20}$$

It can also be used to express sim_{cGIC}, a *new*[1] pairwise measure based on sim_{GIC} (a measure initially designed to compare groups of concepts, Equation 3.40):

$$sim_{cGIC}(u,v) = \frac{\sum_{c \in A(u) \cap A(v)} IC(c)}{\sum_{c \in A(u)} IC(c) + \sum_{c \in A(v)} IC(c) - \sum_{c \in A(u) \cap A(v)} IC(c)} \tag{4.21}$$

An interesting aspect of the modularity provided by the framework is that a component of measures can easily be tuned to generate *new* measures best fitting specific needs. As an example, sim_{Faith} (Equation 4.19) considers the MICA as an estimator of commonality. As we have seen, this estimator can be limiting for the comparison of concepts defined in ontologies in which multiple inheritance is extensively used[2]. Therefore the expression $\Psi(\tilde{u}, \tilde{v})$ can be modified in order to consider the whole information contained in Ω, the set of NCCAs of compared concepts. As an example, we present $sim_{Faith-ex}$, an extended version of sim_{Faith} which considers the whole set of NCCAs according to the mixing strategy defined by GraSM (method introduced in Section 3.5.5.3):

$$sim_{Faith-ex}(u,v) = \frac{\frac{\sum_{c \in \Omega(u,v)} IC(c)}{|\Omega|}}{IC(u) + IC(v) - \frac{\sum_{c \in \Omega(u,v)} IC(c)}{|\Omega|}} \tag{4.22}$$

[1] We here adopt the terminology that has been used in the literature dedicated to semantic measures for decades; however the terms *unpublished* or *expression* are more appropriated.

[2] Empirical evaluations have underlined an improvement of the accuracy of measures in tested context usages [Couto and Silva, 2011].

Other examples of instantiations of semantic measure will be presented throughout this manuscript, in particular in Section 5.1.2.2.

4.4.2.2 Expression of groupwise measures

Groupwise measures are used to compare groups of concepts. As we saw, they can be expressed using an indirect strategy, by aggregating pairwise measures (those used to compare a pair of concepts), or using a direct strategy, by generalising the approaches used for expressing pairwise measures. Groupwise measures which are based on an indirect strategy only require an aggregation strategy to be defined in order to be expressed by the framework presented so far. We therefore focus on those which rely on a direct strategy.

Let us remind that the framework has been designed by defining the domain of the function ρ in order to encompass cases in which we want to represent a set of concepts, i.e. $\rho : \mathcal{P}(C) \rightarrow \mathbb{K}$. Therefore, the framework already implicitly takes into consideration groupwise measures. As an example, we propose to use the framework to express sim_{GIC} (Equation 3.40) [Pesquita et al., 2007], sim_{UI} (Equation 3.38), and sim_{LP}[1].

Considering expressions presented in Table 4.7, comparing two sets of concepts U,V, sim_{LP} is defined only considering $\Psi(\tilde{U}, \tilde{V})$. We can also see that the measures sim_{GIC} and sim_{UI} can be expressed in a straightforward manner from an abstract expression of the Jaccard index ($sim_{Jaccard^*}$, see Section 4.4.2.1).

Elements	sim_{LP}	sim_{UI}	sim_{GIC}
$\rho(U) = \tilde{U}$	$G_T^+(U)$	$\bigcup_{c \in U} A(c)$	$\bigcup_{c \in U} A(c)$
$\Theta(\tilde{U})$	$\underset{c \in U}{\text{argmax}}\ lp(c, isa, \top)$	$\lvert\tilde{U}\rvert$	$\sum_{c \in \tilde{U}} IC(c)$
$\Psi(\tilde{U}, \tilde{V})$	$\underset{c \in C_T^+(U) \cap C_T^+(V)}{\text{argmax}}\ lp(c, isa, \top)$	$\lvert\tilde{U} \cap \tilde{V}\rvert$	$\sum_{c \in \tilde{U} \cap \tilde{V}} IC(c)$
$\Phi(\tilde{U}, \tilde{V})$	$\Theta(\tilde{U}) - \Psi(\tilde{U}, \tilde{V})$	$\Theta(\tilde{U}) - \Psi(\tilde{U}, \tilde{V})$	$\Theta(\tilde{U}) - \Psi(\tilde{U}, \tilde{V})$

TABLE 4.7: Examples of particular expressions of core elements from which groupwise semantic similarity measures can be expressed. With $C_T^+(U) = C(G_T^+(U))$.

[1] "*For simLP the similarity measure is the depth of the longest shared path from the root node* [considering a set of concepts X as the graph induced by the union of the ancestors of each concept of X, i.e. $G_T^+(X) = \bigcup_{c \in X} G_T^+(c)$]" [Gentleman, 2007].

4.5 Chapter conclusion

A large diversity of semantic similarity measures have been proposed over recent decades. Most of them focus on specific applications or domains and have been introduced as *new* formulations unrelated to existing proposals. In this chapter, in continuation of existing works which underline relationships between measures, we unified most well-known approaches through the definition of a theoretical framework dedicated to semantic similarity measures.

The main advantages of the proposed framework rely on the identification of the core elements which are commonly used to design semantic similarity measures. We have indeed underlined that most measures can be expressed considering a limited set of core elements (functions) such as those defining (i) how to represent a concept through a processable canonical form (ρ), (ii) how to estimate its specificity (θ) and the specificity of its representation (Θ), and (iii) how to estimate the degree of commonality (Ψ) and difference (Φ) between two concept representations. In fact, we demonstrate how these core elements can be used to express a large diversity of (existing) measures based on generic parametric measures which can be seen as the backbone of semantic measures. The characterisation of measures through the distinguished core elements can be used to better characterise measures relying on different paradigms and to better understand the large diversity of measures introduced in the literature (both pairwise and groupwise). More generally, this framework opens interesting perspectives for the study of semantic measures as it provides a theoretical tool enabling to drive:

- *Theoretical analysis and the understanding of semantic measures.* Distinguishing the core elements on which semantic similarity measures are based allows us to highlight narrow relationships between existing proposals. Indeed, we found that semantic similarity measures can be easily expressed through the definition of a few intuitive core elements and that most, if not all, measures are just particular expressions of a limited set of abstract measures. We therefore demonstrated that several measures which rely on the same abstract measure (e.g., abstracted *ratio model*), only differ due to a specific set of parameters selected to instantiate them (e.g., strategy used to represent a concept or to assess the commonality/difference between concept representations). This strong result is therefore important for the theoretical analysis of semantic measures. Indeed, most applications in which the measures are not selected through empirical analyses expect the measures to fulfil specific properties, e.g., symmetry, respect of the identity of the indiscernibles. Thanks to the breakdown of measures proposed by the framework, their properties can be analysed, not only regarding specific measure instantiations, but also

focusing on both the abstract measures from which they derived and the properties induced by core elements.

- *Creation and tuning of semantic similarity measures.* The separation of measures from the core elements on which they rely enables researchers to focus not just on new *ad hoc* measures, but also on the design of specific strategies to improve the assessment of those core elements. As an example, we have seen that an accurate estimator of the commonality between two concepts (Ψ function), which depends on the canonical form adopted to represent a concept (ρ function), is of major importance in defining semantic measures. Designers of measures can therefore improve several existing measures by improving the way Ψ is estimated w.r.t a specific representation of a concept. It is therefore important to understand that improving the assessment of core elements distinguished by the framework leads to improvements in multiple measures – not just to a specific measure in a concrete context. By distinguishing the core elements of semantic similarity measures, the theoretical tool proposed in this chapter therefore opens interesting perspectives for the definition and improvement of semantic measures in general.

5

Unifying framework: illustration of applications

Contents

Abstract

This chapter presents two practical applications of the theoretical framework of semantic measures introduced in Chapter 4. Illustrations are provided considering a use case scenario related to the biomedical domain.

First, we propose an evaluation of semantic similarity measures using the insight provided by the framework. To this end, numerous measures are expressed from two parametric abstract functions (sim_{RM^*} and sim_{CM^*}); they are used to instantiate concrete measures from specific expressions of the framework's core elements (e.g., ρ, Ψ, Φ). The accuracy of the concrete measures generated is next discussed – the evaluation is based on a gold-standard benchmark built from physician and coder expectations regarding the semantic similarity of biomedical concepts. This study will help us to discuss the notion of semantic measure accuracy and selection. It also gives us the opportunity to discuss the accuracy of measures at the level of granularity provided by the framework, e.g., to discuss the impact to consider a specific expression of this or that core element. Preliminary results are presented. They highlight the new insights and prospects offered by such studies, in particular for the selection and design of semantic measures.

In the second practical application, we study how to extend the process which is commonly used to evaluate the accuracy of measures in order to incorporate uncertainties in experts' judgement (associated to benchmarks). This allows for the introduction, definition and discussion of the notion of semantic measures robustness w.r.t uncertainty on expected scores. Through a use case example, we present the interesting perspectives offered by this notion, in particular to distinguish semantic measures that best resist aforementioned uncertainties, i.e., measures which guarantee good performance even if the benchmark considered for their selection contains approximations. Despite its importance for the practical use of semantic measures, this is, to our knowledge, an aspect of measures which has never been studied.

Associated references on which this chapter is based:

- **A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain**. Sébastien Harispe*, David Sánchez, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. Journal of Biomedical Informatics 2013.
- **Robust Selection of Domain-specific Semantic Similarity Measures from Uncertain Expertise**. Stefan Janaqi*, Sébastien Harispe, Sylvie Ranwez, Jacky Montmain. IPMU 2014 – Information Processing and Management of Uncertainty in Knowledge-Based Systems (In press).

5.1 Selection and optimisation of semantic measures

5.1.1 Motivation and objectives

The selection of semantic measures is a central question which has not been deeply studied in this thesis. It is, however, one of the main centres of interest to the end-users of semantic measures. It is indeed common that experts in the field are asked for the *best* semantic measure? Today, there are two short answers: (i) There is no best semantic measure, (ii) I don't know. The important thing to understand is that state-of-the-art analyses have proved that domain-specific results cannot be generalised. Indeed, despite the fact that specific approaches tend to often provide *reasonable* results in most cases (e.g., some specific measures based on information theory), there is no guarantee that measures which have been proved to be accurate in a specific context usage will remain accurate in another context of use.

In all cases, to distinguish best suited measures we first have to define what a *good* or the *best* measure is. What aspects of measures must be considered when selecting, e.g., accuracy w.r.t expected scores, computational complexity, specific mathematical properties? Though these questions are not deeply discussed in the literature, matters related to the subject have briefly been proposed in [Harispe et al., 2013c]. In this study, we distinguished four criteria for the evaluation of measures in particular:

- Their *accuracy* and *precision*.
- Their *computational complexity*, i.e. *algorithmic complexity*.
- Their *mathematical properties*.
- Their *semantics*.

Please refer to Appendix B for a brief discussion on these central aspects of measures.

In Chapter 4, we defined a framework which provides the interesting possibility to split semantic similarity measures into two main components: (i) an abstract measure which aggregates (ii) specific expressions of core elements commonly found in semantic measures.

The main aim of this section is to highlight the benefits of the unifying framework to study and select semantic similarity measures. We propose, in particular, an evaluation of semantic similarity measures using the insight provided by the framework. To this end, numerous *concrete* measures are expressed from two parametric abstract measures (sim_{RM^*} and sim_{CM^*}). These measures have been obtained from specific expressions of the framework's core elements (e.g., ρ, Ψ, Φ). The accuracy of measures is next discussed using a gold-standard benchmark built from physician and coder expectations regarding

the semantic similarity of biomedical concepts. This study will therefore help us to discuss the accuracy of semantic measures and the problem associated to the selection of measures in particular contexts of use. Interestingly, the time will be right to tackle these questions at the level of granularity provided by the framework, e.g., to evaluate the impact to consider a specific expression of this or that core element on the accuracy of measures. Preliminary results are presented. They highlight the new insights and prospects offered by such studies, in particular for the selection and design of semantic measures.

This section is structured as follows. First we define the experimental design defined to generate the measures and to compare them. Next, we discuss the results which have been obtained and more generally, the relevance of using the experimental protocol defined to study and select semantic measures.

5.1.2 Experimental design

5.1.2.1 Benchmark

For this experiment, we focused on the evaluation of semantic measure accuracy. To this end, we considered the gold-standard benchmark which was proposed in Pedersen et al. [2007]. This benchmark is dedicated to the evaluation of semantic similarity measures. It is commonly used in the biomedical domain to evaluate semantic similarity measures according to human judgement of similarity. It contains 29 pairs of concepts associated to semantic similarity scores. Similarity scores are obtained by averaging the ratings given by two groups of experts: 9 medical coders and 3 physicians. Finally, for each pair of concepts, three similarity scores are given: average scores of coders, averaged scores of physicians and averaged scores of both physicians and coders. Pairs of concepts which make up the benchmarks, associated similarities and additional information are provided in Appendix B.2.

The evaluation of semantic similarity measures is usually tackled by computing the Pearson correlation against the similarity ratings given by each group of human experts. The biomedical ontology SNOMED-CT[1] [Spackman, 2004] was used to extract the required semantics, i.e. pairs of concepts denoted by labels have been manually associated to pairs of unambiguous concepts defined in SNOMED-CT[2].

[1] `http://systems.hscic.gov.uk/data/uktc/snomed/index_html`
[2] The mapping is provided in Appendix B.2.

5.1.2.2 Measure definitions from the framework

The study focuses on the evaluation of semantic similarity measures derived from two abstracted forms of semantic measures: the *contrast model* (sim_{CM^*}) and the *ratio model* (sim_{CM^*}) (refer to equations Table 4.5). Notice that in sim_{CM^*}, the γ parameter, which tunes the contribution of the commonality, was fixed to 1; α and β parameters, which tune the importance given to the information found in u (resp. v) which is not found in v (resp. u), were set from 0 to 15 with a step of 0.1. As a result of this parameter tuning, 22,500 abstract expressions of both sim_{CM^*} and sim_{RM^*} were instantiated and systematically evaluated. For each abstract expression we further tested the four instantiations of the core elements shown in Table 5.1.

Elements	**1**	**2**	**3**	**4**
$\rho(u) = \tilde{u}$	$G_T^+(u)$	$A(u)$	$A(u)$	$A(u)$
$\Theta(\tilde{u})$	$lp(u, isa, \top)$	$IC(u)$	$\lvert A(u)\rvert$	$\sum_{c \in A(u)} IC(c)$
$\Psi(\tilde{u}, \tilde{v})$	$lp(LCA(u,v), isa, \top)$	$IC(MICA(u,v))$	$\lvert A(u) \cap A(v)\rvert$	$\sum_{c \in A(u) \cap A(v)} IC(c)$
$\Phi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$	$\Theta(\tilde{u}) - \Psi(\tilde{u}, \tilde{v})$

TABLE 5.1: Core element expressions evaluated by the experiments

Thus, for each abstract similarity measure, the four instantiations of the core elements led to 90,000 individual measures (i.e. 22,500×4). Note that IC-dependent configurations used the IC calculus model defined in [Sánchez et al., 2011] (Equation 3.7). The final experiment is thus based on the evaluation of more than half a million measure configurations, i.e. 180,000 measure configurations for each evaluation benchmark: physicians, coders and the average of both ratings.

Some measures available in the literature correspond to particular points in the range of measure instantiations studied in this experiment. Table 5.2 highlights some of these links.

For each measure configuration, the Pearson correlation with the scores provided by the three groups of experts (coders, physicians and both) were computed.

5.1.2.3 Empirical evaluation and dataset

Empirical evaluations were performed using the Semantic Measures Library, a software tool dedicated to the large-scale analysis and computation of semantic measures which will be introduced in Chapter 8. The source code and detailed documentation

Measures	Eq.	Ref	Abstract	Parameters	Case
sim_{Resnik}	3.28	[Resnik, 1995]	sim_{CM^*}	$\alpha = 0, \beta = 0$	2
sim_{WP}	3.16	[Wu and Palmer, 1994]	sim_{RM^*}	$\alpha = 0.5, \beta = 0.5$	1
sim_{Lin}	3.29	[Lin, 1998]	sim_{RM^*}	$\alpha = 0.5, \beta = 0.5$	2
sim_{DIC}	3.35	[Mazandu and Mulder, 2011]	sim_{RM^*}	$\alpha = 0.5, \beta = 0.5$	4
sim_{PS}	3.17	[Pekar and Staab, 2002]	sim_{RM^*}	$\alpha = 1, \beta = 1$	1
sim_{Faith}	3.33	[Pirró and Euzenat, 2010b]	sim_{RM^*}	$\alpha = 1, \beta = 1$	2
sim_{CMatch}	3.22	[Maedche and Staab, 2001]	sim_{RM^*}	$\alpha = 1, \beta = 1$	3
sim_{cGIC}	4.21	[Harispe et al., 2013d]	sim_{RM^*}	$\alpha = 1, \beta = 1$	4

TABLE 5.2: Examples of parametric expressions of existing semantic measures. Examples of links that can be established between existing semantic similarity measures and measure instantiations which derive from (i) the abstracted contrast and ratio models (sim_{CM^*}, sim_{RM^*} – Table 4.5), and (ii) instantiations of the core elements defined in Table 5.1

related to the experiment is open sourced and available at `http://www.lgi2p.ema.fr/~sharispe/publications/JBI2013`.

5.1.3 Results and discussion

5.1.3.1 Results

Tables 5.3, 5.4 and 5.5 summarise the best results which were obtained for each configuration of abstract measures and for the four specific strategies used to express the core elements (cf. Case columns and Table 5.1). These results will be discussed in the next section.

Case	Best tuning sim_{CM^*}		Best tuning sim_{RM^*}		Correlations	
	α	β	α	β	sim_{CM^*}	sim_{RM^*}
1	0.5	1.0	14.9	2.1	0.764	0.849
2	0.2	0.7	13.6	3.3	**0.801**	0.862
3	0.5	0.4	14.9	3.5	0.613	**0.865**
4	0.4	0.3	8.1	1.9	0.714	0.858

TABLE 5.3: Best Pearson correlations – coder ratings

Case	Best tuning sim_{CM^*}		Best tuning sim_{RM^*}		Correlations	
	α	β	α	β	sim_{CM^*}	sim_{RM^*}
1	0.2	1.5	6.6	3.2	**0.779**	0.678
2	0.8	0.1	3.6	2.8	0.752	0.683
3	0.3	0.5	3.8	3.4	0.587	0.710
4	0.4	0.4	1.1	1.7	0.670	**0.715**

TABLE 5.4: Best Pearson correlations – physician ratings

Case	Best tuning sim_{CM^*}		Best tuning sim_{RM^*}		Correlations	
	α	β	α	β	sim_{CM^*}	sim_{RM^*}
1	0.3	1.3	14.9	2.6	0.799	0.789
2	0.5	0.4	6.9	3.2	**0.805**	0.798
3	0.4	0.4	7.9	3.7	0.623	**0.810**
4	0.4	0.4	2.8	2.0	0.719	0.808

TABLE 5.5: Best Pearson correlations – average of physician and coder ratings

Figure 5.1 presents the results associated to the Pearson correlations of similarity measures against the average of physician and coder ratings. In these figures, only instantiations which derive from Case 2 and Case 4 expressions of the core elements are provided – Case 2 (A1, B1) and Case 4 (A2, B2) – these instantiations provide interesting results for both abstract measures evaluated. Points of the surface which correspond to maximal correlations and published measures are specified. Additional figures are proposed in Appendix C, in particular those associated to the results which have been obtained with the other instantiations and benchmarks (coders and physicians alone).

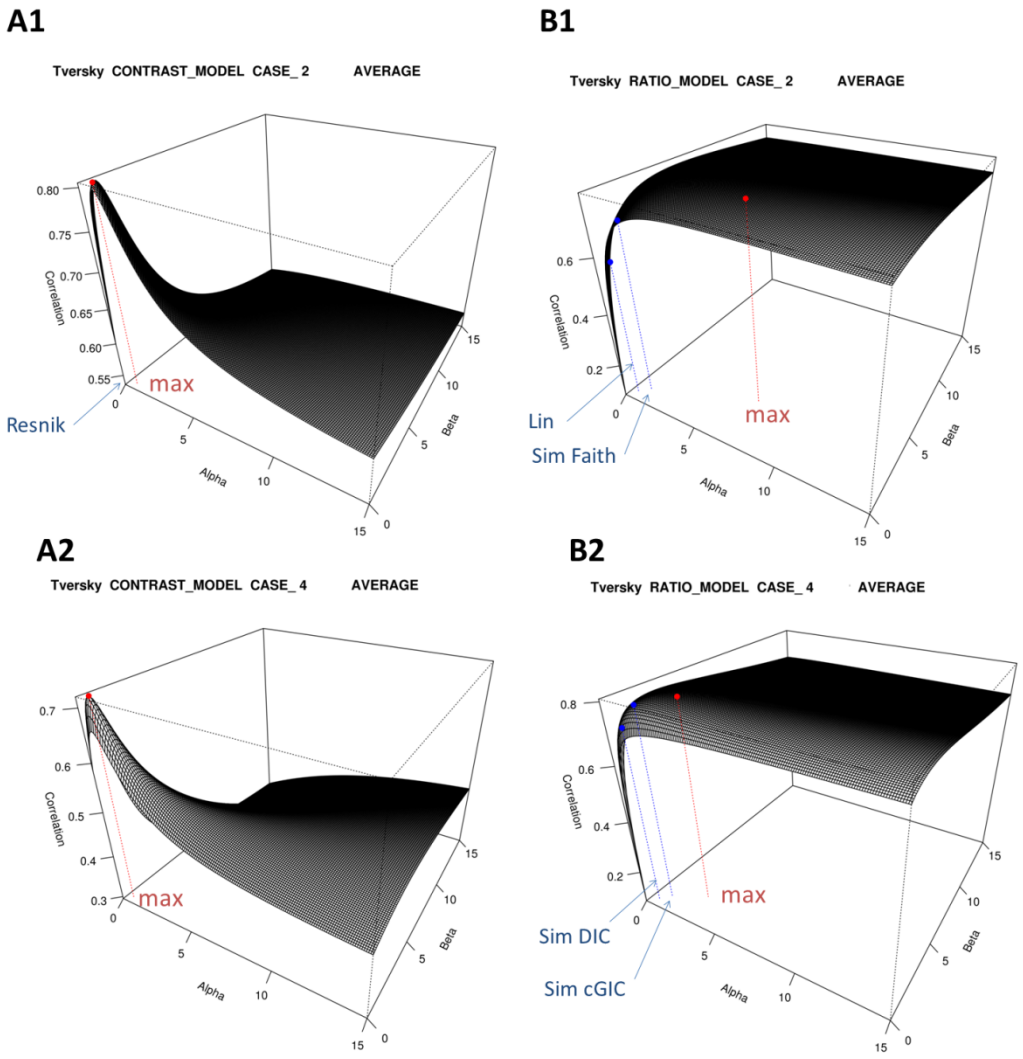

FIGURE 5.1: Surfaces associated to the Pearson correlations of similarity measures against the average physician and coder ratings. Measures have been instantiated from abstract forms of the *contrast model* (A) and the *ratio model* (B) using core elements expressions defined in Table 5.1: Case 2 (A1, B1) and expression Case 4 (A2, B2). Each point making up the surface corresponds to a specific tuning of α and β. For each surface, the dot labelled max corresponds to the maximal value observed. Other dots reflect instantiations that correspond to existing measures, cf. Table 5.2

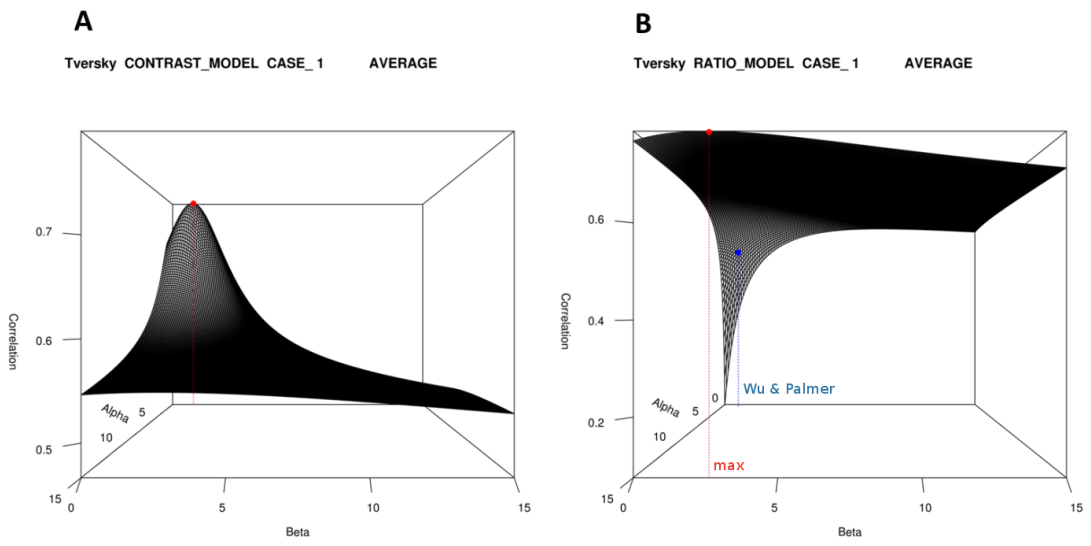

FIGURE 5.2: Plot of the Pearson correlations of the measures obtained from abstract forms of the contrast and ratio models using instantiation Case 1 – resp. (A) and (B) (averaged benchmark including both coder and physician scores). The inclination of the curves enables to discuss the benefits of asymmetric tuning of α and β parameters

5.1.3.2 Discussion

To our knowledge, this is the first large scale analysis of semantic measures to be proposed in the literature. Here we focus on the main conclusions which can be extracted from the analysis of the results:

- *The effect of core elements' expressions on the measures' accuracies depends on the abstract measure considered*: for the *contrast model*, core elements' expressions corresponding to instantiation Case 3 always resulted in the lowest correlations (0.613, 0.587, 0.623 – Tables 5.3, 5.4, 5.5). Conversely, considering the *ratio model*, instantiation Case 3 produced some of the best correlations (0.865, 0.710, 0.810). It is therefore interesting to underline that the suitability of using specific expressions of core elements is a function of the abstract formulation which has been selected – an aspect which must be taken into account for the design of semantic measures using the framework.

- *The accuracy of measures is mainly explained by the selected abstract measure*: indeed, changes in the expression of the core elements only slightly modified the shape of the surface. Moreover, most instantiations associated with well-tuned α and β parameters produced good correlations. The maximum variation between the best correlation observed for the *ratio model* was +/- 0.04 (0.678 - 0.715, see

Table 5.4). However, using the *contrast model*, greater variations were observed: +/- 0.19 (0.587 - 0.779, see Table 5.4). In other words, considering adapted parameters, each instantiation of the core elements can lead to accurate measures. Nevertheless, we also observe that results differ depending on the abstract measure considered. Some abstract measures appear to generate search space solutions with an interesting global maximum regardless of the instantiation of the core elements considered.

- *The variability of scores is mainly due to the selection and tuning of the abstract measure*: by considering the *contrast model*, an important variability of results is observed depending on the values α and β (see Figure 5.1 A1 A2) – very narrow global maximums are observed. However, despite the variability of the results for low α and β, it is also observed for the *ratio model* that the variability significantly decreases with large values of α and β (see Figure 5.1 B1 B2). This is indeed expected since the limit of the correlation function approaches a constant value when $(\alpha, \beta) \to +\infty$, i.e. $\lim_{(x,y)\to+\infty} sim_{RM^*}(u, v) = 0$. Thus $\lim_{(x,y)\to+\infty} corr(\mathbf{s}_{sim_{RM^*}}, \mathbf{s}) = corr([0], \mathbf{s})$, with $\mathbf{s}_{sim_{RM^*}}$ the vector which contains the similarities obtained by an instantiation of sim_{RM^*} for the pairs of concepts which compose the benchmark, $\mathbf{s}$ the vector which contains the expected scores of similarity for a scenario (e.g. coders), and [0] a vector of the same size which contains only 0 values. It is therefore interesting to remark that for certain abstract measures, the accuracy does not depend (or only faintly) on the expression of the abstract operators evaluated, but rather on the selected abstract measure and associated (α, β) configuration[1].

- *Asymmetrical measures tend to provide the best results*: all experiments provided the best correlations by tuning the measures with asymmetric contributions of α and β parameters (see Table 5.3 to Table 5.5 and Figure 5.2). As an example, in Figure 5.2, the asymmetry of the surfaces underlines the benefits of considering asymmetric α and β values. The improvement of an asymmetric tuning of parameters is best outlined in the *ratio model* (Figure 5.2 B). This observation refers to the results obtained in cognitive sciences which underline the necessity of considering an asymmetric estimation to best fit human appreciation of similarity – cf. Section 1.4.2.

[1] This statement obviously only considers *coherent* expressions of abstract elements.

It is tempting to generalise these results and observations. Nevertheless, note that the observations made in this experiment are only driven by the analysis of specific configurations of measures, using a single ontology and a unique benchmark. Therefore, more empirical analyses have to be performed in order to deeply understand and generalise these preliminary results.

However, these results clearly stress the usefulness of such experiments and the added value of the proposed framework to analyse semantic measures using a level of details that have never been obtained. There are numerous applications. This is in particular due to the fact that the adopted experimental protocol, which extensively relies on the unifying framework, both eases and improves the understanding, selection and design of semantic measures. Otherwise stated, this study also highlights that the proposed approach can be of great help to optimise and to select semantic similarity measures for domain-specific usage.

5.2 Estimation of the robustness of semantic measures

5.2.1 Motivation and objectives

We have so far introduced a theoretical framework which provides the ability to design semantic measures by aggregating different core elements commonly used for their definition. In the previous section, we have also indirectly shown that the process of the selection and design of semantic measures can be partially[1] formulated as an optimisation problem: how to select abstract measures and instantiations of the core elements which lead to the most accurate measures?

This process is based on a benchmark from which the accuracy of measures can be evaluated w.r.t expected scores of similarity. In this context, expected scores of similarity provided by *experts* are considered to be the *unquestionable truth*. Therefore, this evaluation/design process does not take into account the uncertainty associated to benchmarks. However, each benchmark is *per se* associated to bias, e.g., due to abnormal sampling in experts and pairs of elements evaluated. This will therefore undeniably impact the selection/design of accurate semantic measures. As we can imagine, considering this bias – which is nothing but uncertainty w.r.t the scores of similarity that make up the benchmark – makes the problem become more complex. Indeed, taking

[1]Note that here we focus on measure accuracy despite the fact that we have underlined other important aspects of semantic measures which could be considered when selecting and designing measures for a specific context of use. Refers to Appendix B.

the uncertainty into account leads to the desire to evaluate semantic measures not only based on their accuracy, but also w.r.t their capacity to be resilient to bias/uncertainty which intrinsically mars benchmarks. This capacity is introduced through the term *robustness*. Therefore, this section proposes to study the consideration of uncertainty for both the selection and design of semantic measures. This is done by formally defining the notion of semantic measure robustness, and by proposing an approach to incorporate uncertainty in the process commonly used to evaluate semantic measure accuracy.

This section is structured as follows. First, considering the unifying framework proposed in Chapter 4, we formalise the process of semantic similarity measure design/selection through an optimisation process. Next, we propose an approach to incorporate uncertainty modelling in the aforementioned optimisation process. This will help us to rigorously define the notion of robustness for semantic measures. The benefits of our proposal is illustrated through a practical use case in which specific semantic measures are evaluated w.r.t their robustness. Finally, we discuss the results which have been obtained and the perspectives opened by the notion of robustness to better analyse semantic measures.

5.2.2 Formalisation of the problem and definition of robustness

5.2.2.1 Design semantic measures through optimisation

Considering a particular abstract expression of a measure, here sim_{RM^*}, the objective is to define the *right* combination of parameters $(\rho, \Theta, \Psi, \Phi, \alpha, \beta)$. Following the framework presented in Chapter 4, this choice proceeds with two steps:

- *Step 1*: Define a finite list $\Pi = \{\pi_l | \pi_l = (\rho_l, \Theta_l, \Psi_l, \Phi_l), l = 1, \ldots, L\}$ of possible instantiations of the core elements $(\rho, \Theta, \Psi, \Phi)$[1], see Table 5.1. This choice can be guided by semantic concerns and application constraints, e.g., based on: (i) the analysis of the assumptions on which specific instantiations of measures rely, (ii) on the desire to respect particular mathematical properties or (iii) the computational complexity of measures. These aspects were discussed in Section 5.1.

- *Step 2*: Choose the couple of parameters $(\alpha_l, \beta_l), l = 1, \ldots, L$ to be associated to π_l in sim_{RM^*}. There are at least two ways to identify the value of (α_l, β_l) that better matches human appreciation of similarity. A couple (α_l, β_l) may be selected in an *ad hoc* manner from a finite list of well-known instantiations (see Table 5.2), e.g., based on the *heavy* assumption that measures performing correctly in other benchmarks are suited for our specific use cases. Alternatively, knowing the

[1] Note that here we consider θ to be defined; it can also be considered as a variable.

expected similarities $s(x, y)$ furnished by domain experts on a learning dataset, (α_l, β_l) can also be obtained from a continuous optimisation process over this dataset. The latter issue is developed hereafter.

For any instantiation $(\pi_l, \alpha_l, \beta_l)$ of the abstract measure sim_{RM^*}, let us denote $s_l(x, y)$ for any couple of concepts:

$$s_l(x, y) \equiv sim_{RM^*(\pi_l, \alpha_l, \beta_l)}(x, y) \tag{5.1}$$

Suppose now that experts have given the expected similarities $s_k = s(x_k, y_k), k = 1, \ldots, N$ for a subset of N couples of concepts (x_k, y_k). Let $\mathbf{s} = [s_1, \ldots, s_N]^T$ be the vector of these expected similarity values. It is possible to estimate the quality of a particular semantic similarity measure tuning s_l with the value of a fitting function. We denote $\mathbf{s}_l$ the vector which contains the similarities obtained by s_l for each pair of concepts evaluated to build $\mathbf{s}$, with: $\mathbf{s}_l = [s_l(x_k, y_k)_{k=1,\ldots,N}]$. Given π_l, the similarities $s_l(x_k, y_k)$ only depend on (α_l, β_l); it is thus possible to find the optimal (α_l^0, β_l^0) values that optimise a fitting function, e.g. the correlation between $\mathbf{s}$ and $\mathbf{s}_l$:

$$\begin{cases} \max_{\alpha,\beta} corr(\mathbf{s}, \mathbf{s}_l) \\ 0 \leq \alpha, \beta \leq M \end{cases} \tag{5.2}$$

The bound constraint of this optimisation problem is reasonable since the case $\alpha \to +\infty$ or $\beta \to +\infty$ should imply $sim_{RM^*}(x, y) = 0$ which has no appeal for us.

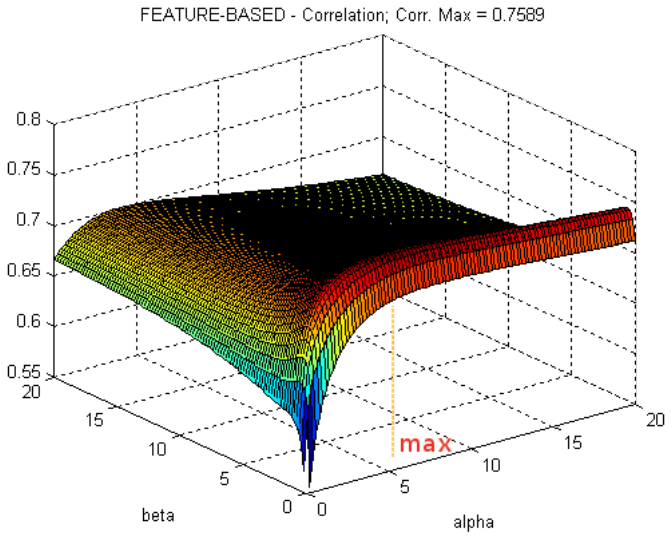

FIGURE 5.3: Fitting function $corr(\mathbf{s}, \mathbf{s}_l(\alpha, \beta))$ for the instantiation of sim_{RM^*} Case 3 (Table 5.1) and Pedersen et al. semantic similarity benchmark

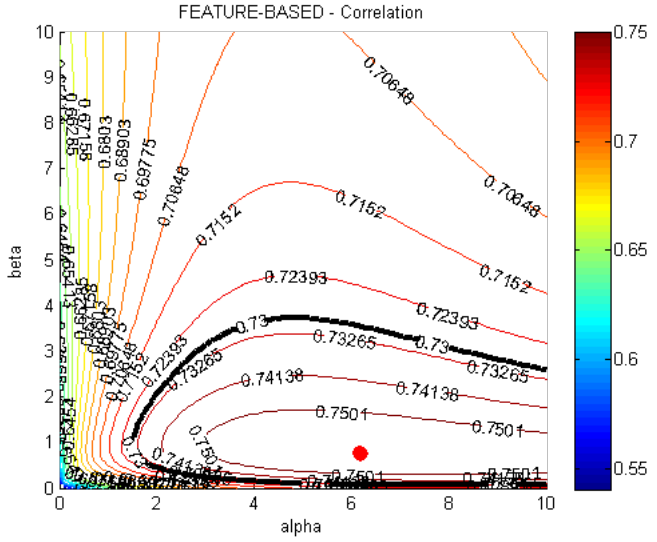

FIGURE 5.4: Level line of the fitting function $corr(\mathbf{s}, \mathbf{s}_l(\alpha, \beta))$ for the instantiation of sim_{RM^*} Case 3 (Table 5.1) and expected semantic similarity scores. The (red) dot refers to the optimal (α, β) configuration, $(\alpha, \beta) \simeq (6.2, 0.8)$

Figures 5.3 and 5.4 present an experiment which has been made to distinguish the optimal (α_l^0, β_l^0) parameters for instantiations of sim_{RM^*} using Case 3 (refer to Table 5.2) and a biomedical benchmark dedicated to semantic similarity (the one proposed by Pedersen et al. [2007] which was presented in Section 5.1.2 and Appendix B.2). The maximal correlation value is 0.759, for $\alpha_l^0 = 6.17$ and $\beta_l^0 = 0.77$ (the dot in Figure 5.4). The strong asymmetry in the contour lines is a consequence of $\Phi(\tilde{u}, \tilde{v}) \neq \Phi(\tilde{v}, \tilde{u})$. As we saw in the previous section, this approach is efficient for deriving the optimal configuration considering a given vector of similarities (i.e., $\mathbf{s}$). Nevertheless, it does not take into account the fact that expert assessments of similarity are inherently marred by uncertainty. We therefore introduce an approach to consider this uncertainty in the process of measure selection.

5.2.2.2 Uncertainty modelling

A classical way to model expert uncertainty is the Gaussian independent noise: $t_k = s_k + \varepsilon_k$ with $\varepsilon_k \sim N(0, \sigma_k^2), k = 1, \ldots, N$. Thus, $\mathbf{t} = \mathbf{s} + \varepsilon$, with $\varepsilon \sim N(0, \Sigma), \Sigma = diag(\sigma_1^2, \ldots, \sigma_N^2), \forall k, \sigma_k^2 \leq \sigma^2$. In our application domain, expected similarities are often provided in a finite ordinal linear scale of type $v_i = v_0 + i\Delta, i = 1 \ldots V$ (e.g., $v_i \in \{1, 2, 3, 4\}$ in the next section). If Δ denotes the difference between two contiguous levels of the scale, next we assume in this case that $\varepsilon_k \in \{-\Delta, 0, \Delta\}$ with probability $p(\varepsilon_k = 0) = q, p(\varepsilon_k = -\Delta) = p(\varepsilon_k = \Delta) = \frac{1-q}{2}$.

This model of uncertainty merely means that expert assessment errors cannot exceed $\pm\Delta$. In addition, it allows for computing the probability distributions of the optimal couples $(\alpha_l(\varepsilon), \beta_l(\varepsilon)) \sim D^u_{\alpha_l,\beta_l}$ with $(\alpha_l(\varepsilon), \beta_l(\varepsilon))$ being the solution of the problem presented in Equation 5.2, with $\mathbf{t} = \mathbf{s} + \varepsilon$ instead of $\mathbf{s}$ as inputs, and u the uncertainty parameter ($u = \sigma$ or q) – note that $(\alpha_l^0, \beta_l^0) = (\alpha_l(0), \beta_l(0))$.

5.2.2.3 Semantic measure robustness

The aim is to quantify the impact of uncertainty in expert assessments on the selection of a measure instantiation, i.e. selection of $(\pi_l, \alpha_l, \beta_l)$. We are interested in the evaluation of semantic similarity measure robustness w.r.t expert uncertainty, and we more particularly focus on the relevance of considering (α_l^0, β_l^0) in case of uncertain expert assessments.

Finding a robust solution to an optimisation problem without knowing the probability density of data is a well-known problem, e.g., [Ben-Tal et al., 2009; Janaqi et al., 2013]. In our case, we do not use any hypothesis in the distribution of $\mathbf{s} - \mathbf{s}_l$. We therefore define a set of *near optimal* (α_l, β_l) using a threshold value r (domain-specific setting). The near optimal solutions are those in the level set:

$$L_r = \{[\alpha_l, \beta_l] \mid corr(s, s_l) \geq r\} \tag{5.3}$$

The robustness is therefore given by:

$$R(u) = \iint_{L_r} D^u_{\alpha_l,\beta_l} d\alpha_l d\beta_l \tag{5.4}$$

The bigger R, the more robust the model (α_l^0, β_l^0). Nevertheless, given that analytical form for the distribution $D^u_{\alpha,\beta}$ cannot be established, even in the normal case $\varepsilon \sim N(0, \Sigma)$, estimation techniques are used for its estimation, e.g., the Monte Carlo method.

The computation of $D^u_{\alpha_l,\beta_l}$ allows for the identification of a robust couple (α_l, β_l) for a given uncertainty level u. An approximation of this point, here denoted (α_l^*, β_l^*), is given by the median of points generated by the Monte Carlo method $(\alpha_l(\varepsilon), \beta_l(\varepsilon))$. Note that (α_l^*, β_l^*) coincides with (α_l^0, β_l^0) for $u = 0$ or little values of u. Therefore (α_l^*, β_l^*) remains inside L_r for most of u and is significantly different from (α_l^0, β_l^0) when u increases.

We have so far (i) formalised the problem of selection of a semantic measure as an optimisation problem, (ii) incorporated uncertainty modelling to it, and (iii) defined the robustness of a semantic measure w.r.t the uncertainty associated to the benchmark on which the optimisation problem relies. The next section is dedicated to a use case

example in which the robustness of semantic measures is discussed in a specific context of use.

5.2.3 Selection of a robust semantic similarity measure: use case

5.2.3.1 Experimental design

As we have seen, most algorithms and treatments based on semantic similarity measures require measures to be highly correlated with human judgement of similarity. Semantic similarity measures are thus commonly evaluated regarding their ability to mimic human appreciation of similarity between domain-specific concepts. In this experiment, similarly to the experiment presented in Section 5.1, we considered the benchmark introduced by Pedersen and collaborators. It can be used to evaluate semantic similarity measures w.r.t similarity scores of pairs of concepts relative to the biomedical domain – similarity scores were provided by medical experts. Nevertheless, conversely to the previous study, the benchmark was used considering pairs of concepts defined in the Medical Subject Headings (MeSH) thesaurus[1] [Rogers, 1963].

In the benchmark considered, the average of expert similarities is given for each pair of concepts; initial ratings are of the form $s_k \in \{1, 2, 3, 4\}$. As a consequence, we considered that the uncertainty is best modelled defining $\varepsilon_k \in \{-1, 0, 1\}$ with probability distribution: $p(\varepsilon_k = 0) = q, p(\varepsilon_k = -1) = p(\varepsilon_k = 1) = \frac{1-q}{2}$.

The approach used to generate measures was defined in the previous section. Remember that the measures were obtained from sim_{RM^*} considering instantiations of the core elements of the framework which was introduced in Table 5.1. Optimal α and β were found by resolving Problem 5.2; the computation of semantic similarity measures were performed by the Semantic Measures Library and the source code related to the resolution of the optimisation problem (i.e. solver) was developed in Matlab[2].

5.2.3.2 Results and discussion

Table 5.6 shows that, considering the average of physicians and coders similarities, one of the best results is obtained using the Case 2 expression. The optimal configuration is obtained with:

$$sim_{RM^*_{C2}} = \frac{\psi}{18.62(IC(u) - \psi) + 4.23(IC(v) - \psi) + \psi} \tag{5.5}$$

[1] SNOMED-CT was used in Section 5.1. Please refer to Appendix B for details on the benchmark.
[2] `www.mathworks.com/products/matlab/`. Source code available on demand (reviewing process).

with $\psi = IC(MICA(u,v))$

	Case 1	**Case 2**	**Case 3**	**Case 4**
Max correlation	0.719	**0.768**	0.759	0.736
Optimal (α_l^0, β_l^0)	(9.89,1.36)	(18.62,4.23)	(6.17,0.77)	(7.26,0.40)

TABLE 5.6: Best Pearson correlations of parametric semantic similarity measures based on sim_{RM^*}, refer to Table 5.1 for details on measures

It can be observed that, naturally, the choice of core elements affects the maximal correlation; instantiations of Cases 2 and 3 always resulted in the highest correlations. Another interesting aspect of the results is that asymmetrical measures provide the best results. All experiments provided the best correlations by tuning the measures with asymmetric contributions of α and β parameters. Note that the best tunings and (α, β) ratio vary depending on the core elements considered. These observations are in accordance with the conclusions of the study presented in the Section 5.1.3.

Setting a threshold of correlation at $r = 0.75$, we now focus on the instantiations which correspond to Cases 2 and 3; they have comparable results (respectively 0.768/0.759, $\Delta = 0.009$). The aim is to evaluate their robustness according to the framework introduced. Considering inter-agreements between pools of experts reported in [Pedersen et al., 2007] (0.68 and 0.78 for physicians and coders respectively), we set the level of near optimality (L_r) to $r = 0.73$. We also choose uncertainty values $q \in \{0.9, 0.8, 0.7, 0.6, 0.5\}$. The probability for the expert(s) to give erroneous values, i.e. their uncertainty, is $1 - q \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For each q-value, a large number of ε-vectors are generated to derive (α_l^*, β_l^*). Estimated values of the robustness $R(u)$ and (α_l^*, β_l^*) for instantiation of measures which derive from Cases 2 and 3 are given in Table 5.7. They are also illustrated in Figure 5.5. Results which have been obtained for the other cases are provided in Appendix C.1.

	$1-q=0.1$	$1-q=0.2$	$1-q=0.3$	$1-q=0.4$	$1-q=0.5$
$R_{C1}(u)$	0.27	0.30	0.36	0.37	0.29
$\alpha_{C1}^*, \beta_{C1}^*$	(10.24,1.22)	(11.07,1.21)	(8.31,1.20)	(8.55,1.21)	(6.24,1.31)
$R_{C2}(u)$	**0.83**	**0.70**	**0.56**	**0.49**	**0.39**
$\alpha_{C2}^*, \beta_{C2}^*$	(18.62,4.23)	(18.62,4.23)	(15.31,4.23)	(16.70,4.07)	(13.71,4.02)
$R_{C3}(u)$	0.76	0.54	0.46	0.39	0.35
$\alpha_{C3}^*, \beta_{C3}^*$	(6.17,0.77)	(6.17,0.76)	(5.52,0.71)	(5.12,0.64)	(4.06,0.70)
$R_{C4}(u)$	0.74	0.57	0.46	0.43	0.35
$\alpha_{C4}^*, \beta_{C4}^*$	(7.26,0.40)	(6.83,0.40)	(4.71,0.39)	(4.98,0.39)	(3.75,0.41)

TABLE 5.7: Robustness of tested parametric semantic similarity measures

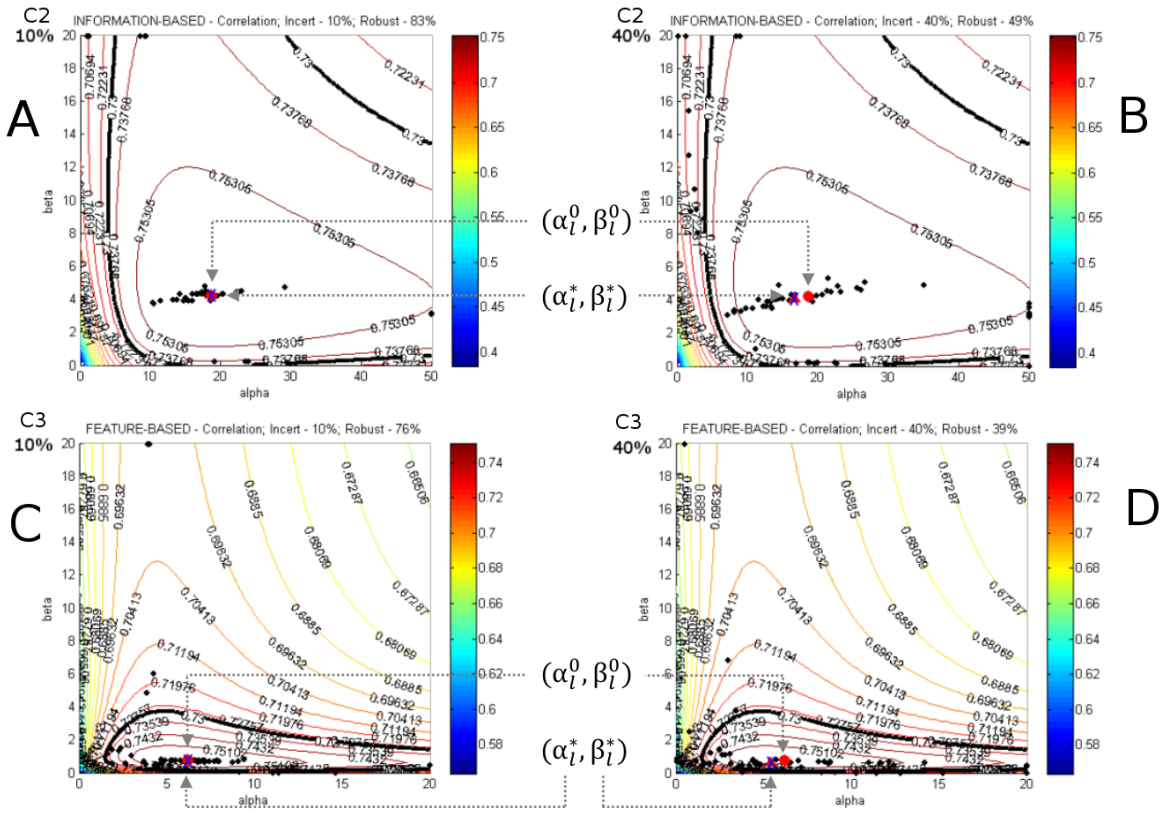

FIGURE 5.5: Plot of robustness of parametric semantic similarity measures based on Case 2 (C2) and Case 3 (C3) for 10% and 40% of uncertainty. In each figure the solutions (α_l^0, β_l^0),(α_l^*, β_l^*) and L_r are plotted. L_r is represented by the area inside the bold black line

Figure 5.5 shows the spread of the couples $(\alpha(\varepsilon), \beta(\varepsilon))$ for Case measures 2 and 3 considering the levels of uncertainty set to 10% and 40%. An interesting aspect of the results is that the robustness is significantly different depending on the case considered: 83% for Case 2 and 76% for Case 3. Therefore, despite the fact that their correlations were comparable ($\Delta = 0.009$), Case 2 is less sensitive to uncertainty w.r.t the learning dataset used to distinguish best-suited parameters. Indeed, only based on correlation analysis, users of semantic similarity measures will generally prefer measures which have been derived from Case 3 since their computational complexity is lower than those derived from Case 2 (computation of the IC and the MICA are more complex). Nevertheless, Case 3 appears to be a more risky choice considering the robustness of the measures and the uncertainty inherently associated to expert evaluations. In this case, one can reasonably conclude that (α_l^0, β_l^0) of optimised Case 2 is robust for an uncertainty lower than 10% $(1 - q = 0.1; R(u) = 0.83)$.

The size of the level set L_r is also a relevant feature for the selection of semantic similarity measures; it indicates the size of the set of parameters (α_l, β_l) that gives high correlations

considering imprecise human expectations. Therefore, both, an analytical and graphical estimator of robustness are introduced.

Another interesting finding of this study is that, even if human observations are marred by uncertainty and the semantic choice of measure parameters, $\pi_l = (\rho_l, \Theta_l, \Psi_l, \Phi_l)$, is not a precise process, the resulting semantic similarity measure is not so sensitive to all these uncertainty factors.

5.2.4 Synthesis of the study and perspectives

Considering the large diversity of measures available, an important contribution for end-users of semantic similarity measures would be to provide tools to select best-suited measures for domain-specific usage. Our approach paves the way to the development of such a tool and can more generally be used to perform detailed evaluations of semantic similarity measures in other contexts and applications (i.e., other ontologies and training data). In this section, we used the unifying framework established in Chapter 4 and a well-established benchmark in order to design semantic similarity measures that fits the objectives of practitioners and designers of semantic measures in a given application context.

We particularly focused on the fact that the selection of the best-suited semantic similarity measures is affected by uncertainties, in particular due to the uncertainty associated to the ratings of human experts used to evaluate measures, etc. To our knowledge, we are the first to propose an approach that finds/creates a best-suited semantic measure which remains robust in the face of these uncertainties. Indeed, contrary to most existing studies which only compare measures based on their accuracy, i.e., correlation with expected scores of similarity (e.g., human appreciation of similarity), our study highlights the fact that robustness of measures is an interesting criteria to better understand measures' behaviour and therefore drive their comparison and selection. We therefore proposed two estimators of robustness, graphical and analytical, which can be used to characterise this important property of semantic measures. Thus, by bringing to light the limits of existing estimator of measures' accuracy, especially when uncertainty is regularly impacting measures (evaluation and definition), we are convinced that our proposals open interesting perspectives for measure characterisation and will therefore ease their accurate selection for domain specific studies.

In addition, results of the real-world example used to illustrate our approach (cf. Section 5.2.3) give us the opportunity to capture new insights about specific types of measures (i.e. particular instantiation of an abstract measure). More benchmarks have to be

studied to derive more general conclusions. This will help to better understand semantic similarity measures and more particularly to better analyse the role and connexions between abstract measures' expressions, core elements instantiations and additional parameters regarding both the accuracy and robustness of semantic similarity measures.

5.3 Chapter conclusion

In this chapter, we have illustrated some of the practical applications offered by the unifying framework of semantic similarity measures presented in Chapter 4. We have, in particular, highlighted the interesting perspectives opened by the framework to study specific and detailed aspects of measures, i.e., role, importance and repercussion associated to the selection of specific components commonly used to build semantic measures (i.e., abstract measure, associated parameters, instantiation of the core elements). At this occasion, through practical use cases related to the biomedical domain, we brought to light some domain-specific and interesting aspects of measures (e.g., asymmetry, limited impact of the choice of specific instantiations of the core elements using particular abstract measures). Despite the importance and implications of such results, our aim was not to derive general conclusions, but rather to demonstrate the suitability and practical feasibility of the proposal. To this end, we defined and formalised an approach which can be used to take advantage of the theoretical framework to evaluate and optimise semantic measures. This approach can now be used to study specific aspects of semantic measures through a degree of granularity previously unseen – there are numerous perspectives to derive new insights on semantic measures.

We also provided a reflection associated to the consideration of uncertainty in protocols which are commonly used to evaluate the accuracy of semantic measures. For this, we defined an approach to incorporate uncertainty modelling in the evaluation process. We also introduced the notion of *robustness*. It can be used to support semantic measure selection w.r.t the degree of uncertainty which can be associated to the benchmark in use. We are convinced that this proposal finds direct applications for the practical use of semantic measures. Indeed, as we have seen, the accuracy of measures is central for their selection, yet it is hard to have *blind confidence* in benchmarks which are *per se* marred by uncertainty.

This led us to another central discussion on semantic measures which have been proposed in this chapter: aspects of semantic measures which have to be discussed for their selection. We have stressed that most studies today (legitimately) focus on the evaluation of semantic measures through their accuracy. Nevertheless, we stressed the limits of this sole criterion and we proposed other aspects of semantic measures which deserve

consideration when discussing the relevance of using a particular measure in a specific application context, i.e., mathematical properties, algorithmic complexity, semantics. The enrichment of this discussion, which could be undeniably facilitated by the proposed theoretical framework, is left in perspective of this thesis but we believe is essential for assisting end-users in the selection of semantic measures w.r.t the plethora of measures proposed in the literature.

6

Semantic measures to compare instances of a semantic graph

Contents

Abstract

Many applications take advantage of both ontologies and the Linked Data paradigm to describe various kinds of resources (e.g., gene products, music bands, documents). To fully exploit this knowledge, not only for an exact search but also for inexact and imprecise search, semantic measures are used to estimate the relatedness of resources regarding their semantic characterisation. Such measures have proved particularly useful for information retrieval based on semantic graphs (e.g. RDF graphs). However, existing proposals mainly focus on specific aspects of the resources (e.g. types) or only partially exploit the semantics expressed in the ontology. To address this limitation, this chapter studies how the unifying framework of semantic similarity measures which has been proposed in Chapter 4, can be extended to define more expressive measures to compare instances defined in a semantic graph. As a result, we introduce a new canonical form of an instance through *projections*. The main proposal relies on the possibility of taking into consideration the complex properties of an instance: properties which are not materialised in the ontology but which can be obtained by aggregating other properties. We then show how this canonical form can be used to easily design semantic measures. The added value of this approach, especially pertaining to recommendation systems, is discussed. In particular, we show how, using semi-supervised techniques, this approach can be used to track the semantics which govern the comparison of instances, and therefore better explain the meaning of high/low scores of similarity between instances. Finally, the practical feasibility of our proposal is illustrated through a prototype of a music band recommender system available at `http://www.lgi2p.ema.fr/kid/tools/bandrec`.

Associated reference on which this chapter is based:

- **Mesures Sémantiques basées sur la notion de projection RDF pour les systèmes de recommandation**. Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. 24es journées francophones d'Ingénierie des Connaissances – IC 2013.

- **Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems**. Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. In On the Move to Meaningful Internet Systems: OTM 2013 Conferences (p. 606–615). Graz (Austria). Springer Berlin Heidelberg. doi:10.1007/978-3-642-41030-7_44

6.1 Motivation and objectives

"*Which music bands are similar to the Rolling Stones?*". It would be quite natural to ask such a question to a friend with some knowledge of music. Most classic search engines, however, will fail to provide an answer. In fact, the answers to such questions constitute resources defined as `MusicBands` (here lies the notion of type) and must be related to the `rollingStones` (notion of semantic relatedness). Answering such questions proves essential for recommendation systems since they extensively rely on similarity evaluations to formulate recommendations: "*If you like the* `rollingStones`*, you might also like...*". So, how is it possible to define whether or not two music bands are related by studying their properties (e.g. `musicGenres`, `dateOfFormation`)? More generally, how can the degree of relatedness of two instances be assessed? Data Retrieval techniques based on an exact (Boolean) search cannot be used herein; the inaccuracy expressed by the query entails the consideration of imprecise results and therefore requires the use of Information Retrieval (IR) techniques.

Many contributions have proposed the use of Semantic Web technologies and the Linked Data paradigm to assess the semantic relatedness of entities based on semantic measures. We have studied these measures in the context of IR based on semantic graphs; for the most part, these measures are suitable in comparing pairs of (groups of) classes[1], though only a few can be used to compare instances described through expressive graph-based representations (i.e. RDF graphs, graphs based on the property graph model). As an example, in IR, an instance is usually represented by a reductive canonical form, e.g. bag of concepts. Only a few approaches actually take advantage of solutions proposed in the context of instance matching, which seeks to determine whether two instance descriptions refer to a single domain instance. Two main approaches have thus been proposed to estimate the degree of relatedness of instances defined in a semantic graph: a *direct* one that controls the semantic model associated with the ontology, and an *indirect* one that does not consider or only slightly considers this semantics, e.g. the use of algorithms based on random walk approaches. By definition however, semantic measures must exploit semantics and enable justifying why a strong/weak semantic relatedness between instances is being assessed. This point is indeed critical for recommendation system design, whereby a user must understand why a recommendation is proposed in order to assign it credit (to avoid the *black box effect*).

Ehrig et al. [2004] were the first to propose a framework for defining semantic measures based on ontologies; this framework specifies how to compare instances through

[1]Since we will often refer to the instances of a class, we will prefer the term class over the term concept in this chapter.

their *direct properties* (e.g. types, labels). It was later extended to introduce the notion of a customised similarity function which was used to develop imprecise SPARQL (iSPARQL), in the aim of integrating imprecise evaluations of direct properties into SPARQL [Kiefer et al., 2007]. Based on the concept of a similarity aggregation operator originally defined by Orozco and Belanche [2004], Kiefer et al. [2007] formally defined the notion of similarity strategy, from which a complex element defined in an RDF graph may be compared on the basis of multiple similarity measures and an aggregation scheme. In some cases however, instances can only be compared by incorporating their *indirect properties*, e.g., information relative to properties characterising the instances to which they are related. As an example, a comparison of artists can only be drawn by considering the properties of their artistic productions, e.g. types, styles. To bridge this gap, Albertoni and De Martino [2006] proposed a primary extension to the framework defined by Ehrig et al. [2004] that included an evaluation of indirect properties as a means of better estimating the relatedness of instances.

The present contribution extends existing frameworks by defining a canonical form based on the notion of *projection*. This approach enables a fine-tuned definition of the representation of instances according to specific use contexts. Moreover, our approach makes it possible to express *complex* (indirect) properties which are not taken into account in existing frameworks. This representation of an instance is ultimately used to define a series of parameterised semantic measures which are well adapted to recommendation system definitions.

Note that throughout this chapter we will consider the example of a semantic graph which is presented in Figure 6.1 (already introduced in Chapter 2). In this representation, classes represent the concepts defined in an ontology related to music: `MusicBand`, `MusicGenre`, etc., while the instances are music bands: `rollingStones`, music genres: `rock`, etc. Moreover, a given instance can also establish specific relationships with other instances or data values (e.g. a *literal* corresponding to the name of the band). A semantic graph can therefore be dissected according to: (i) the intensional layer C (classes, taxonomy), (ii) the extensional layer I (instances), and (iii) the data layer D.

The remainder of this chapter will be structured as follows. Section 6.2 provides an overview of semantic measures for IR and recommendation systems; it summarises the state-of-the-art introduced in Chapter 3 which is of interest herein and briefly discusses the use of semantic measures for recommendation system design. Section 6.3.1 proposes a brief discussion on the generalisation of the theoretical framework presented in Chapter 4 to consider the comparison of instances. Section 6.3 is dedicated to (i) a formal definition of the notion of *projection*, (ii) the introduction of a general semantic measure which can be used to compare instances based on this notion, and (iii) possible extensions of the

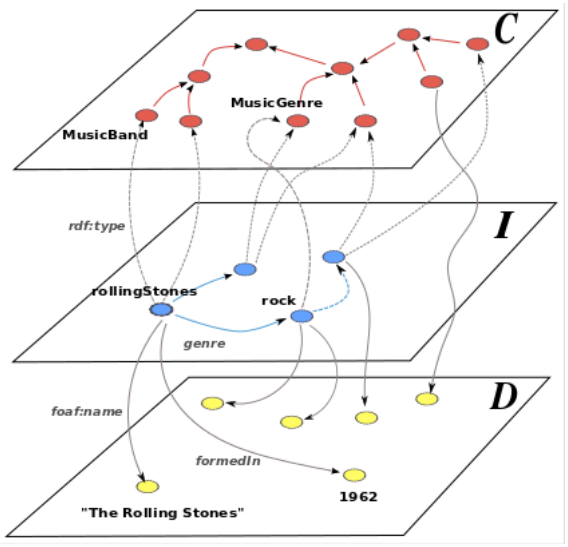

FIGURE 6.1: Partial graphical representation of a semantic graph according to three layers: concepts (C), instances (I), and data (D)

proposal. Section 6.4 discusses the application of the proposed approach for designing a music band recommendation system; a software prototype is also presented. Finally, conclusions are proposed in Section 6.5.

6.2 Overview of related literature

6.2.1 Semantic measures between instances

Semantic measures between instances have been widely studied to perform instance matching in various data/knowledge bases, e.g. RDF and databases [Euzenat and Shvaiko, 2013]. They have also been used to discover relationships between instances [Volz et al., 2009]. In this case, the aim of the measure is to detect duplicated instances in one or more knowledge bases. This is conceptually different to the desire to assess the similarity/proximity of instance representations which do not refer to the same instance.

Evaluating the proximity between instances requires defining a representation (or canonical form) to characterise an instance. Four approaches can be distinguished:

- **Representing an instance as a graph vertex**: When no specific canonical form is adopted, the instance is represented through the vertex of the graph making reference to it. The proximity between two instances is therefore evaluated using measures exploiting graph structure analysis and does not explicitly rely on the semantic carried by the graph, e.g. random walk techniques. Consequently, the

more the compared instances are interconnected, whether directly or indirectly, the more related they will be assumed to be, e.g., [Jeh and Widom, 2002]. The main advantage of this approach is its lack of supervision, while its main drawback is its absence of extensive control over the semantics which are taken into account to estimate the proximity. Indeed, this generates difficulties in justifying and explaining the resulting scores.

- **Representing an instance using a set of classes**: In this case, an instance is associated to the set of possibly weighed concepts[1]. The measures defined to compare sets of concepts can then be used for this canonical representation. Under most conditions, such an approach is adopted whenever the knowledge about instances can be summarised as a set of concepts (e.g. genes or documents are often annotated by concepts defined in a ontology). However, such a canonical form remains too restrictive for representing instances defined in a semantic graph (e.g., RDF graph) since only the types of instances or a very limited set of information will be considered. In Figure 6.1, the instance `roolingStones` would therefore be reduced to its set of affiliated concepts (e.g. $\{$`MusicBand`, ...$\}$). Formally, an instance i is represented by $\{c|\exists(i, \texttt{isA}, c)\}$[2].

- **Representing an instance through a list of properties**: An instance can be evaluated by studying its direct properties, i.e., resources linked to the instance by a single relationship characterised by a specific predicate (e.g. `rdfs:label`). According to Section 3.4.1.2, two types of properties can be distinguished: *non-taxonomic* (object and datatype properties in OWL); and *taxonomic*, i.e. those involving concepts structured into a taxonomy. Non-taxonomic properties corresponding to datatype properties can be compared using measures adapted to the type of properties considered, e.g., using a measure to compare dates of music band formations. Properties associated to instances (i.e. object properties) are, on most occasions, compared using set-based measures, which will evaluate the quantity of instances of shared sets (e.g. the number of music genres two groups have in common). Moreover, taxonomic properties are evaluated using semantic measures adapted to the comparison of concepts. Scores produced by the various measures are thus aggregated in order to obtain a global relatedness score for two instances [Euzenat and Shvaiko, 2013]. Such a representation is commonly adopted in ontology alignment, instance matching or link discovery between instances [Araujo et al., 2011; Oldakowski and Bizer, 2005; Volz et al., 2009]. The study of these measures inherits from early work related to both the comparison

[1] E.g. the classes of which the instance is a member.

[2] Considering that transitive reductions have been performed. In some cases more complex approaches will be used to associate a set of concepts to an instance, e.g., using a SPARQL query.

of objects defined into knowledge bases and the comparison of entities defined in a subset of the first order logic [Bisson, 1992, 1995]. Indeed, these measures have been extensively studied for comparing objects using aggregations of specific measures used to compare each of the properties of the compared objects [Valtchev, 1999a,b; Valtchev and Euzenat, 1997]. As an example, these contributions have formed the basis of several frameworks which are used for comparing instances or concepts in the field of ontology alignment or instance matching, e.g., OWL Lite Alignment (OLA) method has been proposed to compare ontologies based on aggregations of several measures [Euzenat et al., 2004; Euzenat and Valtchev, 2004].

- **Representing an instance through an extended list of properties**: This representation is an extension of the previously presented canonical form. It can be implemented to take into account indirect properties of instances, i.e. properties induced by the resources associated with the represented instances. As an example, two music bands will be compared w.r.t the music genres associated to their music productions.

Several contributions underline the relevance of indirect properties in comparing entities represented through graphs, especially in object models [Bisson, 1995]. On reflecting on our music-related example, such a representation might be used to consider the characteristics (properties) of the music genres for the purpose of comparing two music bands. A formal framework, which extends those proposed by Euzenat et al. [2004] and Ehrig et al. [2004], has thus been proposed to capture some of the indirect properties [Albertoni and De Martino, 2006]. This framework formally defines an indirect property of an instance along a path in the graph. The indirect properties to be taken into account are defined for a class and depend on a specific context, e.g. application context. From a different perspective, Andrejko and Bieliková [2013] suggested an unsupervised approach for comparing a pair of instances by considering their indirect properties. Each direct property shared between the compared instances plays a role in computing the global similarity. When the property corresponds to an object property (i.e. linking one instance to another), the approach combines a taxonomic measure with a recursive process to take into account the properties of instances associated with the instance being processed. Lastly, in estimating the similarity between two instances, the measure aggregates the scores obtained during the recursive process.

6.2.2 Semantic measure specificities for recommendation

The purpose of a recommendation system is to propose relevant resources to users in accordance with a context and their specific interests. Such a system takes a user model into consideration for guiding the recommendation. The underlying model can be built explicitly, e.g., based on queries or satisfaction forms, or in an implicit manner, by analysing the more recent user interactions with the system or else based on a statistical analysis of users. Many websites (e.g. Amazon, Youtube) rely extensively on recommendation systems in order to facilitate both information retrieval and the exploration of the associated knowledge base [Heitmann and Hayes, 2010]. A recommendation system is a specific type of information retrieval that relies on three components: (i) the ontology (intensional and extensional knowledge), (ii) information characterising system users, and (iii) an algorithm for exploiting components (i) and (ii) in order to produce recommendations [Burke, 2002]. Based on a characterisation of the relationships established between the various instances of an ontology, the Linked Data paradigm and ontologies have both been proven to be particularly well suited for defining such recommendation systems, e.g., [Celma and Serra, 2008].

Despite the existence of numerous approaches for defining recommendation systems [Burke, 2002], this contribution focuses on the *content-based* approach that relies on resource properties: as an example, let us seek resources with characteristics related to those of interest to users, without any prior knowledge of user preferences (i.e., *cold start*).

In most cases, recommendation systems are fine-tuned by experts possessing an in-depth understanding of the underlying knowledge model and who are capable of distinguishing the properties that need to be taken into account in order to parameterise the recommendation algorithm. In this context, the representation of an instance based on the extended list of properties therefore seems to be the most appropriate strategy for defining semantic measures in most application contexts, i.e., due to its higher degree of expressiveness.

Though expressed using a high formalism that hampers its applicability, the theoretical framework proposed by Albertoni and De Martino [2006] enables the use of indirect properties of instances in order to define semantic relatedness measures. However, this framework does not take *complex* (indirect) properties into account, i.e. properties that rely on combining various other (indirect) properties - this is also a limitation of other related works [Bisson, 1995; Ehrig et al., 2004; Euzenat et al., 2004; Valtchev and Euzenat, 1997]. It is impossible, for example, to evaluate a `Person` whose `weight` and `size` have been specified through his *body mass index* (which can be computed from

the `weight` and `size` alone), if the property `bodyMassIndex` is not defined in the ontology. Moreover, this framework cannot be used to exploit the characterisation of various types of instances. A comparison of two instances requires specifying which properties are to be taken into account, and it is not possible to use the characterisation of instances related to the ones undergoing comparison. To address this latter limitation, Andrejko and Bieliková [2013] proposed the application of a recursive process on the instances linked to those being evaluated. This solution, however, cannot be used to define the direct and indirect properties which are to be taken into consideration. Moreover, comparing instances by taking all of their shared properties into account leads, in some cases, to treatment sequences requiring long computation times. In addition, performing a recursive treatment without defining the associated stop conditions makes semantic relatedness scores difficult to interpret.

The direct and indirect properties to be considered when comparing two instances depend, to a great extent, on the usage of the semantic measure, i.e., the semantics associated with the measures and the semantic interpretation to be drawn from the scores. These considerations, however, do not challenge the benefits of designing a generic approach for the definition of semantic measures. As previously observed, the expressiveness of existing frameworks merely enables partial characterisation of an instance defined in a semantic graph. The difficulty lies in expressing indirect properties and the impossibility of evaluating complex (in)direct properties limits the definition of semantic measures. To remedy this shortcoming, this chapter introduces a new approach for defining semantic measures.

6.3 Proposal to compare instances of a semantic graph

This section will define our approach to characterise both direct and indirect properties using a canonical form of instances based on the notion of *projection*. We will therefore introduce a generic semantic measure that enables the estimation of the semantic relatedness of two instances based on the notion of projection.

6.3.1 Towards a generalisation of the unifying framework

Let us first discuss the relationships which can be made with the theoretical framework defined in Chapter 4 for semantic similarity measures dedicated to concept comparison. In the previous section, we stressed that the central element of existing approaches designed to compare instances is the canonical form (representation) which has been

adopted to represent an instance, i.e., bag of concepts, direct list of properties. Interestingly, the theoretical framework we proposed relies extensively on the canonical form adopted to represent a concept (function ρ). As we have seen, defining such a function, as well as the approaches used to assess the similarity and differences of two representations, can be used to derive a large variety of measures from abstract ones.

In the context of instance comparison, we have also stressed the importance of controlling the semantics associated to a measure – generally to ensure the coherency of the treatment, but also in some cases to track the semantics of the scores which have been produced. To this end, the commonly adopted approach for comparing instances is to distinguish which properties should be taken into account in order to perform the comparison ([in]direct list of properties), to further aggregate the scores produced by the comparison of each property using a specific measure. In other words, each property is represented by a specific canonical form to further be compared. As an example, the comparison of two instances w.r.t their types leads to the comparison of groups of concepts. This specific case has already been treated in the section dedicated to the framework, i.e. remember that the definition of the function $\rho : \mathcal{P}(C) \rightarrow \mathbb{K}$. In the same vein, an approach could be defined in order to characterise any property which can be used to represent an instance through a canonical form which will enable its processing. Therefore, speaking informally, an instance i could be seen as a set of properties $\rho(i) = \{p_1(i), p_2(i), \ldots, p_n(i)\}$ or more particularly as a set $\rho(i) = \{\rho_1(p_1(i)), \rho_2(p_2(i)) \ldots, \rho_n(p_n(i))\}$. This stresses a potential and interesting break down of the problem through a recursive representation based on ρ functions relying on other representation functions which can be used to characterise specific aspects of the prior layer, and so on. To encompass such cases, the formalisation of the framework should be highly abstracted – the domain of the function ρ should be highly relaxed – so much so that the relevance of such an enterprise would be questionable. In order to ensure its practical application, and to ease its understanding, the proposal presented in this chapter does not rely on an extension of the formalism on which the definition of the unifying framework introduced is based. Nevertheless, as you will notice, this is conceptually the case.

6.3.2 Characterising an instance through projections

A direct or indirect property of instance i corresponds to a partial representation of i. In Figure 6.1 for example, the `rollingStones` instance can be represented by its name (*The Rolling Stones*) or music genres ({`rock`, ...}). A *simple property* of an instance is therefore expressed through resources linked to the particular instance. Representing an instance through its labels is therefore the same as considering all the l labels for

which a path links i to l through the relationship `rdf:label`; in other words, a triplet `i rdf:label l` exists, i.e. formally $\{l|\exists(i, \texttt{rdf:label}, l)\}$.

Generally speaking, the path linking two resources is characterised by an ordered list of relationships of specific predicate, i.e. a path pattern $< r_0, r_1, \ldots, r_n >$, with $r_i \in R$[1]. As with a property, a path is also associated to a range. It is defined by the type of resources specified by the range of r_n, its last relationship. It thus becomes possible to characterise some of the properties of instances of class `X` through a path $p : \mathcal{I}(\texttt{X}) \rightarrow \mathbb{K}'$, with $\mathcal{I}(\texttt{X})$ the instances of class `X` and $\mathbb{K}'$ the range of path p, a set of values that may be included in C, I or composed of values of the type `rdfs:Datatype`, e.g. String[2]. We distinguish three types of paths depending on the range of their last property r_n:

- **Data**: the range is a set of data, e.g. Strings, Dates (Figure 6.2, Case 2).
- **Instances**: the range is a set of instances (Figure 6.2, Case 1).
- **Classes**: the range is a set of classes (Figure 6.2, Case 3).

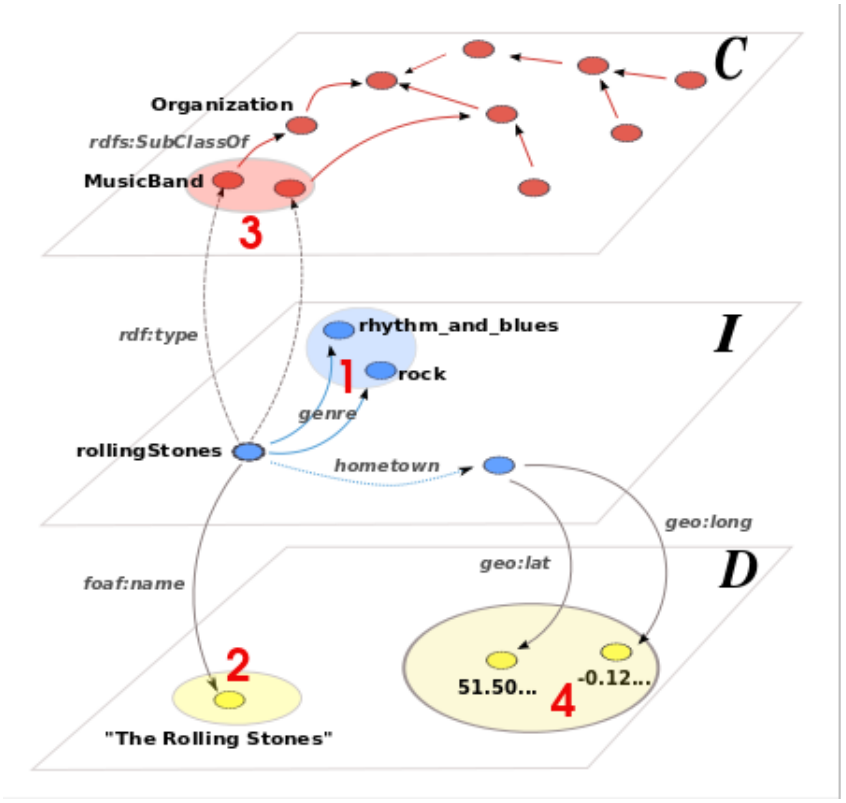

FIGURE 6.2: Graphical representation of projections in a semantic graph

[1]Note that we could also use the property path notation introduced in SPARQL 1.1. For those who are familiar to this latter notation, $< r_0, r_1, \ldots, r_n >$ is equivalent to the notation $/r_0/r_1/.../r_n$.

[2]In this chapter the domain $\mathbb{K}$ and $\mathbb{K}'$ have direct relationship with the domain $\mathbb{K}$ introduced to present the framework in Chapter 4.

A path may be used for characterising simple (either direct or indirect) properties of an instance. Complex properties, however, require several paths in order to be expressed. A comparison of two music bands through the Euclidian distance between their places of origin does indeed involve defining a complex property encompassing the latitude and longitude of a place that requires two paths $\{<$ `hometown`, `geo:lat` $>, <$ `hometown`, `geo:long`$\}$ (Figure 6.2, Case 4). In other words, the information characterising a music band via a property defining its place of origin corresponds to the projection of the instance onto two specific resources capable of being reached through paths in the semantic graph. In order to characterise all properties of an instance, the notion of path can thus be generalised by introducing the notion of projection.

A projection refers to projecting a mathematical structure from one space to another[1] – it is a vision of an entity which is related to the notion of *point de vue* discussed in Ducournau et al. [1998]. In formal terms, a projection P is composed of a set of paths and defined by $P : I \rightarrow \mathbb{K}$, with $\mathbb{K}$ being the set defining the types of projection $k \in \mathbb{K}$, onto which an instance can be *projected*.

The projection type corresponds to the range associated with this projection, i.e. the type of values potentially used to characterise the instance. When simple projections are used, i.e. when the projection is composed of a single path, then the projection range is defined by the path range, i.e. $\mathbb{K} = \mathbb{K}'$. Yet when complex projections involving multiple paths are used, other types of projections can be defined, in yielding $\mathbb{K} = \mathbb{K}' \cup \mathbb{K}''$ with $\mathbb{K}''$ being a set indicating the complex objects available for use in representing the complex properties of an instance. Let us note that complex objects are used to represent properties which are not explicitly expressed in the ontology, e.g. geographic location based on latitude and longitude, body mass index based on weight and size.

Four types of projections can therefore be distinguished: the three capable of being associated with a single path (*Data*, *Instances*, *Classes*), and the *Complex* type used to represent an instance by means of a set of complex objects combining various properties (not necessarily simple, e.g. geolocation). Let us denote P^k the projection of range $k \in \mathbb{K}$ and $P^k(i)$ the type k projection of instance i.

To ease the formalism, a set of projections called the *context of projection* CP^X can be associated with a class X. This context of projection associated with a class serves to define the approach adopted to represent an instance of this class. A context of projection thus allows for distinguishing the various properties of interest to be distinguished

[1] Despite the fact that links can be underlined, this notion is different to the one introduced for conceptual graphs.

for the purpose of characterising an instance. The following section will define a semantic measure for estimating the proximity of two instances relative to their associated projections.

6.3.3 Semantic measures that take advantage of projections

The proximity of two instances is evaluated based on the context of projection associated with the class of affiliated instances. This measure takes into account all projections composing the context of projection for the class. The methods that enable the comparison of two instances relative to a specific projection must therefore first be defined. To this end, each projection is associated with a measure σ^k that enables the comparison of a pair of instance projections of type k, we here assume that $\sigma^k : k \times k \rightarrow [0, 1]$.

Two projections of the *Classes* type can be compared using a semantic measure adapted to a comparison of classes. A comparison of *Data* type projection requires defining a measure adapted to the type of value constituting the values sets produced by the given projection. As an example, two strings may be compared using the Levenshtein distance. *Instance* type projections, associated with a group of instances, can be compared using set-based measures in order to evaluate the number of instances shared by the projections of the two instances being compared. The measure that can be applied in this specific case will be discussed later in this section. *Complex* projections require the definition of a measure to enable the comparison of two complex objects. Let us note that in some cases, complex objects or compared data values will require some data preprocessing prior to use of the proximity function; as an example, such a preprocessing step could consist of computing the body mass index from the `size` and `weight` of an instance of a class `Person`.

Once a measure has been chosen to compare each projection, a general measure σ_X can be defined between two instances u and v of type X w.r.t CP^X. Here we present a simple example based on a weighed sum:

$$\sigma_X(u, v) \sum_{P_i^k \in CP^X, \exists P_i^k(u) \wedge \exists P_i^k(v)} w_i \times \sigma^k(P_i^k(u), P_i^k(v)) \tag{6.1}$$

where w_i is the projection weight associated to the projection P_i^k and the sum of weights equals 1. Such an approach for comparing objects considering instances w.r.t to specific properties and weights is common in the literature, e.g., [Bisson, 1992; Euzenat et al., 2004; Valtchev, 1999b; Valtchev and Euzenat, 1997].The instances of the class are compared based on a specific characterisation of all relevant properties that must be taken into account in order to rigorously conduct the comparison. In some cases,

due to lack of information, specific projections will not be expressible for an instance. This measure exploits each projection shared between the compared instances. Here, we used a weighed sum to aggregate the score of relatedness w.r.t each projection; other aggregation can also be used.

As previously observed, a projection defines a set of resources that characterise a specific property of an instance. To estimate the similarity of two instances relative to a specific projection, a measure σ^k must be specified so as to compare two sets (sometimes singletons) of resources. Various approaches are available for evaluating these two sets, namely:

- *Cardinality*: The measure evaluates the cardinality of both sets, e.g. by comparing two instances of a class `Parent` with respect to the number of children they have.
- *Direct method*: A measure adapted for a set comparison is to be used (e.g. Jaccard index); one example herein would be to compare instances relative to the number of overlapping resources, e.g. the number of common friends.
- *Indirect method*: This method relies on evaluating the proximity of the pair of resources able to be built by considering the compared sets (a Cartesian product of sets), e.g., couples of strings. In this case, an aggregation strategy must be established to aggregate the proximity scores obtained for all resource pairs built from the Cartesian product of the two compared sets. Classic operators such as Min, Max, Average or more refined approaches may be used to aggregate the scores.

As pointed out above, when an indirect method is used to compare two projections, a measure enabling the comparison of two sets of resources needs to be defined. Several approaches are available for comparing sets of classes, strings or numerical values. Note that the relevance of using a measure is once again defined by the context of usage and its semantics, i.e. the meaning scores are required to carry.

Two groups of instances can be compared by using a direct or indirect approach; an example is provided in the next section. When an indirect approach is selected, a strategy to enable the comparison of a couple of instances must be determined. It is therefore possible to use the context of projection defined for the class of the two instances under comparison. This context of projection actually defines the properties that must be taken into account when comparing two instances of this specific type. Applying such a strategy potentially corresponds to a recursive treatment, for which a stop condition is required. In all cases, computing the proximity of two projections should not imply the use of the context of projection containing both projections. A proximity measure can thus be represented through an execution graph highlighting the dependencies occurring between contexts of projection. Consequently, this execution graph must be analysed to

detect cycles for the purpose of ensuring computational feasibility. If a cycle is detected, the measure will not be computable.

6.3.4 Potential extensions

Extensions of the proposed approach have not been discussed in depth; we have merely presented the extensions to be explored while further developing the approach.

The partial ordering of classes can be exploited in order to enhance the characterisation of an instance according to the projections associated with its inferred classes. It might, therefore, be worthwhile to provide projection overloading mechanisms depending on the partial ordering (note the drawback of multiple inheritance), or else to define contexts of projection that characterise subsets of instances not framed into specific classes (e.g. a set of instances returned by a SPARQL query in RDF graphs).

The proposed approach thus enables an easy comparison of instances of a class based on the fine-tuned definition of their properties. The instances of different classes can be compared according to projections shared by the least common ancestors of their classes, i.e. projections characterising the more concrete and similar affiliated classes. Such a strategy, however, features certain drawbacks in the context of a relatedness evaluation since only instances of similar classes will tend to obtain high relatedness scores. This is because the global measure is solely driven by the property (feature) comparison of the targeted instances. In some use contexts, instances of various types are in fact expected to show high relatedness, e.g., in mimicking the human expectation of semantic relatedness, for example the instance `rollingStones` and the `Tongue` concept must be highly related since the tongue is part of the band's popular logo. These specific dimensions of relatedness can *only* be captured by measures evaluating the structural properties of the graph, i.e., the (indirect) interlinking of instances, and moreover must be framed in a graph-theoretic model corresponding to the structural approach, e.g. measures based on random walks. This definition of a context of projection can therefore be relaxed in order to allow for interlinking metrics to be included. Another approach would be to extend the notion of projection to represent an instance through abstract properties which are processed using measures evaluating interlinking, e.g., instances could be represented through their induced graph (weighed according to distance) so as to take greater advantage of measures based on graph diffusion distances and interlinking analysis.

As noted above, the notion of projection can, in some cases, be used to *transcend* the (inferable) information relative to an instance, e.g. characterising an instance through complex properties not shown in the semantic graph (e.g., the body mass index example).

The projection paradigm introduced is thus not limited to representing an instance through *literals* or numerical values but instead may be used to represent an instance through a subset of instances or complex objects. A projection is in fact defined by three core elements: (i) the resource(s) of the semantic graph from which the evaluated property will be captured, (ii) a transformation function that will ultimately preprocess the resource(s) in order to obtain the feature(s) to be evaluated by (iii) a specific measure.

Furthermore, we could also consider that the transformation function can be expected to retrieve an image associated to a String property of an instance, representing a URL or file location, to subsequently apply an image similarity measure as a comparison function to assess the relatedness of two instances regarding their projection. At the dawn of the Web of Things, driven by the use of HTTP as an application protocol, more sensors and applications will receive greater exposure; the representation of an instance through the notion of projection has not been framed into a conceptualised representation of instances (and is thus compatible with inevitable societal evolutions).

6.4 Application to content-based recommendation systems

We have proposed an approach that enables semantic measures to be expressed for the purpose of comparing instances defined in a semantic graph. This approach is particularly well suited to defining semantic measures for the design of content-based recommendation systems. Keep in mind that nothing prevents enhancing the content-based measures by incorporating other metrics in evaluating the importance or popularity of instances or other recommendation system paradigms.

6.4.1 A music band recommendation system

This section will present an example of how to use the proposed approach to define a music band recommendation system. The specific semantic graph employed was built from DBpedia [Auer et al., 2007] and Yago2 [Hoffart et al., 2013]. Other examples of Linked Data use for the purpose of deriving music recommendation systems can be found in Baumann and Schirru [2012]; Celma and Serra [2008]; Passant [2010]. The aim of the system proposed herein is to recommend music bands in considering a particular music band of interest. The user specifies a band, and the recommendation system subsequently proposes a set of bands that had been tagged as related. The relatedness is assessed based on the information contained in the ontology, the strategy adopted to define compared instances, as well as the proximity measures. The discussion will focus

on both the context of projections used to leverage the comparison of bands and the definition of interaction between the system and the user.

This recommendation system relies on a relatedness measure between two instances of the class `MusicBand`. On considering the target band, i.e. the band of interest specified by the user, e.g. `rollingStones`, the recommendation system proposes related music bands to the user. The higher the score of relatedness of one music band with the target band, the more relevant this band becomes for the recommendation. This relatedness measure is defined using two contexts of projection associated with the classes `MusicBand` and `MusicGenre`.

The context of projection associated with the class `MusicBand` is composed of three simple projections, which serve to compare any two bands with respect to: (i) their names, (ii) their types (e.g. Yago2 affiliated classes), and (iii) proximity of their related music genres. Projection (i) corresponds to the maximum similarity obtained using a Levenshtein distance. Projection (ii) is evaluated using a measure that enables the comparison of groups of classes using the taxonomic structure of ontologies. Projection (iii), related to the music genres associated with the music bands, is based on an average type of aggregation strategy; the measure used to compare two music genres relies on the context of projection defined for the class `MusicGenre`.

The context of projection of the class `MusicGenre` is composed of two simple projections, the first of which compares the labels associated with music genres. The second projection enables the comparison of music genres by taking advantage of the structuration defined by the `subgenre` relationship that establishes a partial ordering among the various music genres. The measures adopted to take these projections into account are similar to those used for the context of projection of the class `MusicBand` – projections (i) and (ii), respectively.

The music bands are therefore compared through the context of projection associated with the class `MusicBand`. It relies on the context of projection defined for the class `MusicGenre` in the projection (iii) to be computed. Other projections can be easily added in order to enrich the defined context of projections and therefore refine the comparison of instances, e.g., by taking into account the musical labels associated to the music bands. Since the goal of this experiment is to introduce the proposed approach, only these projections will be considered herein.

To distinguish the relevant music bands when considering the target band, the three projections composing the context of projection associated with the class `MusicBand` need to be evaluated. The aim then is to distinguish those bands which are more highly related to the target band in accordance with the various projections. To proceed,

a vector containing the relatedness of the target band with other music bands must be computed for each projection. The vector associated with projection (i), which in turn is associated with the band names, therefore contains the proximity of the target band with the other bands when only considering the band names. In terms of algorithmic complexity, the computation of these projection vectors constitutes the most time-consuming treatment involved in the approach. This algorithmic complexity depends, to a great extent, on the selected projections and measures. As an example, computing all vectors in order to provide recommendations for a single group takes one second using our (non-optimised) implementation based on the Semantic Measures Library[1].

The aggregation of vectors used to distinguish the more highly related music bands is not time-consuming. Such a treatment does in fact require computing a global relatedness vector based on a weighted sum, in considering the weights associated with each projection making up the context of projection defined for the class `MusicBand`. It is therefore possible to compute the vectors of projections before run time so as to further enable users to set the contribution of each projection by driving the recommendation according to their will.

6.4.2 Online application and discussions

This approach has been adopted in the demonstrator made available at `http://www.lgi2p.ema.fr/kid/tools/bandrec`.

Figure 6.3 presents the music bands which are directly or indirectly associated to the `musicGenre rockMusic`, i.e. groups in red are associated to `rockMusic` and groups in orange are annotated by a subgenre of `rockMusic`. Among the 30K bands characterised in the ontology, around 14K are annotated by this music genre. This visualisation is provided by the prototype developed for this project[2].

[1] `http://www.semantic-measures-library.org` – presented in Chapter 8.
[2] The author of the manuscript designed and developed the prototype by himself.

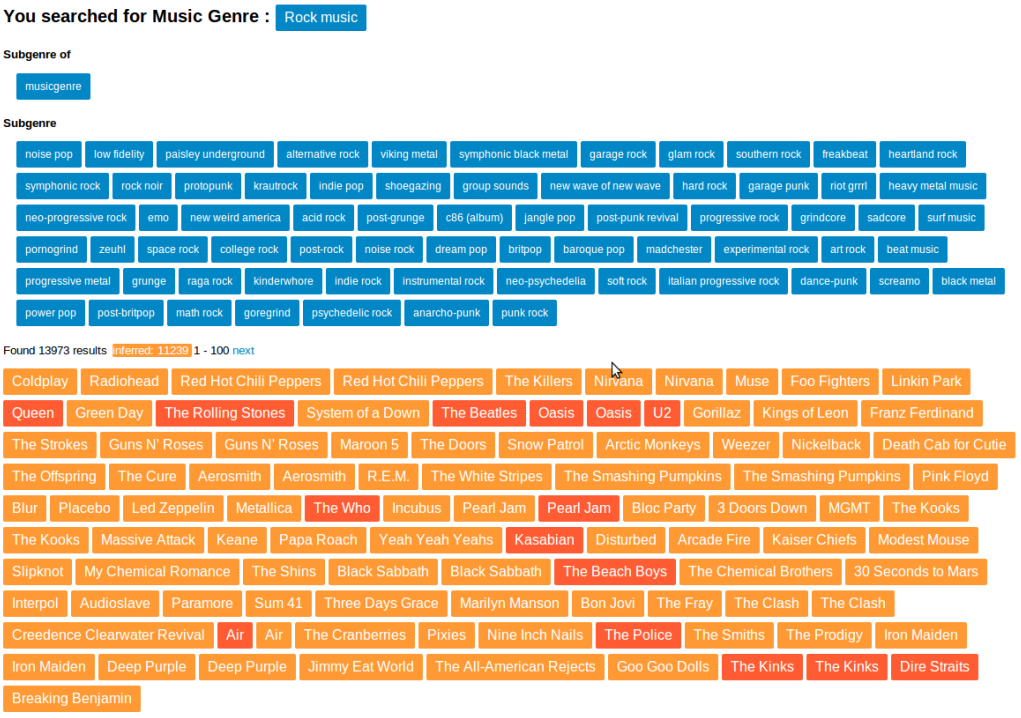

FIGURE 6.3: Screenshot of the prototype of the music band recommender system which has been developed. Information associated to the music genre `rockMusic`, i.e. more general and specific music genres, associated music bands

Figure 6.4 and 6.5 present other screenshots of the prototype. In Figure 6.4 you can see the advanced mode of the search field. The user can specify the importance to be given to each projection through sliders which are horizontal indicators of the weight associated to a projection. These sliders also ensure that the sum of the weights is equal to one. Four sliders can be distinguished in the picture, three of them are for the projections defined for the class `MusicBand` (Music Genres, Tags[1], Name), the last one is used to take the popularity of bands into consideration. The user therefore has fine-grained control over the semantics of the results produced by the system. As an example, in Figure 6.4 you can see that the system specifies (to the users) the semantics associated to their chosen configuration "*You are giving Very High importance to their music genres, Very low importance to their tags, No importance to their names and Medium importance to their popularity.*". Finally, Figure 6.5 presents the results which have been obtained by the system. The user can see details of relatedness obtained using each projection, aggregated score and information related to music bands (music genres, DBpedia URIs...). Visualisation techniques can obviously been used to polish the presentation of the results.

[1]The projection named *Tags* refers to the Yago classes to which the music bands are associated. The name Tags appeared to be more intuitive for users.

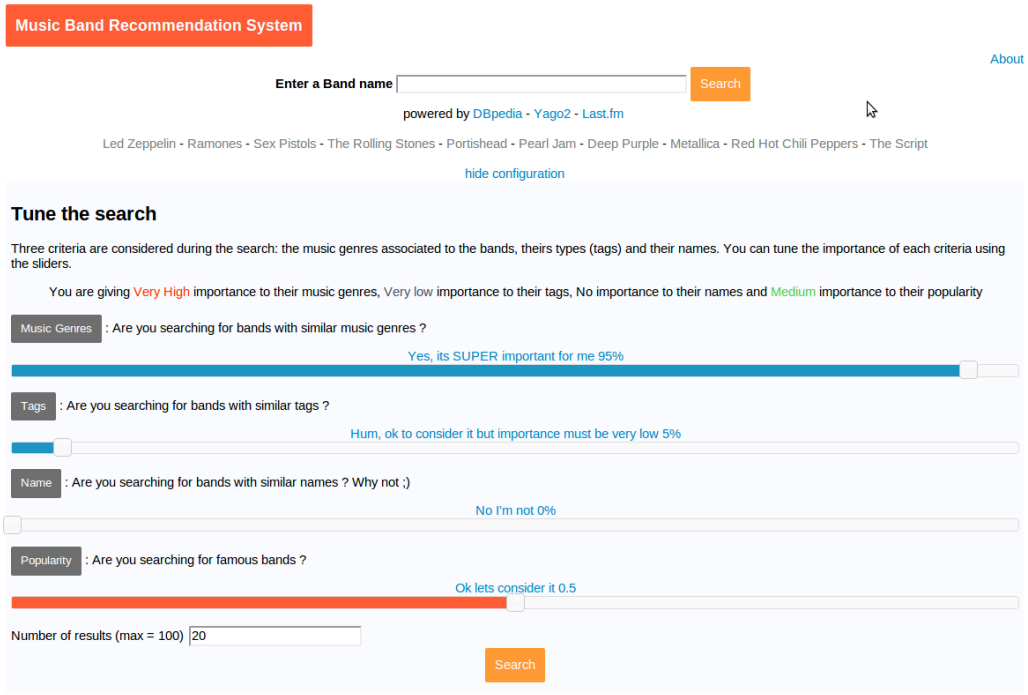

FIGURE 6.4: Interface used for the configuration of the importance to give to each projection. Therefore, using the sliders, the user can express the semantics he wants the measure to have

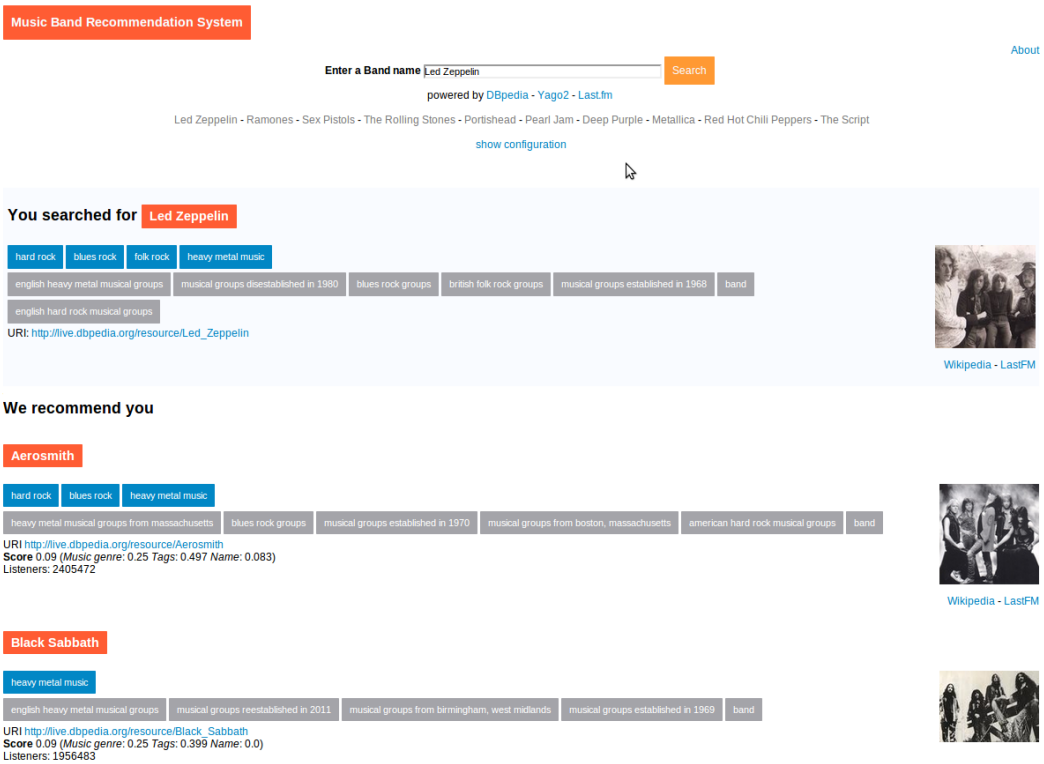

FIGURE 6.5: Examples of results provided by the prototype of the music band recommender system which has been developed: results obtained with the query `LedZeppelin`

In order to discuss the relevance of a recommendation system based on this proposed approach, we compared the results obtained by our demonstrator to those recommended by *Last.fm*[1]. This evaluation has been made for information purposes and do not aim to extensively compare both recommender systems. For each music band, *Last.fm* proposes a set of bands and artists denoted as similar. This recommendation relies both on a large database dedicated to the music and on an analysis of their user preferences. Our demonstrator makes use of a less curated knowledge base (built from DBpedia), although it still relies on a structured representation of knowledge. Our recommendation system does not focus on collaborative filtering but merely exploits a content-based approach since this approach only incorporates some of the music band characteristics (e.g. music genres associated with the bands). We have also added the notion of music band popularity, which enables importance to be assigned to this specific dimension during the search (the popularity ranking was retrieved from *Last.fm*). Finding the music bands output by *Last.fm* using our demonstrator will thus allow us to validate the proposed approach. This evaluation step has relied on 11 queries, whose results obtained by our approach were compared to those proposed by *Last.fm*, in the aim of determining the number of recommendations offered by *Last.fm* that were found by our system. For this evaluation step, we assigned high importance to the projection associated with music genres and group popularity.

Among the 40 bands proposed by *Last.fm* for these 11 queries, 19 were also recommended by our system. Differences between the recommendations mainly rely on the quality of annotations associated to the bands, as well as on the importance assigned to group popularity. It was indeed difficult to know which aspect of the system the experiment was evaluating (ontology, projections considered, weights associated to the projections...). This result is, however, promising since many of the additional recommendations proposed by our system are relevant and coherent according to the semantic characterisation associated to the targeted band (subjective evaluation performed by the designers of the system).

This first evaluation has shown, not only the added value of the proposed approach for defining semantic measures that serves to compare the resources defined in a semantic graph, but also how this approach can be used to design content-based recommendation systems. Let us underline the importance of the selection of both the projections and associated measures in ensuring the relevance of results output by the system. In addition, we underline that this proposed approach requires a high degree of expertise in both the particular field and the underlying ontology. Furthermore, more experiments and comparative studies have to be made to better characterise performance of this proposal compared to other recommendation techniques – the main aim of this first

[1] `http://www.last.fm`

step was to design an approach in which the semantics of the score of relatedness will be traceable. Indeed, the important contributions of this proposal is to provide an easy way to describe instances and to compare instances through these descriptions by enabling end-users to understand the meaning (semantics) associated to the scores of relatedness which have been obtained.

6.5 Chapter conclusion

We have proposed herein a new approach for defining semantic measures between pairs of instances contained in a semantic graph. Based on the intuitive notion of *projection*, this approach allows for an improved characterisation of instances' properties and has thus paved the way for the design of highly specific semantic measures compatible with a wide array of application contexts.

Based on a software prototype which implements our proposal, we have further demonstrated the suitability of our proposal for Information Retrieval, and more particularly, for content-based recommendation system design. More evaluations still have to be discussed to evaluate the accuracy of measures produced using this approach. Nevertheless, an interesting aspect of this approach is that it enables domain experts to explicitly define the aspects of instances that must be taken into account to ensure the relevance of results.

We have also demonstrated the added value of this approach including the user in the recommendation process on providing a means to weigh the importance of the various projections which drive the recommendation algorithm. Such an approach proves valuable in avoiding the *black box effect* of systems that rely on semantic measures and is therefore able to associate a specific semantics to a recommendation, e.g. "*This band is recommended to you because its music genres and date of creation are related to those of the* `rolling Stones`".

7

Algorithmic contributions

Contents

Abstract

This chapter presents two algorithmic contributions related to semantic measures. First, we focus on an algorithm for computing the similarity of all pairs of concepts defined in a taxonomy. This treatment is generally required when using scores of semantic measures in computational intensive applications, e.g., information retrieval systems. However, using large taxonomies, this treatment is challenging given that it generally requires millions of pairs of concepts to be compared. Nevertheless, so far, no solutions have been proposed to tackle this problem. Focusing on specific properties of certain semantic measures – which are here characterised through the framework presented in Chapter 4 – we propose a practical algorithmic solution adapted to a specific class of measures. Finally, we study the problem of assessing the semantic similarity of concepts defined in different taxonomies. Conversely to existing approaches, we propose a measure which is not restricted to using mappings between taxonomies in order to assess the commonalities/differences of compared concepts. To this end, we study in particular how the measure of pointwise mutual information, a well-known measure of association proposed by information theory, can be adapted to analyse existing mappings in order to find pairs of concepts which better estimate commonalities and differences of compared concepts. Using two gold-standard benchmarks related to the biomedical domain, we demonstrate that our proposal outperforms several existing measures, and can therefore be used to better estimate the semantic similarity of concepts defined in different taxonomies.

Associated reference on which this chapter is based:

- **An information theoretic approach to improve the semantic similarity assessment across multiple ontologies**. Batet Montserrat*, Harispe Sébastien, Ranwez Sylvie, Sánchez David, Ranwez Vincent. Information Sciences (Elsevier) 2014 (In press).

7.1 Introduction

This chapter discusses some of the algorithmic problems related to semantic measures which have been studied in this thesis. Two of them are presented in particular.

First, we focus on an algorithm for computing the similarity of all pairs of concepts defined in a taxonomy. This treatment is generally required when using scores of semantic measures in computational intensive applications, e.g., information retrieval systems. Indeed, in these cases, the performed treatments are too complex and time consuming to enable the computation of semantic measures to be done *on-the-fly*. Therefore, scores of semantic measures are precomputed for quick access. However, using large taxonomies, this treatment is challenging given that it generally requires millions of pairs of concepts to be compared. Nevertheless, so far, no solutions have been proposed to tackle this problem. Focusing on specific properties of certain semantic measures, which are here characterised through the framework presented in Chapter 4, we propose a practical algorithmic solution adapted to a specific class of measures.

Finally, we study the problem of assessing the semantic similarity of concepts defined in different taxonomies. This treatment is required in knowledge base systems which integrate several ontologies. In these cases, the comparison of concepts must take into account the information carried by all the taxonomies, and it must, for instance, be possible to assess the semantic similarity of concepts defined in different ontologies. Conversely to existing approaches which have been designed for this purpose, we propose a measure which is not restricted to using mappings between taxonomies in order to assess the commonalities/differences of compared concepts. To this end, we study in particular, how the measure of pointwise mutual information, a well-known measure of association proposed in the domain of information theory, can be adapted to analyse existing mappings in order to find pairs of concepts which better estimate commonalities and differences of compared concepts. Using two gold-standard benchmarks related to the biomedical domain, we demonstrate that our proposal outperforms several existing measures, and can therefore be used to better estimate the semantic similarity of concepts defined in different taxonomies.

7.2 Computing the semantic similarity of all pairs of concepts of a taxonomy using MSCA-based measures

Note: We sincerely thank Professor Vincent Ranwez (Montpellier SupAgro) for his useful comments on this work.

7.2.1 Motivation and objectives

It is common to take advantage of semantic measures in computational intensive applications in which they are used as components of more complex algorithms, e.g., recommendation or information retrieval systems. In these cases, precomputing semantic measure scores for their quick access is generally required. To this end, with focus on the semantic similarity of concepts, this involves estimating the score of semantic similarity for each pair of concepts defined in the considered taxonomy, which often represents a large amount of computation. Indeed, as an example, considering a symmetric semantic measure and a taxonomy composed of $n = |C|$ concepts, this leads to $\binom{n}{2} = (n \times (n-1))/2$ comparisons, e.g., considering the size of the Gene Ontology ($n = 30 \cdot 10^3$), the number of pairs of concepts for comparison is around 450 million. Moreover, in some cases, the ontology is frequently updated (sometimes daily), which requires scores of similarity to be updated. In this context, it is clear that the naive approach, which consists of computing all semantic similarities independently, is not adapted; optimisation techniques have to be used.

This section proposes to study optimised algorithmic solutions which can be used to compute the semantic similarity of all pairs of concepts defined in a taxonomy. To our knowledge, no prior contributions on the topic have been proposed. However, as we will see, analogies can be made with well-known problems tackled by graph theory.

Given that algorithmic optimisation can only be made w.r.t the chosen semantic measure, our proposal does not aim to cover all use cases. We rather take advantage of particular properties of certain semantic measures. The contribution we propose focuses more particularly on semantic measures for which the computational complexity can mainly be explained by the computation of the Most Specific Common Ancestor (MSCA) of the two compared concepts. Some of them are among the most commonly used semantic measures. They are denoted as MSCA-based measures hereafter.

MSCA-based measures can easily be identified using the theoretical framework proposed in Chapter 4 and the associated notations. Indeed, these measures are those which are

based on θ maximisation over Ω. Hence, when comparing two concepts u,v, the commonalities and differences of u and v are mainly assessed as a function of $\omega^*(u,v)$, the concept which maximises a selected θ function over $\Omega(u,v)$, the set of Non Comparable Common Ancestors of u and v (NCCAs). Examples of θ function expressions are presented in Section 3.3.2, e.g., intrinsic and extrinsic information contents. Therefore, MSCA-based measures encompass numerous measures which have been proposed in the literature, for instance, the information theoretical measures presented in Section 3.5.3. More generally, MSCA-based measures refer to all measures which can be derived from an abstracted form of a semantic measure (e.g., ratio/contrast models, σ_α and σ_β), which relies on a max aggregation over Ω. Focusing on information theoretical measures, we can for instance cite: sim_{Resnik}, sim_{Lin} and sim_{Rel} (Equations 3.28,3.29 and 3.34 respectively). The algorithmic contributions which will be introduced hereafter are dedicated to this specific type of measures. Therefore, for the sake of clarity, these algorithms are not generalised to semantic measures which take advantage of an aggregation strategy over Ω other than the maximum[1], e.g., semantic measures based on GraSM or DiShin strategies (refer to Section 3.5.5.3).

7.2.2 Algorithmic proposals

In practice, when semantic measures based on a θ function are computed, the θ value of each concept is assumed to be precomputed. This is because optimisation techniques can be used to efficiently compute all θ values by taking advantage of the partial ordering of concepts. Therefore, to compare two concepts u and v using a MSCA-based measure, the main complexity of the measure is encompassed by the computation of the MSCA of u and v, i.e., the concept $\omega^*(u,v)$ for which $\theta(\omega^*(u,v)) = \underset{c\in\Omega(u,v)}{\arg\max}\ \theta(c)$.

As an example, sim_{Lin} is of the form $2\cdot\theta(\omega^*(u,v))/(\theta(u)+\theta(v))$. Thus, considering that the access of θ values is in $O(1)$ (they have been precomputed), the algorithmic complexity of computing $sim_{Lin}(u,v)$ is defined by the algorithmic complexity associated to the computation of $\omega^*(u,v)$. Optimising of MSCA-based semantic measures, thus requires to optimise the computation of ω^*.

There is, therefore, a clear link with the detection of the Least Common Ancestor/Lowest Common Ancestor (LCA) of two nodes in a tree or in a directed acyclic graph – a well-known problem of graph theory [Bender et al., 2005; Czumaj et al., 2007; Harel and Tarjan, 1984; Schieber and Vishkin, 1988]. Nevertheless, the notion of LCA used by graph theory can be different to the notion of MSCA[2]. Indeed, in graph theory, the

[1] Nevertheless, although not explored hereafter, adaptations may be possible.

[2] As an example, if the $\theta(c)$ function is defined as the depth of the concept c, classical LCA search algorithms can be used.

LCA (in a graph without redundancies) is defined as the common ancestor of two nodes which has the longest shortest path to the root. However, considering cases in which $|\Omega| > 1$, and a θ function which is not only-based on the depth of concepts, the LCA (as defined in graph theory) might not be the MSCA. Therefore, due to this specificity, the numerous algorithms proposed in graph theory are not adapted.

7.2.2.1 First proposal

We introduce an approach which can be used to compute in $O(V^3)$[1] all pairs of concepts defined in a taxonomy using any MSCA-based semantic measure. This approach is based on a simple notion which will be explained hereafter. Let us first introduce or recall some notations:

- $D(c)$ (resp. $D^-(c)$), the descendants (exclusive descendants) of the concept c according to the partial ordering defined by the taxonomy.
- $C^+(c)$ the set of concepts for which $\forall x \in C^+(c) : \exists(c, \texttt{subClassOf}, x)$. Note that in some cases $C^+(c) \neq parent(c)$ since we can have a pair of concepts $(x, y) \in C^+(c) \times C^+(c)$ for which $x \preceq y \vee y \preceq x$.
- T_θ, an ordered list of concepts composed of the elements of C ordered according to a selected θ function. We denote $|T_\theta|$ the size of the list. T_θ is ordered such as $\theta(T_\theta[0]) = \arg\max_{c \in C} \theta(c)$, i.e., $\forall i, 0 < i < |T_\theta| - 1$, and $\theta(T_\theta[i]) > \theta(T_\theta[i+1])$.
- $pos(T_\theta, c)$ is the position of c in T_θ with $\forall c \in C : 0 \leq pos(T_\theta, c) \leq |T_\theta| - 1$.
- $\Omega(u, v)$ the set of NCCAs of the concepts (u, v) and $\omega^*(u, v) \in \Omega(u, v)$, the concept which maximises a select θ function, i.e., $c = \omega^*(u, v) \implies \theta(c) = \arg\max_{x \in \Omega(u,v)} \theta(x)$.
- With X a set, the notation $[X]$ is used to manipulate X as a list (in which all elements are associated to a specific index in the list).
- We use the notation $\sigma_\theta(x, y) = f(\theta(\omega^*) \leftarrow \theta(c))$, to highlight that the similarity of the pair (x, y) w.r.t to the semantic measure σ_θ is made by considering $\theta(\omega^*) = \theta(c)$.

Considering the notations introduced, the computation can be performed using Algorithm 1.

[1] V refers to C and E to E_T.

Algorithm 1: Computation of the similarity of all pairs of concepts of a taxonomy using a MSCA-based semantic measure – naive approach

Data: G_T,θ,σ_θ
Result: Compute $\sigma_\theta(u,v) \forall (u,v) \in C \times C$

```
1  mapDesc ← as a map such as ∀c ∈ C, mapDesc[c] ← {};
2  sim ← as a matrix [|C|][|C|] initialised with -1 values;
3  ∀c ∈ C compute θ(c) ;
4  T_θ ← sort C by increasing value of θ ;
5  for i ← 0; i < |T_θ|; i ← i + 1 do
6  |   c ← T_θ[i] ;
7  |   mapDesc[c] ← mapDesc[c] ∪ {c} ;
8  |   computeSMscoresDesc(c, mapDesc[c], sim) ;
9  |   for y ∈ C⁺(c) do
10 |   |   mapDesc[y] ← mapDesc[y] ∪ mapDesc[c];
11 |   end
12 end
```

Algorithm 2: *computeSMscoresDesc*
Compute the scores of semantic measure for all descendants of the given concept

Data: $c \in C$, *setDc* the set $D(c)$, *sim* the result matrix presented in Algorithm 1.
Result: Compute the scores of semantic measure for all descendants of the given concept c

```
1  dc ← [setDc]
2  for i ← 0; i < |dc|; i ← i + 1 do
3  |   x ← dc[i] ;
4  |   id_x ← pos(T_θ, x) ;
5  |   for j ← i + 1; j < |dc|; j ← j + 1 do
6  |   |   y ← dc[j] ;
7  |   |   id_y ← pos(T_θ, y) ;
8  |   |   if sim[id_x][id_y] = −1 then
9  |   |   |   // ⟹ ω*(x, y) = c
   |   |   |   sim[id_x][id_y] ← σ_θ(x, y) = f(θ(ω*) ← θ(c)) ;
10 |   |   |   sim[id_y][id_x] ← σ_θ(y, x) = f(θ(ω*) ← θ(c)) ;
11 |   |   end
12 |   end
13 end
```

Note that so as not to over-complicate algorithms, this section will not discuss the treatment which can be made to reduce the result matrix (i.e., *sim*). In addition *mapDesc*[*c*] which stores the descendants for the concept c can also be created and removed *on-the-fly* in order to avoid memory consumption (this approach will be used in the next algorithm).

Algorithm 1 can simply be explained by stressing that:

1. We know that $\forall x \in D(c) : pos(T_\theta, x) < pos(T_\theta, c)$. This is ensured by the fact that θ is strictly decreasing from the leaves to the root of the taxonomy and that T_θ is built ordering elements of C according to θ. We therefore have $u \preceq v \implies pos(T_\theta, u) \leq pos(T_\theta, v)$. A bottom-up approach according to the θ ordering of concepts ensures that descendants of each concept can be propagated during the process. By ensuring that $D(c)$ is computed when the concept c is processed we avoid useless computation of $D(c)$ at each iteration.

2. Therefore, according to (1), we have the guarantee that when a concept c is processed in loop Line 5, for each $(x, y) \in D(c) \times D(c)$, $pos(T_\theta, \omega^*(x, y)) \leq pos(T_\theta, c)$. The proof is trivial since for any pair $(x, y) \in D(c) \times D(c)$, if $pos(T_\theta, \omega^*(x, y)) > pos(T_\theta, c)$, it must mean that $\theta(c) > \theta(\omega^*(x, y))$. Nevertheless, since $(x, y) \in D(c) \times D(c) \implies c \in A(u) \cap A(v)$, it would mean that there is a concept c which is a common ancestor of u and v and for which $\theta(c) > \theta(\omega^*(x, y))$. This contradict the definition of $\omega^*(x, y)$. Thus, in Line 8 of Algorithm 2, if the similarity of the pair (x, y) is not already computed, this means that $\omega_*(x, y) = c$.

3. We have the guarantee that the similarity is computed for all pairs of concepts if the taxonomy is rooted by a concept $\top$[1]. This is ensured by the fact that at each iteration in which the concept c is processed, we are ensured that the similarity of all pairs $(x, y) \in D(c) \times D(c)$ is computed. Thus, at the last iteration we are ensured that we process all pairs $(x, y) \in D(\top) \times D(\top) = C \times C$.

Despite the theoretical soundness of this algorithm, its practical use is hampered by the size of the result matrix *sim*. Indeed, recall that $\binom{n}{2}$ similarities must be computed for $n = |C|$. Considering large ontologies, this leads to the manipulation of a matrix corresponding to tens of gigabits, which cannot generally be stored into memory. Persistent storage techniques are therefore used in these cases. Nevertheless, using these techniques, performance of read-write treatments on the *matrix* are highly reduced compared to in-memory preprocessing. This drawback therefore highly impacts the performance and questions the practicability of the algorithm. This is particularly true in Line 8 of Algorithm 2, we check if $sim(x, y)$ has already been computed, and for each iteration of the loop defined in Line 5 Algorithm 1, this process is made $|D(c) \times D(c)|$ times. Therefore, the number of value checkings is quickly important (bounded by $|C|^3$) and make the treatment intractable. As an example, considering the Gene Ontology ($n = 30 \cdot 10^3$), the last iteration in which the root is processed requires checking that all pairs of concepts in $C \times C$ have been computed – recall $\binom{30^3}{2} \simeq 450 \cdot 10^6$ comparisons.

[1] MSCA-based measures expect the taxonomy to contain a unique root.

If we consider that the persistent storage enables us to check if $sim(x, y)$ has been computed in 0.001 sec (which is a good performance), the last iteration alone would take more than five days[1]... We therefore introduce a refinement of the approach to avoid useless checking in the persistent storage and reduce descendant storage, i.e. memory required for maintaining *mapDesc*.

7.2.2.2 Refined approach

The pseudocode presented in Algorithm 3 is similar to Algorithm 1 except that additional data structures are used to avoid useless checking of similarity computation.

Algorithm 3: Computation of the similarity of all pairs of concepts of a taxonomy using a MSCA-based semantic measure

Data: G_T,θ,σ_θ
Result: Compute $\sigma_\theta(u, v) \forall (u, v) \in C \times C$

```
1  mapDesc ← empty map ;
2  previousNotDesc ← empty map ;
3  ∀c ∈ C compute θ(c) ;
4  T_θ ← sort C by increasing value of θ ;
5  for i ← 0; i < |T_θ|; i ← i + 1 do
6  |   c ← T_θ[i] ;
7  |   if not exists(mapDesc[c]) then mapDesc[c] ← {} end
8  |   mapDesc[c] ← mapDesc[c] ∪ {c} ;
9  |   computeSMscoresDesc_Opt(c, mapDesc[c]) ;
10 |   for y ∈ C+(c) do
11 |   |   if not exists(mapDesc[y]) then mapDesc[y] ← {} end
12 |   |   mapDesc[y] ← mapDesc[y] ∪ mapDesc[x]
13 |   end
14 |   remove(mapDesc[c]) ;
15 end
```

[1] $450 \cdot 10^6 \times 1^{-3} = 450 \cdot 10^3$ (s) $/60 = 7500$ (min) $/60 = 125$ (h) $/24 \simeq 5.21$ days.

Algorithm 4: *computeSMscoresDesc_Opt*

Compute the scores of semantic measures for all descendants of the given concept – optimised version

Data: $c \in C$, $setDc$ the set $D(c)$.

Result: Compute the scores of similarity for all descendants of the given concept c

for $x \in setDc$ **do**
  write $\sigma_\theta(c, x) = f(\theta(\omega^*) \leftarrow \theta(c))$[a] ;
  write $\sigma_\theta(x, c) = f(\theta(\omega^*) \leftarrow \theta(c))$;
end
$previousNotDesc[c] \leftarrow T_\theta[0, ..., pos(T_\theta, c)] \setminus setDc$;
for $x \in setDc \setminus \{c\}$ **do**
  for $y \in previousNotDesc[x] \cap setDc$ **do**
    write $\sigma_\theta(x, y) = f(\theta(\omega^*) \leftarrow \theta(c))$;
    write $\sigma_\theta(y, x) = f(\theta(\omega^*) \leftarrow \theta(c))$;
    $previousNotDesc[x] \leftarrow previousNotDesc[x] \setminus \{y\}$
  end
end

[a] Consider that writes in the persistent storage are made by chunks of data, e.g., each 10^3 call of the function write.

The map $previousNotDesc$ is used to store, for each concepts c, the set of concepts preceding c in T_θ and which are not descendants of c.

Once again this algorithm relies on several simple ideas:

1. If for each Iteration i on T_θ, we compute the MSCA of all pairs of concepts of $\{c\} \times \{T_\theta[0, ..., i]\}$ with $c = T_\theta[i]$, the MSCAs required to compare all pairs of concepts in $C \times C$ will be found.
2. In addition, we know that when the concept c is processed, c is either the MSCA or subsumes the MSCA of all pairs of concepts in $D(c) \times D(c)$.
3. Considering (1) and (2), we know that processing a concept c at Iteration i, c is the MSCA of each pair in: $\bigcup_{x \in D(c)} \{x\} \times \{\{T_\theta[0, ..., pos(T_\theta, x)]\} \cap D(c)\}$ for which no MSCA has been found in Iteration $j < i$.

For convenience, we define $W_{c_i} = previousNotDesc[c_i]$. Thus, considering that at each Iteration $i > j$, we store the set of concepts $W_{c_i} = \{T_\theta[0, ..., i]\} \setminus D(c_i)$ associated to the concept $c_i = T_\theta[pos(T_\theta, i)]$, and that each time we compute the similarity for the pair (c_i, c_j) we have $W_{c_i} = W_{c_i} \setminus \{c_j\}$, we can ensure that at each Iteration i, the pairs $\bigcup_{x \in D(c_i)} \{x\} \times \{W_x \cap D(c_i)\}$ will not be computed twice and that c_i will be their MSCA. In addition, given that at each Iteration i we resolve the computation of the pairs of

concepts in $D(c_i) \times D(c_i)$ which have not been resolved, we still ensure that all the results will be computed if the graph is connected.

The process of the algorithm is graphically illustrated in Appendix C.3.

The algorithmic complexity of Algorithms 1 and 3 is in $O(V^3)$[1] even if, in practice, the complexity is much lower (a taxonomy in which a transitive reduction has been applied is sparse *per* definition).

7.2.3 Synthesis

We have presented two $O(V^3)$ algorithms which can be used to compute the score of semantic similarity for each pair of concepts of a taxonomy. To our knowledge, no solution has been proposed so far for this problem. These algorithms can be used with any semantic measures which are based on the Most Specific Common Ancestor (MSCA) of compared concepts – otherwise stated, according to the notations and measure characterisation provided by the framework presented in Chapter 4, any semantic measure which is based on the maximisation of a θ function over Ω. We also proposed an algorithmic optimisation which can be used to reduce the computational time required in practical use cases.

[1] Which is similar to a naive approach in which we compute the similarity for each pair of the taxonomy independently from the others.

7.3 An information theoretic approach to improve semantic similarity assessments across multiple ontologies

From the different methods and paradigms proposed to define knowledge-based semantic similarity measures, those based on quantifying the Information Content (IC) of concepts are the most widespread solutions due to their high accuracy in most evaluations performed in the literature. However, these measures were initially designed to exploit a single ontology. They thus cannot be leveraged in many contexts in which multiple ontologies are considered. In this section, we propose a new approach to achieve accurate IC-based similarity assessments for concept pairs spread throughout several ontologies. Based on information theory, this method defines a strategy to accurately measure the degree of commonality between concepts belonging to different ontologies – a cornerstone for estimating their semantic similarity. Based on this proposal, classic IC-based measures can therefore be directly applied in a multiple ontology setting. Using well-established benchmarks and ontologies related to the biomedical domain, empirical evaluations illustrate the accuracy of our approach. We demonstrate, in particular, that the proposed approach enables similarity estimations that are significantly more correlated with human ratings of similarity than those obtained via evaluated measures.

Associated reference on which this section is based:

- **An information theoretic approach to improve the semantic similarity assessment across multiple ontologies**. Batet Montserrat*, Harispe Sébastien, Ranwez Sylvie, Sánchez David, Ranwez Vincent. Information Sciences (Elsevier) 2014 (In press).

This work has been done in collaboration with Montserrat Batet and David Sànchez from the University of Tarragona, and Professor Vincent Ranwez from Montpellier SupAgro.

7.3.1 Motivation and objectives

Semantic similarity measures coping with multiple ontologies are central in numerous contexts in which information retrieval or knowledge discovery techniques have to be applied in a multiple ontology setting. Nevertheless, these measures have seldom been considered in the literature [Al-mubaid and Nguyen, 2009; Batet et al., 2013; Rodríguez and Egenhofer, 2003; Sánchez and Batet, 2013; Saruladha, 2011] – refer to Chapter 3 for a brief overview. In this section, we study the adaptation of the information theoretical approach which is based on quantifying the Information Content (IC) of concepts. These measures are interesting given that they have been extensively studied in the single ontology setting, and are among the most widespread solutions due to the high accuracy they achieved in most evaluations performed in the literature.

As we saw in Chapter 3 and 4, in the context of IC-based measures, the identification of the Most Informative Common Ancestor (or MICA) is essential for similarity assessments. Indeed, it plays a central role in estimating the commonality between compared concepts (and sometimes in deriving their differences). Therefore, different authors have proposed to define the notion of MICA for concepts defined in different ontologies. In this case, the MICA refers to the pair of concepts of the two ontologies which best summarises the commonality of the compared concepts – we therefore denote it as the Most Informative Mapping among their Ancestors *MIMA*. As an example, existing works based on IC [Sánchez and Batet, 2013; Saruladha, 2011] retrieve the MIMA of a pair of concepts belonging to different ontologies by looking for *equivalences* of concept ancestors. In these cases, *equivalent* concepts are those which share the same labels (i.e., terminological matchings). These approaches have two main drawbacks related to the fact that mappings can only convey a partial amount of the information which should be considered in order to estimate the similarity:

1. *Mappings can be difficult to find, particularly when using simple techniques.* Currently used terminologically-based approaches are naturally hampered by the fact that ontologies do not always model concepts in the same way or refer to them using the same label (e.g., due to synonymy) [Sánchez et al., 2012c]. Indeed, in many cases, current semantic measures are based on techniques which either select ancestors which are too abstract as MIMA or which cannot discover any equivalence at all because they miss suitable concepts which share similar meanings but are referred to with different labels (e.g. *tumor*/*neoplasm*). In other cases, more refined mappings are obtained using elaborated alignment techniques but the problem is not always solved. Indeed, in numerous cases, important mappings are missed. This is particularly true when numerous large ontologies are taken

into consideration and less error-prone semi-supervised mapping techniques can no longer be used.

2. *Even optimal mappings will fail to convey relevant information.* In some cases, a perfect match cannot be defined between two concepts. Nevertheless, they could be linked by a relationship carrying distinct semantics (e.g., `partOf`). It could be relevant to consider these links when assessing the semantic similarity of concepts. As an example, Figure 7.1 illustrates a situation in which we want to compare the two concepts `IntracranialHemorrhage` and `BrainNeoplasm` defined respectively in the SNOMED-CT and the MeSH. Considering the exact mapping which has been found between their ancestors (i.e., *strict equivalence*), only the pair of concepts `Disease-Diseases` will be considered as mapping. Nevertheless, it is clear that the two concepts `DiseaseOfHead` and `BrainDiseases` should also be considered as a clue to capture the commonality of the two compared concepts more finely.

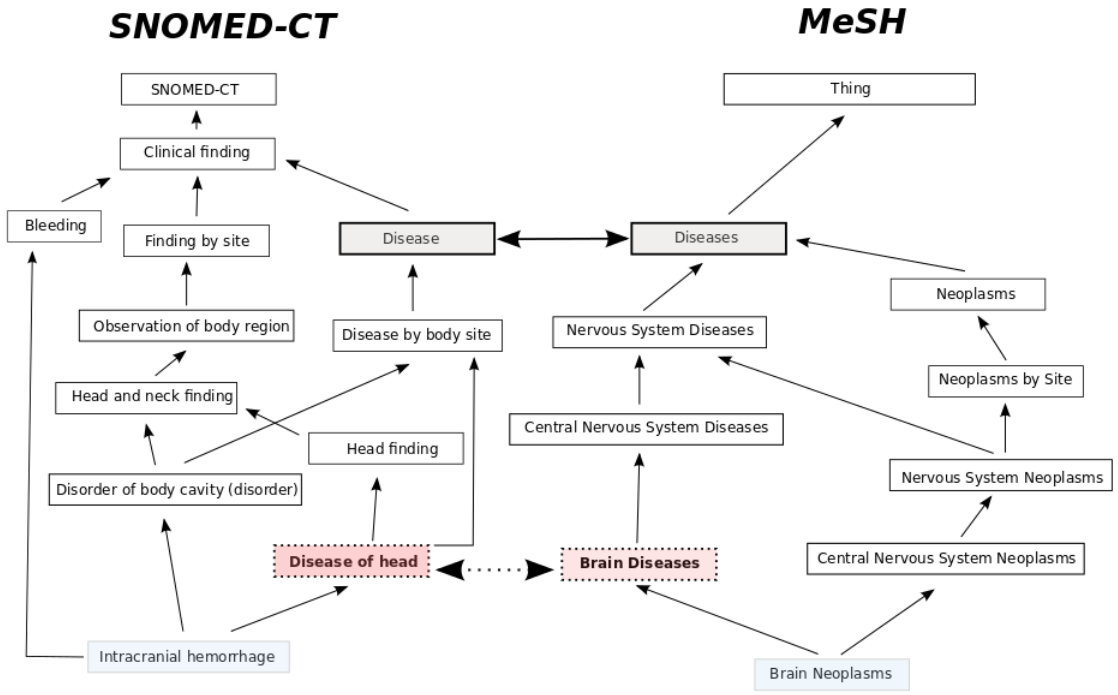

FIGURE 7.1: Comparison of concepts defined in different ontologies

Therefore, in numerous cases, the MIMA which will be found by existing approaches will be too general. This will inevitably lead to largely underestimated concept similarity. Considering the two aforementioned issues, we observe that:

1. *Advanced mapping techniques should be used.* Indeed, it is clear that terminological mapping is not sufficient and that more advanced techniques should be used instead. In this section, we will not tackle the problem of ontology alignment. This is indeed a complex and prolific field of study which has only been touched upon in this thesis. We will therefore consider that a set of concept-to-concept mappings have been found prior to estimating the semantic similarity – whatever the approach used for their computation. These mappings will be extensively used to assess the similarity of two concepts defined in different ontologies.

2. *Techniques have to be designed to overcome inherent limitations of mappings.* As stressed in the example provided in Figure 7.1, we want to avoid the estimation of the commonality of two concepts by only considering the most specific mapping found among their ancestors (MIMA). As we have seen, it can indeed only be a lower bound estimator of their real commonality.

The main objective of this study is therefore to propose a solution to overcome issue 2. To this end, we propose a new method to assess the commonality of two concepts defined in different ontologies. This method aims at not restricting the estimation of the commonality of two concepts to the information carried by their MIMA. Indeed, based on information theory and solely exploiting taxonomic knowledge and a set of mappings between concepts, our approach measures the degree of *taxonomic relationship* between concepts belonging to different ontologies. Next, this notion is used to select a MIMA that is more suitable to assess the commonality of two concepts defined in different ontologies.

The rest of this section is organised as follows. Subsection 7.3.2 presents and formalises our approach. In Subsection 7.3.3, we detail the evaluation process and discuss the results obtained for several benchmarks, ontologies and measures. Finally, this section ends with the synthesis and perspectives of this study.

7.3.2 Improving semantic similarity assessment from multiple ontologies

In this section, we present a method to enable accurate IC-based similarity calculus when concepts belong to different ontologies. Our method goes beyond the terminological matching used in related works and is able to discover semantically similar (but not necessarily terminologically identical) subsumers of two concepts between different ontologies. To do so, and in Line with the notion of IC-based similarity, our approach takes advantage of the notion of *mutual information* to quantify the degree of *taxonomic relationship* between pairs of concept from different ontologies. Once two concepts are

compared, the Most Informative Link among their Ancestors (*MILA*) is computed. Note that the notion of MILA is different from the previously discussed notion of MIMA. Indeed, as we will see, contrary to the MIMA, the MILA does not obviously refer to a concrete mapping – yet rather, conceptually speaking, refers to the notion of MICA, which is commonly used in single ontology setting. Thus, this MILA will be used to estimate the semantic similarity between the two concepts using standard IC-based measures. Note that for convenience, the definition of compared concepts in different ontologies will systematically be considered.

7.3.2.1 Estimating the commonality of two concepts

To lighten the formalism, we denote the taxonomy G instead of G_T. In addition, since we will manipulate multiple taxonomies, we will use subscripts to ease the reading. For instance, we define $C_i = C(G_i)$. Thus, if $u \in G_1$, $\forall c \in A(u)$ we are certain that $c \in C_1$. Nevertheless, to highlight which taxonomy is associated to a notation, we will denote $A_i(u)$ the ancestors of u in G_i. We denote $\langle u, v \rangle$ the mapping between u and v and $M(U, V)$ the set of mappings defined between the two sets of concepts (U,V). As an example, the mapping defined between the ancestors of u and v will be denoted as $M(A_1(u), A_2(v))$.

Considering two concepts defined in different taxonomies, $u \in G_1$ and $v \in G_2$, the notion of $MIMA(u, v)$ doesn't refer to a single concept, but rather to a pair of concepts (x, y), with $x \in A_1(u)$ and $y \in A_2(v)$[1]. We consider that at least one mapping exists among them i.e. $M(A_1(u), A_2(v)) \neq \emptyset$. Thus, to ensure that a MIMA always exists between two concepts, we consider that all taxonomies are rooted by a concept which corresponds to the more abstract notion commonly defined in knowledge modelling (e.g., `Thing`). Thus, for each taxonomy G_i we consider a root $\top_i$, and for any pair of taxonomies (G_i, G_j), we admit the mapping $\langle \top_i, \top_j \rangle$.

We now formally define the notions used to introduce our proposal.

Definition $MIMA : C_i \times C_j \rightarrow C_i \times C_j$: The $MIMA$ of two concepts u, v defined in different taxonomies refers to the Most Informative Mapping among their Ancestors, i.e., the mapping found in $M(A_1(u), A_2(v))$ which respects:

$$MIMA(u, v) = \underset{\langle x,y \rangle \in M(A_1(u), A_2(v))}{\arg\max} IC(x) + IC(y) \tag{7.1}$$

[1]Strictly speaking, this is different from the original definition of the MICA function (e.g. different co-domains), but its role is similar.

Naturally, the maximisation of the sum of ICs of matched ancestors ensures that, in case of multiple matches in $M(A_1(u), A_2(v))$, the most informative pair is taken. For convenience, we will systematically denote the mapping $MIMA(u, v)$ as $\langle bm_1(u), bm_2(v)\rangle$ (bm refers to best mapping).

The $MIMA$ only refers to a lower bound of the commonality of two concepts. Indeed, nothing ensures that all matchings have been obtained and that no more informative mappings can be considered. Indeed, in some cases, a more specific pair of concepts defined in $A_1(u) \times A_2(v)$ should be considered as $MIMA$, nevertheless, the mapping has not been found. We model this aspect by introducing the notion of $MILA$, with $MILA(u, v)$, the Most Informative Link among the Ancestors of u and v. In the best case, if the mappings are very good, the $MILA$ is the $MIMA$. Otherwise stated, the $MILA$ must be more specific or as equally specific as the $MIMA$. Therefore, we consider that the $MILA$ is defined in the set $cMILA$, which refers to the set of pairs of concepts which are possible MILA candidates. Considering the two concepts u and v, $cMILA(u, v)$ refers to the set of pairs of concepts from $A_1(u) \times A_2(v)$, which are taxonomically equal or below their $MIMA(u, v) = \langle bm_1(u), bm_2(v)\rangle$. Formally, the set $cMILA$ of two concepts is defined by:

$$cMILA(u, v) = \{(x, y) \in \{A_1(u) \times A_2(v)\} | (x \preceq bm_1(u)) \wedge (y \preceq bm_2(v))\} \tag{7.2}$$

Considering the given definition of $cMILA$ for two concept u, v, let us precise that we do not consider the pairs $(x, y) \in \{A_1(u) \times A_2(v)\}$ which are not subsumed by the pair of concepts defined by the $MIMA(u, v)$ - this has been highlighted by Jérôme Euzenat in personal communication. We therefore may miss interesting MILA candidates. Nevertheless, the consideration of such pairs highly complicate the approach and we let the evaluation of such a strategy to another study.

Note that, according to the subsumption relation $\preceq$, $cMILA(u, v)$ contains the pairs (u, v) and $(bm_1(u), bm_2(v))$, they are therefore considered as potential $MILA$ for assessing the semantic similarity of u,v.

Among the candidate pairs included in $cMILA(u, v)$, we define the pair with the highest degree of *taxonomic relationship* as $MILA(u, v)$, i.e., the pair $(x, y) \in cMILA(u, v)$ such as x subsumes most of the semantics of y and *vice-versa*. The rationale for this criterion is that we consider that the more two concepts subsume concepts for which mappings have been found, the more similar the semantics they subsume. We can therefore assume that they are interchangeable and correspond to a relevant semantic link worth considering when assessing the similarity of compared concepts.

We therefore need to design an estimator of the strength of the *taxonomic relationship* which links two concepts defined in different ontologies, w.r.t the mappings defined between their descendants. To this end, we propose to adapt the notion of Mutual Information (MI) and more particularly the notion of Pointwise Mutual Information (PMI) [Church and Hanks, 1990]. Similar to the IC calculus, the PMI between two concepts can be computed according to their probabilities of (co-)occurrence. Formally, considering two concepts u and v, the PMI quantifies the difference between the probability that u and v co-occur given their joint and marginal probabilities.

$$PMI(u,v) = log\frac{p(u,v)}{p(u)p(v)} \tag{7.3}$$

The PMI function is symmetric. Given the above expression, $PMI(u,v) = 0$ means that u, v are completely independent, i.e. $p(u,v) = 0$, whereas increasing positive values indicate an increasing degree of association between the concepts. On the contrary, negative values reflect mutual exclusion, which is quite uncommon among concepts or words, since most of them tend to be semantically related to a certain degree [Anandan and Clifton, 2011].

A common criticism concerning PMI is that it tends to provide relatively high scores for rare events [Bouma, 2009]. For example, we have $p(u) = p(v) = p(u,v)$ when two terms only occur together and it then follows from Equation 7.3 that $PMI(u,v) = -log(p(u,v))$. This means that, for perfectly correlated concepts, their PMI value will be higher when they appear less frequently. Moreover, PMI has no fixed upper bounds, which complicates its interpretation since it is thus hard to know from a given PMI value, if two concepts are almost perfectly associated (respectively almost independent). These problems may be partly solved by using the Normalised Pointwise Mutual Information (NPMI). Indeed, NPMI values are bound within the interval $[-1, 1]$ and are less impacted by low frequency data [Bouma, 2009]. Therefore, NPMI normalisation is done by dividing the PMI ratio by the actual probability of co-occurrence between the two terms:

$$NPMI(u,v) = \frac{PMI(u,v)}{-log(p(u,v))} = \frac{log\frac{p(u,v)}{p(u)p(v)}}{-log(p(u,v))} \tag{7.4}$$

NPMI results in a maximum value of 1 for perfect correlation, a minimal value of -1 for mutually exclusive concepts, and a value of 0 for independent concepts since their PMI is null.

Given the above arguments and properties, NPMI provides a sound way to measure concept mutuality. In the next section, we detail how the NPMI is computed considering only the topology of taxonomies and the set of mappings between their concepts. We

will also show how it can be used to derive the MILA of two concepts defined in different ontologies.

7.3.2.2 Adaptation of the NPMI

The estimator selected for estimating the co-occurrence probability $p(u, v)$ is crucial to ensure that $NPMI(u, v)$ correctly reflects the strength of the taxonomic relationship between two concepts. As for the IC estimation, this probability can be estimated using suitable corpora, by counting the number of simultaneous appearances of those two concepts (this can also be done with instances). In our setting, this results in two main issues:

1. Corpora-based probability calculus is hampered by data sparseness and restricted by corpora availability (refer to Section 2.2.2.2).

2. Term co-occurrences are not usually disambiguated, i.e., the kind of semantic relationship, taxonomic or non-taxonomic, on which the co-occurrences rely are not defined. Therefore, the co-occurrence frequency only reflects the semantic relatedness between concepts [Bollegala et al., 2009]. Indeed, co-occurrence frequencies and (N)PMI measures based on them have already been applied to evaluate several types of semantic association, e.g., word collocation [Bouma, 2009; Sánchez and Isern, 2009], taxonomic subsumption [Vicient et al., 2013], and a variety of non-taxonomic relationships [Sánchez, 2010; Sánchez et al., 2012b]. However, in our case, the use of co-occurrence frequency to estimate $p(u, v)$ may result in high NMPI scores although the strength of the taxonomic relationship between u and v may be weak. For example, let us consider the concepts: `Cancer` and `Chemotherapy`. Assuming the availability of an appropriate corpus, their degree of textual co-occurrence is likely to be high, resulting in an equally high NPMI value. However, since `Chemotherapy` is a common treatment for `Cancer`, those numerous co-occurrences reflect a semantic relatedness rather than a degree of taxonomic relationship. This issue becomes problematic since we are interested here in unravelling concepts with strong taxonomic relationship in order to find the suitable MILA of two concepts.

We propose to tackle these problems by using probability estimations derived from the taxonomies in which concepts are modelled. The probabilities of individual concepts are computed intrinsically, according to the premises of the intrinsic IC calculus discussed in Section 3.3.2. Specifically, as in some intrinsic IC models, we rely on the fact that the meaning of a concept is partially defined and bound by its set of descendants [Sánchez et al., 2011; Seco et al., 2004]. Hence, from the leaves to the roots of taxonomies,

the number of shared descendants between two concepts gives us a good idea of their common *taxonomic trigger potential*, i.e., the concepts they might both refer to when they are mentioned/encountered. Indeed, shared descendants of two concepts are the concepts among their descendants which partially refer to the same concepts when they are mentioned. Therefore, as the overlap between the descendant sets of two concepts increases, their meanings become more equivalent. Note that this notion, which is the core of our subsumer matching method, differs from that of similarity quantified by IC-based measures. Indeed, two sibling concepts (e.g. `BreastCancer` and `LungCancer`) may be highly similar (according to their IC-based similarity) while sharing no descendants and, hence, being completely different w.r.t their taxonomic trigger potential.

Considering the concepts u in G_1 and v in G_2, our desire is to approximate $p(u, v)$, the joint probability of (u, v), as a function of the number of mappings founds between the descendants of u and v (i.e., $|M(D_1(u), D_2(v))|$), and the number of mappings founds between C_1 and C_2, i.e., $|M(C_1, C_2)|$. Nevertheless, some considerations have to be made. Indeed, comparing two concepts u,v, we can theoretically obtain a number of mappings between u,v which is completely different to their number of descendants, e.g., $max(|D_1(u)|, |D_2(v)|) \ll |M(D_1(u), D_2(v))|$. Otherwise stated, it is not the number of mappings which matters but rather the number of concepts involved in these mappings. Thus, the notions of intersection and union of concepts defined in different ontologies have to be redefined. To this end, we propose considering the size of the intersection between a set S_1 of concepts of G_1 and a set S_2 of concepts of G_2 as:

$$|S_1 \cap S_2| = \frac{|\{c_1 \in S_1 | \exists (c_2 \in S_2 \wedge \langle c_1, c_2 \rangle)\}| + |\{c_2 \in S_2 | \exists (c_1 \in S_1 \wedge \langle c_1, c_2 \rangle)\}|}{2} \quad (7.5)$$

The size of the union is then simply defined as the complement of the defined intersection: $|S_1 \cup S_2| = |S_1| + |S_2| - |S_1 \cap S_2|$.

Note that the defined union and intersection are not necessarily associated to integer values. However, they ensure consistent results even with ontologies that have heterogeneous granularities, i.e., for which 1:M or even N:M matchings between concepts can be found. For instance, if $S_1 = \{u, v\}$; $S_2 = \{x\}$ and the mappings $\langle u, x \rangle$ and $\langle v, x \rangle$ have been defined, then $|S_1 \cap S_2| = 1.5$ and $|S_1 \cup S_2| = 1.5$. Therefore, in accordance to our desire, thanks to the higher cardinality of descendant sets associated to general concepts, and the reduction of the ambiguity which increases with the specificity of concepts, the chance of obtaining a representative number of matchings increases in comparison with subsumer sets.

Based on the union and intersection operators which have been introduced, we can formally define the *joint probability* of u and v, $p(u,v)$, by:

$$p(u,v) \simeq \frac{|D_1(u) \cap D_2(v)|}{|C_1 \cup C_2|} \tag{7.6}$$

We also define the *marginal probability* of an individual subsumer u as:

$$p(u) \simeq \frac{|D_1(u)|}{|C_1 \cup C_2|} \tag{7.7}$$

Given the above instantiations of joint and marginal probabilities, we define the intrinsic NPMI of a pair of concepts as follows. Given $u \in G_1$ and $v \in G_2$, their *intrinsic NPMI* (or *iNPMI*) is defined as:

$$\begin{aligned} iNPMI(u,v) &= \frac{log\frac{p(u,v)}{p(u)p(v)}}{-log(p(u,v))} \\ &= \frac{log\frac{\frac{|D_1(u)\cap D_2(v)|}{|C_1\cup C_2|}}{\frac{|D_1(u)|}{|C_1\cup C_2|}\times\frac{|D_2(v)|}{|C_1\cup C_2|}}}{-log\frac{|D_1(u)\cap D_2(v)|}{|C_1\cup C_2|}} \end{aligned} \tag{7.8}$$

which can for instance be simplified by:

$$iNPMI(u,v) = \frac{log\frac{|D_1(u)\cap D_2(v)|}{|D_1(u)|\times|D_2(v)|}}{log\frac{|C_1\cup C_2|}{|D_1(u)\cap D_2(v)|}} \tag{7.9}$$

Numerically, $iNPMI(u,v) = 0$ indicates that u and v have no overlapping and, therefore, that these two concepts cannot serve as MILA. On the contrary, an $iNPMI(u,v)$ value close to 1 indicates that there is a strong taxonomic link between u and v since they share most of their descendants. In our approach, the *MILA* of two concepts is thus the candidate with the highest $iNPMI$ value.

Therefore, the MILA for $u \in G_1$ and $v \in G_2$ is a pair of concepts from $cMILA(u,v)$. First, we consider:

$$MILA_*(u,v) = \underset{(x,y)\in cMILA(u,v)}{\arg\max} iNPMI(x,y) \tag{7.10}$$

In rare cases, Equation 7.10 will lead to multiple pairs of concepts. In these cases, the MILA is the pair with the maximum sum of IC values (i.e. the most informative one, in coherency with the notion of MICA):

$$MILA(u,v) = \underset{(x,y)\in MILA_*(u,v)}{\arg\max} \{IC(x) + IC(y)\} \tag{7.11}$$

The above-described method can be generalised for encompassing cases in which u and/or v belong to several disjoint sets of ontologies (e.g., u belongs to G_1 and G_3, and v belongs to G_2 and G_4). In that case, the proposed method is applied for each combination of pairs of ontologies (e.g., $G_1 - G_2$, $G_1 - G_4$, $G_3 - G_2$ and $G_3 - G_4$). The MILA of two concepts is selected as the pair of their subsumers, from the different pairs of ontologies, that produces the highest iNPMI value. The rationale is that, comparing two concepts u and v, the more the iNPMI increases among a pair $(x, y) \in cMILA(u, v)$, the higher the number of mappings between the descendants of (x, y) and hence, the pair (x, y) will be suitable for comparing concepts u and v. Remember that the iNPMI is normalised and has bound outputs. This is quite convenient for comparing NPMI values computed from different ontologies, regardless of their size and degree of granularity.

7.3.2.3 IC-based similarity calculus

The $MILA$ of two concepts is a pair of concepts. However, for comparing two concepts, most IC-based measures have been designed for aggregating the IC of their $MICA$, i.e. a single IC value. Therefore, to be used in a straightforward manner with existing measures, our notion of $MILA$ should be associated to a single IC value. As we saw in Section 3.3.2, the IC of the MICA should always be lower than any of its descendants; this is required to ensure the consistency of IC-based similarity measures [Resnik, 1995]. Therefore, to ensure that this property will be fulfilled in our setting, we define the intrinsic IC of the MILA as the minimum IC value of its concept (computed from their respective taxonomy). Thus, considering $MILA(u, v) = (x, y)$ with $x \in G_1$ and $y \in G_2$, we define $IC'(x, y)$[1] by:

$$IC'(MILA(u, v)) = min(IC(x), IC(y)) \tag{7.12}$$

Note that using this definition we ensure that:

$$IC'(MILA(u, v)) \leq IC(x) \wedge IC'(MILA(u, v)) \leq IC(y)$$

The proof is trivial considering that: (i) $MILA(u, v) = (x, y)$, (ii) $u \preceq x$ and $v \preceq y$, (iii) IC decreases monotonically from the leaves to the root on a taxonomy, i.e. $IC(x) \leq IC(u)$ and $IC(y) \leq IC(v)$. Thus, we have $IC'(MILA(u, v)) = min(IC(x), IC(y)) \leq IC(x) \leq IC(u)$ and $IC'(MILA(u, v)) = min(IC(x), IC(y)) \leq IC(y) \leq IC(v)$.

[1] Denote IC' since the domain of the function is $C \times C$ rather than C for the IC.

Thanks to the definition of the MILA proposed in this section, all IC-based measures which rely on the notion of IC and MICA can now be used to compare pairs of concepts defined in different taxonomies.

7.3.3 Evaluation

In this section, we evaluate the proposed method in comparison with related works. Since our final goal is to enable a precise IC-based assessment of similarity in a multiple ontologies setting, we focused on cases where each concept that is to be evaluated belongs to a different ontology. In such scenarios, similarity assessments directly depend on the adequacy of the subsumer pair selected as MILA and the subsequent IC calculus, as detailed in the previous subsection. Hence, by quantifying the accuracy of the similarity assessment, we also indirectly test the relevance of our MILA identification strategy – which is finally a substitute for the MICA identification strategy. The accuracy of the proposed method is compared with those of related works also focusing on multiple ontologies IC-based similarity calculus [Sánchez and Batet, 2013; Saruladha, 2011]. We also compare our approach with results obtained in an *ideal* single ontology setting, i.e., when similarities are computed from a single ontology.

To enable an objective evaluation, the accuracy of similarity assessments was quantified by comparing them with human judgements of similarity for two well-known term pair benchmarks [Pakhomov et al., 2010; Pedersen et al., 2007]. The accuracy of alternative similarity estimations has been measured through their correlation with human ratings via the Pearson's correlation coefficient, as done in many similar studies [Al-mubaid and Nguyen, 2009; Bollegala et al., 2009; Pirró, 2009; Sánchez and Batet, 2011; Sánchez et al., 2012a]. A correlation value near 1 indicates that both ratings are very close and, hence, that the computerised assessment accurately reflects human judgement of similarity.

Tested methods propose different solutions to identify the MICA of concepts from different ontologies and to estimate its IC, which can next be used to compute the semantic similarity using IC-based measures like those introduced in Section 3.5.3. In our tests, the IC of individual concepts was computed intrinsically according to the equation defined in [Sánchez et al., 2011] (Equation 3.7). Moreover, since the accuracies of tested methods may depend on the IC-based measure chosen for similarity calculus, we tested them with several measures: Resnik's, Lin's and Jiang and Conrath's (please refer to Section 3.5.3 for equations and references).

The evaluation was conducted on biomedical datasets because of the availability of different ontologies and similarity benchmarks in this field. In particular, SNOMED-CT and the MeSH ontologies have been used. Alternative methods have been compared

using two biomedical benchmarks, i.e. the one proposed by Pedersen et al. [2007] and the one proposed by Pakhomov et al. [2010]. The former consists of a set of medical term pairs whose similarity was assessed by a group of medical experts from the Mayo Clinic: 9 medical coders who were introduced to the notion of semantic similarity and 3 physicians who rated terms without any special training. Term pairs included in the benchmark were specifically selected by the authors to maximise the inter-rating agreement, resulting in a correlation value of 0.68 obtained for physicians and of 0.78 for coders. The benchmark built by Pakhomov et al. consists of one set of concept pairs associated with similarity and relatedness ratings given by four medical residents from the University of Minnesota. We took the similarity ratings since we were focusing specifically on semantic similarity. Note that for the Pakhomov et al. benchmark, the inter-rating agreement (0.47) is significantly lower than that obtained for the dataset of Pedersen et al.

Even though such benchmarks are intended to evaluate similarity measures in a single ontology setting, related works have already used them in a multiple ontology framework by artificially considering that each term of each pair belongs to a different ontology. According to the same protocol as in Batet et al. [2013]; Sánchez and Batet [2013], we took the 25 term pairs from the Pedersen et al. benchmark and the 150 concept pairs from the Pakhomov et al. benchmark, such that both elements of the pair could be found in SNOMED-CT as well as in MeSH. Hence, it was possible for the 175 pairs considered in our evaluation procedure to assess their pairwise similarity in a single ontology setting. Since similarity estimations are obviously harder and more precarious in a multiple ontology setting, we can consider that the single ontology results give us a reasonable approximated upper bounds of the best accuracies we can expect in a multiple ontology setting. For each benchmark, we conducted two different multiple ontology evaluations. In the first case, referred to as SNOMED-CT + MeSH, the first concept of each pair was retrieved from SNOMED-CT and the second one from MeSH. Whereas in the second case, referred to as MeSH + SNOMED-CT, the first concept was retrieved from MeSH and the second one from SNOMED-CT.

The SNOMED-CT release of July 2012 (20120731) and the MeSH 2013 release were used for the evaluation. semantic measures were implemented using the Semantic Measures Library[1]. Mappings between ontologies were computed using a terminological comparison of the labels associated to concepts. Details on the evaluation and on the computation of mappings, as well as the source code and associated datasets used in this experiment, can be checked and downloaded from the dedicated webpage: `http://www.lgi2p.ema.fr/~sharispe/publications/IS2013/`.

[1] Open source library dedicated to semantic measures which will be introduced in the next chapter.

Measure	ontologies	MICA discovery	Physicians	Coders	Both
Resnik	SNOMED-CT	None	0.553	0.598	0.602
	MeSH	None	0.608	0.668	0.670
	SNOMED-CT + MeSH	Sánchez and Batet	0.489	0.544	0.542
	SNOMED-CT + MeSH	Saruladha et al.	0.474	0.546	0.535
	SNOMED-CT + MeSH	**This work**	**0.617**	**0.624**	**0.649**
	MeSH + SNOMED-CT	Sánchez and Batet	0.444	0.534	0.512
	MeSH + SNOMED-CT	Saruladha et al.	0.432	0.536	0.508
	MeSH + SNOMED-CT	**This work**	**0.562**	**0.639**	**0.632**
Lin	SNOMED-CT	None	0.566	0.628	0.625
	MeSH	None	0.614	0.674	0.676
	SNOMED-CT + MeSH	Sánchez and Batet	0.512	0.561	0.561
	SNOMED-CT + MeSH	Saruladha et al.	0.501	0.569	0.560
	SNOMED-CT + MeSH	**This work**	**0.637**	**0.654**	**0.674**
	MeSH + SNOMED-CT	Sánchez and Batet	0.446	0.542	0.517
	MeSH + SNOMED-CT	Saruladha et al.	0.432	0.543	0.511
	MeSH + SNOMED-CT	**This work**	**0.561**	**0.648**	**0.637**
JC	SNOMED-CT	None	0.538	0.612	0.602
	MeSH	None	0.618	0.670	0.676
	SNOMED-CT + MeSH	Sánchez and Batet	0.514	0.573	0.569
	SNOMED-CT + MeSH	Saruladha et al.	0.505	0.580	0.569
	SNOMED-CT + MeSH	**This work**	**0.637**	**0.651**	**0.673**
	MeSH + SNOMED-CT	Sánchez and Batet	0.423	0.527	0.498
	MeSH + SNOMED-CT	Saruladha et al.	0.404	0.524	0.487
	MeSH + SNOMED-CT	**This work**	**0.542**	**0.638**	**0.622**

TABLE 7.1: Correlation values of different IC-based measures against human ratings for term pairs extracted from the Pedersen et al. benchmark in single and multiple ontology scenarios. Rows in boldface show the results of our proposal

Tables 7.1 and 7.2 show the correlation values for the two benchmarks for each IC-based similarity measure. Tables show the cases in which both concepts are retrieved from SNOMED-CT, when both are evaluated in MeSH, and when each concept is considered in a different ontology (SNOMED-CT + MeSH and MeSH + SNOMED-CT), using the MICA discovery and calculus strategies of Sánchez and Batet [2013], Saruladha [2011] and the one presented in this section.

Methods based only on terminological matchings resulted in correlation values that were below the worst single ontology setting, i.e. those of SNOMED-CT in these tests. Given that, in our testing protocol, all methods relied on the same IC calculus [Sánchez et al., 2011], the differences between the method of Sánchez and Batet and that of Saruladha et al. were minor. In fact, since both methods look for the most specific pair of matching subsumers, the only practical difference for the evaluated scenarios regards the criterion of the IC calculus for the discovered pair. Indeed, Saruladha et al. select the minimum IC from the matched pair, whereas Sanchez and Batet take the maximum value[1].

[1]Note that, as discussed with the authors, selecting the higher IC of the two concepts is problematic as it can lead to incoherent results (since $IC(MILA(u,v))$ might be greater than $IC(u)$ or $IC(v)$).

Measure	ontologies	MICA discovery	Experts
Resnik	SNOMED-CT	None	0.513
	MeSH	None	0.511
	SNOMED-CT + MeSH	Sánchez and Batet	0.315
	SNOMED-CT + MeSH	Saruladha et al.	0.305
	SNOMED-CT + MeSH	**This work**	**0.493**
	MeSH + SNOMED-CT	Sánchez and Batet	0.260
	MeSH + SNOMED-CT	Saruladha et al.	0.243
	MeSH + SNOMED-CT	**This work**	**0.429**
Lin	SNOMED-CT	None	0.505
	MeSH	None	0.519
	SNOMED-CT + MeSH	Sánchez and Batet	0.320
	SNOMED-CT + MeSH	Saruladha et al.	0.310
	SNOMED-CT + MeSH	**This work**	**0.505**
	MeSH + SNOMED-CT	Sánchez and Batet	0.257
	MeSH + SNOMED-CT	Saruladha et al.	0.244
	MeSH + SNOMED-CT	**This work**	**0.447**
JC	SNOMED-CT	None	0.456
	MeSH	None	0.520
	SNOMED-CT + MeSH	Sánchez and Batet	0.313
	SNOMED-CT + MeSH	Saruladha et al.	0.232
	SNOMED-CT + MeSH	**This work**	**0.505**
	MeSH + SNOMED-CT	Sánchez and Batet	0.257
	MeSH + SNOMED-CT	Saruladha et al.	0.199
	MeSH + SNOMED-CT	**This work**	**0.448**

TABLE 7.2: Correlation values of different IC-based measures against human ratings for term pairs extracted from the Pakhomov et al. benchmark in single and multiple ontology scenarios. Rows in boldface show the results of our proposal

The difference in performance between those two methods and the single ontology settings strongly depends on the considered dataset and IC-based measures. In some cases, this difference could be small (e.g. 0.534-0.546 vs. 0.598-0.668 for Resnik's measure and the Pedersen et al. coder ratings) or significantly large (e.g. 0.199-0.313 vs. 0.456-0.520 for Jiang and Conrath's measure and the Pakhamov et al. expert ratings, which represents a more challenging dataset). As discussed, those two methods are hampered by the fact that, in many cases, the matching subsumer pair is more abstract than it should be, and these results in an underestimation of the true similarity between the compared concepts. This issue is specifically tackled by our approach, which looks for a pair of subsumers with a higher degree of taxonomic relationship than those involved in the best matching. A more suitable pair of ancestors can therefore be found, hence improving similarity assessments. As an example, Figure 7.1 page 244, presents the result which were obtained using the proposed approach. The two concepts `IntracranialHemorrhage` and `BrainNeoplasms` were compared. In this case, the iNPMI function defined the pair of concepts (`DiseaseOfHead`, `BrainDiseases`) as MILA – note that their MIMA was initially $\langle$`Disease`, `Diseases`$\rangle$. Note also that not all results appear as appealing as this one, even if, finally, the correlation with expected scores was improved.

Indeed, in the performed evaluation, correlations obtained using our method noticeably improve those of related works (e.g., 0.624-0.639 vs. 0.534-0.546 for Resnik's measure and the Pedersen et al. coder ratings, and 0.448-0.505 vs. 0.199-0.313 for Jiang and Conrath's measure and the Pakhomov et al. expert ratings). In fact, these results are close to those obtained in single ontology contexts, e.g. 0.624-0.639 vs. 0.598-0.668 for Resnik's measure and the Pedersen et al. coder ratings, and 0.448-0.505 vs. 0.456-0.520 for Jiang and Conrath's measure and the Pakhomov et al. expert ratings. These differences are also more uniform for all measures and datasets than those of related works. Recall that, as stated above, correlation values reported in both tables for single ontology settings give us a reasonable approximation of the best correlation that can be achieved in multiple ontology scenarios.

Regarding IC-based similarity, our results confirm that Lin's and Jiang and Conrath's measures tend to lead to better results than those of Resnik. This is expected as the Resnik measure, unlike the two other measures, associates the same similarity to pairs of concepts with identical MICA, regardless of the IC of the compared concepts.

Note, finally, that computed similarities are more congruent with human ratings for the Pedersen et al. benchmark than for those of Pakhomov et al. This result is coherent with the difference in inter-human agreement figures for the two benchmarks: 0.68-0.78 for Pedersen et al. compared to 0.47 for Pakhamov et al. The influence of the reliability of human ratings is also evident with the Pedersen et al. benchmark, where computed similarities are better correlated with the coders' ratings (which are more consistent, i.e. inter-rating agreement of 0.78) than with the physicians' ratings (which are less consistent, i.e. inter-rating agreement of 0.68). The higher inter-rating agreement among coders is certainly related to the fact that they were trained on the notion of semantic similarity, whereas the physicians rated pairs of terms without previous training [Pedersen et al., 2007].

7.3.4 Discussion

The applicability of IC-based semantic measures is hampered by the fact that they were designed to deal with a single ontology. This constitutes a serious limitation given the increasing importance of scenarios in which multiple heterogeneous ontologies have to be used [Batet, 2011a; Coates et al., 2010; M.C. Lange, D.G. Lemay, 2007].

In this section, a method is proposed to enable accurate IC-based similarity assessments from multiple ontologies. It proposed to overcome shortcomings of existing proposals and in particular to revisit the way the MICA of two concepts defined in different ontologies can be computed. Our approach, grounded on the foundations of the information theory,

and on an intrinsic redefinition of the NPMI, looks for the available pair of concepts that can act as the best estimator of the commonality of the compared concepts – even if no mapping has been found between them. This proposal overcomes the limitation of related works which solely rely on an existing set of mappings. Conversely, our method proposes a way to measure the strength of the taxonomic link between two concepts defined in different taxonomies by analysing the topology of the taxonomies and associated mappings. As a result, we discover pairs of ancestors that better represent the commonalities of the compared concepts and that therefore enable more accurate semantic similarity assessments.

The empirical evaluation, carried out on several well-established benchmarks, ontologies and measures, sustained the theoretical hypothesis: our method achieved similarity results that correlated significantly better with human ratings than those of tested related works. In addition, the results obtained were very close to those obtained in the "*optimal*" single ontology setting. More evaluations using other benchmarks, ontologies and mapping acquisition techniques have to be performed in order to generalise these encouraging results.

7.4 Chapter conclusion

This section has presented two algorithmic contributions related to semantic measures. First, we presented an algorithm for computing the semantic similarity of all pairs of concepts defined in a taxonomy. Using the framework proposed in Chapter 4, we characterised specific properties of semantic measures which can be used to design efficient algorithm to tackle this problem. Based on these findings, we proposed an efficient and practical algorithmic solution.

Finally, in the last section, jointly with Montserrat Batet, David Sànchez and Vincent Ranwez, we proposed a new approach to designing semantic similarity measures for comparing concepts defined in different taxonomies. Based on well-known notions of information theory, we proposed a new approach to finding estimators of the commonalities and differences of compared concepts. Interestingly, this approach has proved to increase the accuracy of several semantic measures in two gold-standard benchmarks related to biomedicine.

8

The Semantic Measures Library

Contents

Abstract

The Semantic Measures Library and Toolkit are robust open source software solutions dedicated to semantic measures. They can be used for large-scale computation and analysis of semantic similarities, proximities or distances between terms or concepts defined in ontologies, e.g., structured vocabularies, taxonomies, RDF graphs. The comparison of instances (e.g., documents, patient records, genes) annotated by concepts is also supported. An important aspect of these new solutions is that they are generic and are therefore not tailored to a specific application context. They can thus be used with various controlled vocabularies and knowledge representation languages (e.g. OBO, RDF, OWL). The project targets both designers and practitioners of semantic measures providing a Java source code library, as well as a command-line toolkit which can be used on personal computers or computer clusters.

The library implements a large collection of state-of-the-art measures and several parametric measures provide fine-grained tuning capabilities for specific usage contexts. The Application Programming Interface associated to the library, and the numerous algorithms and metrics implemented, equip developers with an extensive framework for the development of new measures. It also provides researchers with a development platform particularly suited for the comparison and evaluation of semantic measures. In addition, it also enables developers to easily take advantage of semantic measures and to use the functionalities of the library in their development projects. The library and toolkit have been extensively used for several use case scenarios in which fast computation of semantic measures were required, e.g., large-scale analysis and computation of semantic measures, development of conceptual information retrieval systems. Interestingly, despite their generic aspect, they have proved to compete or even outperform the performances of domain-specific solutions. In short, the Semantic Measures Library and Toolkit aim at equipping communities studying and using semantic measures with robust, reliable and efficient, open source, generic and tools dedicated to them. Downloads, documentations, updates and community support are available at `http://www.semantic-measures-library.org`

Associated references on which this chapter is based:

- **The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies**. Sébastien Harispe*, Sylvie Ranwez, Stefan Janaqi, Jacky Montmain. Oxford Bioinformatics 2013.
- **From Theoretical Framework to Generic Semantic Measures Library**. Sébastien Harispe*, Stefan Janaqi, Sylvie Ranwez, Jacky Montmain. On the Move to Meaningful Internet Systems: OTM 2013 Workshops Lecture Notes in Computer Science Volume 8186, 2013, pp 739-742; `http://dx.doi.org/10.1007/978-3-642-41033-8_98`

8.1 Motivation

Throughout this manuscript we have stressed that numerous communities are involved in the study of semantic measures, e.g. Natural Language Processing, Artificial Intelligence, Semantic Web and Bioinformatics, to mention a few. Moreover, we also underlined that, due to their popularity, a large number of semantic measures have been proposed for a wide range of applications and ontologies. In addition, we stressed that the evaluation of semantic measures, more often than not, relies extensively on empirical analysis aiming to support the added value of specific proposals for a specific task, e.g. disambiguation, clustering, correlation with human expectations regarding semantic similarity.

Nevertheless, no extensive software tools dedicated to semantic measures were available at the start of this thesis. A state-of-the-art outlining the software solutions available for the computation and analysis of semantic measures was published in [Harispe et al., 2013c]. In summary, most software solutions focusing on those dedicated to knowledge-based semantic measures, have been developed for a specific usage context and are dedicated to a specific ontology, e.g., Wordnet [Pedersen et al., 2004], UMLS [McInnes et al., 2009], the Gene Ontology[1](GO) or the Disease Ontology (DO) [Li et al., 2011]. Table 8.1 summarises some characteristics of existing libraries/tools.

As an example, a large number of tools have been developed for the GO alone (only some of them are presented in the Table 8.1). Nevertheless, besides some particular aspects requiring *ad hoc* or domain specific tuning, all measures defined for particular usage contexts can be expressed using the same graph-based formalism. This was particularly underlined in Chapter 4 which is dedicated to the unification of semantic similarity measures. Thus, from a theoretical point of view, nothing prevents the definition of a generic software or library dedicated to semantic measures. However, to our knowledge, only four attempts to develop generic libraries related to semantic measures[1] exist, i.e., OWLSim, SimPack, OntoSim and SemMF, the rest being dedicated to particular ontologies. We briefly discuss some of the characteristics of these generic libraries.

SemMF[2] is a library which can be used to evaluate the similarity of instances represented as RDF graphs [Oldakowski and Bizer, 2005]. The library proposes, among other matching techniques, some taxonomic matchers relying on Lin's measure or the shortest weighted path restricted by the Least Common Ancestor. SemMF is nevertheless no

[1]Note that half a dozen libraries/tools are dedicated to the Gene Ontology: `http://www.geneontology.org/GO.tools_by_type.semantic_similarity.shtml`

[1]Note that semantic measures here only refers to semantic measures relying on an ontology.

[2]`http://semmf.ag-nbi.de`

Name	ontology	Types	Measures	Language
FastSemSim[1]	*GO*	CLI, LIB	P, G	Python
Similarity Library [Pirró and Euzenat, 2010a]	*Wordnet,MeSH,GO*	CLI, LIB	P, G	Java
DOSim [Li et al., 2011]	*DO*	CLI, LIB	P, G	R
WordNet-Similarity [Pedersen et al., 2004]	*WordNet*	CLI, LIB	P, G	Perl
UMLS-Similarity [McInnes et al., 2009]	*UMLS*	LIB	P	Perl
OWLSim [Washington et al., 2009]	OWL, RDF, OBO	LIB	P	Java
SimPack [Bernstein et al., 2005]	OWL, RDF	LIB	P	Java
SemMF [Oldakowski and Bizer, 2005]	OWL, RDF	LIB	P, G	Java
OntoSim [David and Euzenat, 2008]	OWL, RDF	LIB	P	Java
SML [Harispe et al., 2014]	OWL, RDF, OBO	CLI, LIB	P, G	Java

TABLE 8.1: Some characteristics of a selection of libraries/software enabling the computation of knowledge-based semantic measures. Types: Command Line Interface (CLI), Library (LIB). Measures: pair of concepts (Pairwise – P), pair of groups of concepts (Groupwise – G).

longer supported (last version 2008). OWLSim[1] proposes a reduced implementation of semantic measures to be applied to ontologies and focuses mainly on the comparison of OWL objects [Washington et al., 2009].

Among the three generic solutions evaluated, SimPack[2] is probably the most extensive library providing numerous types of semantic measures [Bernstein et al., 2005]. Focusing on measures dedicated to the comparison of concepts, some information theoretical measures (e.g. Lin's measure), or measures based on the structural approach, are proposed. However, this library does not provide ways to compare sets of concepts, and, more importantly, SimPack is no longer under active development (last version published in 2008). Finally, our tests reveal that this library cannot be used in numerous use cases since it is impossible to load relatively large ontologies, such as those available in the biomedical domain today. As an example, during our evaluations, we failed to load the Disease Ontology (*only* 8656 concepts) despite the 6Go memory allocated to the process. This aspect is today highly problematic since ontologies tend to grow in size – let us note for instance that the GO structures more than $30 \cdot 10^3$ concepts. The SimPack source code analysis which was been performed revealed that the underlying data structure manipulated by the library was not adapted to handle large ontologies.

[1] `https://code.google.com/p/owltools/wiki/OwlSim`
[2] `https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack`

Unfortunately, the modification of this data structure implied too many profound modifications to be conceivable. Finally, SimPack is only suited to experienced developers as no command-line interface is provided for end-users.

OntoSim provides a generic framework for defining various similarities for comparing entities defined in an ontology. It can therefore be used to compare instances or concepts defined in an ontology, as well as for comparing ontologies. This software solution provides several measures which can be aggregated. It also relies on external libraries for specific types of measures, e.g., SimPack for comparing two concepts defined in a taxonomy. Note that OntoSim has not been evaluated and compared to the other solutions. However, strictly speaking OntoSim does not focus on measures for comparing a pair of (groups of) concepts and is therefore commonly used for other usage context, e.g. to compare ontologies. Nevertheless, we strongly encourage the reader to visit the dedicated website, to refer to the documentation and to test the solution: `http://ontosim.gforge.inria.fr` .

Not only focusing on generic solutions, another limit of existing software solutions is that they only give access to a limited set of measures which is not representative of the large numbers of measures available today. It is worth noting that most measures which have been proposed in the literature have generally not been implemented in a software solution. This situation challenged both the use and the study of semantic measures. Indeed, semantic measures users were constrained to using domain-specific tools which often only propose few measures (generally no more than five), and do not include theoretical findings made by other communities, e.g., more accurate measures or algorithmic optimisations. Thus, semantic measure users were limited to available ontology-specific implementations; if no software solutions had been developed (and were supported) for their ontology of interest, it often meant that you had to develop your own source code, often from scratch.

Semantic measure designers and more generally semantic measure studies were also limited by the lack of extensive and efficient generic software solutions dedicated to semantic measures. Indeed, to date, most evaluations of semantic measures have been made using *private* and *closed* source code, a situation which highly challenges comparisons of semantic measure and experiment reproducibility[1]. The situation was very limiting as experiment reproducibility, one of the main tenets of the scientific method, is the only way to validate empirical results.

[1] We personally spent hours trying to reproduce published results before, most often, giving up...

Therefore, to federate efforts related to the design and analysis of semantic measures, and to respond to the need for a generic and extensive open source software tool dedicated to them, this chapter introduces the Semantic Measures Library, a software solution dedicated to the computation and analysis of semantic measures which has been developed during this thesis[1].

8.2 The Semantic Measure Library: generic software solutions for the computation and analysis of semantic measures

The Semantic Measures Library (SML) is a source code library dedicated to the computation, development and analysis of semantic measures. Numerous functionalities provided by the SML are also available within the SML-Toolkit, a command-line programme which can be used by non-developers to easily compute semantic measures on personal computers or computer clusters. The SML and the toolkit are distributed under the open source CeCILL license[2] (GPL-compatible). They both use cross-platform Java programming language which is available for most operating systems (version 1.7).

The SML and the toolkit can be used to compute semantic similarities of concepts or structured terms defined in ontologies. They can also be used to assess the semantic similarity of pairs of instances annotated by concepts, e.g., patient records annotated by groups of concepts, genes annotated by GO terms, PubMed articles annotated by MeSH descriptors... Considering a pair of entities (concepts/instances), these tools provide an easy way to compute a score of semantic similarity, relatedness or distance depending on the measure considered. The SML and associated toolkit can be downloaded from their dedicated website: `http://www.semantic-measures-library.org`.

We briefly present the SML, associated toolkit and other contributions related to this project.

[1] Note that the development, maintenance, packaging, support, and documentation writing of the solutions presented in this chapter have only been supported by the author of the manuscript.

[2] http://www.cecill.info/index.en.html

8.2.1 SML: a source code library dedicated to semantic measures

The SML is a generic Java source code library dedicated to semantic measures. Developers can easily embed source code referring to the library to compute semantic measures in their own algorithms and applications. The library supports various ontology formats and specifications (e.g., OBO[1], RDF, OWL). It takes advantage of the Sesame library to handle RDF graphs [Broekstra et al., 2002]. The SML relies on a graph-based data model and is therefore adapted to compute semantic measures on any ontologies relying on graph representations or which can be reduced to such a representation. Specific ontology loaders are provided to handle several widely used domain-specific ontologies distributed using specific formats. As an example, specific loaders for biomedical terminologies such as the MeSH and SNOMED-CT, or for other ontologies such as WordNet, are available. Custom ontology loaders can also be easily added for processing specific data formats.

Low-level access to the library enables developers to finely control the underlying graph-based data model in order to apply specific treatments which are sometimes required for the computation of semantic measures (e.g. transitive reduction to remove taxonomic redundancies or annotation redundancies). This aspect is often essential to ensure the coherency of the computation of semantic measures.

A large collection of semantic measures are provided out-of-the-box, version 0.7 of the SML supports numerous state-of-the-art semantic measures relying on different strategies (e.g., information theoretical, structure-based, feature-based). Thanks to the fine-grained control provided by the library, this leads to about 1500 specific measure configurations that can be specified for context-specific applications (considering the various ICs, measures and aggregation strategies implemented). The library also gives access to several parametric measures which can be used by developers for fine-grained tuning in specific usage contexts.

In addition, the algorithms developed in the SML provide designers of semantic measures an extensive Application Programming Interface (API) and framework to easily develop, test and evaluate new measures. Interestingly, due to its generic underlying graph data model, semantic measures developed using the SML will benefit a large audience. Indeed, measures developed using the SML are not restricted to a specific ontology and can therefore be used with the various ontologies supported by the library. Note that measures based on complex DL constructs are currently not supported.

The SML has been developed for the large-scale computation and analysis of semantic measures. It supports multi-threaded processes for fast parallel computation on

[1] Open Biomedical Ontologies format [OBO, 2013].

multi-core processors. The library can therefore be used to compute semantic measures between entities characterised in graphs composed of millions of triplets. Nevertheless, the in-memory data model on which the library now relies may be limiting when processing large collections of triplets on classic computers (e.g. hundreds of millions/billions of triplets)[1].

Figure 8.1 presents a source code example which shows how semantic measure scores can easily be computed using the SML – in this simple example, the semantic similarity between two concepts structured in a taxonomy is computed using the measure defined by [Lin, 1998] and the information content proposed by Sánchez et al. [2011] (Equations 3.29 and 3.7 respectively).

```java
public class SMComputation {

    public static void main(String[] params) throws SLIB_Exception{

        // Init
        URIFactory factory = URIFactoryMemory.getSingleton();

        // Load the graph
        URI graph_uri = factory.createURI("http://graph/");
        G graph = new GraphMemory(graph_uri);

        GDataConf graphconf = new GDataConf(GFormat.NTRIPLES, "/tmp/graph_test.nt");
        GraphLoaderGeneric.populate(graphconf, graph);

        // Semantic measure configuration (Lin + topological IC)
        ICconf icConf = new IC_Conf_Topo(SMConstants.FLAG_ICI_SANCHEZ_2011_a);
        SMconf smConf = new SMconf(SMConstants.FLAG_SIM_PAIRWISE_DAG_NODE_LIN_1998);
        smConf.setICconf(icConf);

        // URIs of the concepts to compare
        URI whale = factory.createURI("http://graph/class/Whale");
        URI horse = factory.createURI("http://graph/class/Horse");

        // Load the engine used to compute semantic measures
        SM_Engine engine = new SM_Engine(graph);

        // Computation
        double sim = engine.computePairwiseSim(smConf, whale, horse);
    }
}
```

FIGURE 8.1: Example of Java source code which relies on the SML to compute the semantic similarity between two concepts (`Whale` and `Horse`)

[1]Reflections have been initiated on this specific aspect and prototyping (which appeared unsatisfactory) has been made during the thesis – efficient handling of large ontologies through graph traversals is still an open problem (note that triplestores are not adapted in this case).

8.2.2 SML-toolkit for non-developers

Numerous functionalities provided by the SML are also available within the SML-Toolkit, a command-line programme which can be used by non-developers to easily compute semantic measures on personal computers or computer clusters. Indeed, the toolkit provides access to some functionalities of the library through command-line software. This is particularly important since most users of semantic measures are not developers and only use semantic measures for knowledge-based data analysis, e.g., in bioinformatics users analyse gene products through their GO annotations, for instance based on the analyse of clusters computed using semantic measures.

The SML-Toolkit is highly tuneable and enables context specific configurations to be specified depending on the experiment performed: knowledge base to use (ontologies, conceptual annotations), required data pre-processing (e.g., the removal of taxonomic or annotation redundancies), measure constraints (e.g., algorithmic complexity, mathematical properties), set of queries to perform (i.e. concept or instance identifiers), and other (optional) parameters (e.g. output file, computer resources allocated).

Detailed configurations can be specified using an XML file. An example is provided in Figure 8.2; the configuration specifies how to compute the semantic similarity of pairs of gene products considering their GO annotations and more particularly the annotations related to the biological processes in which they are involved. We briefly detail the meaning of this XML configuration:

- **A** – The user can specify **global configurations**, such as the number of threads to use during the computation. *Variables* can also be defined in order to reduce the size of the configuration file and to ease its modification. *Namespaces* are sometimes required to load prefixed URIs used in certain data files.

- **B** – The **knowledge base** used during the process is composed of the ontology, here the GO, and the annotations which specify the GO terms associated to each gene. As you can see, the user can specify pre-processing treatments to perform prior to semantic measure computation. In this example, the user defines that only the biological process aspect of the GO must be considered; the taxonomy of the ontology is modified using the `REROOTING` command, other concepts and associated annotations will be removed. Note that gene annotations associated to GO terms not found in the ontology will be excluded. Advanced configurations can also be used to define the behaviour to adopt if, for instance, annotations are not found or compared genes have no associated annotations, e.g., to set the score to a specific value, to throw an error/a warning... – this can be done in **D**.

- **C** – The **configuration of semantic measures**. Multiple measure configurations can also be specified if one wishes to compute multiple scores of semantic measures in the same run, i.e., the toolkit will compute the score of each semantic measure configuration for each pair of gene products specified. In this example, the configuration of the measure corresponds to an indirect groupwise measure. Comparing two genes u, v, which corresponds here to the comparison of two sets of GO terms U and V, the maximal score of similarity which has been obtained by comparing each pair of GO terms composing the Cartesian product $U \times V$ will be considered. The semantic similarity of each pair of GO terms of the Cartesian product is computed using an information theoretical measure (Resnik's measure considering an information content computed according to the annotation usage in the set of annotations provided w.r.t the partial ordering of concepts defined in the GO).

- **D** – The user finally specifies the **input file** which contains the pairs of gene products for which the semantic similarity must be computed; the **output file** and **extra configurations** are also defined.

Considering that the configuration file presented in Figure 8.2 is saved in the file named `sm_conf_human_bp.xml`, the execution can easily be launched using the following command-line:

```
java -jar sml-toolkit-[version].jar -t sm -xmlconf sm_conf_human_bp.xml
```

The XML interface provides advanced possibilities for tuning the SML-Toolkit configuration. Nevertheless, such an interface may still appear too complex, or unnecessarily complex for numerous users and use cases. We have therefore also developed kinds of domain-specific command-line interfaces which can be used to take advantage of the toolkit in specific use cases. This aspect is particularly important given that the usage of semantic measures is largely interdisciplinary. Indeed, most of the time, users do not understand what the documentation means when we talk, for instance, about conceptually annotated entities or instances of an RDF graph; a molecular biologist wants to read: gene products annotated by GO terms. Thus, specific command line interfaces, called *profiles*, are also developed to ease the use of the SML-Toolkit in specific application contexts.

As an example, a profile has been developed to estimate the similarity of genes regarding their GO term annotations. Users can therefore easily compute the semantic similarity between GO terms structured in the GO or between genes which are conceptually characterised by GO terms. The following command-line can therefore be used to compute

```
<sglib>
    <opt  threads = "4"  /> <!-- Global options configuration we only use multiple threads-->

    <variables> <!-- Global variables which will be used to shorten the configuration file -->
            <var key   = "HOME"         value = "/tmp/sml/demo" />
            <var key   = "GRAPH_GO"     value = "{HOME}/go_tutorial/data/gene_ontology_ext.obo" />
            <var key   = "ANNOT_HUMAN" value = "{HOME}/go_tutorial/data/gene_association.goa_human" />

A           <var key = "URI_GRAPH"           value = "http://biograph/" />
            <var key = "URI_GO"              value = "{URI_GRAPH}go/" />
            <var key = "URI_UNIPROTKB"       value = "{URI_GRAPH}uniprot/" />
            <var key = "BIOLOGICAL_PROCESS" value = "{URI_GO}0008150"/>
    </variables>

    <!-- Namespaces to consider e.g. GO:0008150 will be loaded as http://biograph/go/0008150 -->
    <namespaces> <nm prefix="GO" ref="{URI_GO}" /> </namespaces>

    <graphs>
            <graph uri="{URI_GRAPH}"  >
                <data>
                    <file format="obo" path="{GRAPH_GO}" />
                    <file format="gaf_2" prefix = "{URI_UNIPROTKB}" path="{ANNOT_HUMAN}" />
B               </data>
                <actions> <!-- Pre-processing to perform after graph loading -->
                    <!-- Reduce the graph considering GO:0008150 as root -->
                    <action type="REROOTING" root_uri="{BIOLOGICAL_PROCESS}" />
                </actions >
            </graph>
    </graphs>

    <!-- Configuration of the SML-toolkit module dedicated to semantic measures computation -->
    <sml module="sm" graph="{URI_GRAPH}" >

            <ics> <ic id   = "icCorpus" flag = "IC_ANNOT_RESNIK_1995_NORMALIZED"  /> </ics>

            <measures type = "pairwise">
                    <measure id = "Resnik" flag = "SIM_PAIRWISE_DAG_NODE_RESNIK_1995" ic = "icCorpus" />
C           </measures>

            <measures type = "groupwise">
                    <measure id   = "MAX"
                             flag  = "SIM_GROUPWISE_MAX"
                             pairwise_measure = "Resnik" />
            </measures>

             <queries
                    id      = "queries_human_genes"
                    type    = "oTOo"
D                   file    = "{HOME}/go_tutorial/data/queries/queries.csv"
                    output  = "{HOME}/go_tutorial/data/results/queries_results.csv"
                    uri_prefix = "{URI_UNIPROTKB}"
            />
    </sml>
</sglib>
```

FIGURE 8.2: Example of SML-Toolkit XML configuration file which can be used to compute semantic similarities of pairs of gene products annotated by Gene Ontology terms. Please consult last release documentation

the semantic similarity between pairs of GO terms specified in the file (`query.tsv`[1]). Providing the GO (`go.obo`), the output file (`results.tsv`) and a measure configuration (`-pm schlicker -ic sanchez`), the following command-line can be used:

```
java -jar sml-toolkit-<version>.jar -t sm -profile GO -go go.obo
                                    -mtype p -queries query.tsv
                                    -output results.tsv
                                    -pm schlicker -ic sanchez -aspect BP
```

Such profiles are interesting since they hide the advanced capabilities of the library and must be associated to domain-specific documentation. They therefore enable users to focus on the important aspects of the domain use case. It therefore improves the experience for users who are only interested in computing semantic measure scores in a specific context of use (e.g. gene or disease analysis). The development of more profiles is a short term objective.

8.2.3 Website & other contributions

The development of the SML and associated toolkit goes alongside several initiatives to both promote them and to ease their use. As an example, the website of the SML project is available at: `http://www.semantic-measures-library.org`.

It gives access to:

- Downloads (library, toolkit, javadoc, tutorials).
- Extensive documentations associated to the library and the toolkit, as well as technical documentations associated to semantic measures. Tutorials which show how to use the toolkit are also provided.
- Community support associated to a Google group `sml-support`[2] and a mailing list.
- An extensive literature related to the field. A Mendeley group has also been created in order to share references associated to semantic measures[3].
- News & updates related to the project.

[1] Tabular separated file (can be customised), e.g.:

```
P16591 Q00839
Q00839 E2QRD5
E2QRD5 Q9H9B1
Q9H9B1 Q9H9B4
...
```

[2] `https://groups.google.com/forum/#!forum/sml-support`

[3] `http://www.mendeley.com/groups/2907161/semantic-measures-library-bibliography`

8.3 Comparison with domain specific tools

8.3.1 Aim of the comparison

This section presents an evaluation which has been performed in order to compare the performance of the SML w.r.t other domain-specific solutions dedicated to semantic measures. Here, we focus on the comparison of the SML-Toolkit with other solutions developed to compute semantic measures using the GO.

This section has been written according to the documentation and results of the sm-tools-evaluation project which was made during this thesis. Documentation and material associated to the project are available at `https://github.com/sharispe/sm-tools-evaluation`.

> **Important**: this evaluation does not aim to criticise tools or denigrate the work made by their developers – we only define a strict evaluation protocol in order to provide objective metrics which can be relevant when comparing tools. Please keep in mind that tools which do not perform well on the tests defined herein may have other advantages that are not discussed in this evaluation. In addition, this evaluation does not pretend to cover all aspects which could be useful to consider in order to evaluate software solutions. Here, we focus on objective metrics and mainly aim at evaluating the speed of the programme given specific resource constraints (memory allocated to the tool, computational time).

We only provide results which are *strictly reproducible* given the source code and information considered during the evaluation. The aim is not to discuss aspects relative to the (subjective) individual user experience or other important aspects such as documentation and code quality, usability, overall sustainability, community support, release updates... We do, however, encourage users to refer to corresponding tool documentation and websites to evaluate these aspects. More general information related to software evaluation can be found at `http://www.software.ac.uk/software-evaluation-guide` (other references are provided in the project documentation).

The source code used to perform this evaluation is open source; it can therefore be used to reproduce the results presented herein by simply downloading the repository and following the instructions. Note that results may vary considering the hardware configuration of the machine on which the test is performed; nevertheless, rankings must be the same.

The tools which have been compared are:

- The `Semantic Measures Library Toolkit` (SML) – version 0.7[1]
- `GOSim` – version 1.2.7.7[2]
- `GOSemSim` – version 1.18.0[3]

[1] `http://www.semantic-measures-library.org`
[2] `http://cran.r-project.org/src/contrib/Archive/GOSim`
[3] `http://www.bioconductor.org/packages/release/bioc/html/GOSemSim.html`

- `FastSemSim` (`FSS`) – version 0.7.1[1]

8.3.2 Evaluation protocol

Tools are compared regarding their computation time. Two tests are presented in this section:

- Computation of the semantic similarity between GO terms.
- Computation of the semantic similarity between gene products annotated by GO terms.

For both tests the following datasets have been used:

- Gene Ontology – lite version of 2013 03 02, so as to be in accordance with GOSim and GOSemSim which both rely on Bioconductor[2] `R` package [Gentleman et al., 2004].
- Gene annotations – Human gene annotations provided in Bioconductor version 2.12.

8.3.2.1 Semantic similarity between Gene Ontology terms

This test aims to compare the tools for the computation of semantic similarities between a pair of GO terms. Four tests of different sizes were generated: 1K, 10K, 1M and 100M pairs of GO terms[3]. Each test is therefore composed of a set of pairs of terms for which we want the semantic similarity to be computed. All the samples can be downloaded at project webpage.

For each test of size x (e.g. 1M), three random samples of size x were generated in order to reduce the probability that the evaluation of the performance is biased by abnormal sampling. As an example, the test composed of 1M pairs of terms is composed of three different samples r_0, r_1, r_2 which each contains 1M pairs of GO terms. For each sample (e.g., r_1), three runs ($r_{1.0}$, $r_{1.1}$, $r_{1.2}$) were performed. This is to reduce the probability of results being biased by abnormal operating system behaviour or material lags.

The sets of pairs of terms which make up the 3 samples of each test were generated using the tool provided in the project. Both the tool and its source code are open-sourced and publicly available. This tool is used to generate benchmarks composed of pairs of GO terms. As we said, for each size of benchmarks (1K, 10K, 1M and 100M), three samples are generated. These benchmarks were built selecting random pairs of terms specified in the Biological Process aspect of the GO (all pairs of terms are composed of terms subsumed by the term `GO:0008150`). In addition, all terms which appear in the test are used to annotate at least one gene defined in the gene annotations considered. Indeed, some libraries cannot compute the similarity of terms which are not used to annotate at least one gene – this is due to the computation of Resnik's Information Content (IC).

We selected Lin IC-based measure (Equation 3.29) to evaluate the tools performance, the formula is presented in Chapter 3. Lin's measure is commonly used to compare two concepts/terms

[1] `http://sourceforge.net/projects/fastsemsim`
[2] `http://www.bioconductor.org/packages/2.12`
[3] K= 10^3, M= 10^6

defined in a taxonomy. It requires the Most Informative Common Ancestor of the compared terms and (by default) Resnik IC to be computed. This choice of measure configuration was made given that (i) IC-based measures are the most commonly used measures, and (ii) MICA determination and IC computation are the two most time consuming treatments of all IC-based measures.

Specific constraints were specified in order to simulate user expectations using the tools:

- *Memory consumption*: processes cannot use more than 6Go of memory – we expect the tools to be used on personal computers.
- *Time constraint*: processes cannot take more than two hours.

If these constraints are not respected, the execution of the program stops.

Note that GOSim and GOSemSim do not have command line interfaces. We therefore developed scripts which can be used to compute all the similarities for the pairs of entries (terms or gene products) contained in a file. The scripts are provided in the source code associated to the project.

8.3.2.2 Semantic similarity between gene products

This test aims to compare tools for the computation of semantic similarities between pairs of gene products annotated by GO terms. The protocol is similar to the one used for the comparison based on the computation of GO term semantic similarity. However, in this case, the comparison of a pair of gene products (groups of concepts) was made using a indirect groupwise measure, i.e., comparing to group of concepts U and V, the maximal similarity of the pairs of concepts in $U \times V$ was computed (using the measure used for the previous test, i.e. Lin's proposal). Four tests were designed. Each test is composed of a set of pairs of gene products for which we want the semantic similarity to be computed. Similarly to the other tests, four sizes were considered: 10k, 100k, 1M and 100M pairs of gene products.

The sets of pairs of gene products were generated using the open source tool used to generate the aforementioned tests. Note that no restriction is applied on the Evidence Code associated to the annotations linked to the considered gene products (e.g., inferred electronically annotations [IEA] have been considered). In addition, only annotations related to gene products' Biological Process (BP) were used during this test.

In this test the constraints considered are: Memory consumption – 6Go of memory, and time constraint – four hours.

8.3.3 Results

First we discuss the correlation between the semantic similarity results provided by software solutions evaluated. We then present the computational performance obtained for the two tests.

8.3.3.1 Result correlations and associated discussion

We evaluated the Pearson correlations between the results obtained by the tested tools. The correlations were computed taking GO term to GO term 10K $r_{0.0}$ sample into consideration. Remember that the semantic similarities were computed using Lin's measure. The Pearson correlations between the results produced by the tools are presented in Table 8.2. The details can be found in the project webpage[1].

	FastSemSim	**FSS ISA**	**SML**	**GOSim**	**GOSemSim**
FastSemSim	1	0.68	0.69	0.85	0.86
FSS ISA		1	0.99	0.58	0.58
SML			1	0.57	0.58
GOSim				1	0.99
GOSemSim					1

TABLE 8.2: Correlation of pairwise similarity results obtained using various tools

FSS ISA corresponds to the results obtained using a special build of the FastSemSim library only considering `subClassOf`[2] relationships, version 0.7.1.1. This version is not an official release supported by Marco Mina, the developer of FastSemSim. This build was made in order to change undesired behaviour relative to the way version 0.7.1 compute parents/ancestors. Indeed, version 7.1 considers all types of relationships as `subClassOf` relationships when parents are computed. This behaviour changes the common ancestors or the MICA of the two terms which will be considered by the measures.

Both GOSIM and GOSemSim rely on GO.db R package[3]. They also consider concepts which *subsumes* a concept x, not only according to `subClassOf` relationships, as ancestors of x. See GO.db documentation[4], and more particularly details of the function GOBPPARENTS on which the tools rely: "*Each GO BP term is mapped to a named vector of GO BP terms. The name associated with the parent term will be either* `is-a`*,* `has-a` *or* `part-of`*, where* `is-a` *indicates that the child term is a more specific version of the parent, and* `has-a` *and* `part-of` *indicate that the child term is a part of the parent. For example, a telomere is part of a chromosome.*". We therefore suspect that GOSim and GOSemSim do not differentiate the type of relationships when the common ancestors are computed.

[1] `https://github.com/sharispe/sm-tools-evaluation`

[2] Note that in the OBO format specification, the taxonomic relationship (denoted `subClassOf` in this manuscript) is denoted `is-a`.

[3] http://www.bioconductor.org/packages/2.12/data/annotation/html/GO.db.html

[4] `http://www.bioconductor.org/packages/2.13/data/annotation/manuals/GO.db/man/GO.db.pdf`

We observe that GOSIM and GOSemSim have a maximal Pearson correlation (0.99). This was expected since both tools rely on GO.db package. They also both have a strong correlation with FastSemSim (0.85). The differences between GOSIM/GOSemSim and FastSemSim can be potentially explained by the way the tools compute the information content.

The SML however produces scores which are faintly correlated to FastSemSim, GOSIM and GOSemSim. We investigated the results to understand the causes of the differences. We found that FastSemSim, GOSIM and GOSemSim perform treatments which are not in accordance with the original definition of Information Content based measures. Indeed, IC-based measures clearly rely on the taxonomic graph in order to be computed. The taxonomic graph is the subgraph of the ontology which only contains taxonomic relationships[1]. This graph is considered to compute the ancestors of a term and is therefore important to compute the MICA (or NCCAs) in information content based measures. FastSemSim, GOSIM and GOSemSim consider relationships other than taxonomic ones to compute the ancestors, which explains the variation obtained. They also consider `part-of` relationships to define ancestors (`regulates` is even used in the tested version of FastSemSim).

To ensure that the poor correlations were down to this difference, we built a modified version of the FastSemSim library (available at project webpage). This version can be used to compute the similarities using FastSemSim source code and only considering taxonomic relationships when ancestors are computed. Considering this modification we obtained the expected correlation between FastSemSim and the SML (0.99 – Table 8.2). Therefore, the results produced by the SML appeared to be in accordance with the original definition of the evaluated measure.

Nevertheless, an important (and worrying) finding highlighted in this experiment is that high variations can be observed between the results produced by available software solutions. These variations appear to stem from the differences between the various interpretations and implementations of measures proposed by tested libraries.

8.3.3.2 Evaluation of computational performances

The tests were performed on a personal computer with an Intel(R) Core(TM) i5 CPU M 560 @ 2.67GHz with 6Go allocated to the tools. The obtained results are presented in Tables 8.3 and 8.4. They correspond to the average computational times obtained for each sample associated to each evaluated set (e.g., 1M) – in each case the variation between the samples was low; they are therefore not presented in the results. Complete results can be consulted from the dedicated repository. Also remember that they can be reproduced following the instructions detailed in the documentation.

[1] I.e. `is-a` relationship in this case.

	1K	10K	1M	100M
FastSemSim	0m12.3	0m12.83	0m31.68	X
GoSim	0m49.46	3m21.5	X	X
GoSemSim	1m34.69	16m21.34	X	X
SML	0m9.23	0m9.76	0m19.55	16m30.24
SML parallel	0m9.22	0m9.56	0m14.47	8m58.29

TABLE 8.3: Running times of tools dedicated to the computation of GO terms semantic similarity. Four tests were performed considering random samples of pairs of GO terms with fixed sizes (see columns, K= 10^3, M= 10^6). SML parallel corresponds to the SML configured with four threads. 'X' specifies that the process required more than 6Go of RAM or took more than 2 hours

	1K	10K	1M	100M
FastSemSim	0m13.36	0m16.79	7m8.14	X
GoSim	X	X	X	X
GoSemSim	27m02.66	X	X	X
SML	0m10.01	0m11.18	1m38.87	133m27.44
SML parallel	0m9.80	0m10.24	0m47.62	58m

TABLE 8.4: Running times of tools dedicated to the computation of gene products semantic similarity. Four tests were performed considering random samples of gene pairs with fixed sizes (see columns, K= 10^3, M= 10^6). SML parallel corresponds to the SML configured with four threads. 'X' specifies that the process required more than 6Go of RAM or took more than 4 hours

8.3.4 Discussion

The results of the two tests presented in Table 8.3 and Table 8.4 stress that the SML is perfectly adapted for the large-scale computation of semantic measures. Indeed, although the SML is generic and not tailored to a specific ontology and usage (contrary to the other solutions), it outperformed domain-specific solutions in all the evaluated cases.

The *poor* performance obtained using GOSim and GOSemSim can be explained by the fact that these libraries manipulate persistent data (*via* Bioconductor), which requires more computation time[1]. This is indeed extremely limiting for the fast computation of semantic measure scores, e.g., in our evaluation, GoSemSim took more than 27 minutes to compute the semantic similarities between 1000 pairs of gene products (Table 8.4). Nevertheless, this also means that these libraries

[1]Contrary to other tools which work *in-memory*.

can theoretically handle larger datasets than the current implementation of FastSemSim and SML. However, in most use cases involving large datasets, the performance obtained by GOSim and GOSemSim cannot be admitted and severely hampers their use.

Another insight provided by these experiments is that the SML gives an important reduction of computation time for extensive computation – any bigger than 1 million computations. As an example the SML took 1 minute and 39 seconds to compare 1 million pairs of genes despite FastSemSim taking 7 minutes 8 seconds (Table 8.4). More importantly, the SML enables computations which were not possible using other software solutions, e.g., 100 million computation.

Therefore, simply put, despite its generic layer, the SML appears to be an efficient and reliable software library for the computation of semantic measures.

8.4 The Semantic Measures Library in action

The SML and the toolkit are not limited to a specific ontology and can therefore be used in a broad field of application, (scientific) projects and software solutions. We present some of the applications which have been made, focusing on those which are tightly linked to the contributions presented in this manuscript:

- Analysis of semantic measures.
- Large-scale computation of semantic similarities.
- Application to a content-based recommendation system.
- Integration to a conceptual information retrieval system.

8.4.1 Analysis of semantic measures

Throughout this thesis, the SML has been used to analyse semantic measures. The analyses performed were related to the comparison of semantic measures for specific usage contexts. The SML was also extensively used in the contribution related to the practical application of the abstract framework which was presented in Chapter 5. In both studies, the generic aspect of the library, the large number of (parametric) measures implemented and its performance were required. More broadly, throughout this thesis, the generic aspect of the library gives us the opportunity to study and use semantic measures with several ontologies: the Gene Ontology, the Disease Ontology, the MeSH, SNOMED-CT, Wordnet, the DBpedia Ontology and Yago, to mention a few.

8.4.2 Large-scale computation of semantic similarity

The SML has often been used for large-scale computation of semantic measure scores. This is particularly true in the collaboration initiated with Clément Jonquet (LIRMM Montpellier) for

the SIFR project (Semantic Indexing of French Biomedical Data Resources)[1]. We have been solicited in order to take advantage of the capabilities offered by the SML to face large-scale computation of semantic measure scores. One of the aims of this collaboration is to compute semantic measure scores using the collection of ontologies provided by BioPortal – a portal dedicated to ontologies related to the biomedical domain which contains more than one hundred ontologies[2]. The main aim is to give access *via* a web service to semantic similarities of pairs of concepts computed using several semantic measures. The astronomical number of computation required, i.e., hundreds of billions of concept-to-concept semantic similarity computation, challenged the SML but the computational part of the objective was reached. This project has also been supported by two master's students. The students used the library to compute the scores of similarity and therefore integrate the results onto SIFR platform.

More information about this project can be found at: `http://www.lirmm.fr/sifr/positions/2014_TER_M1_Jonquet_semsim_web_service.html`.

8.4.3 Application to the design of a content-based music recommendation system

The library has also been used in the development of the projection-based approach which has been proposed for the comparison of instances characterised through a semantic graph – please refer to the contribution presented in Chapter 6. This framework has been used to develop a simple recommender system which was implemented using the SML. Despite the fact that numerous utility functions provided by the SML have been used in this project, the SML has mainly been used to compute semantic measures involving the comparison of (groups of) concepts.

A prototype of the recommender system applied to music band recommendation is available at: `http://www.lgi2p.ema.fr/kid/tools/bandrec`.

8.4.4 Use in the design of an conceptual information retrieval system

During this thesis, the SML was also integrated into OBIRS – Ontology-Based Information Retrieval System (version 2) [Sy et al., 2012]. Given a set of concepts as a query and a collection of instances semantically characterised by groups of concepts (e.g., documents annotated by MeSH descriptors, genes annotated by GO terms), OBIRS returns the more relevant instances w.r.t the query. In this context, the capabilities and the performance offered by the SML were essential to ensure the performance of the information retrieval system. The development of this new version of OBIRS was performed by an engineer with the assistance of the SML support team for the technical aspects related to semantic measure computation. An instance of the new version of OBIRS, which is based on the SML, and which enables PubMed documents

[1] `http://www.lirmm.fr/sifr`
[2] `http://bioportal.bioontology.org`

related to Cancer to be queried w.r.t their associated MeSH descriptors, is available at: `http://obirs.itcancer.mines-ales.fr`.

8.5 Chapter conclusion

This chapter presented the Semantic Measures Library (SML) and associated toolkit, which were developed to respond to the need of efficient, extensive and robust open source software tools dedicated to semantic measures.

The first major benefit of these two tools lies in their generic aspect which enable them to be used to compute semantic measures over a large diversity of ontologies and in a wide array of applications. In addition, by conducting reproducible evaluations using a rigorous test protocol, we demonstrate that the generic aspect of these tools does not hamper their computational performance as both the SML and the toolkit have proved to outperformed domain-specific tools in several use cases. This aspect is essential for the adoption of semantic measures and their use in practical applications; this has for instance been shown through the several applications of the SML which have already been made (e.g., recommendation and information retrieval systems). Finally, these tools give access to a large collection of measures and related metrics/algorithms and thus offer a development platform of choice for the comparison and selection of semantic measures. Therefore, these contributions, targeting both users and designers of measures, open interesting perspectives for the large adoption of semantic measures, as well as their large-scale computation and analysis.

The main challenge for the SML is to federate semantic measure developers and users (i) by providing extensive updated documentation, (ii) by ensuring constant development/improvement of the tools, and (iii) by stimulating community support – something we already initiated through the website and the mailing list. We also think that the evaluation of the several tools we made, stresses the importance of performing more extensive comparisons of existing solutions. Thus, we are convinced that discussions and collaborations have to be initiated between developers of tools related to semantic measures in order to define a standardised and recognised evaluation campaign. This is necessary to reduce the differences we observed between the results provided by the different tools as much as possible. This could also be the occasion to formalise ontology handling to limit the different strategies today adopted by the different implementations. Among others, another important challenge is to test and design internal persistent and/or distributed data models of ontologies. They are prerequisites for developing tools which can be used to compute semantic measures over very large ontologies composed of hundreds of millions of triplets (e.g., complete DBpedia or Freebase ontologies).

General Conclusion

This section concludes our work. We summarise the contributions which have been presented in this manuscript by emphasising their added-value, as well as their limitations. This will help us in particular to define several perspectives of this work.

[A] A broad overview of the landscape of semantic measures, and an in-depth analysis of knowledge-based semantic measures. Chapters 2 and 3 are dedicated to the introduction of the notion of semantic measure. To this end, we have presented a digested version of the extensive, interdisciplinary and sometimes disrupted literature related to the field. Several definitions and properties related to semantic measures have been introduced; they can be used for better characterising semantic measures and more specifically their semantics. As an example, we used some of these definition and properties to bring to light a classification of the different types of semantic measures which have been proposed in the literature so far. Next, we focused more particularly on knowledge-based semantic measures, and to be more precise, we focused on those which rely on the analysis of network-based ontologies. Many technical details have been introduced for this type of measures, and a large collection of measures have been identified, classified and analysed. Although, this work is only partial, and does not cover important topics such as the selection of semantic measures in detail, we are convinced that we give practitioners and designers of semantic measures access to a better understanding of the field as a whole.

An important aspect of this work has been to federate efforts made by several distinct communities. Focusing on semantic measures, our desire has been to emphasise that domain-specific contributions generally have an interdisciplinary scope as they can benefit other communities facing different problems. This is, for instance, the case of cognitive models which were initially proposed by cognitive scientists to study human appreciation of similarity. As we have seen, they are now used by designers of semantic measures to define semantic similarity models. By extension, they will therefore be used to define approaches in the aim of analysing a broad variety of entities, e.g., conceptually annotated genes or diseases.

The detailed vision of the field provided by our preliminary analysis helped us to derive several perspectives and goals related to the study of semantic measures. Therefore, Section 3.7 was dedicated to underlining some of the teachings of this work. As a result, six goals have been identified in particular:

1. *To better characterise semantic measures and their semantics.* It is clear that measures have to be analysed through specific mathematical properties, and that the semantics associated to their scores must be clearly understood, i.e., end-users must understand the implications associated to a score of semantic similarity/relatedness. As we have seen, this impacts the benefits of using a specific approach w.r.t a particular usage context.

2. *To provide tools for the study of semantic measures.* Both theoretical and practical tools must be proposed in order to gain new insight of semantic measures. This is necessary in order to (formally) characterise the large diversity of measures defined in the literature. To this end, we underlined that theoretical frameworks must be used in order to both classify measures and highlight relationships between existing proposals. We have also stressed that more benchmarks and evaluation protocols used for analysing and comparing semantic measures have to be developed. In addition, we highlighted the fact that the development of software solutions which enable a larger adoption and analysis of measures must be encouraged. In this regard, we underline the potential limitations of restricted and domain-specific implementations, to further highlight the benefits of developing generic open source solutions dedicated to semantic measures.

3. *To standardise ontology handling.* We also discuss some of the practical limitations induced by the lack of standardisation in the way ontologies are processed prior to semantic measure computation. We stress in particular the fact that a too large degree of freedom is given to developers of measures. It often creates a gap between theoretical definitions and practical implementations of measures. This has been exemplified by the comparison of the results produced by the SML to those produced by other tools. Correlations have proven to be particularly low, despite the deterministic nature of the measure in use. Excluding implementation errors, this worrying result is due to the fact that specific treatments are not clearly standardised and defined, which forces developers to select particular strategies, e.g., in the way ontologies are reduced prior to being used as semantic graphs: are the taxonomic redundancies removed (even if they do not directly impact the coherence of the measure)?; are redundant annotations considered for the computation of extrinsic information content?; If you define that `propB subPropertyOf propA`, how many relationships do you consider in the graph defined by the triplet `X propB Y`? There are numerous examples of interpretation in the way ontologies have to be handled. Nevertheless, they have to be clearly defined and if possible standardised, to ensure that scores of semantic measures are not dependant on the implementation used for their computation.

4. *To promote interdisciplinary studies.* As we have seen, numerous communities are involved in the study of semantic measures. We underline that narrower bridges must be created between them. The development of interdisciplinary theoretical and software tools is a step in this direction. In addition, we also mention some of the communities that are currently not involved in the study of semantic measures but whom it could be of interest to work with in the future.

5. *To study algorithmic complexity of measures.* We underline that only few studies focus on the algorithmic complexity of semantic measures. This is a clear limitation since it clearly impacts the practical use of semantic measures. It is therefore of major importance in order to help end-users of semantic measures which feel the need to select a semantic measures.

6. *To support context-specific selection of semantic measures.* It is, at present, difficult to select a semantic measure w.r.t a specific usage context. Thus, most users select measures

according to their popularity and availability in tools, generally without considering the properties which characterise the measure and the semantics associated to the scores it produces. Therefore, more studies are needed on this topic, in particular to better define what a context of use is and which aspects of measures are important w.r.t them. Moreover, empirical comparative analyses of a representative sample of the diversity of available measures have to be performed in different domains. These analyses are essential to identify the potential existence of classes of measures which tend to perform better than others, and if these results can be generalised.

In the scope of Goal 2, we chose to focus our efforts in this thesis on the development of theoretical and practical tools for the analysis of knowledge-based semantic measures. Our choice was incited by the growing adoption of ontologies and associated semantic measures, and by the large impact and perspectives which can potentially arise from the resolution of this challenge. Indeed, focusing on knowledge-based semantic measures, theoretical and practical tools are central to their analysis. They can have a clear impact on each of the goals outlined above. In this context the two main contributions of our work are the following.

[B] A unifying framework for knowledge-based semantic similarity measures. As we demonstrated in Chapter 4, most of these measures can be broken down through parametric functions which rely on a limited set of abstract elements. Thus, we highlight the fact that most measures are only specific expressions of generic abstract measures. This finding, which stems from the detailed analysis of prior works on relationships between measures, provides a new insight into the diversity of measure proposals. It opens interesting perspectives for characterising central elements in assessing semantic similarity and more generally for designing semantic measures.

Some examples of the practical usage of this framework were presented in Chapter 4 and 5. We showed, in particular, how it can be used to study the accuracy of measures, to support context-specific design of measures through parametric optimisations, and more generally, to identify potential rooms for improvement of these measures. We also proposed a new angle of analysis for semantic measures through the study of their robustness, i.e., their degree of resilience w.r.t the uncertainty which intrinsically hampers benchmarks used for evaluating their accuracy. Through these studies, we underlined how the proposed framework appears to be particularly adapted for fine-grained analyses of semantic measures. We then drew special attention to the fact that the framework can be used to analyse properties of measures and to classify them according to these properties. This was used in Chapter 7 to characterise some properties of MSCA-based measures, i.e., measures which compare pairs of concepts by mainly exploiting their Most Specific Common Ancestor. Finally, based on these properties, we proposed an optimised algorithm to compute the semantic similarity of all pairs of concepts defined in a taxonomy.

Despite the fact that the proposed framework has proven to be particularly useful for analysing semantic measures, as underlined by the multiple examples provided in this manuscript, some potential hesitations deserve to be discussed. The main limitation is surely due to the main strength of the framework: its degree of abstraction. Most designers of semantic measures are governed by practical applications in specific usage contexts, which explains the large diversity of

semantic measures defined by numerous communities. Therefore, it is clear that it will be difficult to federate future contributions related to the field through a common formalism introducing a layer of abstraction. Indeed, such an abstraction can limit the expression of domain-specific measures in a way that can be understood by members of communities who are most often non-experts in the design of semantic measures. Nevertheless, we are convinced that the approach adopted in this thesis proposes a solution to regulate and better understand the incessant flow of new measures published in the communities directly or indirectly related to semantic measures. This is particularly important given that a growing number of specialised communities are adopting semantic measures to support data analysis or algorithm designs.

Therefore, to ensure that the efforts made in this thesis are not futile, we strongly believe that further studies of the theoretical framework must be made, in particular to further exemplify its added value for our communities. This can be made by doing extensive studies of measures in particular usage contexts, e.g. by analysing the impact of selecting specific measures core elements on the performance of measures. In addition, more efforts have to be made to encompass measures which are now difficult to study through the insight provided by the framework, e.g., some graph-based measures which rely on random-walk approaches. To this end, adaptations and extensions of the framework may be required. Nevertheless, as underlined by two recent publications related to the unification of knowledge-based semantic similarity measures [Cross et al., 2013; Mazandu and Mulder, 2013] (published independently and after the design of our proposal), we are convinced that such an initiative for unifying semantic measures had to be initiated. Thus, with sincere humility, we are pleased to lay one of the stones composing the base which will support this enterprise.

[C] The Semantic Measures Library: fast, open-source and generic software solutions dedicated to semantic measures. Throughout this manuscript we have stressed the importance of empirical evaluations for assessing the accuracy of semantic measures. We have underlined that most software solutions dedicated to semantic measures were dedicated to domain- specific ontologies. Thus, despite some initiatives a few years ago proposing software solutions which were independent of a specific usage context (i.e., SimPack), no extensive software solutions dedicated to large-scale computation of knowledge-based semantic measures were available at the beginning of our study. Therefore, as highlighted in Chapter 8, we invested a lot of effort in studying the feasibility of developing such a solution, and subsequently designing, developing, promoting, supporting and maintaining the Semantic Measure Library (SML).

The SML provides fast and robust open software tools for computing and analysing knowledge-based semantic measures. It is compatible with numerous ontologies, and with standardised ontology languages (RDF[S], OWL). By providing a source code library which implements numerous semantic measures and associated algorithms, it can be used for designing and studying semantic measures in a large variety of usage contexts. Its suitability for these tasks has been shown through multiple experiments presented in this manuscript. We also mentioned that the SML has already been used in several projects, e.g., OBIRS, an Ontology-based Information Retrieval System, the semantic-based recommender system presented in Chapter 6, or even for

the large-scale computation of semantic similarities between concepts related to BioPortal ontologies. This shows the broad-spectrum of applications of the SML and that it can already be used in demanding applications. In addition, and this is an important aspect, efforts have been made to give non-developers access to some of the functionalities provided by the library, in particular to compute scores of semantic measures.

Though it is still early days for the SML, the increasingly high numbers of solicitations it has generated over the last month underlines the need for such a contribution. Similarly to the unifying framework, efforts have to be made to improve user experience, as well as the functionalities and computational performance of the SML. To this end, we proposed to decline domain-specific tools in order to facilitate usage of the command-line interface. Documentation must also be improved to clearly explain the capabilities and limits of these tools; this is a continuous work.

In Chapter 2, we showed also how, associated to the proposed theoretical framework, the library can be used for the detailed analysis of semantic measures, and to better understand the importance of each component of semantic measures w.r.t a specific usage context. Thus, once again, we are pleased to provide tools for studying and analysing semantic measures: the main aim of this thesis. Such tools are prerequisite to better understanding the landscape of measures proposed in the literature, and to tackling the complex subject of semantic measure selection. The important challenge now is to federate users and developers of the library in order to ensure a long life for both the library and associated software solutions. Initiatives are being taken to this end, an example being an introductive session to the SML in a national workshop which has already been planned to ensure that this goal will be reached[1] – similar events will be proposed in international conferences.

[D] Algorithmic contributions related to semantic measures. Concurrently to the main contributions which have been summarised so far, we also studied other problems related to semantic measures.

In Chapter 6, we presented a new approach for characterising and comparing instances defined in a semantic graph, e.g., an RDF graph. This work extends existing proposals by defining the notion of projection of an instance into a semantic graph. Based on this notion, we then proposed a new canonical form which can be used to better characterise instances, in particular by taking into account some of their properties which are not expressed in the ontology *per se* (remember the example of the body mass index). In addition, we showed how this approach can be used for comparing instances by explaining the meaning of results, i.e., by ensuring that the semantics of the scores of relatedness will be traceable. Finally, using a prototype of a music band recommender, we underlined the practical feasibility of the approach, and we showed how it can be used for designing a semi-supervised recommender system which takes advantage of Open Linked Data. More work has to be done in this field, in particular to better characterise the performance of this approach w.r.t related works and other datasets. We also plan to study how machine learning techniques could be used to learn which relevant projections and weights should be considered in specific usage contexts. This opens the door to personalised information

[1] 25èmes Journées Francophones d'Ingénierie des Connaissances (IC 2014), Monday 12 May.

retrieval and knowledge discovery based on ontology analysis, two fields of study in which our team plans to invest time and energy.

Complementary to the contributions presented so far, we also studied some algorithmic aspects of semantic measures, three of which have been mentioned in this manuscript (Chapter 7). We therefore defined algorithms to: (i) compute the semantic similarity of all pairs of concepts defined in a taxonomy by using a specific type of semantic measures, and (ii) extend information theoretical measures for comparing concepts defined in different ontologies, without being restricted to the mappings defined between ontologies.

All PhD theses come to an end. This one will conclude with a quote from Isaac Asimov:

"*The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' but 'That's funny...*"

Between a perpetual fight against an overwhelming literature, and difficulties in convincing others of the relevance of measure unification, this thesis wasn't funny every day, but we did have our share, thanks to the devoted team who contributed to this work. I hope our contributions will serve our communities and I wish the readers a lot of *fun* in their career.

Appendices

A

From ontologies to semantic graphs

A.1 Ontologies: a brief introduction

This section briefly introduces the reader to the field of knowledge representation to define the notion of ontology considered in this manuscript. We use the general term ontology to refer to any computational artefact used to express knowledge in a machine understandable form [Davis et al., 1993]. Indeed, as stressed in [Guarino et al., 2009], "*For AI Systems, what 'exists' is that which can be represented*". It is therefore commonly stressed that ontologies should be considered as surrogates, enabling things to be manipulated by computers, and, by extension, give the opportunity to study a domain without acting on its constitutive elements. Ontologies express how a domain must be understood and what types of logical reasoning can be applied to it. This is done by defining (i) its key elements, (ii) the formal ontological commitments on which it relies, and (iii) the interpretations which can be made on it. The different goals which can motivate the development of ontologies are well summarised in the literature, e.g., [Noy et al., 2001; Uschold and Gruninger, 1996]:

- *To describe non-ambiguous information and knowledge which can be understood and reused among people or software agents.* Non-ambiguous characterisation of things is central for human and human-machine communication and therefore interaction. This is also essential for existing knowledge to be reused and for systems using ontologies to be interlinked and aggregated. Simple examples are classifications; for instance, the International Classification of Diseases[1] (ICD) encodes diseases and symptoms which can be used to track diagnostics in a formal way.

- *To make domain assumptions explicit and to separate domain knowledge from the operational knowledge.* To be able to provide explicit expressions of the assumptions governing a domain is central to ensure that the notions which are manipulated represent a consensus among domain experts. Indeed, non-explicit expressions of domain assumptions, such as source code, highly reduce the amount of people who will be able to understand the representation of the domain.

[1] `http://www.who.int/classifications/icd`

- *To analyse and automatically take advantage of domain knowledge.* The explicit and non-ambiguous character of ontologies enables domain-knowledge to be studied, shared and better characterised. It also enables the emphasis on central elements of the domain or even incoherences regarding its current understanding. Therefore, probably the most important aspects of ontologies is that they enable computers to process an expression of our knowledge, automatically check its consistency and reason on it. As we have said, ontologies are surrogates which enable domains to be manipulated by computers by interacting in an abstract manner with their main constitutive elements and the rules which define their interactions.

Note that, despite the fact that ontologies are also essential for people to share knowledge, we will mainly consider ontologies as a way to convey knowledge to machines. For more information related to the large field of study of knowledge representation, the reader can refer to some of the seminal contributions on which this brief introduction is based, e.g., [Baader et al., 2010; Borst, 1997; Davis et al., 1993; Gruber, 1993; Guarino et al., 2009; Hitzler et al., 2011; Minsky, 1975; Noy et al., 2001; Robinson and Bauer, 2011; Sowa, 1984; Studer et al., 1998]

This section is structured as follows. (1) We first clarify the notions of data, information and knowledge by presenting their common definitions. (2) We informally discuss the process of defining forms of knowledge which can be understood by computers and the implications for computer science. (3) From simple taxonomies to expressive logic-based ontologies, several types of ontologies are briefly introduced. (4) We discuss the notion of semantics which will be considered throughout this manuscript. (5) A technical section briefly introducing the reader to the languages and specifications used to express ontologies is also proposed, and finally, (6) we introduce reasoning techniques which can be made on ontologies.

A.1.1 From data to knowledge... and beyond

We will often refer to the notions of data, information and knowledge. They have been extensively discussed in the literature and alternative definitions have been proposed. They are generally structured in a bottom-up fashion; data can be processed to obtain information, which can be further analysed to derive knowledge, which in turn leads to wisdom. Figure A.1 presents the relationships between the various notions proposed in [Bellinger et al., 2004]. The figure represents the different levels of understanding required to derive knowledge and wisdom from data.

Data: "*The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media*" [Oxford Dict., 2012].

Data is often considered as *raw*, signs, stimuli or signals [Bellinger et al., 2004]. It is also commonly admitted that "*data is* [...] *discrete, atomistic, tiny packets that have no inherent structure or necessary relationship between them*" [Hey, 2004]. They correspond to elementary facts which can be captured by a device, stored and shared for reuse and analysis. Moreover,

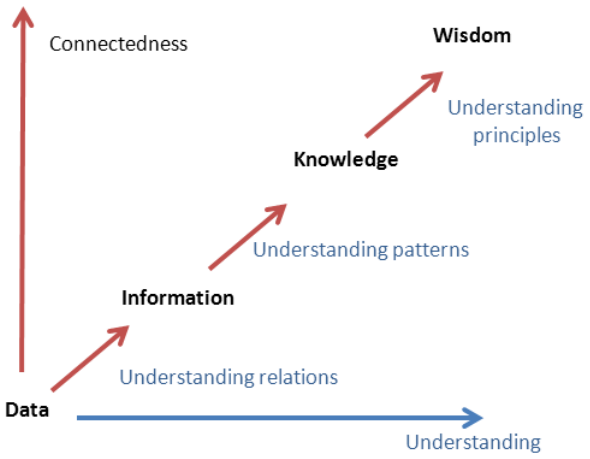

FIGURE A.1: Representation of the relationships between Data, Information, Knowledge and Wisdom. Reproduction from [Bellinger et al., 2004].

data are generally considered to be useless without the context in which they have been obtained. An example of data could be:

```
...0,1,0,0,1;1,0,1,0,0;2,0,0,1,0...
```

Information: "*What is conveyed or represented by a particular arrangement or sequence of things*" [Oxford Dict., 2012].

By definition, information is something which informs; it is generally defined as "*data that has been given meaning by way of relational connection*" [Bellinger et al., 2004]. Information is therefore obtained by giving meaning to aggregation of data processed in a given context. The raw data previously presented can, for instance, be processed to obtain the following information:

user id	Drug A	Drug B	Placebo	Cured
0	yes	no	no	yes
1	no	yes	no	no
2	no	no	yes	no

Knowledge:"*Facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject*" [Oxford Dict., 2012].

Knowledge emerges when patterns are understood from information. Knowledge is therefore any understanding which has been gained by means of study and analysis of experiment outcomes represented by information. It generally refers to conclusions which lead to a proper understanding of problems and domains of study. As an example, the previous information can be analysed to extract a piece of knowledge, for instance, the fact that:

```
Drug A seems to cure the disease.
```

Wisdom: "*The quality of having experience, knowledge, and good judgement; the quality of being wise*" [Oxford Dict., 2012].

Wisdom can be associated to the understanding of the principles explaining knowledge and the use of judgement to discern relevant pieces of knowledge.

```
Drug A seems to cure the disease by altering vital organs;
                it must therefore not be used.
```

Data and information are easy to store on computer. Numerous mathematical techniques and theories have been developed to extract knowledge from them, e.g., Information Theory, Data Mining techniques. Knowledge and wisdom are abstract notions and are therefore more complex to manipulate through computers. Knowledge is assimilated to facts derived from experiences and can therefore be conveyed through language. This implies that formal languages which are sufficiently expressive can be used to express knowledge in a machine understandable form. Conversely, wisdom refers to the existence of conscience and requires forms of judgement, notions with which computers are currently unequipped. In this manuscript, we will mainly manipulate the notions of information and knowledge.

A.1.2 Communicating knowledge to computers

The challenging problem tackled by the field of knowledge representation, i.e., how to formally express knowledge, has received a lot of attention in AI given that numerous processes require the modelling of complex domains in order to be performed, e.g., medical diagnosis. Such an enthusiasm for knowledge modelling is therefore naturally explained by the large perspectives opened by formal expressions of knowledge in computer science, i.e., to give computers and algorithms access to our knowledge.

Language is an essential ingredient to communication; it enables the transmission of messages which carry information in order to reach a specific goal, e.g., to explain, to convince, to give orders. Nevertheless, not all forms of language have the interesting property of being formal and unambiguous, i.e., to ensure that messages are conveyed without being distorted during the communication process. As an example, the complex human language is subject to subjective interpretation, which explains that communication between humans are sometimes challenging. It is, for instance, common to think that an agreement has been reached only to subsequently realise that the result doesn't conform to your original expectations.

Most human-machine communication protocols are (obviously) not based on ambiguous languages which are subject to potentially different interpretations. As an example, satellite behaviour is defined by on-board computers which control and monitor their speed and altitude in order to achieve a predefined mission, e.g., to point to a specific position in space. Therefore, excluding material or software problems, there is no chance that the instructions communicated to satellites will lead to unpredicted behaviour. Indeed, when software developers write source code in a specific programming language, the instructions executed by machines are non-ambiguous;

the machines will therefore execute them according to their exact definitions, which explains that heretic machine behaviour can only be a direct result of human or material errors.

A cornerstone of human-machine interaction is therefore the ability to communicate in a non-ambiguous fashion. As an example, to express that *Tigers are Animals*, we need a vocabulary to disambiguate what we understand by the strings of characters *Tigers*, *Animals* and *are*. We also need to clearly define the implications associated to this specific association/ordering of the three words. Otherwise stated, we need to define the meaning, i.e., the semantics, of the terms, as well as the meaning of the given relationship established between them. As we will see in the following overview of ontologies, formal expressive ontologies require the definition of complete logic: vocabulary, syntax, semantics, as well as the interpretation of the syntax (the rules of inferences).

A.1.3 An overview of the diversity of ontologies

From simple controlled vocabularies and taxonomies to complex ontologies based on description logics, a large range of ontologies of increasing expressiveness and complexity have been introduced in the literature. See Figure A.2 for a graphical representation. Two broad and non-distinct categories are generally distinguished depending on the level of formalism used to model the knowledge: *network-based* (or *structure-based*) and *logic-based* ontologies, e.g., [Baader et al., 2010]. These ontologies can also be distinguished w.r.t their degree of expressivity, interoperability and standardisation [Studer et al., 1998].

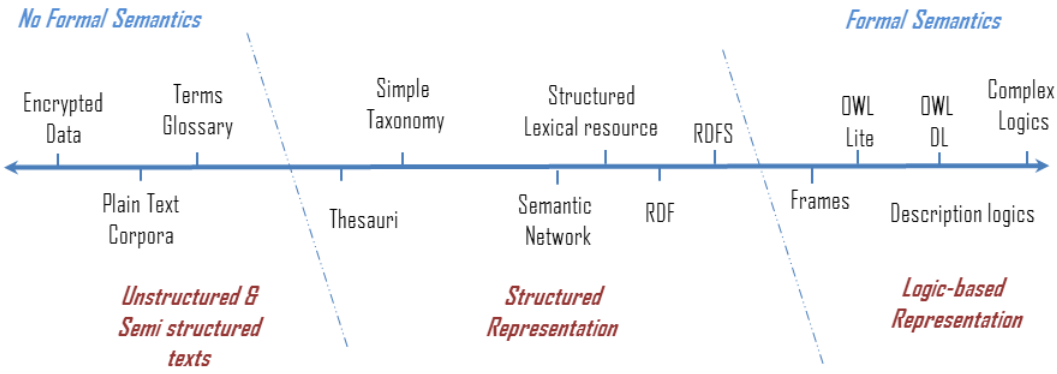

FIGURE A.2: Overview of the diversity of ontologies from non-formal network-based (i.e., structure-based) ontologies to logic-based ontologies – adapted from [Jimeno-Yepes et al., 2009].

A.1.3.1 Network-based ontologies

Network-based ontologies do not rely on logic-based formalisms and are commonly used in natural language processing and computational linguistics. In their simplest forms, they are generally used to characterise domain knowledge through *semantic networks*: graphs composed of nodes and oriented edges. Nodes refer to terms, concepts or instances, and edges, which are associated to a specific label, define relationships between pairs of nodes.

Among the first contributions related to network-based ontologies, we can cite the work of [Collins and Quillian, 1969] in which semantic networks are built by studying retrieval time from semantic memory. The relationships between elements were defined as a function of the response time people took to correctly answer questions involving two elements, e.g., *Is a Canary a Bird? – Is a Canary an Animal?*. Approaches used to define such ontologies are generally derived from cognition; they often rely on non-formal textual descriptions and simply correspond to structured and controlled vocabularies, e.g., thesaurus, non-formal taxonomies. In these ontologies, terms with similar meaning or groups of similar objects are characterised by a unique preferred name; they are next structured through linguistic relationships without formally defining the interpretations associated to a specific relationship, i.e., the semantics of the relationship is implicitly defined by its name or a textual description.

As an example, WordNet models the lexical knowledge of native English speakers in a lexical database [Fellbaum, 2010; Miller, 1998]. It is defined through a semantic network composed of sets of synonyms (called *synsets*) which are linked by semantic and lexical relationships, e.g., hyperonymy, hyponymy, meronymy. Synsets are associated to a unique preferred name and the semantics of both synsets and semantic relationships are defined by a short description (i.e., *gloss*). Figure A.3 presents a graphical representation of a simple semantic network similar to Wordnet.

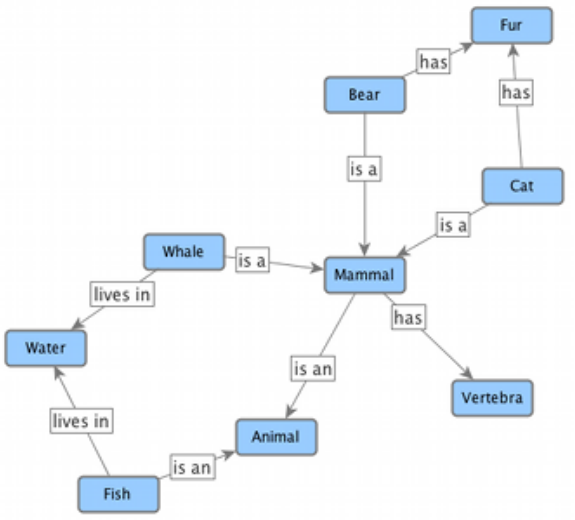

FIGURE A.3: Example of a semantic network.
source: `http://docs.yworks.com/yfilesdotnet/developers-guide/figures/semantic.png`

Another example of a non-formal ontology is the MeSH (Medical Subject Header) [Rogers, 1963][1], a medical thesaurus which provides a structured and controlled vocabulary composed of hierarchies of biological and medical terms. Figure A.4 presents a graphical representation of the MeSH.

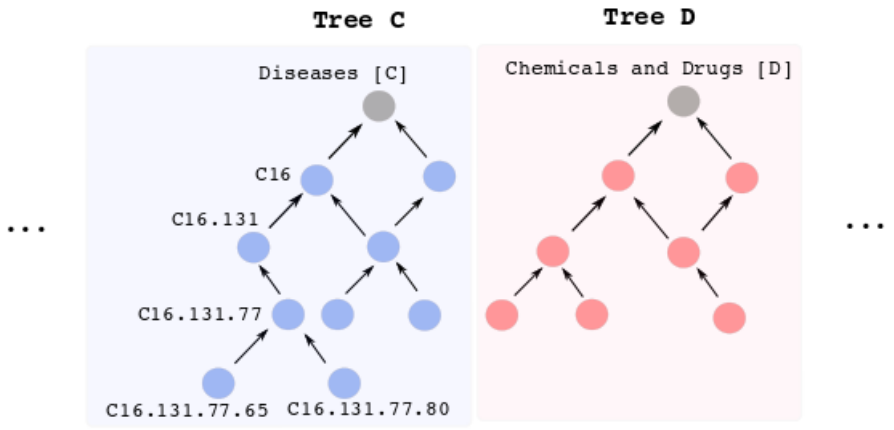

FIGURE A.4: Partial representation of the Medical Subject Header thesaurus (MeSH). The structuring of the vocabulary is given by means of several trees (nodes refer to *concepts*). `C16` refers to the concept "*congenital, hereditary, and neonatal diseases and abnormalities*", `C16.131` to "*abnormalities*", `C16.131.077` to "*abnormalities, multiple*", and `C16.131.077.065` refers to a particular genetic disorder

More complex semantic networks can also be defined by enabling nodes to refer to predicates (types of relationships) or complete statements, e.g. to define properties associated to a specific statement.

Network-based ontologies have encountered a large success and are widely used to model extensive domain knowledge. This can be partially explained by the fact that they provide an intuitive and graphical way to represent and structure knowledge. Nevertheless, network-based ontologies were originally structured by poorly characterised semantic relationships which are not understood, *per se*, by computers. For instance, *for a computer*, taxonomies defined in a non-formal language are only graphs with the specific property of being acyclic. In order to utilise them for inferences, developers have to programmatically define the expected behaviour induced by taxonomic relationships, i.e., considering that the statement `Human subClassOf Mammal` has been specified in a taxonomy, the program will consider that all instances of the class `Human` are also instances of the class `Mammal`. In other words, the semantics of the predicate `subClassOf`, in this case its implications, are hard-coded in a program. Therefore, if a knowledge designer express another taxonomy using `subClass-Of`, `isA` or `aKindOf` as the taxonomic relationship (instead of `subClassOf`), the program will no longer work as expected. In other cases, ambiguity will be explained by the fact that predicates with the same label will not have the same intended semantics across ontologies (e.g., several semantics can be associated to the predicate `partOf`). Therefore, one of the limits of early network-based ontologies is that their semantics were often defined at implementation level. They only defined approaches to express knowledge by means

[1] `http://www.ncbi.nlm.nih.gov/mesh`

of a simple graph structure without defining common vocabularies which can be unambiguously reused in ontologies. Another important aspect is that the semantics of the element defined is informally defined through descriptions.

Several limitations are associated to the ambiguous expression of semantics. The first is obviously the fact that heretic behaviour can be observed using ontologies associated to semantic interpretations which are software dependant. This is not compatible with the desire to define an explicit specification of a domain. The second drawback is the lack of interoperability between this type of ontologies. Indeed, since the elements used to express knowledge are not formally expressed, it is difficult to reuse ontologies without collaborating with knowledge modellers, and without carefully analysing ontologies.

To overcome the limitations of early network-based ontologies relying on weak semantics, languages and vocabularies have been proposed to add formal semantics to graph structures. As an example, the Resource Description Framework (RDF) provides a graph data model and a vocabulary which enable the unambiguous characterisation of resources through graphs. In addition, the semantics of RDF graphs can be enriched using RDF-Schema (RDFS) which provides a vocabulary to define and structure concepts. RDFS also defines the interpretations which can be made in order to reason over the vocabulary. RDF(S) can therefore be used to formally express simple forms of domain specific knowledge. As an example, defining a taxonomy of concepts using RDFS, the semantics of the taxonomic relationship used to order classes, denoted `rdfs:subClassOf`, is formally defined by standardised entailment rules [W3C, 2004]. These rules define how to interpret RDFS vocabulary and therefore standardise expected behaviour at implementation level, e.g., RDFS rule number 11 states that any relationship associated to the predicate `rdfs:subClassOf` is transitive, which means that:

$$\texttt{(Human rdfs:subClassOf Mammal)} \wedge \texttt{(Mammal rdfs:subClassOf Animal)} \Rightarrow \texttt{Human rdfs:subClassOf Animal}$$

In this case, the meaning carried by the relationship `rdfs:subClassOf` is not ambiguous and, by defining statements using this predicate, one can easily express a formal taxonomy of concepts.

Several extensions based on RDF have been proposed to express specific types of knowledge. As an example, SKOS (Simple Knowledge Organization System) can be used to express thesauri, taxonomies and classifications. We will later introduce RDF and RDFS in more detail; what is important to understand for now is that by defining a graph data model, vocabularies, and associated entailment regimes, solutions have been proposed to express non-ambiguous and formal knowledge expressions through graph structures.

Other network-based ontologies have been derived from semantic networks. For instance, it has been proposed to represent knowledge through interlinked frames which define facts about particular objects [Minsky, 1975]. Conceptual graphs also correspond to another type of ontologies based on a graph formalism [Sowa, 1984], they are logically founded, framed in first-order logic, and still extensively studied [Chein and Mugnier, 2009]. These ontologies have not been covered in this thesis. In accordance with [Baader et al., 2010], we therefore consider that taxonomies,

thesaurus, semantic networks, frames and conceptual graphs can be seen as network-based ontologies.

Formal graph expressions of knowledge, such as RDF(S) graphs, are based on ontology languages which only provide a limited set of semantic constructs and therefore do not allow the definition of certain complex forms of knowledge. As an example, these ontology languages cannot be used to express that the concept `Father` refers to any `Person` who is not a `Women` and who `hasForChild` at least one `Person`. Neither can they be used to define that a predicate implies symmetry, e.g., that the statement `Mike isMarriedTo Lora` implies `Lora isMarriedTo Mike`, or that classes are disjoint together, e.g., that you cannot find a person which is member of both classes `Rich` and `Poor`[1]. To this end, more refined ontology languages have therefore been proposed to formally express knowledge through logic-based languages. These ontology languages rely on variants of first-order predicate calculus and are generally defined by a description logic [Baader et al., 2010]; they are used to express ontologies which cannot be expressed by simple graph structures[2].

A.1.3.2 Logic-based ontologies

As of yet, no distinction has been made between the different types of knowledge which can be represented in an ontology. Nevertheless, in knowledge modelling, a rather *conceptual* distinction is considered most of the time [De Giacomo and Lenzerini, 1996]:

- The *TBox* (Terminological Box), i.e., the general, abstract and generally static knowledge relative to a domain. This encompasses the statements relative to concepts, predicates, and their respective taxonomies, i.e., `Mammal subClassOf Animal`. The analogy with schema data encountered in the database world is often encountered.

- The *ABox* (Assertional Box), i.e., knowledge relative to instances which is generally more specific and more tied to a specific context of used, e.g., `bob isA Man`. Instance definitions are expected to be compliant with the TBox. As an example, if it is defined that `Man` and `Women` are two disjoint concepts, the conceptualisation is violated by the definition of both statements `bob isA Man` and `bob isA Women`.

Therefore an ontology can also be seen as a pair composed of a TBox and an ABox. In some cases, the ontology only encompasses the TBox and the association of both is denoted as a knowledge base.

Logic-based ontologies have been introduced to overcome the limitations of non-formal network-based ontologies and more generally to enhance the expressivity of network-based ontologies. Note that conceptually speaking they are not distinct from network-based ontologies as they can be used to express formal network-based ontologies. Logic-based ontologies are mainly based on Description Logics (DLs), a family of languages which can be used to formulate expressive

[1] E.g., only considering the amount of money it has.

[2] For the sake of clarity, it is nevertheless important to stress that complex ontologies can be expressed using graphs such as RDF graphs, but that the semantics of these graphs is no longer only carried by their structures. In other words, the graph here is used as a way to serialise a complex ontology which cannot be represented as a graph *per se*.

ontologies (TBox and ABox), through the definition of complex concepts and predicates and instances[1].

Most DLs can be seen as decidable and expressive fragments of first order logic, they enable definitions of concepts, predicates, instances and axioms based on a large variety of logical constructs: Boolean constructs, e.g., conjunction ($\sqcap$), disjunction ($\sqcup$), negation ($\neg$), as well as existential or value restrictions [Baader et al., 2010]. Hereafter, we only briefly present some of the statements which can be expressed based on a selection of logical constructs:

- $\texttt{Man} \sqcap \texttt{Woman} \equiv \bot$, the concepts `Man` and `Women` are disjoint, that is to say, there is no `Man` which are also `Women`, i.e., by considering $\mathcal{I}(\texttt{Man})$ the instances of the concept `Man` we obtain $\mathcal{I}(\texttt{Man}) \cap \mathcal{I}(\texttt{Woman}) = \emptyset$.
- $\texttt{Man} \equiv \texttt{Person} \sqcap \texttt{Male}$, the concept `Man` refers to `Person` which are also `Male`.
- $\texttt{Man} \equiv \texttt{Person} \sqcap \neg\texttt{Woman}$, `Man` refers to `Person` which are not `Women`.
- $\texttt{Father} \equiv \texttt{Man} \sqcap \geq 1.\texttt{hasForChild}.\texttt{Person}$, `Father` refers to any `Person` which is not a `Women` and which `hasForChild` at least one `Person`.
- $\texttt{Man} \sqsubseteq \texttt{Person}$, the concept `Person` subsumes the concept `Man` which implies $\mathcal{I}(\texttt{Man}) \subseteq \mathcal{I}(\texttt{Person})$.

These constructs can also be used to express statements which will constrain the possible interpretation of concepts or predicates. Therefore, they have been used to define numerous DL syntaxes with various degrees of expressivity and complexity. The presentation of the various logical constructors proposed in DLs, and the DLs syntaxes which can be formed from them, is out of the scope of this manuscript.

A.1.4 Definition of ontologies: RDF(S) and OWL

We briefly present some of the standard languages which can be used to semantically describe resources and to define ontologies. These standards have been proposed by the W3C and derive from other works which will not be discussed hereafter, e.g., DAML (+) OIL.

A.1.4.1 RDF – Describing resources through graphs

The Resource Description Framework (RDF) was initially proposed by the W3C as a graph-based data model to expose and exchange information in the Web and more particularly to express metadata [W3C, 2004]. RDF is, however, not restricted to use on the Web and several ontology languages are based on it. Being an abstract data model, RDF can be expressed and exchanged using several notations and serialisation formats (e.g., XML, Turtle).

[1] Note that predicates are called roles and instances are denoted individuals.

Similarly to other data models such as the entity-relationship model, RDF provides a way to describe resources through intuitive `subject predicate object` (`spo`) statements; we already used them to introduce ontologies, e.g., `Human rdfs:subClassOf Animal`. A set of statements forms an RDF graph which is a labelled directed graph.

An important aspect of RDF is that resources are identified by Uniform Resource Identifiers (URIs)[1]. In short, URIs generalise URLs by providing a way to unambiguously denote resources. As an example, the unique URI `http://purl.uniprot.org/taxonomy/9685` will replace the different lexical identifiers which may be used to refer to the concept `Cat` (e.g., *Cat*, *Chat*, *Felis catus*). A short prefix can be defined to shorten URIs, e.g., defining `tax` as a prefix for `http://purl.uniprot.org/taxonomy/`, the URI associated to the concept `Cat` can also be written as `tax:9695`. In this manuscript, the prefixes will be removed as much as possible in order to facilitate reading.

Using URIs, the sentence "*Bob is a cat*" can be expressed by the RDF statement `bob rdf:type tax:9695` (`Cat`). The meaning associated to URIs is therefore defined through `spo` statements. In this case, the URI `rdf:type` is used to denote the membership of an instance to a class[2] – as formally defined by the semantics of RDF(S).

Any `spo` statement must respect the following restrictions:

- The *subject* can be a URI or a *blank node*. A blank node is a reference to an anonymous resource: it unambiguously refers to something for which we don't want to define a specific URI.
- The *predicate* is always a URI.
- The *object* can be a URI, a blank node or a string literal. Literals can be used to represent typed data values by specifying a datatype, e.g., `"2013-09-09"^^xsd:date` specifies that the literal `"2013-09-09"` must be understood as a date, i.e. according to the definition of a date unambiguously defined by the URI `xsd:date` (`http://www.w3.org/2001/XMLSchema#date`).

By providing a graph (meta)data model and a built-in vocabulary, RDF can be used to characterise resources through simple `spo` statements. In addition, *reification* techniques can be used to define properties about a statement. RDF provides the vocabulary dedicated to this purpose. As an example, to model the knowledge associated to the statement that "*Luc Bar sent an email to his friend Marc Foo the 2013-09-09*", we can define the RDF graph presented in Figure A.5.

Among the various alternatives which have been proposed to query RDF, SPARQL[3] is the W3C recommendation defined to manipulate and query RDF [W3C, 2013b].

RDF also provides a framework for the definition of ontology languages, we further presents two of them: RDF Schema (RDFS) and the Web Ontology Language (OWL).

[1] In accordance with the literature we will mainly refer to the term URI in this manuscript despite the fact that Internationalized Resource Identifiers (IRIs) would be more appropriate.

[2] `http://www.w3.org/1999/02/22-rdf-syntax-ns`

[3] Recursive acronym, SPARQL Protocol and RDF Query Language – current version 1.1.

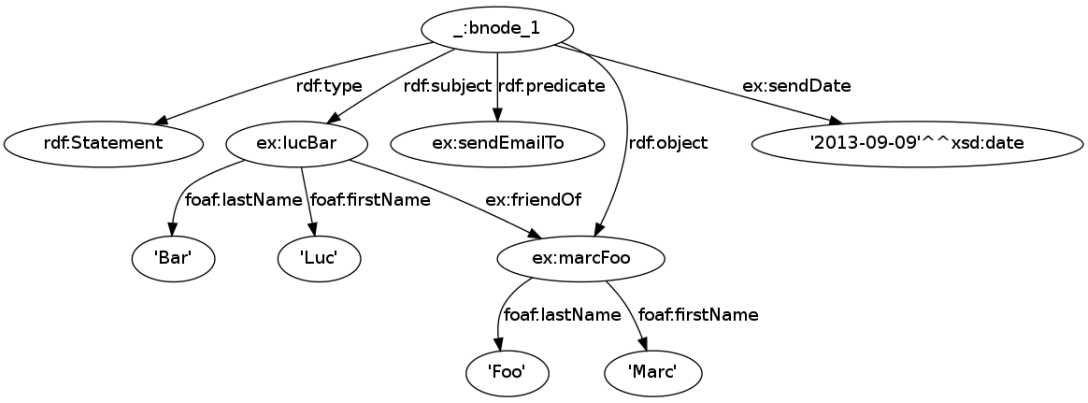

FIGURE A.5: Example of a *reification* in an RDF graph. The graph models the statement that "*Luc Bar sent an email to his friend Marc Foo the 2013-09-09*".

A.1.4.2 RDFS – Add formal semantics to RDF

We have seen that RDF defines a vocabulary to express statement with specific meaning, e.g., `rdf:type`, `rdf:statement`. RDF Schema (RDFS) provides a vocabulary to extend the semantics of RDF graphs [W3C, 2004], and defines how this vocabulary must be interpreted for reasoning. It can be used to define simple ontologies by means of taxonomies of concepts and predicates, as well as predicate restrictions, i.e., the domain and the co-domain (range) which must be associated to a specific predicate. The vocabulary and semantics provided by RDFS can be used to define simple terminological knowledge of ontologies.

RDFS is associated to an entailment regime which specifies the semantics of its constructs. This semantics is defined through deductive rules, e.g., the implications of the taxonomic relationship. Let us consider the semantics carried by predicate restrictions: using RDFS, it is possible to define the type of instances which are involved in a specific statement. For this, the domain and the range of a specific predicate can be specified. As an example, it can be defined that the predicate `hasFather` has `Person` for domain and `Man` for range. In other words, this means that only `Person` have fathers and that fathers can only be `Man`, i.e., members of the class `Man`. Therefore, defining that `jean hasFather marc`, we can infer that `jean` is a `Person` and that `marc` is a `Man`. In addition, by defining that `Man` is a subclass of `Person` and that `hasFather` is a subproperty of `hasParent`, we can also infer that `marc` is a `Person` and that the statement `jean hasParent marc` holds. A graphical representation of this example is presented in Figure A.6, red dotted relationships correspond to some of the `spo` statements which can be inferred from the RDF graph defined by bold relationships considering RDFS semantics.

Notice that RDFS defines the semantics of RDF graphs by constraining the interpretations which can be made from them, i.e., by defining how the vocabulary has to be understood. It is, however, important to understand that in most cases RDFS cannot be used to evaluate the validity of a specific statement. Indeed, considering the statements `bobJunior rdf:type Cat` and `bobJunior hasFather bob`, considering the domain associated to the predicate `hasFather` (i.e., `Person`), a reasoner will infer `bobJunior rdf:type Person`, even if we consider that an instance cannot be both a member of the classes `Cat` and `Person`. Indeed, using RDFS, it is

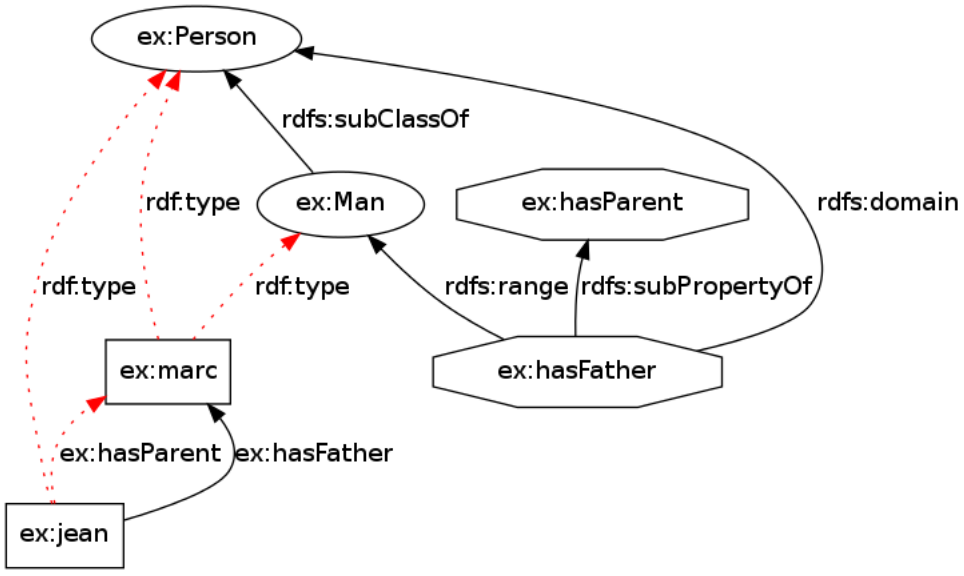

FIGURE A.6: RDF graph in which the semantics defined by RDFS is considered for statement inference. Red dotted relationships correspond to some of the statements which can be inferred from the rest of the graph.

impossible to express that an instance cannot be a member of two classes, i.e., that two classes are disjoint together. Thus, the reduced vocabulary and semantics (entailment regime) provided by RDFS can be limiting to model certain domains. More expressive languages, also based on RDF, have therefore been proposed.

A.1.4.3 OWL – Web Ontology Language

OWL (now version 2[1]) is the Web Ontology Language proposed as a W3C recommendation for the definition of rich ontologies based on Description Logic (DL)[2] [W3C, 2013a]. OWL is a family of languages which can be used to define ontologies with various degree of expressivity. Ordered by increasing degree of expressiveness, the sublanguages OWL Lite, OWL DL and OWL Full have been distinguished[3]. OWL DL is the most commonly used as it ensures completeness, decidability, and also provides an interesting threshold between expressivity and reasoning efficiency [Nardi and Brachman, 2003]. Indeed, it is worth noting that expressivity has a price as it negatively impacts efficiency of reasoning procedures in term of computational complexity.

OWL provides a vocabulary and model theory to define expressive ontologies which cannot be defined using RDF(S). Among the various capabilities offered by the OWL vocabulary, it is possible to define extra relationships between concepts (e.g., disjointness), to restrict predicates using cardinality (e.g., to express statement such as *people have exactly one brain*), to define properties of predicates (e.g., symmetry) and between predicates (e.g., inverse), etc. We will not cover OWL in detail in this thesis.

[1] As with RDF, no versioning are mentioned using acronyms.
[2] OWL is compatible with the description logic $\mathcal{SROIQ}$.
[3] New profiles have been proposed in OWL 2.

A.2 Building a semantic graph from an ontology

In this manuscript, since most knowledge-based semantic measures consider ontologies as network-based structures, we will mainly manipulate ontologies through their representation into semantic graphs. Nevertheless, even if most ontologies can be expressed, stored and exchanged using graph-based formalisms such as RDF, it doesn't mean that they can be processed as graph in a straightforward manner. Indeed, as an example, since graph syntax based on triplets are limiting for expressing certain facts (e.g., restrictions), many OWL constructs are encoded into a set of triplets. Therefore, otherwise stated, the semantics of the graph is not carried by its structure.

Therefore, for manipulating expressive ontologies as semantic graphs, specific transformations and sometimes reductions have to be performed, e.g., for materialising knowledge implicitly defined in ontologies into their corresponding semantic graphs. As an example, if a specific domain `C` is associated to a predicate `p`, any triplet of the form `u p v` means that `u` is a `C`. The membership of `u` into the class `C` (i.e. relationship `u isA C`), is implicitly defined into the ontology. However, if we consider this ontology as a semantic graph without pre-processing, this knowledge cannot be inferred by means of traversal, i.e., no semantically coherent traversal links `u` to the class `C`. Therefore, in order to be processed as a semantic graph, a specific relationship linking `u` to `d` must therefore be explicitly defined.

Due to the complexity and vast extend of this topic, we will not propose a systematic way to convert any ontologies into a semantic graph. Nevertheless, this appendix discusses specific aspects of this issue and defines how expressive ontologies can be reduced into a semantic graph in order to be processed by semantic measures. In particular, we present the main steps which have to be considered to process an ontology as a semantic graph.

A general methodology can be defined to model the main steps which can be applied to obtain a semantic graph from any ontology. Figure A.7 illustrates this general process.

The main steps are:

1. *Knowledge modelling*: Steps 1 and 2 represent the modelling of a piece of knowledge to a machine understandable and computational representation. Step 2 defines the expression of an ontology in a specific language, e.g., OWL, RDF(S), OBO. The language which is used conditions the expressivity of the language constructs and therefore the possibility to represent the knowledge defined in the ontology into a semantic graph.

2. *Knowledge inference*: Step 3 represents the optional use of a reasoner to infer knowledge implicitly defined in the ontology. As an example, in an ontology expressed in RDF(S), this step may correspond to the entailment of the RDF graph according to the semantics defined by RDFS, i.e., the use of a reasoner to infer triplets according to RDFS entailment rules. Reasoners may also be used to build the taxonomy of concepts from complex logic-based ontologies.

3. *Mapping to a graph representation*: Step 4 is of major importance. It corresponds to the mapping of the ontology to a graph representation which can be processed by most

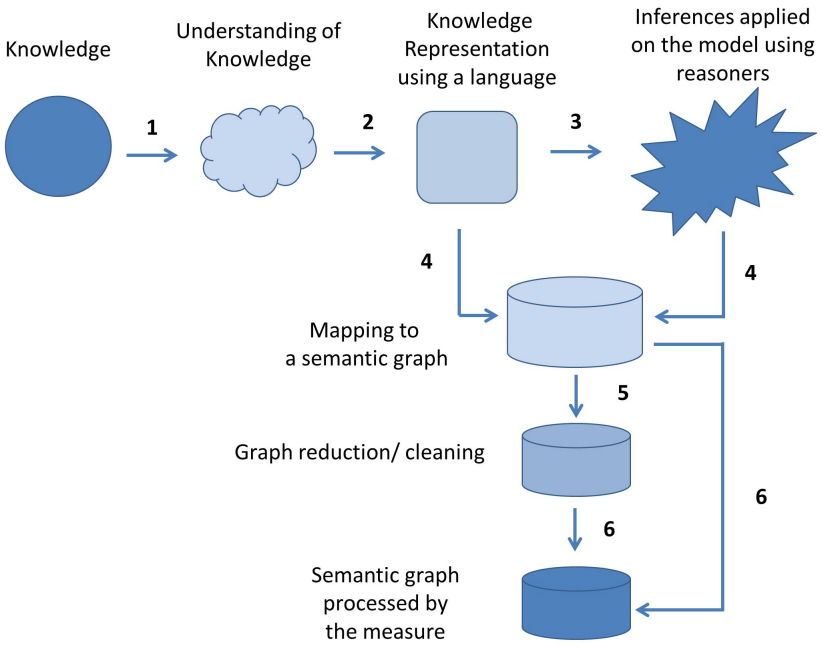

FIGURE A.7: Main steps which can be applied for building a semantic graph from any ontology

semantic measures. In some cases, this step is implicit since the ontology is already expressed through a network-based structure, e.g., graphs corresponding to taxonomies, WordNet lexical database. Depending on the language used to express the ontology, this phase may imply a loss of knowledge and must therefore be carefully considered. It is therefore important to understand that some ontology defined using expressive ontology languages, such as some logic-based ontology language, may only partially be modelled in a graph structure as expected by most semantic measures.

4. *Graph reduction / cleaning*: Step 5 corresponds to the reduction of the semantic graph in order to focus on specific knowledge. As an example, in some cases, only the taxonomy of concepts will be considered. In other cases, this is the semantic graph induced by both concepts and instances which will be considered. After the reduction a cleaning phase may also be required, it corresponds of the removal of some relationships or concepts defined in the graph. It may be required for ensuring the coherency of semantic measures.

We further discuss the notion of graph reduction and graph cleaning.

Formally, we denote $G(O)$, shorten G if there is no ambiguity, the reduction of the ontology O to a semantic graph G. In addition, we denote $G_{R'}(O)$, also shorten $G_{R'}$ if there is no ambiguity, the reduction of O as a semantic graph only considering the relationships having as predicate $r \in R' \subseteq R$. A common reduction of an ontology as a graph is $G_{\texttt{subClassOf}}$, shortened by G_T and named the taxonomic reduction (to be more precise, this is the taxonomic reduction of

the ontology which is made only considering the concepts defined in O). In other words, G_T corresponds to the taxonomy $\preceq_C$ represented as a graph, and therefore only contains concepts, i.e., the vertices of the graph only refer to concepts. This reduction is widely used for computing the semantic similarity between concepts.

Graph reductions can naturally be more complex. The graph $G_{Rx}(O)$, with $Rx = \{\texttt{subClassOf,isA}\}$, refers to the reduction which is composed of the relationships having for predicate `subClassOf` or `isA`. Conversely to G_T, the vertices of this graph refer to both instances and concepts. We denote such a graph G_{TI} (T stands for Taxonomic and I for `isA` relationships).

Studies relying on semantic graphs can be conducted taking the full semantic graph into account or focusing on a particular subgraph. Depending on the amount of information considered, some properties of the graph may change (e.g., acyclicity), along with the strategies and algorithmic treatments used for their processing. Since most semantic measures require the graph to fulfil specific properties, we briefly discuss the link between the properties of the graph structure and semantic measures.

Considering all types of semantic relationships, a semantic graph generally forms a connected directed graph which can contain cycles, i.e. path from a node to itself. The taxonomic reduction (G_T), also leads to a graph given that a concept can inherit from multiple concepts. Nevertheless, due to the transitivity of taxonomic relationships, G_T is expected to be acyclic. Taxonomic reductions composed of a unique concept which subsumes the others form a Rooted Directed and Acyclic Graph (RDAG). DAG properties enable efficient graph treatments to be performed; numerous semantic measures take advantage of them. The graph G_{TI} is also a RDAG.

Figure A.8 presents some of the reductions of a semantic graph which are usually performed prior to consider semantic measures treatments. This example is based on the reduction of the Gene Ontology (GO) in order to extract the taxonomic knowledge which is related to a specific aspect of the GO. Such a reduction is generally performed before comparing pairs of concepts. The figure shows the GO, which is composed of three subparts (sub-graphs): Molecular Function (MF), Biological Processes (BP), and Cellular Component (CC). The GO originally forms a cyclic graph composed of concepts linked by various semantic relationships. The first reduction shows the isolation of the MF subgraph. Only concepts composing the MF subpart and the relationships involving a pair of MF concepts are considered. The resulting graph can be cyclic. The final reduction only contains MF concepts linked by taxonomic relationships, which corresponds to a RDAG (Rooted Directed Acyclic Graph).

The accuracy of treatments relying on semantic measures and the semantics of their results highly depends on the semantic graph which is processed. In this context, the quality of semantic graphs (w.r.t semantic measures) relies on the way knowledge is defined. As an example, a semantic graph may contain relationship redundancies. Such redundancies can impact semantic measures' results and thus have to be removed, e.g., documented in [Park et al., 2011]. They appear when a direct semantic relationship between two elements can be inferred (explained) by an indirect

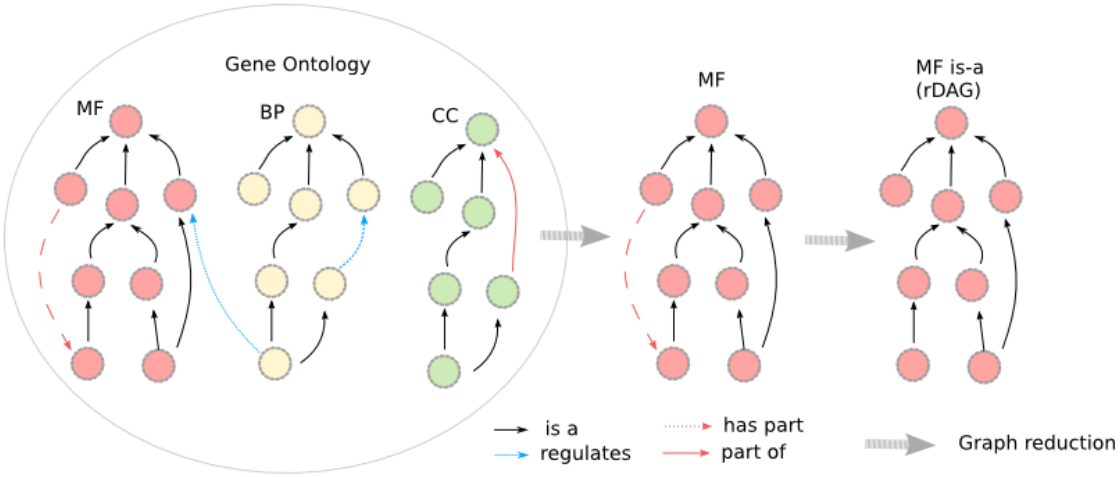

FIGURE A.8: Example of a reduction of an ontology and its effects on graph properties

one, i.e., expressed in term of graph traversal[1]. Redundancies involve transitive relationships. As an example, since the taxonomic relationship is transitive, if the semantic graph defines that `Human subClassOf Mammal` and `Mammal subClassOf Animal` a semantic reasoner can infer that `Human subClassOf Animal`. In this case, a redundancy occurs when an explicit (non-inferred) relationship defines `Human subClassOf Animal`.

Most knowledge-based semantic measures proposed in the literature have been defined for semantic graphs. However, they are generally presented as if they were suited for all "*ontologies*". Nevertheless, despite this aspect is generally not mentioned in the literature, not all ontologies can be used *per se* for computing semantic measures. This leads developers to face complex problems for implementing semantic measures, and therefore hampers practical usages of measures. Indeed, some semantic measures expect processed knowledge to be expressed into a semantic graph. Therefore, this requires expressive ontologies to be expressed as semantic graphs. In this appendix, we have underlined that the transformation of ontologies into semantic graphs is not a trivial process. We stressed that this process must be carefully considered and, as an initial contribution, we distinguished its main steps. These steps have to be considered for modifying an ontology into a semantic graph, and are therefore required for processing any ontology by taking advantage of the large diversity of knowledge-based semantic measures relying on semantic graph analysis.

[1]For those familiar to RDF(S), the domain and the range (co-domain) of a predicate, even if represented as a relationship, cannot induce redundancies, e.g. the triplet `isAParentOf rdfs:domain Human` doesn't mean that the triplet `Jean rdf:type Human` is redundant considering that `Jean isAParentOf Louise` is specified in the ontology. Here redundancies are evaluated by mean of graph traversal.

B

A discussion on the evaluation of semantic measures

This appendix discusses information relative to the selection of semantic measures. The aim is not to provide an exhaustive state-of-the-art related to reflections on the subject but rather to distinguish central aspects of measures which it may be of interest to discuss in order to guide both selection and comparison of semantic measures. More information about the subject can be found in [Harispe et al., 2013c].

Evaluation protocols and benchmarks are essential for the analysis of the benefits and drawbacks of existing or newly proposed semantic measures. They are of major importance in objectively evaluating new contributions and in guiding users of semantic measures in the selection of best suited measures w.r.t their needs (e.g., application context). Nevertheless, despite the vast literature related to semantic measures, only few contributions focus on this specific topic, e.g., [Al-Mubaid and Nagar, 2008; Lee et al., 2008; Petrakis and Varelas, 2006; Slimani, 2013].

Generally, any evaluation aims to distinguish the benefits and drawbacks of compared alternatives according to specific criteria. Such comparisons are generally used to rank the *goodness* of measures regarding the selected criteria. Therefore, to be compared, three important questions deserve to be answered:

1. What criteria can be used to compare semantic measures?
2. How can the *goodness* of a measure be evaluated w.r.t a specific set of criteria?
3. Which criteria must be considered in order to evaluate semantic measures for a specific usage?

This appendix mainly focuses on the criteria which can be considered to compare semantic measures. We will nevertheless also present some benchmarks which are commonly used to evaluate semantic measure accuracy.

B.1 Criteria for the evaluation of semantic measures

Several criteria can be used to evaluate measures. Among them, we distinguish:

- Their *accuracy* and *precision*.
- Their *computational complexity*, i.e., algorithmic complexity.
- Their *mathematical properties*.
- Their *semantics*.

As we will see, these (non-disjoint) criteria can be used to evaluate several aspects of measures. Each one is briefly introduced.

Accuracy and Precision

The accuracy of a measure can only be discussed according to predefined expectations regarding the results produced by the measure. Indeed, as defined in metrology, the science of measurement, the accuracy of a measurement must be understood as the closeness of the measurement of a quantity regarding the *true* value of that quantity [BIPM et al., 2012].

The precision of a measure (system of measurement) corresponds to the degree of reproducibility or repeatability of the score produced by the measure under unchanged conditions. Since most semantic measures are based on deterministic algorithms, i.e., they produce the same result given a specific input, here we focus on the notion of accuracy. Note that the precision of a measure can be regarded as a mathematical property since some semantic measures are non-deterministic (e.g., semantic measures based on random-walk approaches). Given that most measures are deterministic, the precision of semantic measures will not be discussed hereafter.

The notion of accuracy of a measure is compulsory tight to a context, e.g., benchmark, semantic proxy (specific corpus, ontology, etc.), tuning of measure parameters (if any). Indeed, there is no guarantee that a measure which has been proved accurate in a specific context, will be accurate in all contexts. As we will see, the accuracy of semantic measures is therefore evaluated according to expected results.

Computational complexity

The computational complexity or algorithmic complexity of semantic measures is of major importance in most applications. It is indeed worth noting that given the growing volumes of datasets processed in semantic analysis (large corpus of texts and ontologies), the algorithmic complexity of measures plays an important role towards their large adoption.

Considering equivalent accuracy in a specific context, most users of semantic measures will prefer to make concessions on measure accuracy for a significant reduction of computational time. However, the literature relative to semantic measures is very limited on this subject. It

is therefore difficult to discuss algorithmic implications of current proposals; this hampers non-empirical evaluations and burdens the selection of measures. It is, however, difficult to blame designer of semantic measures for not providing detailed algorithmic analyses of their proposal. Indeed, computational complexity analyses of measures are both technical and difficult to make. In addition, most of the time, these analyses depend on the specific data structure which is used to represent the semantic proxy taken into account by measures (e.g., ontology), a degree of detail which is generally not discussed in contributions related to semantic measures – note that this sometimes creates a gap between theoretical possibilities and practical implementations.

Despite its major importance, the evaluation of semantic measures regarding their computational complexity is still difficult today.

Mathematical properties

Several mathematical properties of interest for semantic measures were distinguished in Chapter 2, e.g., symmetry, identity of the indiscernibles, normalisation. These mathematical properties are of particular importance for the selection of semantic measures. They are, for instance, essential for the application of specific optimisation techniques (e.g., based on the normalisation of measures). They also play an important role in better understanding the semantics carried by measures, i.e., the meaning carried by their results.

Mathematical properties are central for the comparison of measures since they are generally required to ensure the coherence of treatments which rely on semantic measures. This, for instance, is the case when inferences have to be made based on scores produced by semantic measures. As an example, the implication of the non-respect of the identity of the indiscernibles has to be carefully considered; it can be conceptually disturbing that the comparison of a concept to itself produces non-maximal or even low similarity scores. It is, however, the case using some measures in specific contexts[1].

Analyses of mathematical properties of measures are thus required to deeply understand their expected behaviour and to evaluate their relevance for domain-specific applications.

Semantics of measures

The meaning (semantics) of semantic measure results deserves to be thoroughly understood by end-users. This aspect is central for the selection of a measure. The semantics of semantic measures is defined by the assumptions on which their algorithmic design relies. Some of these assumptions can be understood through the mathematical properties of the measures. The semantics is also defined by the cognitive model on which the measure relies, the semantic proxy in use and the semantic evidences analysed. As we saw in Section 3.3, semantic evidence taken

[1] As an example, using Resnik's measure based on the notion of information content of concepts (Equation 3.28), the semantic similarity of a general concept (near to the root – low θ) to itself will be low.

into account by the measure generally defines its type/general semantics (e.g., the measure evaluates semantic similarity, relatedness...).

It is difficult to compare measures regarding the semantics they carry. It is, however, essential for semantic measure users to understand that measure selection may in some cases strongly impact the conclusions which can be supported by the measurement, e.g., it for instance not adapted to perform substitutions (for example in the context of recommendation) which are supported by semantic relatedness instead of semantic similarity.

Existing protocols to evaluate accuracy of semantic measures

The accuracy of semantic measures is today considered as the *de-facto* criterion to evaluate measures. It can be evaluated using a direct or an indirect approach. In most cases, measures are evaluated using a direct approach, i.e., based on expected scores of measurement of pairs of elements (e.g., similarity, relatedness). In all cases, the evaluation is performed w.r.t specific expectations/assumptions:

- *Direct evaluation*: based on the correlation of semantic measures with expected scores or results produced by other metrics. Measures are, for instance, evaluated regarding their capacity to mimic human rating of semantic similarity/relatedness. In this case, the accuracy of measures is discussed based on their correlations with gold-standard benchmarks composed of pairs of terms/concepts associated to expected ratings. For domain-specific studies, a set of experts is used to assess expected scores which will make up the benchmark (e.g., physicians in biomedical studies). In other cases, measures will be evaluated regarding their capacity to produce scores highly correlated to specific metrics. These metrics are expected to summarise our knowledge of compared elements. This strategy is adopted in bioinformatics to evaluate semantic measures which have been designed to compare gene products according to their conceptual annotations, i.e., the evaluation can be based on their correlation with other measures which are commonly used to compare genes (e.g., sequence similarity), e.g., [Lord, 2003].
- *Indirect evaluation*: The evaluation of measures relies on the analysis of the performance of applications or algorithms which take advantage of semantic measures. The treatment considered is domain-specific, e.g., accuracy of terms' disambiguation techniques, performance of classifiers, clustering techniques or synonymy detection systems which rely on semantic measures.

B.2 Benchmarks for semantic measures evaluation

This section presents the benchmark of [Pedersen et al., 2007] which was used in this manuscript to evaluate measures. Other benchmarks are presented in [Harispe et al., 2013c].

Pedersen benchmark for semantic similarity

Term A	Term B	Physician	Coder	Avg
Renal failure	Kidney Failure	4	4	4
Heart	Myocardium	3.3	3	3.15
Stroke	Infarct	3	2.8	2.9
Abortion	Miscarriage	3	3.3	3.15
Delusion	Schizophrenia	3	2.2	2.6
Congestive Heart Failure	Pulmonary Edema	3	1.4	2.2
Metastasis	Adenocarcinoma	2.7	1.8	2.25
Calcification	Stenosis	2.7	2	2.35
Diarrhea	Stomach cramps	2.3	1.3	1.8
Mitral Stenosis	Atrial Fibrillation	2.3	1.3	1.8
Chronic obstructive pulmonary disease	Lung infiltrates	2.3	1.9	2.1
Rheumatoid Arthritis	Lupus	2	1.1	1.55
Brain tumor	Intracranial Hemorrhages	2	1.3	1.65
Carpel Tunnel Syndrome	Osteoarthritis	2	1.1	1.55
Diabetes Mellitus	Hypertension	2	1	1.5
Acne	Syringes	2	1	1.5
Antibiotic	Allergy	1.7	1.2	1.45
Cortisone	Total knee replacement	1.7	1	1.35
Pulmonary fibrosis	Lung cancer	1.7	1.4	1.55
Cholangiocarcinoma	Colonoscopy	1.3	1	1.15
Lymphoid hyperplasia	Laryngeal Cancer	1.3	1	1.15
Multiple Sclerosis	Psychosis	1	1	1
Appendicitis	Osteoporosis	1	1	1
Rectal polyp	Aorta	1	1	1
Xerostomia	Alcoholic Cirrhosis	1	1	1
Peptic Ulcer disease	Myopia	1	1	1
Depression	Cellulites	1	1	1
Varicose vein	Entire knee meniscus	1	1	1
Hyperlipidemia	Metastasis	1	1	1

TABLE B.1: Pedersen et al. [2007] benchmark for semantic similarity. Scores of semantic similarity for pairs of terms related to the biomedical domain. Scores of similarities are provided for Physicians – Coders – Physicians + Coders (Avg). Mappings to concepts defined in the MeSH and SNOMED-CT are provided in Table B.2

Term A	**Term B**	**Concept A MeSH**	**Concept B MeSH**	**Concept A SNOMED**	**Concept B SNOMED**
Renal Insufficiency	Kidney Failure	D051437	D051437	42399005	42399005
Heart	Myocardium	D006321	D009206	80891009	74281007
Stroke	Infarction	D020521	D007238	230690007	55641003
Abortion, Habitual	Miscarriage	D000026	D000022	70317007	17369002
Delusions	Schizophrenia	D003702	D012559	48500005	58214004
Heart Failure	Pulmonary Edema	D006333	D011654	42343007	19242006
Neoplasm Metastasis	Adenocarcinoma	D009362	D000230	128462008	443961001
Calcinosis	Constriction, Pathologic	D002114	D003251	125369001	415582006
Mitral Valve Stenosis	Atrial Fibrillation	D008946	D001281	79619009	49436004
Arthritis, Rheumatoid	Lupus Erythematosus	D001172	D008180	69896004	200936003
Brain Neoplasms	Intracranial Hemorrhages	D001932	D020300	254935002	1386000
Carpal Tunnel Syndrome	Osteoarthritis	D002349	D010003	57406009	396275006
Diabetes Mellitus	Hypertension	D003920	D006973	73211009	38341003
Acne Vulgaris	Syringes	D000152	D013594	11381005	61968008
Anti-Bacterial Agents	Hypersensitivity	D000900	D006967	255631004	106190000
Cortisone	Arthroplasty, Replacement, Knee	D003348	D019645	32498003	179344006
Pulmonary Fibrosis	Lung Neoplasms	D011658	D008175	51615001	363358000
Cholangiocarcinoma	Colonoscopy	D018281	D003113	70179006	73761001
Pseudolymphoma	Laryngeal Neoplasms	D019310	D007822	128863005	363429002
Multiple Sclerosis	Psychotic Disorders	D009103	D011618	24700007	69322001
Appendicitis	Osteoporosis	D001064	D010024	74400008	64859006
Xerostomia	Liver Cirrhosis, Alcoholic	D014987	D008104	87715008	420054005
Peptic Ulcer	Myopia	D010437	D009216	13200003	57190000
Depression	Cellulitis	D003863	D002481	35489007	128045006
Hyperlipidemias	Neoplasm Metastasis	D006949	D009362	55822004	363346000

TABLE B.2: Mapping between terms used in Pedersen et al. [2007] semantic similarity benchmark (Table B.1) and MeSH descriptors and SNOMED-CT concepts

C

Empirical analysis: supplementary results

This appendix provides additional results for several experiments and discussions which are presented in the manuscript.

C.1 Study of semantic measures in the biomedical domain: additional results

Supplementary results of Section 5.1. The figures show the surfaces of correlation obtained w.r.t Pedersen et al. [2007] benchmark (average coder-physician) and semantic similarity measures derived from abstract formulations of the contrast and ratio models (Figures C.1 and C.2 respectively). In each figure, the red dot refers to the maximal correlation. Refer to Table 5.1 for details on measures (Case i corresponds to a specific column in the table). In each figure, four pairs of surfaces are shown, i.e., one per instantiation (case). Note that, even if it is not specified in these figures, numerous points of the surfaces refer to specific measures proposed in the literature.

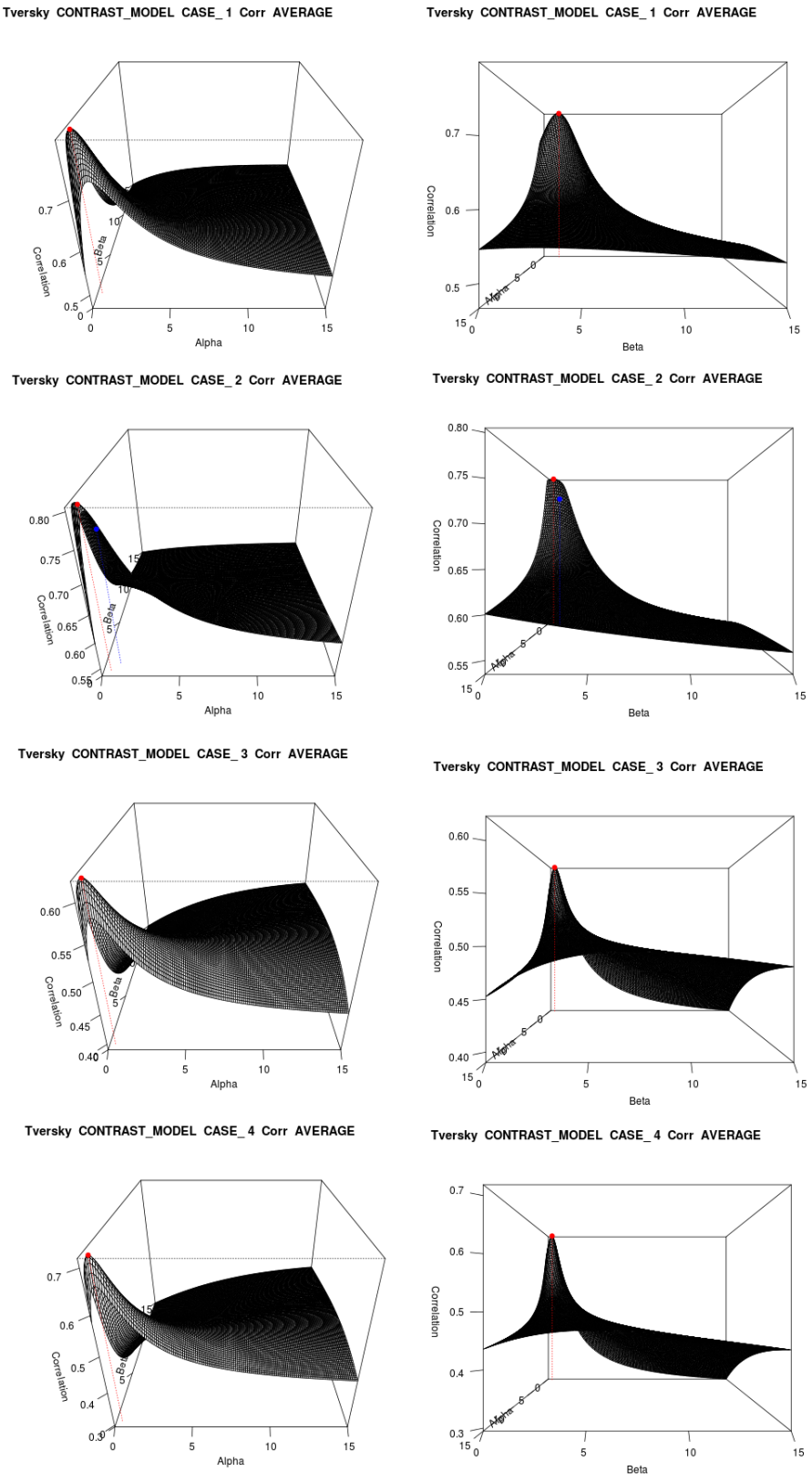

FIGURE C.1: Surface of correlation considering Pedersen et al. [2007] benchmark (average Coders – Physicians) and semantic similarity measures derived from an abstract formulation of the *contrast model* ($\gamma = 1$) – Table 4.5 (sim_{CM^*}). The red dot represents the maximum correlation value.

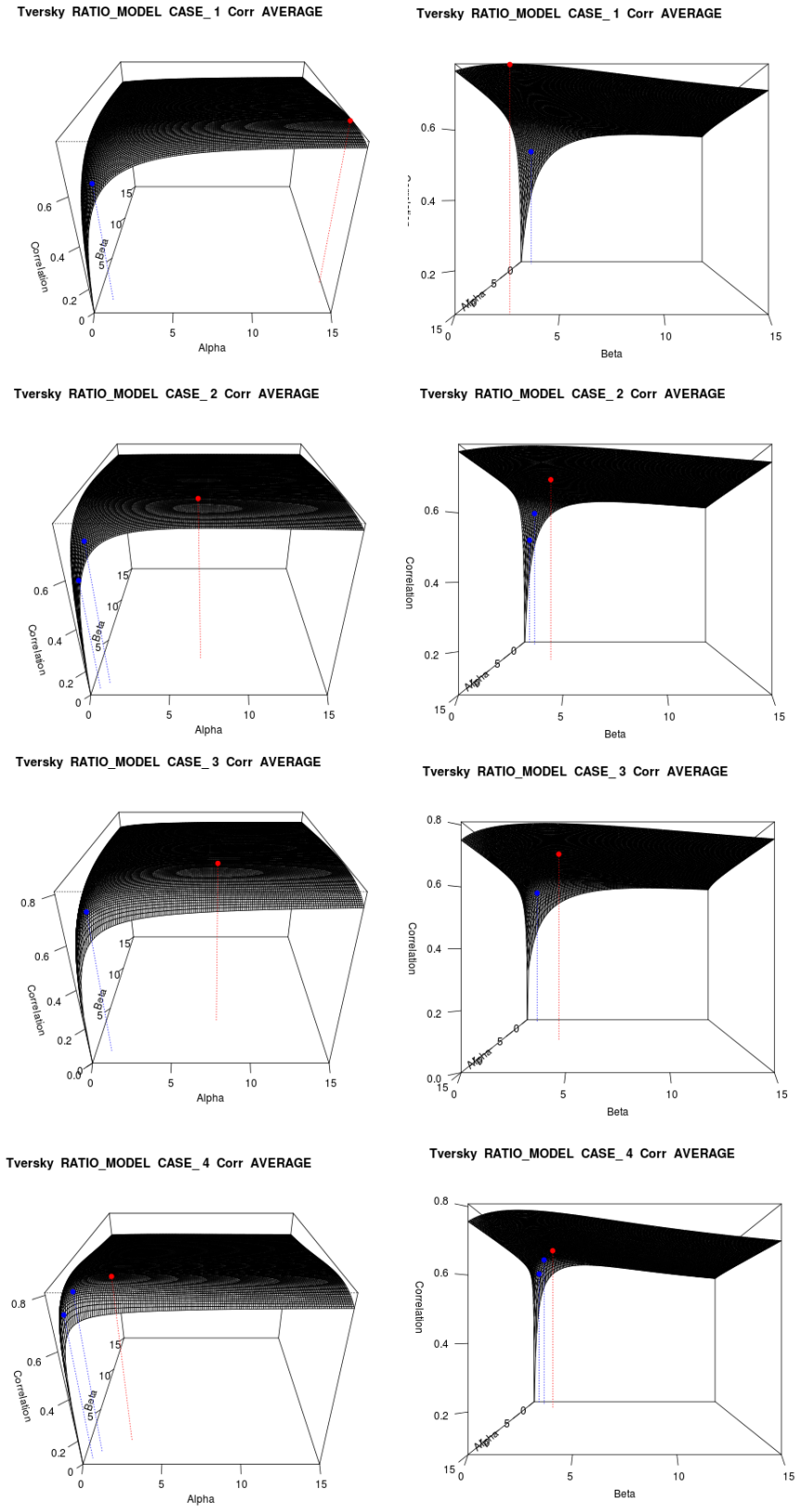

FIGURE C.2: Surface of correlation considering Pedersen et al. [2007] benchmark (average Coders – Physicians) and semantic similarity measures derived from an abstract formulation of the *ratio model* – Table 4.5 (sim_{RM^*}). The red dot represents the maximum correlation value.

C.2 Reflection on the robustness of semantic measures: additional results

Supplementary results of the Section 5.2. In each figure, the red dot corresponds to (α_l^0, β_l^0) and the red triangle refers to (α_l^*, β_l^*). Refer to Table 5.1 for details on measures (Ci corresponds to case i). Note also that in some cases the two points overlap and that L_r is represented by the area inside the bold black line. Finally, recall that semantic similarity measures refers to knowledge-based semantic similarity measures.

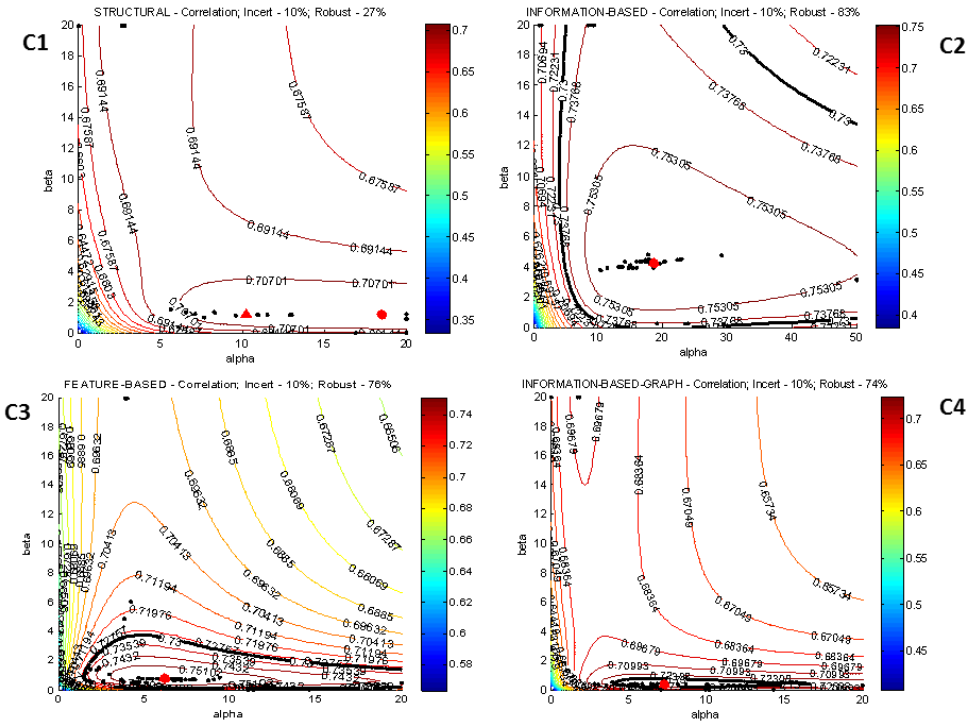

FIGURE C.3: Plot of robustness of parametric semantic similarity measures considering 10% of uncertainty.

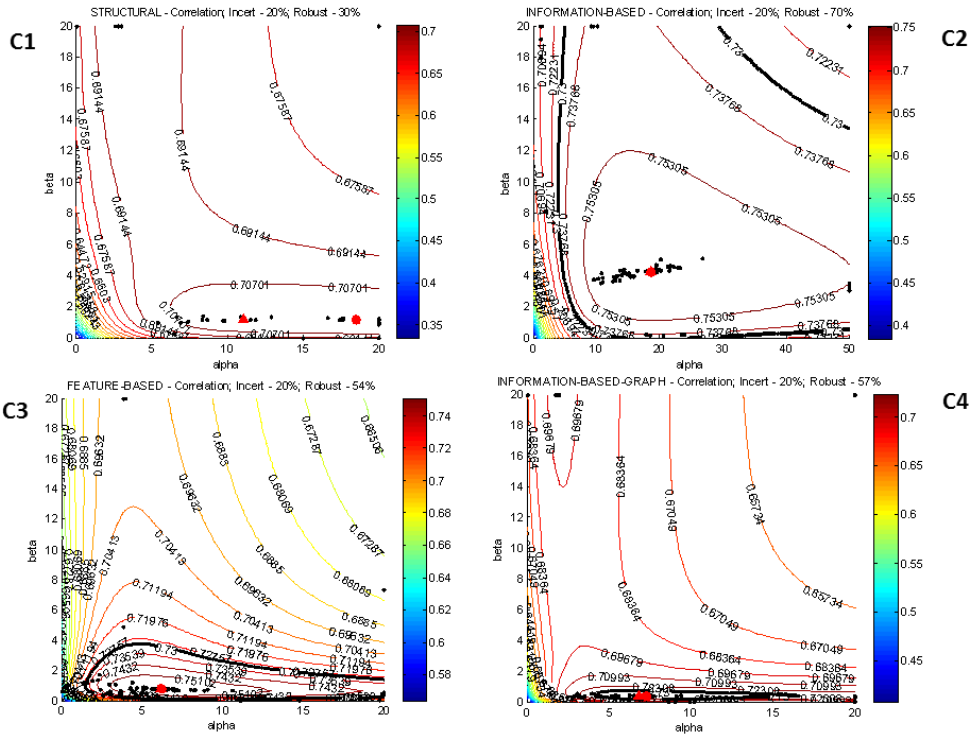

FIGURE C.4: Plot of robustness of parametric semantic similarity measures considering 20% of uncertainty

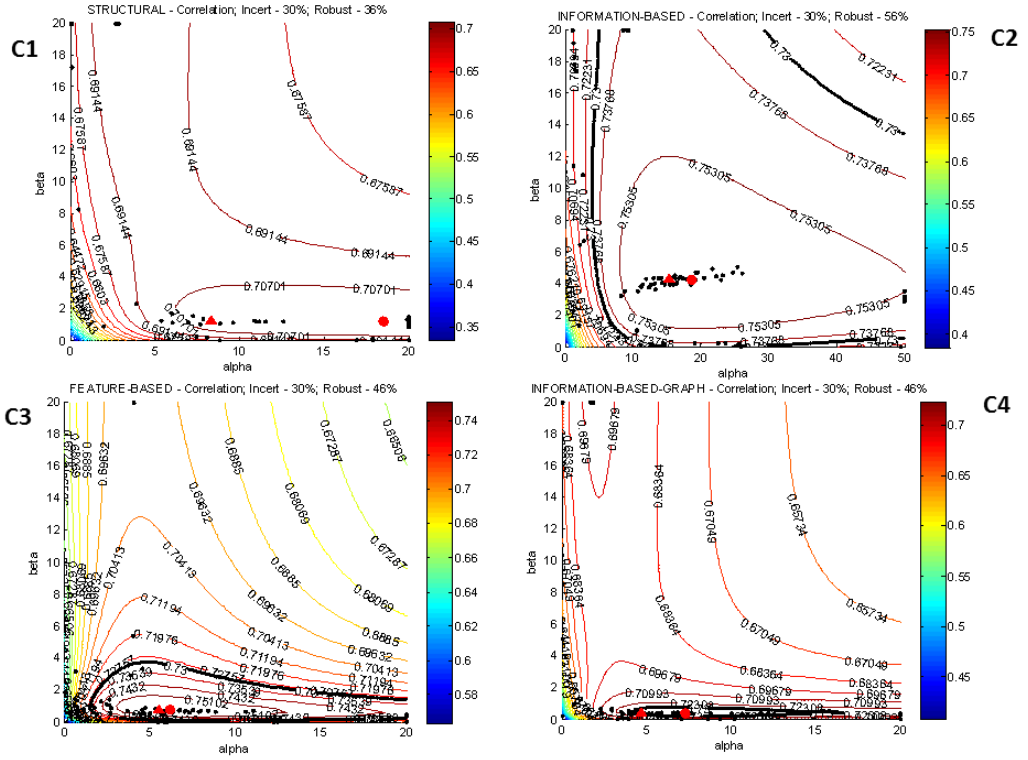

FIGURE C.5: Plot of robustness of parametric semantic similarity measures considering 30% of uncertainty

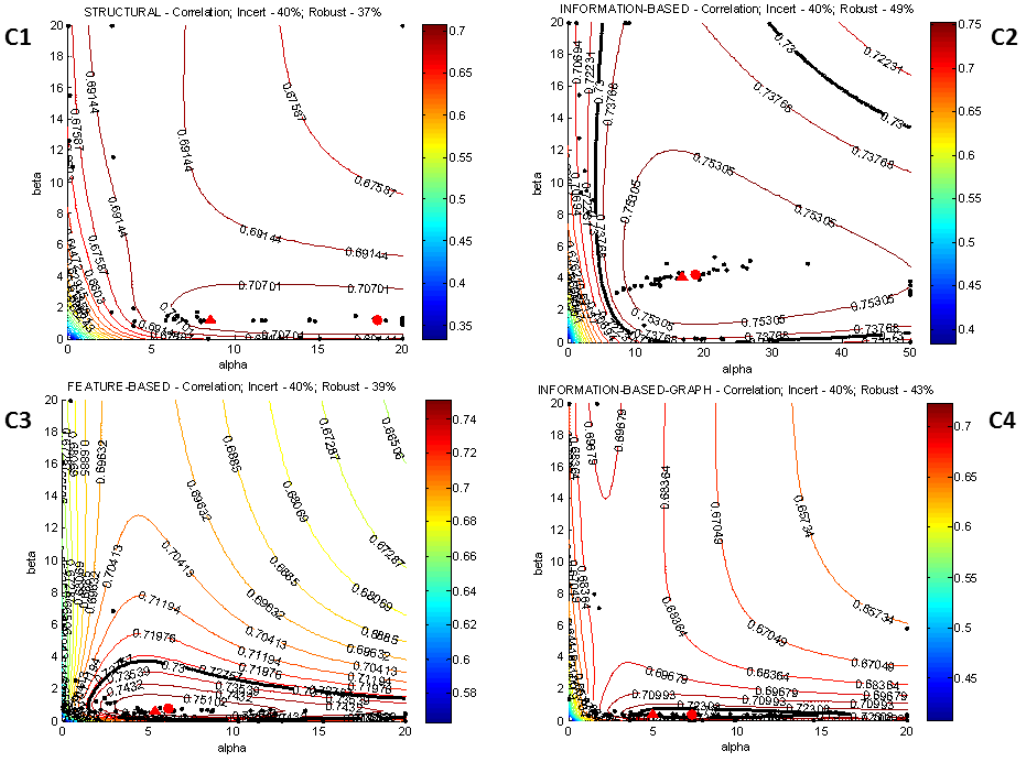

FIGURE C.6: Plot of robustness of parametric semantic similarity measures considering 40% of uncertainty

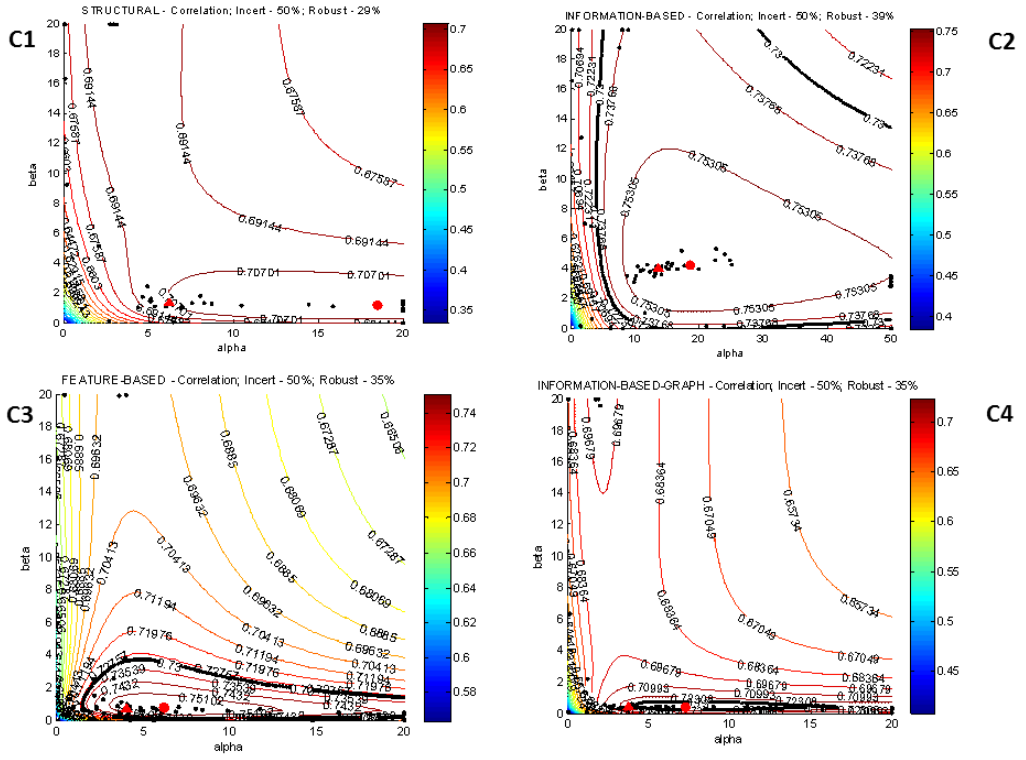

FIGURE C.7: Plot of robustness of parametric semantic similarity measures considering 50% of uncertainty

C.3 Illustration the algorithm presented in Chapter 7

The process of the algorithm presented in Section 7.2.2.2 is graphically illustrated based on the taxonomy presented in Figure C.8. The concept $\omega^*(u, v)$ refers to MSCA of the concepts u, v. In this example, we consider the following T_θ ordering:

$$T_\theta = [C, B, A, D, E, F, R]$$

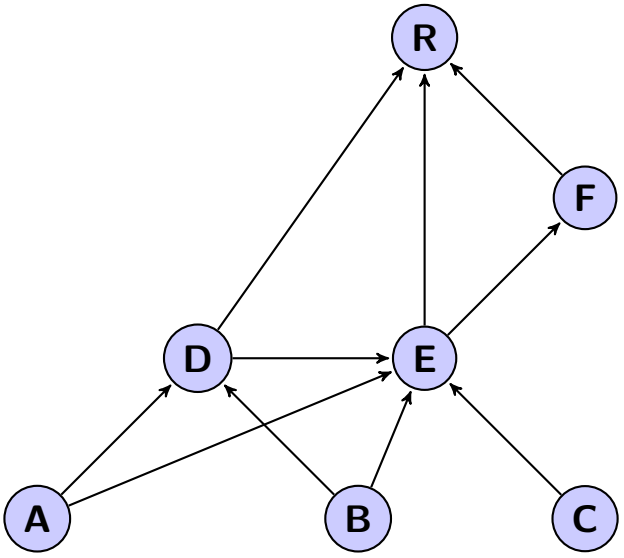

FIGURE C.8: Taxonomy used to illustrate the algorithm

Iteration $i = 0$; $c = C$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	$\{C\}$	$\{\}$
1	B	*null*	*null*
2	A	*null*	*null*
3	D	*null*	*null*
4	E	$\{C\}$	*null*
5	F	*null*	*null*
6	R	*null*	*null*

$$\omega^* \begin{matrix} \\ \cancel{\mathbf{C}} \\ B \\ A \\ D \\ E \\ F \\ R \end{matrix} \begin{matrix} C & B & A & D & E & F & R \end{matrix} \begin{pmatrix} C & . & . & . & . & . & . \\ ? & ? & . & . & . & . & . \\ ? & ? & ? & . & . & . & . \\ ? & ? & ? & ? & . & . & . \\ ? & ? & ? & ? & ? & . & . \\ ? & ? & ? & ? & ? & ? & . \\ ? & ? & ? & ? & ? & ? & ? \end{pmatrix}$$

Iteration $i = 1$; $c = B$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	*null*	$\{\}$
1	**B**	$\{B\}$	$\{C\}$
2	A	*null*	*null*
3	D	$\{B\}$	*null*
4	E	$\{C, B\}$	*null*
5	F	*null*	*null*
6	R	*null*	*null*

$$\omega^* \begin{matrix} \\ \cancel{\mathbf{C}} \\ \mathbf{B} \\ A \\ D \\ E \\ F \\ R \end{matrix} \begin{matrix} C & B & A & D & E & F & R \end{matrix} \begin{pmatrix} C & . & . & . & . & . & . \\ ? & B & . & . & . & . & . \\ ? & ? & ? & . & . & . & . \\ ? & ? & ? & ? & . & . & . \\ ? & ? & ? & ? & ? & . & . \\ ? & ? & ? & ? & ? & ? & . \\ ? & ? & ? & ? & ? & ? & ? \end{pmatrix}$$

Iteration $i = 2$; $c = A$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	*null*	$\{\}$
1	**B**	*null*	$\{C\}$
2	**A**	$\{A\}$	$\{C, B\}$
3	D	$\{B, A\}$	*null*
4	E	$\{C, B, A\}$	*null*
5	F	*null*	*null*
6	R	*null*	*null*

$$\omega^* \begin{matrix} \\ \cancel{\mathbf{C}} \\ \mathbf{B} \\ \mathbf{A} \\ D \\ E \\ F \\ R \end{matrix} \begin{matrix} C & B & A & D & E & F & R \end{matrix} \begin{pmatrix} C & . & . & . & . & . & . \\ ? & B & . & . & . & . & . \\ ? & ? & A & . & . & . & . \\ ? & ? & ? & ? & . & . & . \\ ? & ? & ? & ? & ? & . & . \\ ? & ? & ? & ? & ? & ? & . \\ ? & ? & ? & ? & ? & ? & ? \end{pmatrix}$$

Iteration $i = 3$ **;** $c = D$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	$null$	$\{\}$
1	**B**	$null$	$\{C\}$
2	**A**	$null$	$\{C\}$
3	**D**	$\{B, A, D\}$	$\{C\}$
4	E	$\{C, B, A, D\}$	$null$
5	F	$null$	$null$
6	R	$\{B, A, D\}$	$null$

ω^*	C	B	A	D	E	F	R
~~**C**~~	C	.	.	.	.	.	.
B	?	B	.	.	.	.	.
A	?	D	A	.	.	.	.
D	?	D	D	D	.	.	.
E	?	?	?	?	?	.	.
F	?	?	?	?	?	?	.
R	?	?	?	?	?	?	?

Iteration $i = 4$ **;** $c = E$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	$null$	$\{\}$
1	**B**	$null$	$\{\}$
2	**A**	$null$	$\{\}$
3	**D**	$null$	$\{\}$
4	**E**	$\{C, B, A, D, E\}$	$\{\}$
5	F	$\{C, B, A, D, E\}$	$null$
6	R	$\{B, A, D, C, E\}$	$null$

ω^*	C	B	A	D	E	F	R
~~**C**~~	C	.	.	.	.	.	.
~~**B**~~	E	B	.	.	.	.	.
~~**A**~~	E	D	A	.	.	.	.
~~**D**~~	E	D	D	D	.	.	.
~~**E**~~	E	E	E	E	E	.	.
F	?	?	?	?	?	?	.
R	?	?	?	?	?	?	?

Iteration $i = 5$ **;** $c = F$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	$null$	$\{\}$
1	**B**	$null$	$\{\}$
2	**A**	$null$	$\{\}$
3	**D**	$null$	$\{\}$
4	**E**	$null$	$\{\}$
5	**F**	$\{C, B, A, D, E, F\}$	$\{\}$
6	R	$\{B, A, D, C, E, F\}$	$null$

ω^*	C	B	A	D	E	F	R
~~**C**~~	C	.	.	.	.	.	.
~~**B**~~	E	B	.	.	.	.	.
~~**A**~~	E	D	A	.	.	.	.
~~**D**~~	E	D	D	D	.	.	.
~~**E**~~	E	E	E	E	E	.	.
~~**F**~~	F	F	F	F	F	F	.
R	?	?	?	?	?	?	?

Iteration $i = 6$ **;** $c = R$

i	T_θ	$mapDesc$	$previousNotDesc$
0	**C**	$null$	$\{\}$
1	**B**	$null$	$\{\}$
2	**A**	$null$	$\{\}$
3	**D**	$null$	$\{\}$
4	**E**	$null$	$\{\}$
5	**F**	$null$	$\{\}$
6	**R**	$\{B, A, D, C, E, F, R\}$	$\{\}$

ω^*	C	B	A	D	E	F	R
~~**C**~~	C	.	.	.	.	.	.
~~**B**~~	E	B	.	.	.	.	.
~~**A**~~	E	D	A	.	.	.	.
~~**D**~~	E	D	D	D	.	.	.
~~**E**~~	E	E	E	E	E	.	.
~~**F**~~	F	F	F	F	F	F	.
~~**R**~~	R	R	R	R	R	R	R

Iteration $i = 7$

$i > |T_\theta| \implies$ **Algorithm complete**.

Bibliography

ACL (2013). http://www.aclweb.org/aclwiki/index.php?title=Distributional_Hypothesis.

Aimé, X. (2011). *Gradients de prototypicalité, mesures de similarité et de proximité sémantique: une contribution à l'Ingénierie des Ontologies*. PhD thesis, Université de Nantes (France).

Akoka, J., Liddle, S. W., Song, I.-Y., Bertolotto, M., Comyn-Wattiau, I., Heuvel, W.-J., Kolp, M., Trujillo, J., Kop, C., and Mayr, H. C., editors (2005). *Perspectives in Conceptual Modeling*, volume 3770 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Al-Mubaid, H. and Nagar, A. (2008). Comparison of four similarity measures based on GO annotations for Gene Clustering. In *2008 IEEE Symposium on Computers and Communications*, number 3, pages 531–536. IEEE.

Al-Mubaid, H. and Nguyen, H. A. (2006). A cluster-based approach for semantic similarity in the biomedical domain. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society.*, volume 1, pages 2713–7.

Al-mubaid, H. and Nguyen, H. A. (2009). Measuring Semantic Similarity Between Biomedical Concepts Within Multiple Ontologies. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions*, 39(4):389–398.

Albertoni, R. and De Martino, M. (2006). Semantic similarity of ontology instances tailored on the application context. In *On the Move to Meaningful Internet Systems 2006: CoopIS, DOA, GADA, and ODBASE*, volume 4275 of *Lecture Notes in Computer Science*, pages 1020–1038, Berlin, Heidelberg. Springer.

Ali, W. and Deane, C. M. (2009). Functionally guided alignment of protein interaction networks for module detection. *Bioinformatics (Oxford, England)*, 25(23):3166–73.

Alvarez, M., Qi, X., and Yan, C. (2011). A shortest-path graph kernel for estimating gene product semantic similarity. *Journal of biomedical semantics*, 2:3.

Alvarez, M. and Yan, C. (2011). A graph-based semantic similarity measure for the gene ontology. *Journal of bioinformatics and computational biology*, 9(6):681–95.

Anandan, B. and Clifton, C. (2011). Significance of Term Relationships on Anonymization. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, volume 3, pages 253–256. IEEE.

Andrea Rodríguez, M. and Egenhofer, M. J. (2004). Comparing geospatial entity classes: an asymmetric and context-dependent similarity measure. *International Journal of Geographical Information Science*, 18(3):229–256.

Andrejko, A. and Bieliková, M. (2013). Comparing Instances of Ontological Concepts for Personalized Recommendation in Large Information Spaces. *Computing and Informatics*, 28(4):429–452.

Anna, F. (2008). Concept similarity in Formal Concept Analysis: An information content approach. *Knowledge-Based Systems*, 21(1):80–87.

Araújo, R. and Pinto, H. S. (2007). SEMilarity: Towards a Model-Driven Approach to Similarity. In *International Workshop on Description Logics (DL)*, pages 155–162. Bolzano University Press.

Araujo, S., Hidders, J., Schwabe, D., and de Vries, A. P. (2011). SERIMI - Resource Description Similarity, RDF Instance Matching and Interlinking. *arXiv preprint arXiv:1107.1104*.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., and ... (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29.

Ashby, F. G. and Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological review*, 95(1):124–150.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: a nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference*, pages 722–735. Springer-Verlag.

Azuaje, F., Wang, H., and Bodenreider, O. (2005). Ontology-driven similarity approaches to supporting gene functional assessment. In *Proceedings of the ISMB'2005 SIG meeting on Bio-ontologies*, pages 9–10.

Baader, F., Calvanese, D., McGuinness, D., Nardi, D., and Patel-Schneider, P., editors (2010). *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2nd edition.

Ballatore, A., Bertolotto, M., and Wilson, D. C. (2012). Geographic knowledge extraction and semantic similarity in OpenStreetMap. *Knowledge and Information Systems*, 37(1):61–81.

Banerjee, S. and Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, volume 2276, pages 136–145. Springer Berlin Heidelberg.

Banerjee, S. and Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence*, pages 805–810, Acapulco (Mexico).

Batet, M. (2011a). *A study on semantic similarity and its application to clustering*. VDM Verlag Dr. Müller.

Batet, M. (2011b). Ontology-based semantic clustering. *AI Commun.*, 24(3):291–292.

Batet, M., Sánchez, D., and Valls, A. (2010a). An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*, 4(1):39–52.

Batet, M., Sánchez, D., Valls, A., and Gibert, K. (2010b). Exploiting Taxonomical Knowledge to Compute Semantic Similarity: An Evaluation in the Biomedical Domain. In *Lecture Notes in Computer Science*, volume 6096/2010, pages 274–283.

Batet, M., Sánchez, D., Valls, A., and Gibert, K. (2013). Semantic similarity estimation from multiple ontologies. *Applied Intelligence*, 38(1):29–44.

Baumann, S. and Schirru, R. (2012). Using Linked Open Data for Novel Artist Recommendations. In *13th Internal Society for Music Information Retrival Conference*, Porto (Portugal).

Baziz, M., Boughanem, M., Pasi, G., and Prade, H. (2007). An Information Retrieval Driven by Ontology: from Query to Document Expansion. In *RIAO '07 Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 301–313.

Bell, D. E., Raiffa, H., and Tversky, A., editors (1988). *Decision making: Descriptive, normative, and prescriptive interactions*. Cambridge University Press.

Bellinger, G., Castro, D., and Mills, A. (2004). Data, information, knowledge, and Wisdom. http://www.systems-thinking.org/dikw/dikw.htm.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). Robust optimization. Princeton series in applied mathematics. *Daubechies, E. Weinan, JK Lenstra, and E. Suli, Editors*, page 568.

Benabderrahmane, S., Devignes, M.-D., Smail-Tabbone, M., Poch, O., and Napoli, A. (2010a). IntelliGO: Towards a New Synthetic Semantic Similarity Measure by Considering Metadata of Gene Functional Annotations Quality. Technical report, Université Henri Poincaré - Nancy I.

Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., and Devignes, M.-D. (2010b). IntelliGO: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics*, 11(1):588.

Bender, M. A., Farach-Colton, M., Pemmasani, G., Skiena, S., and Sumazin, P. (2005). Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms*, 57(2):75–94.

Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, pages 29–37.

Bernstein, A., Kaufmann, E., Kiefer, C., and Bürki, C. (2005). SimPack: A Generic Java Library for Similarity Measures in Ontologies. Technical report, University of Zurich Department of Informatics.

Bin, S., Liying, F., Jianzhuo, Y., Pu, W., and Zhongcheng, Z. (2009). Ontology-based Measure of Semantic Similarity between Concepts. In *Software Engineering, 2009. WCSE'09. WRI World Congress*, pages 109–112. IEEE.

BIPM, IEC, IFCC, ILAC, IUPAC, IUPAP, ISO, and OIML (2012). International vocabulary of metrology – Basic and general concepts and associated terms (VIM) JCGM 200:2012.

Bisson, G. (1992). Conceptual Clustering in a First Order Logic Representation. In *ECAI*, volume 92, pages 458–462.

Bisson, G. (1995). Why and How to Define a Similarity Measure for Object Based Representation Systems. In Mars, N. J. I., editor, *Towards Very Large Knowledge Bases*, pages 236–246. IOS Press, 1st edition.

Blanchard, E. (2008). *Exploitation d'une hiérarchie de subsomption par le biais de mesures sémantiques*. PhD thesis, Université de Nantes.

Blanchard, E. and Harzallah, M. (2005). A typology of ontology-based semantic measures. *EMOI-INTEROP'05, Proc. Open Interop Workshop on Enterprise Modelling and Ontologies for Interoperability*.

Blanchard, E., Harzallah, M., and Kuntz, P. (2008). A generic framework for comparing semantic similarities on a subsumption hierarchy. *18th European Conference on Artificial Intelligence*, pages 20–24.

Blanchard, E., Kuntz, P., Harzallah, M., and Briand, H. (2006). A tree-based similarity for evaluating concept proximities in an ontology. In *10th conference of the International Fedederation of Classification Societies*, pages 3–11. Springer, Ljubljana, Slovenia.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Bodenreider, O., Aubry, M., and Burgun, A. (2005). Non-lexical approaches to identifying associative relations in the gene ontology. In *Pacific Symposium on Biocomputing.*, volume 102, page 91. NIH Public Access.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2009). A relational model of semantic similarity between words using automatically extracted lexical pattern clusters from the web. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 2 - EMNLP '09*, page 803.

Borgida, A., Walsh, T., and Hirsh, H. (2005). Towards measuring similarity in description logics. In *International Workshop on Description Logics (DL2005)*, volume 147, Edinburgh, Scotland.

Borst, W. N. (1997). *Construction of engineering ontologies for knowledge sharing and reuse*. Phd thesis, University of Twente (Enschede, Netherlands).

Botafogo, R. A., Rivlin, E., and Shneiderman, B. E. N. (1992). Structural Analysis of Hypertexts: Hierarchies and Useful Metrics Identifying. *Human-Computer Interaction*, 10(2):142–180.

Bouchon-Meunier, B., Rifqi, M., and Bothorel, S. (1996). Towards general measures of comparison of objects. *Fuzzy Sets and Systems*, 84(2):143–153.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen.

Bradshaw, J. (1997). Introduction to Tversky similarity measure. In *11th Annual Daylight User Group Meeting (MUG'97)*, Verona (Italy).

Broekstra, J., Kampman, A., and Harmelen, F. V. (2002). Sesame: A generic architecture for storing and querying rdf and rdf schema. In Horrocks, I. and Hendler, J., editors, *The Semantic Web — ISWC 2002*, pages 54–68. Springer Berlin Heidelberg.

Bulskov, H., Knappe, R., and Andreasen, T. (2002). On Measuring Similarity for Conceptual Querying. In *Proceedings of the 5th International Conference on Flexible Query Answering Systems*, volume 1, pages 100–111, London, UK. Springer-Verlag.

Burke, R. (2002). Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

Caillez, F. and Kuntz, P. (1996). A contribution to the study of the metric and Euclidean structures of dissimilarities. *Psychometrika*, 61(2):241–253.

Cazzanti, L. and Gupta, M. (2006). Information-theoretic and Set-theoretic Similarity. In *2006 IEEE International Symposium on Information Theory*, pages 1836–1840. IEEE.

Celma, O. and Serra, X. (2008). FOAFing the music: Bridging the semantic gap in music recommendation. *Web Semantics Science Services and Agents on the World Wide Web*, 6(4):250–256.

Cha, S. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):300–307.

Chabalier, J., Mosser, J., and Burgun, A. (2007). A transversal approach to predict gene product networks from ontology-based similarity. *BMC Bioinformatics*, 8:235.

Chater, N. and Hahn, U. (1997). Representational distortion, similarity and the universal law of generalization. In *SimCat97: Proceedings of the Interdisciplinary Workshop on Similarity and Categorization*, pages 31–36, Edinburgh University, Edinburgh. Department of Artificial Intelligence, Edinburgh University.

Chebotarev, P. and Shamis, E. (2006a). A Matrix-Forest Theorem and Measuring Relations in Small Social Group. *Autom. Remote Control*, 58(9):1505–1514.

Chebotarev, P. and Shamis, E. (2006b). On proximity measures for graph vertices. *Autom. Remote Control*, 59(10):1443–1459.

Chein, M. and Mugnier, M.-L. (2009). *Graph-based knowledge representation: computational foundations of conceptual graphs*. Springer; 2009 edition (October 21, 2008).

Cheng, G., Ge, W., and Qu, Y. (2008). Falcons: searching and browsing entities on the semantic web. In *Proceedings of the 17th international conference on World Wide Web*, pages 1101–1102. ACM.

Cho, M., Choi, J., and Kim, P. (2003). An efficient computational method for measuring similarity between two conceptual entities. *Lecture notes in computer science*, pages 381–388.

Cho, Y.-R., Hwang, W., Ramanathan, M., and Zhang, A. (2007). Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinformatics*, 8:265.

Choi, S.-s., Cha, S.-h., and Tappert, C. C. (2010). A Survey of Binary Similarity and Distance Measures. *Journal on Systemics, Cybernetics and Informatics*, 0(1):43–48.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Coates, J., Durance, P., Godet, M., Aaltonen, M., and Holmström, J. (2010). Multi-ontology topology of the strategic landscape in three practical cases. *Technological Forecasting and Social Change*, 77(9):1519–1526.

Collins, A. M. and Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407.

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2):240–247.

Corby, O., Dieng-Kuntz, R., Gandon, F., and Faron-Zucker, C. (2006). Searching the semantic web: Approximate query processing based on ontologies. *Intelligent Systems, IEEE*, 21(1):20–27.

Couto, F. M., Silva, M., and Coutinho, P. M. (2003). Implementation of a Functional Semantic Similarity Measure between Gene-Products. Technical Report DI/FCUL TR 03–29, Department of Informatics, University of Lisbon.

Couto, F. M. and Silva, M. J. (2011). Disjunctive Shared Information between Ontology Concepts: application to Gene Ontology. *Journal of Biomedical Semantics*, 2(1):5.

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2005). Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors. In *Conference in Information and Knowledge Management*, pages 343–344. ACM.

Couto, F. M., Silva, M. J., and Coutinho, P. M. (2007). Measuring semantic similarity between Gene Ontology terms. *Data & Knowledge Engineering*, 61(1):137–152.

Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E., Apweiler, R., Kirsch, H., and Rebholz-Schuhmann, D. (2006). GOAnnotator: linking protein GO annotations to evidence text. *Journal of Biomedical Discovery and Collaboration*, 1(1):19.

Cross, V. (2004). Fuzzy semantic distance measures between ontological concepts. In *IEEE Annual Meeting of the Fuzzy Information, 2004. NAFIPS'04*, volume 2, pages 635–640 Vol.2. Ieee.

Cross, V. (2006). Tversky's Parameterized Similarity Ratio Model: A Basis for Semantic Relatedness. In *Fuzzy Information Processing Society, 2006. NAFIPS 2006. Annual meeting of the North American*, pages 541–546, Montreal, Quebec.

Cross, V. and Sun, Y. (2007). Semantic, fuzzy set and fuzzy measure similarity for the gene ontology. *2007 IEEE International Conference on Fuzzy Systems Vols 14*, pages 1951–1956 2075.

Cross, V. and Yu, X. (2010). A Fuzzy Set Framework for Ontological Similarity Measures. *Computational Intelligence*, pages 18–23.

Cross, V. and Yu, X. (2011). Investigating Ontological Similarity Theoretically with Fuzzy Set Theory, Information Content, and Tversky Similarity and Empirically with the Gene Ontology. In Benferhat, S. and Grant, J., editors, *Scalable Uncertainty Management*, volume 6929 of *Lecture Notes in Computer Science*, pages 387–400. Springer Berlin Heidelberg.

Cross, V., Yu, X., and Hu, X. (2013). Unifying ontological similarity measures: A theoretical and empirical investigation. *International Journal of Approximate Reasoning*, 54(7):861–875.

Curran, J. R. (2004). *From distributional to semantic similarity.* Phd thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

Cyganiak, R. and Jentzsch, A. (2011). http://lod-cloud.net/.

Czumaj, A., Kowaluk, M., and Lingas, A. (2007). Faster algorithms for finding lowest common ancestors in directed acyclic graphs. *Theoretical Computer Science*, 380(1):37–46.

D'Amato, C. (2007). *Similarity-based Learning Methods for the Semantic Web.* Phd thesis, Università degli Studi di Bari (Italy).

D'Amato, C., Fanizzi, N., and Esposito, F. (2005a). A Semantic Dissimilarity Measure for Concept Descriptions in Ontological Knowledge Bases in. *The Second International Workshop on Knowledge Discovery and Ontologies.*

D'Amato, C., Fanizzi, N., and Esposito, F. (2005b). A semantic similarity measure for expressive description logics. *Proceedings of Convegno Italiano di Logica Computazionale CILC05.*

D'Amato, C., Staab, S., and Fanizzi, N. (2008). On the influence of description logics ontologies on conceptual similarity. In *Knowledge Engineering: Practice and Patterns*, pages 48–63. Springer.

D'Aquin, M. and Motta, E. (2011). Watson, more than a semantic web search engine. *Semantic Web*, 2(1):55–63.

D'Aquin, M., Motta, E., Sabou, M., Angeletou, S., Gridinoc, L., Lopez, V., and Guidi, D. (2008). Toward a new generation of semantic web applications. *Intelligent Systems, IEEE*, 23(3):20–28.

David, J. and Euzenat, J. (2008). Comparison between Ontology Distances (Preliminary Results). pages 245–260.

Davis, R., Shrobe, H., and Szolovits, P. (1993). What is a knowledge representation? *AI magazine*, 14(1):17.

De Giacomo, G. and Lenzerini, M. (1996). TBox and ABox reasoning in expressive description logics. In *Proceedings of the Fifth International Conference on Principles of Knowledge Representation and Reasoning (KR'96)*, pages 316–327, Cambridge, Massachusetts, USA. Morgan Kaufmann.

Delugach, H. (1993). An exploration into semantic distance. In Pfeiffer, H. and Nagle, T., editors, *Conceptual Structures: Theory and Implementation*, volume 754 of *Lecture Notes in Computer Science*, pages 119–124. Springer Berlin Heidelberg.

Deza, M. M. and Deza, E. (2013). *Encyclopedia of distances*. Springer Berlin Heidelberg, 2nd edition.

Diaz-Diaz, N. and Aguilar-Ruiz, J. S. (2011). GO-based Functional Dissimilarity of Gene Sets. *BMC Bioinformatics*, 12(1):360.

Dinu, G. (2011). *Word meaning in context: A probabilistic model and its application to Question Answering*. PhD thesis, Saarlan University.

Du, Z., Li, L., Chen, C.-F., Yu, P. S., and Wang, J. Z. (2009). G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic acids research*, 37(Web Server issue):W345–9.

Ducournau, R., Euzenat, J., Masini, G., and Napoli, A. (1998). *Langages et modèles à objets État des recherches et perspectives*, volume 1.

Eco, U. (1989). *Foucault's Pendulum*. Secker & Warburg.

Ehrig, M., Haase, P., Hefke, M., and Stojanovic, N. (2004). Similarity for Ontologies - a Comprehensive Framework. In *Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM*.

Eronen, L. M. and Toivonen, H. T. (2012). Biomine: predicting links between biological entities using network models of heterogeneous databases. *BMC bioinformatics*, 13(1):119.

Euzenat, J., Loup, D., Touzani, M., and Valtchev, P. (2004). Ontology alignment with OLA. In *Proc. 3rd ISWC2004 workshop on Evaluation of Ontology-based tools (EON)*, pages 59–68.

Euzenat, J. and Shvaiko, P. (2013). *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2nd edition.

Euzenat, J. and Valtchev, P. (2004). Similarity-based ontology alignment in OWL-Lite. In *ECAI*, pages 333–337.

Fanizzi, N. and D'Amato, C. (2006). A similarity measure for the aln description logic. *Proceedings of CILC 2006 - Italian Conference on Computational Logic*, pages 26–27.

Fellbaum, C. (2010). *WordNet*. Springer.

Fernando, S. and Stevenson, M. (2008). A semantic similarity approach to paraphrase detection. *Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloqium*.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Foo, N., Garner, B. J., Rao, A., and Tsui, E. (1992). Semantic distance in conceptual graphs. *Conceptual Structures: Current Research and Practice*, pages 149–154.

Fouss, F., Pirotte, A., Renders, J.-m., and Saerens, M. (2007). Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369.

Freitas, A., Oliveira, J. a. G., O'Riain, S., Curry, E., and Da Silva, J. a. C. P. (2011). Querying Linked Data Using Semantic Relatedness: A Vocabulary Independent Approach. In Muñoz, Rafael and Montoyo, Andrés and Métais, E., editor, *Natural Language Processing and Information Systems*, pages 40–51. Springer Berlin Heidelberg.

Gabbay, D. M., Hogger, C. J., Robinson, J. A., Siekmann, J., Nute, D., and Galton, A. (1998). *Handbook of logic in artificial intelligence and logic programming*. Clarendon Press.

Gandon, F., Corby, O., Dieng-Kuntz, R., and Giboin, A. (2005). Proximité conceptuelle et distances de graphes. *Proc. Raisonner le Web Sémantique avec des Graphes - Journée thématique de la plate-forme AFIA*.

Gandon, F., Corby, O., and Faron-Zucker, C. (2012). *Le Web sémantique: comment lier les données et les schémas sur le web?* Dunod.

Ganesan, P., Garcia-Molina, H., and Widom, J. (2003). Exploiting hierarchical domain structure to compute similarity. *ACM Transactions on Information Systems*, 21(1):64–93.

Ganesan, V., Swaminathan, R., and Thenmozhi, M. (2012). Similarity Measure Based On Edge Counting Using Ontology. *International Journal of Engineering Research and Development*, 3(3):40–44.

Gardner, B., Lukose, D., and Tsui, E. (1987). Parsing Natural Language through Pattern Correlation and Modification. In *Proceedings of the 7th International Workshop on Expert Systems & Their Applications*, pages 1285–1299, Avignon (France).

Garla, V. N. and Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics*, 13(1):261.

Gentleman, R. (2007). Visualizing and distances using GO. *Retrieved Jan. 10th*.

Gentleman, R., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):R80.

Gentner, D. (2001). Exhuming similarity. *Behavioral and Brain Sciences*, 24(04):669.

Gentner, D. and Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological science*, 5(3):152–158.

Gentner, D. and Markman, A. B. (1997). Structure mapping in analogy and similarity. *American psychologist*, 52(1):45.

Goldstone, R. L. (1994a). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(1):3.

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, 52(2):125–157.

Goldstone, R. L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(4):988.

Goldstone, R. L. and Son, J. Y. (2004). Similarity. *Psychological Review*, 100:254–278.

Goodman, N. (1972). *Problems and projects*. Bobbs-Merrill Indianapolis.

Gower, J. C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3(1):5–48.

Grau, B. C., Dragisic, Z., Eckert, K., Euzenat, J., Ferrara, A., Granada, R., Ivanova, V., Jiménez-Ruiz, E., Kempf, A. O., and Lambrix, P. (2013). Results of the Ontology Alignment Evaluation Initiative 2013. In *Proc. 8th ISWC workshop on ontology matching (OM)*, pages 61–100.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5.2(April):199–220.

Guarino, N. and Giaretta, P. (1995). Ontologies and knowledge bases - towards a terminological clarification. *Towards very large knowledge bases: knowledge building & knowledge sharing 1995*, pages 25–32.

Guarino, N., Oberle, D., and Staab, S. (2009). What is an Ontology? In *Handbook on ontologies*, pages 1–17. Springer.

Guo, X., Liu, R., Shriver, C. D., Hu, H., and Liebman, M. N. (2006). Assessing semantic similarity measures for the characterization of human regulatory pathways. *Bioinformatics (Oxford, England)*, 22(8):967–73.

Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M. (2012). Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in Bioinformatics*, 13(5):569–585.

Hagedoorn, M. (2000). *Pattern matching using similarity measures*. PhD thesis, Utrecht University (The Netherlands).

Hahn, U. (2011). What makes things similar? (Invited speacker): http://videolectures.net/simbad2011_hahn_similar/. In *1st International Workshop on Similarity-Based Pattern Analysis and Recognition*.

Hahn, U., Chater, N., and Richardson, L. B. (2003). Similarity as transformation. *Cognition*, 87(1):1–32.

Hahn, U. and Ramscar, M. (2001). Conclusion: Mere similarity. *Similarity and categorization*, pages 257–272.

Hall, M. (2006). A Semantic Similarity Measure for Formal Ontologies. Technical report, Fakultat fur Wirtschaftswissenschaften und Informatik, Alpen-Adria Universitat Klagenfurt, Klagenfurt, Austria.

Harary, F., Norman, R. Z., and Cartwright, D. (1965). *Structural models: An introduction to the theory of directed graphs*. John Willey & Sons, Inc.

Harel, D. and Tarjan, R. E. (1984). Fast algorithms for finding nearest common ancestors. *SIAM Journal on Computing*, 13(2):338–355.

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013a). Mesures Sémantiques basées sur la notion de projection RDF pour les systèmes de recommandation. In *IC 2013: journées francophones d'ingénierie des connaissances - In press.*

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013b). Semantic Measures Based on RDF Projections: Application to Content-Based Recommendation Systems. In *On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, pages 606–615. Springer Berlin Heidelberg, Graz (Austria).

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2013c). Semantic Measures for the Comparison of Units of Language, Concepts or Entities from Text and Knowledge Base Analysis. *ArXiv*, 1310.1285.

Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2014). The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5):740–742.

Harispe, S., Sánchez, D., Ranwez, S., Janaqi, S., and Montmain, J. (2013d). A Framework for Unifying Ontology-based Semantic Similarity Measures: a Study in the Biomedical Domain. *Journal of Biomedical Informatics*, In press.

Harris, Z. S. (1981). *Distributional structure*. Springer.

Heath, T. and Bizer, C. (2011). Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136.

Heitmann, B. and Hayes, C. (2010). Using linked data to build open, collaborative recommender systems. *Artificial Intelligence*, pages 76–81.

Hey, J. (2004). The Data, Information, Knowledge, Wisdom Chain: The Metaphorical link. Technical report.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An electronic lexical database.*, pages 305–332, Cambridge, MA. MIT Press.

Hitzler, P., Krotzsch, M., and Rudolph, S. (2011). *Foundations of semantic web technologies*. Chapman and Hall/CRC.

Hliaoutakis, A. (2005). *Semantic similarity measures in MeSH ontology and their application to information retrieval on Medline*. Master's thesis, Technical University of Crete, Greek.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E., and Milios, E. (2006). Information retrieval by semantic similarity. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2(3):55–73.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194:28–61.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., and Thagard, P. R. (1989). *Induction: Processes of inference, learning, and discovery*. A Bradford Book.

Holyoak, K. J. and Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15(4):332–340.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C., and Lempicki, R. a. (2007). The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome biology*, 8(9):R183.

Hughes, T. and Ramage, D. (2007). Lexical Semantic Relatedness with Random Graph Walks. *Computational Linguistics*, 7(June):581–589.

Iordanskaja, L., Kittredge, R., and Polguere, A. (1991). Lexical selection and paraphrase in a meaning-text generation model. In *Natural language generation in artificial intelligence and computational linguistics*, pages 293–312. Springer.

Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.

Jain, S. and Bader, G. D. (2010). An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562.

Janaqi, S., Aguilera, J., and Chèbre, M. (2013). Robust real-time optimization for the linear oil blending. *RAIRO - Operations Research*, 2013(47):465–479.

Janowicz, K. (2006). Sim-dl: Towards a semantic similarity measurement theory for the description logic ALCNR in geographic information retrieval. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 1681–1692. Springer.

Janowicz, K., Raubal, M., and Kuhn, W. (2011). The semantics of similarity in geographic information retrieval. *Journal of Spatial Information Science*, 2:29–57.

Janowicz, K., Raubal, M., Schwering, A., and Kuhn, W. (2008). Semantic Similarity Measurement and Geospatial Applications. *Transactions in GIS*, 12(6):651–659.

Janowicz, K. and Wilkes, M. (2009). Sim-dla: A novel semantic similarity measure for description logics reducing inter-concept to inter-instance similarity. In *The Semantic Web: Research and Applications*, pages 353–367. Springer.

Jeh, G. and Widom, J. (2002). SimRank. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data mining*, page 538, New York, USA. ACM Press.

Jiang, J. and Conrath, D. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *In International Conference Research on Computational Linguistics (ROCLING X)*, pages 19–33.

Jimeno-Yepes, A., Jiménez-Ruiz, E., Berlanga-Llavori, R., and Rebholz-Schuhmann, D. (2009). Reuse of terminological resources for efficient ontological engineering in Life Sciences. *BMC bioinformatics*, 10(Suppl 10):S4.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Keeney, R. L. (1993). *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge University Press, Cambridge, UK.

Kiefer, C., Bernstein, A., and Stocker, M. (2007). The Fundamentals of iSPARQL - A Virtual Triple Approach For Similarity-Based Semantic Web Tasks. *The Semantic Web*, 4825:295–309.

Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., and Lee, R. (2009). Media meets semantic web–how the bbc uses dbpedia and linked data to make connections. In *The semantic web: research and applications*, pages 723–737. Springer.

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*, 85(4):457–64.

Kondor, R. I. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, pages 315–322.

Kozima, H. (1993). Text segmentation based on similarity between words. *ACL*, pages 286–288.

Lao, N. (2012). *Efficient Random Walk Inference with Knowledge Bases*. PhD thesis, Pennsylvania State University.

Leacock, C. and Chodorow, M. (1994). Filling in a sparse training space for word sense identification.

Leacock, C. and Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In Fellbaum, C., editor, *WordNet: An electronic lexical database.*, chapter 13, pages 265–283. MIT Press.

Lee, J. H., Kim, M. H., and Lee, Y. J. (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of Documentation*, 49(2):188–207.

Lee, W.-N., Shah, N., Sundlass, K., and Musen, M. (2008). Comparison of ontology-based semantic-similarity measures. In *AMIA - Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, pages 384–8.

Lehmann, K. and Turhan, A.-Y. (2012). A Framework for Semantic-Based Similarity Measures for ELH-Concepts. In *Logics in Artificial Intelligence*, pages 307–319. Springer.

Li, B., Wang, J. Z., Feltus, F. A., Zhou, J., and Luo, F. (2010). Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins. *Gene*, pages 1–54.

Li, J., Gong, B., Chen, X., Liu, T., Wu, C., Zhang, F., Li, C., Li, X., Rao, S., and Li, X. (2011). DOSim: An R package for similarity between diseases based on Disease Ontology. *BMC Bioinformatics*, 12(1):266.

Li, M. and Vitányi, P. (1993). *An introduction to Kolmogorov complexity and its applications. (Texts in Computer Science)*. Springer, New York, 3rd edition.

Li, Y., Bandar, Z. A., and McLean, D. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):871–882.

Li, Y., McLean, D., Bandar, Z., O'Shea, J., and Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, 18(8):1138–1150.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. In *15th International Conference of Machine Learning*, pages 296–304, Madison,WI.

Lintean, M. C., Moldovan, C., Rus, V., and McNamara, D. S. (2010). The Role of Local and Global Weighting in Assessing the Semantic Similarity of Texts Using Latent Semantic Analysis. In *FLAIRS Conference*.

Lord, P. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Maedche, A. and Staab, S. (2001). Comparing Ontologies - Similarity Measures and a Comparison Study (Internal Report). Technical report, Institute AIFB, University of Karlsruhe, Germany, Karlsruhe.

Maguitman, A. G. and Menczer, F. (2005). Algorithmic detection of semantic similarity. In *WWW '05: Proceedings of the 14th International Conference on World Wide Web*, pages 107–116, New York, USA. ACM Press.

Maguitman, A. G., Menczer, F., Erdinc, F., Roinestad, H., and Vespignani, A. (2006). Algorithmic Computation and Approximation of Semantic Similarity. *World Wide Web*, 9(4):431–456.

Mao, W. and Chu, W. W. (2002). Free-text medical document retrieval via phrase-based vector space model. In *AMIA Symposium. American Medical Informatics Association*, pages 489–93.

Markman, A. B. and Gentner, D. (1990). Analogical mapping during similarity judgments. In *In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society.*

Markman, A. B. and Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, 25(4):431–467.

Mazandu, G. K. and Mulder, N. J. (2011). IT-GOM: An Integrative Tool for IC-based GO Semantic Similarity Measures. Technical report, University of Cape Town (South Africa).

Mazandu, G. K. and Mulder, N. J. (2013). Information content-based Gene Ontology semantic similarity approaches: Toward a unified framework theory. *BioMed Research International*, 2013.

M.C. Lange, D.G. Lemay, J. G. (2007). A multi-ontology framework to guide agriculture and food towards diet and health. *Journal of the science of food and agriculture*, 87:1427–1434.

McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4):12.

McInnes, B. T., Pedersen, T., and Pakhomov, S. V. S. (2009). UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, 2009:431–435.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

Miller, G. A. (1998). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language & Cognitive Processes*, 6(1):1–28.

Minsky, M. (1975). A framework for representing knowledge. *The Psychology of Computer Vision.*

Mistry, M. and Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327.

Mohammad, S. and Hirst, G. (2012a). Distributional Measures as Proxies for Semantic Relatedness. *CoRR*, abs/1203.1.

Mohammad, S. and Hirst, G. (2012b). Distributional Measures of Semantic Distance: A Survey. *ArXiv*, 1203.1889.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning.* The MIT Press.

Moss, C. S. (1960). Current and projected status of semantic differential research. *Psychological Record*, 10:47–54.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, 92(3):289.

Nagar, A. and Al-Mubaid, H. (2008). A New Path Length Measure Based on GO for Gene Similarity with Evaluation using SGD Pathways. In *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 590–595. IEEE.

Nardi, D. and Brachman, R. J. (2003). An Introduction to Description Logics. In *Description Logic Handbook*, pages 1–40.

NIST (2012). TREC 2012. In *The Twenty-First Text REtrieval Conference*.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual review of Psychology*, 43(1):25–53.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):510.

Noy, N. F., McGuinness, D. L., and Others (2001). Ontology development 101: A guide to creating your first ontology.

OBO (2013). http://www.geneontology.org/GO.format.obo-1_2.shtml.

Oldakowski, R. and Bizer, C. (2005). SemMF: A Framework for Calculating Semantic Similarity of Objects Represented as RDF Graphs. *Poster at the 4th International Semantic Web Conference*.

Olsson, C., Petrov, P., Sherman, J., and Perez-Lopez, A. (2011). Finding and Explaining Similarities in Linked Data. In *Semantic Technology for Intelligence, Defense, and Security*.

Orozco, J. and Belanche, L. (2004). On aggregation operators of transitive similarity and dissimilarity relations. In *IEEE International Conference on Fuzzy Systems*, volume 3, pages 1373–1377. IEEE.

Othman, R. M., Deris, S., and Illias, R. M. (2008). A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences. *Journal of biomedical informatics*, 41(1):65–81.

Oxford Dict., editor (2012). *Oxford Dictionaries*. Oxford University Press, 7 edition.

Pakhomov, S., McInnes, B., Adam, T., Liu, Y., Pedersen, T., and Melton, G. B. (2010). Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2010:572–576.

Pakhomov, S. V. S., Pedersen, T., McInnes, B., Melton, G. B., Ruggieri, A., and Chute, C. G. (2011). Towards a framework for developing semantic relatedness reference standards. *Journal of biomedical informatics*, 44(2):251–65.

Panchenko, A. (2013). *Similarity Measures for Semantic Relation Extraction*. PhD thesis, Université catholique de Louvain.

Panchenko, A. and Morozova, O. (2012). A study of hybrid similarity measures for semantic relation extraction. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 10–18. Association for Computational Linguistics.

Park, Y. R., Kim, J. J. H., Lee, H. W., and Yoon, Y. J. (2011). GOChase-II: correcting semantic inconsistencies from Gene Ontology-based annotations for gene products. *BMC Bioinformatics*, 12 Suppl 1(Suppl 1):S40.

Passant, A. (2010). Dbrec - music recommendations using DBpedia. In *The Semantic Web–ISWC 2010*, pages 209–224. Springer.

Patwardhan, S. (2003). *Incorporating dictionary and corpus information into a context vector measure of semantic relatedness*. Master thesis, Minnesota (USA).

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational*, pages 241–257.

Patwardhan, S. and Pedersen, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *EACL Workshop Making Sense of Sense — Bringing Computational Linguistics and Psycholinguistics TogetherWorkshop Making Sense of Sense— Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8.

Paul, R., Groza, T., Zankl, A., and Hunter, J. (2012). Semantic similarity-driven decision support in the skeletal dysplasia domain. *The Semantic Web–ISWC 2012*.

Pedersen, T., Pakhomov, S. V. S., Patwardhan, S., and Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–99.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity: measuring the relatedness of concepts. In *HLT-NAACL, Demonstration Papers*, pages 38–41, Stroudsburg, PA, USA.

Pekar, V. and Staab, S. (2002). Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In *COLING'02 Proceedings of the 19th international conference on Computational linguistics*, volume 2, pages 1–7. Association for Computational Linguistics.

Pesquita, C., Faria, D., and Bastos, H. (2007). Evaluating gobased semantic similarity measures. *Proc. 10th Annual Bio-*, 2007:37–40.

Pesquita, C., Faria, D., Bastos, H., Ferreira, A. E. N., Falcão, A. O., and Couto, F. M. (2008). Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9 Suppl 5:S4.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., and Couto, F. M. (2009a). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*, 5(7):12.

Pesquita, C., Pessoa, D., Faria, D., and Couto, F. M. (2009b). CESSM: collaborative evaluation of semantic similarity measures. In *JB2009: Challenges in ...*.

Petrakis, E. and Varelas, G. (2006). Design and evaluation of semantic similarity measures for concepts stemming from the same or different ontologies. *... Multimedia Semantics (...*.

Petrakis, E., Varelas, G., Hliaoutakis, A., and Raftopoulou, P. (2006). X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. *Journal of Digital Information Management*, 4(4):233.

Phillips, J. and Buchanan, B. G. (2001). Ontology-guided knowledge discovery in databases. In *Proceedings of the 1st international conference on Knowledge capture*, pages 123–130. ACM.

Pirró, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data & Knowledge Engineering*, 68(11):1289–1308.

Pirró, G. (2012). REWOrD: Semantic Relatedness in the Web of Data. In *AAAI Conference on Artificial Intelligence*.

Pirró, G. and Euzenat, J. (2010a). A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness. In *Proceedings of the 9th International Semantic Web Conference ISWC 2010*, pages 615–630. Springer.

Pirró, G. and Euzenat, J. (2010b). A Semantic Similarity Framework Exploiting Multiple Parts-of Speech. In Meersman, R., Dillon, T., and Herrero, P., editors, *On the Move to Meaningful Internet Systems, OTM 2010*, volume 6427 of *Lecture Notes in Computer Science*, pages 1118–1125. Springer Berlin Heidelberg.

Pirró, G. and Seco, N. (2008). Design , Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. *Lecture Notes in Computer Science Volume 5332*, 5332:1271–1288.

Popescu, M., Keller, J. M., and Mitchell, J. A. (2006). Fuzzy Measures on the Gene Ontology for Gene Product Similarity. *IEEEACM Transactions on Computational Biology and Bioinformatics*, 3(3):263–274.

Pothos, E. M., Busemeyer, J. R., and Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, 120(3):679–696.

Rada, R. and Bicknell, E. (1989). Ranking documents with a thesaurus. *Journal of the American Society for Information Science. American Society for Information Science*, 40(5):304–10.

Rada, R., Mili, H., Bicknell, E., and Blettner, M. (1989). Development and application of a metric on semantic nets. *Ieee Transactions On Systems Man And Cybernetics*, 19(1):17–30.

Ramage, D., Rafferty, A. N., and Manning, C. D. (2009). Random Walks for Text Semantic Similarity. In Linguistics, A. f. C., editor, *Workshop on Graph-based Methods for Natural Language Processing*, pages 23–31.

Ranwez, S. (2013). Les ontologies comme support à l'interaction et à la personnalisation dans un processus décisionnel. Exploitation de la sémantique pour favoriser l'automatisation cognitive.

Ranwez, S., Ranwez, V., Villerd, J., and Crampes, M. (2006). Ontological Distance Measures for Information Visualisation on Conceptual Maps. *Lecture notes in computer science*, 4278/2006(On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops):1050–1061.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAI*, volume 1, pages 448–453.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130.

Richardson, R., Smeaton, A. F., and Murphy, J. (1994). Using WordNet as a Knowledge Base for Measuring Semantic Similarity Between Words. Technical report, Dublin City University, School of Computer Applications, Dublin (Ireland).

Rissland, E. (2006). AI and Similarity. *IEEE Intelligent Systems*, 21(3):39–49.

Robinson, I., Webber, J., and Eifrem, E. (2013). *Graph Databases*. O'Reilly.

Robinson, P. N. and Bauer, S. (2011). *Introduction to Bio-ontologies*. Taylor & Francis US.

Roddick, J. F., Hornsby, K., and Vries, D. D. (2003). A Unifying Semantic Distance Model for Determining the Similarity of Attribute Values. *Reproduction*, 16.

Rodríguez, A. and Egenhofer, M. J. (2003). Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456.

Rodríguez, M. A., Cruz, I., Levashkin, S., and Egenhofer, M. J., editors (2005). *GeoSpatial Semantics*, volume 3799 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rogers, F. B. (1963). Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–6.

Ross, B. H. (1987). This is like that: The use of earlier problems and the separation of similarity effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(4):629.

Ross, B. H. (1989). Distinguishing types of superficial similarities: Different effects on the access and use of earlier problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3):456.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Russell, S. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Prentice Hall Series In Artificial Intelligence. Prentice Hall, 3 edition.

Saerens, M., Fouss, F., Yen, L., and Dupont, P. (2004). The principal components analysis of a graph, and its relationships to spectral clustering. In *Machine Learning: ECML 2004*, pages 371–383. Springer.

Sahlgren, M. (2006). *The Word-space model*. PhD thesis, University of Stockholm (Sweden).

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Linguistics*, 20(1):33–54.

Salton, G. (1968). *Automatic Information Organization and Retrieval*. McGraw - Hill Book Company.

Salton, G. and McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw - Hill Book Company, New York.

Sánchez, D. (2010). A methodology to learn ontological attributes from the Web. *Data & Knowledge Engineering*, 69(6):573–597.

Sánchez, D. and Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of biomedical informatics*, 44(5):749–759.

Sánchez, D. and Batet, M. (2013). A semantic similarity method based on information content exploiting multiple ontologies. *Expert Systems with Applications*, 40(4):1393–1399.

Sánchez, D., Batet, M., and Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297–303.

Sánchez, D., Batet, M., Isern, D., and Valls, A. (2012a). Ontology-based semantic similarity: a new feature-based approach. *Expert Systems with Applications*, 39(9):7718–7728.

Sánchez, D. and Isern, D. (2009). Automatic extraction of acronym definitions from the Web. *Applied Intelligence*, 34(2):311–327.

Sánchez, D., Moreno, A., and Del Vasto-Terrientes, L. (2012b). Learning relation axioms from text: An automatic Web-based approach. *Expert Systems with Applications*, 39(5):5792–5805.

Sánchez, D., Solé-Ribalta, A., Batet, M., and Serratosa, F. (2012c). Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain. *Journal of Biomedical Informatics*, 45(1):141–55.

Sarkar, P., Moore, A. W., and Prakash, A. (2008). Fast incremental proximity search in large graphs. In *Proceedings of the 25th international conference on Machine learning - ICML'08*, pages 896–903, New York, New York, USA. ACM Press.

Saruladha, K. (2011). Information content based semantic similarity for cross ontological concepts. *Science And Technology*, 3(6):5132–5140.

Saruladha, K. and Aghila, G. (2011). COSS : Cross Ontology Semantic Similarity Measure-An Information Content Based Approach. *Trends in Information*, pages 485–490.

Saruladha, K., Aghila, G., and Bhuvaneswary, A. (2010a). Computation of Semantic Similarity among Cross Ontological Concepts for Biomedical Domain. *Methods*, 2(8):111–118.

Saruladha, K., Aghila, G., and Raj, S. (2010b). A Survey of Semantic Similarity Methods for Ontology Based Information Retrieval. *2010 Second International Conference on Machine Learning and Computing*, pages 297–301.

Schickel-Zuber, V. and Faltings, B. (2007). OSS : A Semantic Similarity Function based on Hierarchical Ontologies. *Artificial Intelligence*, pages 551–556.

Schieber, B. and Vishkin, U. (1988). On finding lowest common ancestors: simplification and parallelization. *SIAM Journal on Computing*, 17(6):1253–1262.

Schlicker, A., Domingues, F. S., Rahnenführer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302.

Schwartz, H. A. and Gomez, F. (2011). Evaluating Semantic Metrics on Tasks of Concept Similarity. In *FLAIRS Conference*.

Schwering, A. (2008). Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey. *Transactions in GIS*, 12(1):5–29.

Seco, N. (2005). *Computational Models of Similarity in Lexical Ontologies*. Master thesis, University College, Dublin, Ireland.

Seco, N., Veale, T., and Hayes, J. (2004). An Intrinsic Information Content Metric for Semantic Similarity in WordNet. In *16th European Conference on Artificial Intelligence*, pages 1–5. IOS Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Sheehan, B., Quigley, A., Gaudin, B., and Dobson, S. (2008). A relation based measure of semantic similarity for Gene Ontology annotations. *BMC Bioinformatics*, 9:468.

Shenoy, M., Shet, K. C., Acharya, D., Shenoy K, M., and Dinesh Acharya, U. (2012). A New Similarity Measure for Taxonomy Based on Edge Counting. *CoRR*, abs/1211.4.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323.

Shvaiko, P. and Euzenat, J. (2013). Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):25.

Singh, J., Saini, M., and Siddiqi, S. (2013). Graph Based Computational Model for Computing Semantic Similarity. In *Emerging Research in Computing, Information, Communication and Applications, ERCICA 2013*, pages 501–507. Elsevier.

Singhal, A. (2012). Introducing the Knowledge Graph: things, not strings. *Official Google Blog, May*.

Slimani, T. (2013). Description and Evaluation of Semantic Similarity Measures Approaches. *International Journal of Computer Applications*, 80(10):25–33.

Slimani, T., Boutheina, B. Y., and Mellouli, K. (2006). A New Similarity Measure based on Edge Counting. In *World academy of science, engineering and technology*, pages 34–38.

Smith, B. (2004). Beyond concepts: ontology as reality representation. In *Proceedings of the third international conference on formal ontology in information systems (FOIS 2004)*, pages 73–84.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25(11):1251–1255.

Solé-Ribalta, A., Sánchez, D., Batet, M., and Serratosa, F. (2014). Towards the estimation of feature-based semantic similarity using multiple ontologies. *Knowledge-Based Systems*, 55(0):101 – 113.

Sowa, J. F. (1984). *Conceptual structures: information processing in mind and machine*. Addison-Wesley.

Spackman, K. A. (2004). SNOMED CT milestones: endorsements are added to already-impressive standards credentials. *Healthcare informatics: the business magazine for information and communication systems*, 21(9):54–56.

Spitzer, F. (1964). *Principles of random walk*. Springer, 2001.

Stojanovic, N., Alexander, M., Staab, S., Rudi, S., and York, S. (2001). SEAL - A Framework for Developing SEmantic PortALs. In *Proceedings of the 1st international conference on Knowledge capture*, volume 2097/2001,, pages 155–162. ACM.

Stuckenschmidt, H. (2009). A Semantic Similarity Measure for Ontology-Based. *Springer*, pages 406–417.

Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1):161–197.

Su, S. Y. W., Puranik, S., and Lam, H. (1990). Heuristic algorithms for path determination in a semantic network. In *Computer Software and Applications Conference, 1990. COMPSAC 90. Proceedings., Fourteenth Annual International*, pages 587–592. IEEE.

Sussna, M. (1993). Word Sense Disambiguation Using a Massive of Computer for Free-text Semantic Indexing Network. In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)*, Arlington, Virginia. ACM.

Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., and Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC bioinformatics*, 13 Suppl 1(Suppl 1):S4.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, 24(4):629–640.

Teng, Z., Guo, M., Liu, X., Dai, Q., Wang, C., and Xuan, P. (2013). Measuring gene functional similarity based on groupwise comparison of GO terms. *Bioinformatics (Oxford, England)*, 29(11):1424–32.

Thomas, P. D., Wood, V., Mungall, C. J., Lewis, S. E., and Blake, J. A. (2012). On the Use of Gene Ontology Annotations to Assess Functional Similarity among Orthologs and Paralogs: A Short Report. *PLoS computational biology*, 8(2):e1002386.

Tummarello, G., Delbru, R., and Oren, E. (2007). Sindice. com: Weaving the open linked data. In *The Semantic Web*, pages 552–565. Springer.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327–352.

Tversky, A. (2004). *Preference, belief, and similarity: selected writings*. MIT Press.

Tversky, A. and Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, 89(2):123.

Tversky, A. and Itamar, G. (1978). Studies of Similarity. In Rosh, E. and Lloyd, B., editors, *Cognition and categorization*, pages 79–98. Lawrence Erlbaum, Hillsdale, NJ.

UniProt Consortium (2013). Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic acids research*, 41(D1):D43–D47.

Uschold, M. and Gruninger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge engineering review*, 11(2):93–136.

Valtchev, P. (1999a). Building classes in object-based languages by automatic clustering. In *Advances in Intelligent Data Analysis*, pages 303–314. Springer.

Valtchev, P. (1999b). *Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets*. PhD thesis, Joseph Fourier - Grenoble 1.

Valtchev, P. and Euzenat, J. (1997). Dissimilarity measure for collections of objects and values. In *Advances in Intelligent Data Analysis Reasoning about Data*, pages 259–272. Springer.

Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E., and Milios, E. E. (2005). Semantic similarity methods in wordNet and their application to information retrieval on the web. *Proceedings of the seventh ACM international workshop on Web information and data management - WIDM'05*, page 10.

Vicient, C., Sánchez, D., and Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textualresources. *Engineering Applications of Artificial Intelligence*, 26(3):1092–1106.

Volz, J., Bizer, C., Gaedke, M., and Kobilarov, G. (2009). Silk – A Link Discovery Framework for the Web of Data. In *Proceedings of the 2nd Linked Data on the Web Workshop*, pages 1–6, Madrid, Spain.

von Luxburg, U., Radl, A., and Hein, M. (2011). Hitting and commute times in large graphs are often misleading. *ArXiv*, 1003.1266.

Voorhees, E. and Harman, D. K. (2005). *TREC: Experiment and evaluation in information retrieval*, volume 63. MIT press Cambridge.

Vosniadou, S. and Ortony, A. (1989). *Similarity and analogical reasoning*. Cambridge University Press.

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 1063–1064. ACM.

W3C (2004). RDF Vocabulary Description Language 1.0: RDF Schema.

W3C (2009). http://www.w3.org/2009/08/skos-reference/skos.rdf.

W3C (2013a). http://www.w3.org/TR/owl2-overview/.

W3C (2013b). http://www.w3.org/TR/sparql11-overview/.

Wang, C., Kalyanpur, A., Fan, J., Boguraev, B. K., and Gondek, D. C. (2012a). Relation extraction and scoring in DeepQA. *IBM Journal of Research and Development*, 56(3.4):1–9.

Wang, J., Xie, D., Lin, H., Yang, Z., and Zhang, Y. (2012b). Filtering Gene Ontology semantic similarity for identifying protein complexes in large protein interaction networks. *Proteome Science*, 10(Suppl 1):S18.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., and Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics (Oxford, England)*, 23(10):1274–81.

Wang, T. and Hirst, G. (2011). Refining the Notions of Depth and Density in WordNet-based Semantic Similarity Measures. *Computational Linguistics*, pages 1003–1011.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M., and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS biology*, 7(11):e1000247.

Watanabe, S. and Donovan, S. K. (1969). *Knowing and guessing: A quantitative study of inference and information*. Wiley New York.

Weaver, W. (1955). Translation. *Machine translation of languages*, 14:15–23.

Webster, R. (1994). *Convexity*. Oxford University Press, USA.

Weeds, J. E. (2003). *Measures and applications of lexical distributional similarity.* PhD thesis, University of Sussex.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., and Musen, M. A. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Research*, 39(Web Server issue):541–545.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques.* Elsevier.

Wu, X., Zhu, L., Guo, J., Zhang, D.-Y., and Lin, K. (2006). Prediction of yeast protein-protein interaction network: insights from the Gene Ontology and annotations. *Nucleic acids research*, 34(7):2137–50.

Wu, Z. and Palmer, M. (1994). Verb semantics and lexical selection. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 133–138.

Xiao, H. and Cruz, I. (2005). A Multi-Ontology Approach for Personal Information Management. In *proceedings. of Semantic Desktop Workshop at the ISWC*, Galway, Ireland.

Xu, T., Du, L., and Zhou, Y. (2008). Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics*, 9(1):472.

Xu, T., Gu, J., Zhou, Y., and Du, L. (2009). Improving detection of differentially expressed gene sets by applying cluster enrichment analysis to Gene Ontology. *BMC Bioinformatics*, 10(1):240.

Yang, H., Nepusz, T., and Paccanaro, A. (2012). Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics (Oxford, England)*, 28(10):1383–1389.

Young Whan, K. and Kim, J. H. (1990). A model of knowledge based information retrieval with hierarchical concept graphs. *Journal of documentation*, 46(2):113–136.

Yu, H., Gao, L., Tu, K., and Guo, Z. (2005). Broadly predicting specific gene functions with expression similarity and taxonomy similarity. *Gene*, 352:75–81.

Yu, H., Jansen, R., and Gerstein, M. (2007a). Developing a similarity measure in biological function space. *Bioinformatics*, online:1–18.

Yu, H., Jansen, R., Stolovitzky, G., and Gerstein, M. (2007b). Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications. *Bioinformatics (Oxford, England)*, 23(16):2163–73.

Yu, X. (2010). *Mathematical and Experimental Investigation of Ontological Similarity Measures and Their Use in Biomedical Domains.* Master thesis, Miami University.

Zhong, J., Zhu, H., Li, J., and Yu, Y. (2002). Conceptual Graph Matching for Semantic Search. In *ICCS'02 Proceedings of the 10th International Conference on Conceptual Structures: Integration and Interfaces*, pages 92–196. Springer-Verlag.

Zhou, Z., Wang, Y., and Gu, J. (2008). A New Model of Information Content for Semantic Similarity in WordNet. In *FGCNS'08 Proceedings of the 2008 Second International Conference on Future Generation Communication and Networking Symposia - Volume 03*, pages 85–89. IEEE Computer Society.

Zieliński, M. (2014). http://www.pilsudski.org/portal/pl/nowosci/blog/432-unikalne-identyfikatory-w-archiwach-i-bibliotekach.

Abstract

The notions of semantic proximity, distance, and similarity have long been considered essential for the elaboration of numerous cognitive processes, and are therefore of major importance for the communities involved in the development of artificial intelligence. This thesis studies the diversity of semantic measures which can be used to compare lexical entities, concepts and instances by analysing corpora of texts and ontologies. Strengthened by the development of Knowledge Representation and Semantic Web technologies, these measures are arousing increasing interest in both academic and industrial fields.

This manuscript begins with an **extensive state-of-the-art** which presents numerous contributions proposed by several communities, and underlines the diversity and interdisciplinary nature of this domain. Thanks to this work, despite the apparent heterogeneity of semantic measures, we were able to distinguish common properties and therefore propose a general classification of existing approaches. Our work goes on to look more specifically at measures which take advantage of ontologies expressed by means of semantic graphs, e.g. RDF(S) graphs. We show that these measures rely on a reduced set of abstract primitives and that, even if they have generally been defined independently in the literature, most of them are only specific expressions of generic parametrised measures. This result leads us to the definition of **a unifying theoretical framework for semantic measures**, which can be used to: (i) design new measures, (ii) study theoretical properties of measures, (iii) guide end-users in the selection of measures adapted to their usage context. The relevance of this framework is demonstrated in its first practical applications which show, for instance, how it can be used to perform theoretical and empirical analyses of measures with a previously unattained level of detail. Interestingly, this framework provides a new insight into semantic measures and opens interesting perspectives for their analysis.

Having uncovered a flagrant lack of generic and efficient software solutions dedicated to (knowledge-based) semantic measures, a lack which clearly hampers both the use and analysis of semantic measures, we consequently developed **the Semantic Measures Library (SML): a generic software library dedicated to the computation and analysis of semantic measures**. The SML can be used to take advantage of hundreds of measures defined in the literature or those derived from the parametrised functions introduced by the proposed unifying framework. These measures can be analysed and compared using the functionalities provided by the library. The SML is accompanied by extensive documentation, community support and software solutions which enable non-developers to take full advantage of the library. In broader terms, this project proposes to federate the several communities involved in this domain in order to create an interdisciplinary synergy around the notion of semantic measures: `http://www.semantic-measures-library.org`

This thesis also presents **several algorithmic and theoretical contributions** related to semantic measures: (i) an innovative method for the comparison of instances defined in a semantic graph – we underline in particular its benefits in the definition of content-based recommendation systems, (ii) a new approach to compare concepts defined in overlapping taxonomies, (iii) algorithmic optimisation for the computation of a specific type of semantic measure, and (iv) a semi-supervised learning-technique which can be used to identify semantic measures adapted to a specific usage context, while simultaneously taking into account the uncertainty associated to the benchmark in use. These contributions have been validated by several international and national publications.

Résumé

Les notions de proximité, de distance et de similarité sémantiques sont depuis longtemps jugées essentielles dans l'élaboration de nombreux processus cognitifs et revêtent donc un intérêt majeur pour les communautés intéressées au développement d'intelligences artificielles. Cette thèse s'intéresse aux différentes mesures sémantiques permettant de comparer des unités lexicales, des concepts ou des instances par l'analyse de corpus de textes ou de représentations de connaissance (i.e. ontologies). Encouragées par l'essor des technologies liées à l'Ingénierie des Connaissances et au Web sémantique, ces mesures suscitent de plus en plus d'intérêt à la fois dans le monde académique et industriel.

Ce manuscrit débute par un **vaste état de l'art** qui met en regard des travaux publiés dans différentes communautés et souligne l'aspect interdisciplinaire et la diversité des recherches actuelles dans ce domaine. Cela nous a permis, sous l'apparente hétérogénéité des mesures existantes, de distinguer certaines propriétés communes et de présenter une classification générale des approches proposées. Par la suite, ces travaux se concentrent sur les mesures qui s'appuient sur une structuration de la connaissance sous forme de graphes sémantiques, e.g. graphes RDF(S). Nous montrons que ces mesures reposent sur un ensemble réduit de primitives abstraites, et que la plupart d'entre elles, bien que définies indépendamment dans la littérature, ne sont que des expressions particulières de mesures paramétriques génériques. Ce résultat nous a conduits à définir **un cadre théorique unificateur pour les mesures sémantiques**. Il permet notamment : (i) d'exprimer de nouvelles mesures, (ii) d'étudier les propriétés théoriques des mesures et (iii) d'orienter l'utilisateur dans le choix d'une mesure adaptée à sa problématique. Les premiers cas concrets d'utilisation de ce cadre démontrent son intérêt en soulignant notamment qu'il permet l'analyse théorique et empirique des mesures avec un degré de détail particulièrement fin, jamais atteint jusque-là. Plus généralement, ce cadre théorique permet de poser un regard neuf sur ce domaine et ouvre de nombreuses perspectives prometteuses pour l'analyse des mesures sémantiques.

Le domaine des mesures sémantiques souffre d'un réel manque d'outils logiciels génériques et performants ce qui complique à la fois l'étude et l'utilisation de ces mesures. En réponse à ce manque, nous avons développé **la Semantic Measures Library (SML), une librairie logicielle dédiée au calcul et à l'analyse des mesures sémantiques**. Elle permet d'utiliser des centaines de mesures issues à la fois de la littérature et des fonctions paramétriques étudiées dans le cadre unificateur introduit. Celles-ci peuvent être analysées et comparées à l'aide des différentes fonctionnalités proposées par la librairie. La SML s'accompagne d'une large documentation, d'outils logiciels permettant son utilisation par des non informaticiens, d'une liste de diffusion, et de façon plus large, se propose de fédérer les différentes communautés du domaine afin de créer une synergie interdisciplinaire autour la notion de mesures sémantiques : `http://www.semantic-measures-library.org`

Cette étude a également conduit à différentes **contributions algorithmiques et théoriques**, dont (i) la définition d'une méthode innovante pour la comparaison d'instances définies dans un graphe sémantique – nous montrons son intérêt pour la mise en place de système de recommandation à base de contenu, (ii) une nouvelle approche pour comparer des concepts représentés dans des taxonomies chevauchantes, (iii) des optimisations algorithmiques pour le calcul de certaines mesures sémantiques, et (iv) une technique d'apprentissage semi-supervisée permettant de cibler les mesures sémantiques adaptées à un contexte applicatif particulier en prenant en compte l'incertitude associée au jeu de test utilisé. Ces travaux ont été validés par plusieurs publications et communications nationales et internationales.