Vers l'Intuition Artificielle HABILITATION À DIRIGER LES RECHERCHES DE L'UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Aurélie Bertaux

HABILITATION À DIRIGER LES RECHERCHES
DE L'UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE BOURGOGNE

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

par

AURÉLIE BERTAUX

Vers l'Intuition Artificielle

Thèse présentée et soutenue à Dijon, le 25/11/2019

Composition du Jury :

NICOLLE CHRISTOPHE	Professeur à l'Université de Bourgogne	Parrain
KOUKAM ABDERRAFIAA	Professeur à l'Université de Technologie de Belfort Montbéliard	Président
SEDES FLORENCE	Professeur à l'Université Paul Sabatier Toulouse III	Rapporteur
WEMMERT CÉDRIC	Professeur à l'Université de Strasbourg	Rapporteur
TEISSEIRE MAGUELONNE	Directrice de Recherche à l'Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture	Rapporteur

N° | X | X | X |

SPIM
école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Towards Artificial Intuition

Keywords: Data Mining, Formal Concepts Analysis, Graph, Ontology, Big Data, Multidimensional, Fuzzy, Labeling

Abstract:

This document presents an exploration of the concepts of Artificial Intuition and proposes a positioning not interested in the cybernetization of human intuition but in the development of a machine-specific intuition.

The approach is broken down into four stages by presenting the locks, the constraints and the tools to implement it. These steps are: (i) the collection and management of data in consideration of their volume, heterogeneity and relevance; (ii) the creation of a knowledge base from these data by adapted Formal Concepts Analysis methods to handle volume and multidimensional locks; (iii) the definition and algorithms to allow Artificial Intuition to be formed from this knowledge and (iv) the means to understand and express this Intuition to the user in a form that he can understand by a qualification process of knowledge, the proof of its veracity and the automation of the process adapted to the considered knowledge.

Titre : Vers L'Intuition Artificielle

Mots-clés : Fouille de données, Analyse Formelle de Concepts, Graphe, Ontologie, Big Data, Multidimensionnel, Flou, Labelisation

Résumé :

Ce document présente une exploration des concepts de l'Intuition Artificielle et propose un positionnement s'intéressant non pas à la cybernétisation de l'intuition humaine mais au développement d'une intuition propre à la machine.

L'approche est décomposée en quatre étapes en présentant les verrous, les contraintes et les outils pour la mettre en oeuvre. Ces étapes sont : (i) la collecte et la gestion des données en considération de leur volume, de leur hétérogénéité et de leur pertinence; (ii) la création d'un socle de connaissances à partir de ces données par des méthodes d'Analyse Formelle de Concepts modifiées pour traiter des verrous de volume et de multidimensionalité; (iii) la définition et des algorithmes pour permettre à l'Intuition Artificielle de se former à partir de ces connaissances et (iv) le moyen de comprendre et d'exprimer cette Intuition à l'utilisateur dans une forme qu'il peut comprendre par un processus de qualification de la connaissance, de preuve de sa véracité et l'automatisation du processus adapté à la connaissance considérée.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ
Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France

Contents

REMERCIEMENTS

La présentation de cette HDR est le résultat d'un long processus de recherche qui se construit brique à brique depuis ma thèse et de plus en plus activement.

J'ai eu la chance d'être recrutée dans une équipe qui m'a offert tous les moyens de m'épanouir et de mener ma recherche. C'est un cadre de travail très rare avec des collègues agréables qui s'intéressent à l'ambiance générale et aux façons de collaborer mutuellement pour que tout le monde progresse.

Forts de ces relations et sachant vers où mener notre recherche, nous avons pu monter un nouveau laboratoire depuis le 1er janvier 2019 : le CIAD (Connaissance et Intelligence Artificielle Distribuées) où j'ai pu placer ma recherche comme un axe fondateur du laboratoire. La Business Unit du laboratoire qui abrite les ingénieurs de nos contrats est une fourmilière qui permet la concrétisation très rapide des avancées faites en recherche. Pour tout cela je voudrais remercier mes collègues, et notamment Ouassila Narsis Labbani, autre MCF avec qui nous échangeons quotidiennement sur tous les domaines; et Sébastien Gerin qui supervise la Business Unit et toutes les parties techniques des contrats que nous menons.

Mais tout particulièrement je voudrais remercier Christophe Nicolle qui dirige notre laboratoire. Il a su instaurer un cadre de travail très riche en opportunités et humainement il m'a toujours encouragée et motivée à poursuivre ma voie en me laissant libre sur ma façon d'intégrer ma spécialité au sein de notre équipe pluridisciplinaire et m'a offert tous les moyens d'y parvenir. Il a toujours été très disponible et réactif malgré une charge de travail colossale.

Je souhaite également remercier les rapporteurs de cette HDR, car je les ai sollicités parce qu'ils sont intervenus dans ma vie de recherche et que c'est à eux que je voulais tout d'abord exposer ces travaux dans leur globalité.

INTRODUCTION

L'ordinateur ne peut que restituer, sous une forme plus ou moins élaborée, les concepts que le chercheur y a introduits. Il est incapable de faire preuve d'intuition, démarche subtile encore mal comprise qui seule peut conduire à la découverte.

Pierre Joliot-Curie

Notre société actuelle, très orientée vers les réseaux sociaux et la communication aisée génère beaucoup de données hétérogènes à évolution rapide. Le besoin d'outils pour les prendre en considération est à l'origine du domaine du Big Data. Ces outils ont permis d'alimenter en données l'Intelligence Artificielle et lui donner la portée qu'elle a aujourd'hui. Puis, au delà de l'analyse des données, l'intelligence artificielle s'est tournée vers des axes de prédiction et prescription.

Prévoir le futur est un domaine qui touche à l'incertain et c'est donc tout naturellement que l'Intuition Artificielle est devenue une source d'intérêt.

On trouve beaucoup d'écrits, de citations et de références à l'intuition. En effet, c'est une chose à la fois très commune, dont chacun est pourvu et pourtant globalement mal connue. Chacun y projette donc ce qu'il en comprend, ce qu'il en a vécu.

Pour intégrer l'intuition à la machine la première question qui s'est posée est de la comprendre puis l'implémenter. Mais des divergences s'opèrent quant à sa définition sur le plan philosophique et psychologique, quant à la motivation sur ce à quoi doit servir la version artificielle et enfin quant à quelles approches informatiques développer pour l'implémenter.

La problématique de ce domaine est le manque de consensus sur ce qu'il est ou devrait être et majoritairement le domaine en reste au stade du débat d'idées.

J'ai choisi de trancher ce débat, de fixer une définition et de proposer les définitions et les outils pour sa concrétisation. Pour cela, je me positionne sur ce à quoi l'intuition devrait correspondre pour une machine, à mon sens, en considérant qu'il s'agit d'une faculté permettant à une machine de percevoir des éléments manquants, des incohérences... au sein d'un environnement.

Ce manuscrit présente donc une approche permettant de doter la machine de sa propre faculté à subodorer.

Pour cela, je pars du principe que l'intuition prend racine dans une expérience fondamentale de façon à pouvoir se faire une idée de ce qui semble et ne semble pas "normal". Plus l'expérience est vaste et correcte plus l'intuition sera fiable. Pour constituer cette expérience, il faut commencer par collecter les données concernées, puis par un processus de fouille de données, en extraire les connaissances. Forte de cette expérience, l'Intuition

peut s'opérer. Enfin, la machine n'a plus qu'à exprimer sa compréhension du domaine et expliquer les intuitions qu'elle a.

La figure 1 présente, l'ensemble de ces quatre étapes en précisant pour chacune d'entre elle les verrous à lever et les moyens mis en oeuvre pour y parvenir.

Dans ce manuscrit, je consacre un chapitre à chacune de ces quatre étapes.

Tout d'abord, après une introduction au domaine de l'Intuition Artificielle, je présente l'étape de constitution des connaissances issue de mon coeur de recherche depuis ma thèse : la fouille de données et plus spécifiquement l'Analyse Formelle de Concepts (FCA). Cependant les contraintes de l'Intuition Artificielle en terme *volume* de données et de leur *multidimensionnalité* ont imposé l'adaptation des méthodes de FCA pour lever ces verrous. Ces travaux ont pu être menés grâce au montage de deux projets de recherche (Eurostars et FUI).

Ensuite, j'aborde le coeur de l'Intuition Artificielle initiée durant mon post doctorat au LIG. Pour fonctionner, contrairement à ce que l'on trouve essentiellement dans le domaine, j'ai voulu un domaine clairement *défini*, et surtout *formel* de façon à ce que chacun puisse se l'approprier aisément. Je présente également des algorithmes pour améliorer l'*efficacité* de l'approche qui doit gérer des volumes conséquents.

L'Intuition Artificielle se nourrit des connaissances de la FCA qui les construit à partir des données. La collecte et la gestion de ces données, présentées dans le chapitre III doivent répondre à quatre verrous : leur *volume* car le système doit collecter un ensemble conséquent de données *pertinentes* et être en mesure d'en gérer le volume. Ces données seront forcément *hétérogènes* donc le système doit aussi pouvoir palier cette contrainte. Ces verrous s'imposent essentiellement dans le cadre ontologique de la démarche. Pour les lever, j'ai co-encadré les thèses de Thomas Hassan (sur les questions du volume et de la pertinence) et de David Werner (sur la problématique de l'hétérogénéité).

Enfin, le système s'appuie sur un outil sémantique permettant de comprendre la connaissance qui lui est fournie et de pouvoir expliquer l'Intuition qui est émise. Cette étape est présentée en chapitre IV. Les verrous qui se présentent à ce niveau sont de plusieurs ordres. Tout d'abord, il y a une nécessité de construire une ontologie qui corresponde aux données issues de la FCA. Pour cela il faut *qualifier*, i.e. nommer ces connaissances. Vient ensuite la *construction automatique* de l'ontologie, la *véracité* des conclusions émises par le système, et enfin, dans l'idée où le système inclura des données floues, en plus des données multidimensionnelles, il faudra développer un moteur d'*inférence multidimensionnel flou*.
Cette étape est colossale. Pour la prendre en charge j'ai encadré deux thèses, un stage de thèse et monté deux projets dont l'un commencera en 2020 pour la question du moteur d'inférence.

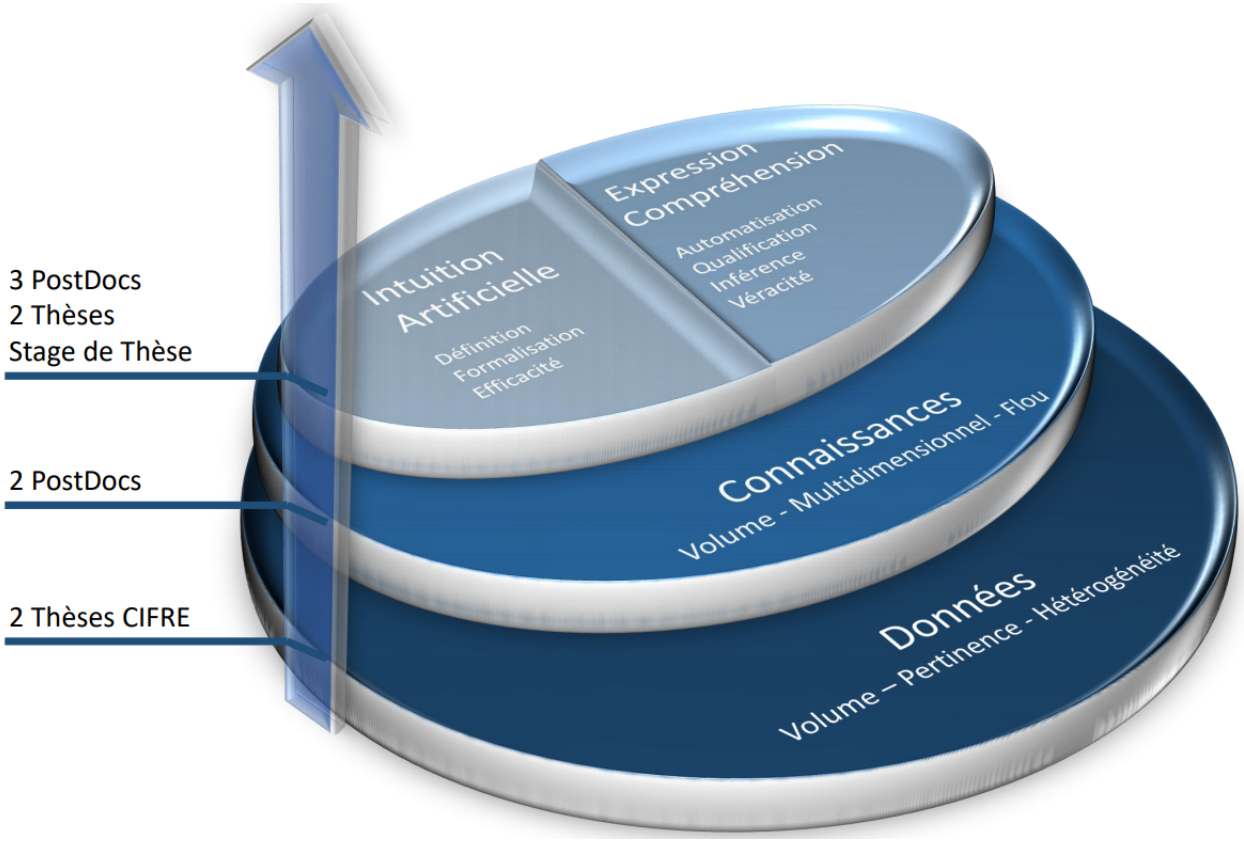

Figure 1 – Les étapes vers l'Intuition Artificielle

Enfin je conclue et présente les prochains travaux à réaliser dans la continuité de ceux-ci, comme l'intégration du flou dans les données initiales, l'apprentissage dynamique et la suite de la construction automatique d'une ontologie (et surtout son moteur d'inférence) multidimensionnelle floue à partir des données.
Dans les annexes je présente les thèses que j'ai encadrées, celles que j'ai supervisées et les autres encadrements de master, stages, post doctorants que j'ai menés. Je fournis également un résumé des projets de recherche montés pour répondre aux verrous mentionnés dans ce travail. Et enfin j'établis un bilan rapide de mes enseignements.

QU'EST CE QUE L'INTUITION ARTIFICIELLE ?

L'Intuition Artificielle aussi appelée "technologie informatique cognitive" est un domaine émergent qui prend de l'ampleur. L'ACM ne l'a pas encore entrée dans sa classification et n'ayant pas de consensus sur ce dont il s'agit, chacun la conçoit comme il l'estime. Notons que c'est aussi le cas de l'intelligence artificielle, mais cependant beaucoup de choses pertinentes et intéressantes sont faites dans ce domaine même si leurs concepteurs ne partagent pas les mêmes définitions.

Dans la littérature, on trouve peu d'articles scientifiques sur l'Intuition Artificielle, une vingtaine ont été publiés dont la moitié n'ont jamais été cités ou sont écrits en allemand [Oosterhuis, 1990] ou en russe [Prokopchuk, 2017] ne favorisant pas la diffusion du domaine. Par ailleurs, on trouve une multitude d'articles philosophiques sur l'intuition,

et même l'Intuition Artificielle, puisque pour beaucoup l'Intuition Artificielle correspond à la cybernétisation de l'intuition humaine et que ce domaine étant mal défini, l'Intuition Artificielle cherche également sa définition.

Dans ce chapitre, j'aborde les grandes approches de l'Intuition Artificielle issues des sciences cognitives, avant d'en décrire les approches scientifiques. Enfin je positionne ma vision du domaine que je décline dans ce manuscrit.

EN SCIENCES COGNITIVES

De manière logique, lorsqu'il a fallu définir l'Intuition Artificielle, la première question qui s'est posée était de savoir ce qu'était l'intuition. De nombreuses études et débats ont eu et ont toujours cours sur la question au sein des sciences cognitives. Quatre grandes tendances ressortent de ces études :

- L'intuition se caractérise par son côté inconscient / instinctif, non rationnel et immédiat [Jung, 1923, Benner et al., 1987, Damasio, 1994, Klein, 1998, Gigerenzer, 2007, Wilson et al., 2008].
- L'intuition est liée à l'affect car elle est liée une charge émotionnelle [Damasio, 1999, Dane et al., 2007, Morgado et al., 2008, Kahneman, 2011].
- L'intuition est "une perception via l'inconscient" selon Carl Jung qui la sépare en deux types (appelés "attitudes") : un type extraverti orienté vers la découverte de possibilités nouvelles mais non prouvées. Et un type introverti guidé par les images de l'inconscient par l'exploration d'archétypes à la recherche de sens.
- L'intuition est une "sagesse biologique ancienne" [Myers, 2002].

Outre savoir ce dont il s'agit, certains se sont intéressés à comprendre son fonctionnement. Ainsi [Minsky, 2006] a examiné les raisonnements non logiques pour traiter les fonctions humaines. [Frantz, 2003], quant à lui parle de "reconnaissance de modèles subconscients" pour expliquer que l'intuition fonctionne selon des schémas acquis par l'expérience.

Quelque soit la vision adoptée, tous s'accordent sur le fait qu'il n'y a aucune définition commune car le sujet manque de cohérence et de méthodes, voire même qu'une définition de l'intuition n'est pas possible en raison du panel de propriétés du concept [Hodgkinson et al., 2008] et des différences de points de vue entre les pionniers du domaine [Kolata, 1982].

EN INFORMATIQUE

LES APPROCHES POUR L'IMPLÉMENTATION DE L'INTUITION ARTIFICIELLE

Au sein des sciences cognitives, de nombreuses visions philosophiques sont envisagées. Mais lorsqu'il s'agit de les implémenter il faut être plus concret, même si certains défendent carrément que la machine ne peut pas simuler l'intuition humaine [Dreyfus et al., 1989, Penrose, 1989, Searle, 1990]. Plusieurs aspects sont cependant défendus.

Le *sens commun* est une notion présentée comme élément majeur de l'Intuition Artificielle ayant nécessité à être implémenté. Il est défini par [Sloman, 1971] comme la capacité de percevoir les conséquences possibles rapidement parmi un large panel de solutions, sans proposition mathématique ou logique. C'est [McCarthy, 1989] qui avance vers la formalisation en tentant de l'expliquer (à défaut de l'implémenter) par la logique et l'inférence.

Du point de vue concrétisation et implémentation de l'Intuition Artificielle, les démarches s'appuient plutôt sur le domaine de l'apprentissage (automatique ou non). Certains intègrent l'intuition à leur système pour l'améliorer, d'autres développent des systèmes pour créer l'intuition. Les deux se mêlant par moment.

Intégration de l'intuition. Dans les outils qui intègrent l'intuition, on peut citer [Heryaningsih et al., 2018] qui a développé l'Intuition-based Learning (IBL) dans l'objectif d'exploiter l'intuition pour aider l'homme à apprendre. Ces méthodes sont reprises dans plusieurs secteurs. Cependant [Tao et al., 2009] déplore le manque de critères de confiance qui se baserait sur l'expérience et les connaissances. Il propose un apprentissage coopératif entre l'utilisateur et le développeur en utilisant les méthodes d'apprentissage : IBL et les réseaux d'Intuition Artificielle[1]. L'intérêt de cette approche est l'amélioration de la théorie de l'apprentissage par l'incertitude en proposant un réseau d'intuition de confiance (TIN) capable de mesurer de manière fiable et précise une intuition de confiance dans le contexte des systèmes d'apprentissage intuitif.

1. ou réseaux intuitifs, sur lesquels CISCO communique beaucoup : ce nouveau réseau décline l'apprentissage automatique (machine learning) à grande échelle qui s'appuie sur les volumes considérables de data qui circulent sur ses réseaux mondiaux, à apprentissage automatique intégré, pour transformer la donnée en perspectives prédictives et exploitables. https://www.globalsecuritymag.fr/Cisco-presente-le-reseau-de-demain,20170621,72017.html

[Kolotygin et al., 2017] propose également l'intervention humaine. Son approche s'écarte de la similarité entre des propriétés pour exploiter plutôt les actions externes. Il propose la construction d'un modèle consistant et cohérent décrit par l'ensemble des liens entre les éléments de l'objet concerné. Ces liens sont construits par l'intervention d'un expert. Ces approches soulignent essentiellement le besoin :

- d'un socle large de connaissances car même l'intuition humaine n'est pas forcément juste mais s'améliore lorsqu'elle est s'appuie sur des expériences passées [Kahneman, 2003, Tao et al., 2009].
- de relations entre ces connaissances
- de mesures de confiance dans l'intuition.

Développement de l'intuition. Même dans le cadre informatique, l'Intuition Artificielle reste souvent au niveau du débat d'idées plus qu'à celui de la concrétisation. Cependant, en s'appuyant sur la définition la plus largement répandue de l'intuition (les caractères insconscient, irrationnel et immédiat), des implémentations ont été proposées.
Généralement elles rencontrent un certain nombre de difficultés telles que le manque d'adaptation : elles n'arrivent pas à gérer les complications ou les détournements et les facteurs influençant l'Intuition Artificielle, les résultats (et leur précision) ne sont pas expliqués [Dundas et al., 2011]. La démarche adoptée commence généralement par une modélisation de l'intuition humaine sous forme de patterns puis applique un pattern matching pour se rapprocher de la solution qu'aurait fournit un expert [Klein, 1998, Gobet et al., 2009, Dundas et al., 2011, McCormick, 2004]. Une autre approche plus globale, considère le cerveau comme un processeur parallèle ce qui introduit les théories dites "Dual-Process" qui sont considérées comme le fondement théorique le plus approprié à ce jour pour la cybernétisation de l'intuition humaine. Ces théories consistent à considérer que l'esprit humain fonctionne avec des modes de traitement interactifs parallèles de l'information; [Epstein et al., 1996] considèrant le modes rationnel et expérimental, alors que [Stanovich et al., 2000] se tournant vers les modes intuitif et logique. Cependant, quelques soient les modes considérés,le raisonnement globalement admis dans la communauté est **d'un coté l'intelligence qui est précise mais lente et l'intuition qui est imprécise mais très réactive**.

Implémentations concrètes de l'intuition. Monica Anderson qui travaille en intelligence artificielle depuis plus de 30 ans, s'est penchée sur l'Intuition Artificielle depuis 2001. Elle réfute l'idée que le cerveau soit logique , mais plutôt une somme d'intuitions, et donc, rapporté à l'informatique : l'intelligence artificielle aurait besoin de l'Intuition Artificielle. Dans son approche, l'intuition opère sur des événements et pas sur des théories[2]. Elle définit l'Intuition Artificielle en tant que "monde bizarre" pour lequel elle a développé des algorithmes ayant de faibles coûts de calcul pour répondre à des problématiques de l'intelligence artificielle en recherchant des mécanismes sous-jacents [Anderson, 2007]. Elle a mis en exergue un ensemble de rôles que l'Intuition Artificielle devrait remplir :

2. http://artificial-intuition.com/

- **Découverte de solutions sans théorie.** il s'agit de la capacité de trouver des réponses même si nous ne comprenons pas complètement le problème par des conclusions plausibles même lorsque nous recevons des données partiellement incomplètes ou incohérentes.
- **Nouveauté.** Personne ne parle de nouveauté ou d'innovation (fournis de manière triviales par l'Intuition Artificielle) en intelligence artificielle ou ne tente de la mettre en œuvre car personne ne sait même comment l'aborder.
- **Prédiction.** La prédiction est le but et l'origine de l'intelligence.
- **Ambiguïté, diversité et points de vue multiples.** Le cerveau (ou l'Intuition Artificielle) génère des dizaines de nouvelles idées. La capacité à traiter avec des concepts multiples et éventuellement conflictuels implique la capacité à traiter avec des entrées ambiguës et contradictoires.
- **Fiabilité.** Le système devrait être résistant aux erreurs internes mineures.
- **La robustesse.** Les données d'entrée peuvent être incomplètes, contenir des erreurs factuelles, être ambiguës, etc. Le système résoudra de tels problèmes au même degré qu'un humain. La capacité à prendre en compte plusieurs points de vue contradictoires à la fois, et la possibilité de fonctionner si une entrée incomplète ou ambiguë est attribuée au même mécanisme que la capacité de fonctionner si elle est mise en œuvre sur un substrat peu fiable.
- **Auto-organisation et auto-réparation.** Si la connaissance du domaine est bien établie - "suffisamment appris pour être considéré comme une compétence", alors nous pouvons nous attendre à un comportement acceptable ... Tout comme nous le ferions avec un humain compétent. L'auto-organisation consiste à ajouter des éléments à des parties incomplètes et «mal comportantes» du système pour le rendre plus compétent.

Intégrations retentissantes de l'Intuition à l'Intelligence

Les publications scientifiques sont rares sur le domaine de l'Intuition Articielle et plus encore sur les outils qui intègrent à la fois intelligence artificielle et Intuition Artificielle. Paradoxalement c'est un sujet dont il est souvent question dans les articles de vulgarisation en ligne. On y avance beaucoup que l'Intuition Artificielle est l'avenir de l'intelligence artificielle et qu'à elles deux, elles forment l'Intelligence Machine. Certains grands acteurs ont développé des outils et notamment deux acteurs majeurs ont récemment fait connaître par une abondante communication dans les médias vulgaires ou spécialisés, être parvenus à créer des intuitions artificielles. Ces acteurs sont le MIT et Google.

Démarche du MIT

Les scientifiques du Massachusetts Institute of Technology ont demandé à un échantillon d'étudiants intelligents du MIT de résoudre des problèmes liés aux algorithmes de planification tels que des itinéraires aériens. Il s'agissait d'optimiser une flotte d'avions afin que tous les passagers puissent se rendre où ils veulent. Les contraintes sont qu'un avion ne doit pas voler à vide ni visiter une ville plus d'une fois au cours d'une période donnée.

Un élève a réussi à battre l'algorithme existant par une stratégie de logique temporelle linéraire qui postule qu'un axiome est vrai tant qu'il n'est pas mis en défaut. C'est en implémentant cette stratégie qu'ils ont créé une Intuition Artificielle.

DÉMARCHE DE GOOGLE

Contrairement au MIT et à toutes les autres approches, Google s'abstient d'intégrer des éléments humains. En effet, Google est arrivé à la création d'une Intuition Artificielle par la confrontation de deux intelligences artificielles : AlphaGo et AlphaGo Zero[3]. AlphaGo développée par DeepMind avant son rachat par Google en 2014 s'est fait connaître en battant des joueurs professionnels du jeu de go. AlphaGo a été battu par une nouvelle version : AlphaGo Zero qui a appris à jouer au jeu de go en jouant seulement des parties contre lui-même évitant l'entraînement avec des milliers de parties humaines. Cet auto-entraînement est appelé "apprentissage par renforcement" et a duré 40 jours pour permettre à AlphaGo Zero de surpasser la précédente version d'AlphaGo. Dans ce cas l'Intuition Artificielle développée chez Google avec l'apprentissage par renforcement entre dans le cadre d'un apprentissage non supervisé. Les méthodes appliquées sont l'algorithme Monte Carlo qui effectue des recherche dans des arbres afin d'effectuer le prochain mouvement pour le jeu; et les règles d'association.

MON POSITIONNEMENT

INTUITION VERSUS LOGIQUE ?

Nombre de chercheurs (et d'industriels) travaillant sur l'Intuition Artificielle considèrent son utilité à la résolution de problèmes où l'on pose une question et l'on obtient une réponse très rapide. Cependant, même dans le cas de l'être humain, parfois, nous n'avons pas de réponses, ni même d'idée sur une question. Ma vision de l'intuition se situe plus dans le domaine de l'alerte pour nous dire que quelque chose n'est pas cohérent, comme un sixième sens, un instinct.

D'un point de vue implémentation, les chercheurs et développeurs opposent majoritairement logique et intuition. Je ne partage pas cette approche. L'intuition humaine est une très belle chose, mais nous ne la comprenons pas suffisamment pour l'intégrer correctement à la machine. Je pense que vouloir pleinement simuler l'intuition humaine par la machine est voué à l'échec à cause de cette méconnaissance. Dans ma vision des choses, je ne souhaite pas mettre en opposition intuition et logique car la machine n'est que logique et ce serait donc aller contre le principe même de l'artificialité.

Les sciences fondamentales reposent sur l'exactitude ce qui en font des sciences dites dures. L'informatique reposant sur des programmes corrects et la logique est donc aussi une science dure et l'intelligence artificielle qui en découle l'est donc également. Même si l'on s'intéresse à l'incertitude (le flou, les probabilités...) elle est traitée en informatique de manière exacte. Si la probabilité d'une éventualité est calculée à 81% ce n'est ni 82% ni 80%. Donc même l'incertitude est exacte.

On pourrait y opposer les ontologies qui raisonnent selon des règles d'experts et parfois

3. https://becominghuman.ai/artificial-intuition-and-reinforcement-learning-the-next-steps-in-machine-learning-6f2abeb9926b

sujette à interprétation. Mais là encore, même si la règle est fausse, l'ordinateur va raisonner faussement mais de manière stricte, conformément aux règles fausses fournies. Ainsi, quoi qu'il arrive, l'Intuition Artificielle , puisqu'elle est émise par une machine sera toujours conforme à son programme, i.e. son raisonnement. Il s'agit donc de trouver la bonne forme de raisonnement, mais elle passera forcément par la logique à mon sens.

A défaut d'opposer intuition et logique, je préfère opposer intuition humaine et Intuition Artificielle. L'intuition humaine est faite pour l'humain, avec ses neurones, sa physique, sa chimie... La machine n'est pas équipée de tout cet arsenal pour remplacer l'humain. Elle doit donc avoir sa propre intuition selon son propre paradigme. De même que lorsque l'homme a voulu voler, il a développé des machines qui copiaient les oiseaux en battant des ailes. Jusqu'à ce qu'il se rende compte qu'avec une morphologie adaptée, l'aile rigide qui ne bat pas génère la portance nécessaire au vol.

Ainsi pour que l'Intuition Artificielle trouve un cadre solide et prenne de l'ampleur, je pense important de rester dans le paradigme de la logique (et de ses 7 règles), et la définir selon ces contraintes :

- Optimalité : Obtenir la meilleure réponse possible.
- Complétude : Obtenir toutes les réponses.
- Répétabilité : Obtenir le même résultat chaque fois que nous répétons une expérience dans les mêmes conditions.
- Promptitude : Obtenir le résultat dans le temps imparti.
- Parcimonie : Découvrir la théorie la plus simple qui explique complètement les données disponibles.
- Transparence : Comprendre comment nous sommes arrivés au résultat.
- Explicabilité : Comprendre le résultat.

Caractéristiques et verrous de l'Intuition Artificielle

Ma considération de l'Intuition Artificielle fait que j'exclurais de la liste de Monica Anderson la découverte de solutions sans théorie car même l'intuition doit être calculée et donc il faut une méthode adéquate basée sur une théorie. J'exclurais aussi la prédiction. Pour la majorité des personnes du domaine, l'Intuition Artificielle n'a de raison d'être que pour cela. Cependant, même dans le cadre philosophique, l'intuition est un sentiment, une sensation immédiate d'une incohérence mais pas forcément à des fins de prédiction. Bien que je ne conserve pas la prédiction dans les éléments fondamentaux de l'Intuition Artificielle, je n'exclus pas de pouvoir l'inclure dans ses applications possibles. Enfin je remplacerais les notions d'ambiguïté, de diversité et de points de vue multiples, de fiabilité, de robustesse et d'auto-organisation par les notions d'hétérogénéité, de cohérence, de consistance et la reconsidération. Ainsi les critères auxquels l'Intuition Artificielle doit répondre sont à mon sens :

- Nouveauté : faculté de proposer des connaissances nouvelles
- Hétérogénéité : faculté de prendre en charge des sources diverses

- Cohérence : faculté de justifier de ses assertions
- Consistance : faculté du système à ne pas se dénaturer, à rester solide et conforme
- Reconsidération : faculté du système à se remettre en question, émettre des doutes quant à la vérité des informations qu'il contient ou l'absence ou l'incohérence dans ses connaissances.

Pour cela, un certain nombre de verrous devront être levés pour concrétiser ce domaine :

- Sémantique. Pour lever les problèmes d'interprétabilité et de désambiguisation. Identifier et comprendre les signaux faibles qui mènent à la conclusion de l'intuition.
- Formel. Définir mathématiquement l'Intuition Artificielle.
- Généralité. Développer un système adaptable à n'importe quel domaine.
- Apprentissage. Rendre le système auto-évolutif.
- Passage à l'échelle. Prendre en compte un nombre de données sans limitation de volume.
- Temps de calcul. Idéalement le système devrait avoir des performance de calcul en temps quasi constant.

I

L'Analyse Formelle de Concepts et ses évolutions

L'Analyse Formelle de Concepts et ses évolutions

Le point de départ de ma recherche trouve ses origines dans ma thèse effectuée en Analyse Formelle de Concepts (FCA). Ce domaine de la fouille de données s'intéresse à extraire des connaissances à partir des données en les regroupant en concepts grâce aux propriétés communes qu'elles partagent.
Ce domaine est très lié aux règles d'association qui permettent d'exprimer que si quelque chose est vrai alors dans un certain pourcentage de cas une autre l'est aussi. Dès ma thèse, menée conjointement au sein de l'équipe de fouille de données du LSIIT de Strasbourg (aujourd'hui iCube) et du laboratoire ENGEES (aujourd'hui LHyGeS) s'est centrée sur l'Analyse Formelle de Concepts (FCA). Je me suis positionnée sur des problématiques de gestion de données **réelles**, autrement dit des données complexes qui sortaient du cas binaire classiquement traité et notamment le traitement de données multivaluées (un attribut pouvant se découper en plusieurs attributs) et floues (une propriété pouvant ou non être portée par un objet, mais pouvant également l'être partiellement). L'importance de travailler sur des cas concrets basés sur des données réelles a toujours était important dans mes travaux. Le travail mené en Intuition Artificielle a confirmé cette nécessité, car le besoin d'une grande masse de données issues d'horizons variés pour former le socle de connaissances du système est impératif. Or dans ma thèse je n'ai travaillé que sur des données multivaluées ce qui se distingue de la multidimensionnalité. Dans le premier cas il n'y a qu'une source de données dont les attributs se découpent en plusieurs sous attributs, dans le second les sources de données sont multiples et hétérogènes.

Ce chapitre traite de la constitution du socle de connaissances nécessaire à l'Intuition Artificielle. Dans l'approche globale, cette étape succède à celle de collecte des données. Le socle de connaissances devant être large les données collectées sont d'origines diverses et impliquent 3 verrous pour un traitement par la FCA : le volume, la multidimensionnalité et le flou. La FCA n'étant pas adaptée à ces traitements, j'ai donc monté un projet de recherche (Eurostars PSDP, cf section 14.1.2) pour le recrutement d'un post doctorant avec qui nous avons démontré et développé un outil appelé MARM pour lever le verrou de la multidimensionnalité. Un second projet (FUI WineCloud, cf section 14.1), en cours, nous permet de travailler actuellement sur le problème de volume des données. Le flou est un verrou identifié qui sera levé dans un projet ultérieur. Dans un premier temps je présente le domaine de la fouille de données, j'explique ce que sont la FCA et des règles d'association. Puis je décris rapidement les verrous levés dans ma thèse avant de présenter MARM.

1

La fouille de données et l'Analyse Formelle de Concepts

L'analyse de données est un domaine très vaste ayant développé des sous domaines dont l'Analyse Formelle de Concepts (FCA) fait partie.
Dans cette partie je présente ce qu'est la fouille de données et plus précisément la FCA ainsi que les règles d'association qui lui sont liées.

1.1/ Data Mining

Le Data Mining (ou exploration de données, fouille de données, extraction de connaissances ou encore knowledge discovery in databases (KDD)) est un domaine de l'informatique dont le but est de proposer des méthodes et des techniques permettant d'extraire des connaissances à partir des données. Extraire des connaissances peut signifier mettre en évidence des données cachées, découvrir des corrélations significatives entre elles ou encore prédire de nouvelles connaissances. Pour résumer, le Data Mining est le procédé permettant de convertir les données en connaissances ("extraire les pépites d'informations de la gangue des données"). Le Data Mining utilise des techniques variées issues de l'intelligence artificielle ou encore des statistiques. Grâce à des capacités de calculs toujours plus importantes et pour traiter des informations toujours plus nombreuses (augmentation de la capacité des supports de stockage, optimisation des moyens de productions de l'information et plus grand nombre de producteurs. . .) et interconnectées, de nombreuses acteurs sont amenés à utiliser les algorithmes issus de ce domaine. Parmi les exemples d'utilisation les plus courants, on trouve notamment des applications dans le marketing, la détection de fraude ou encore certaines expériences scientifiques. L'attrait du monde industriel pour le Data Mining a permis de faire de ce domaine l'un des plus actifs de la communauté scientifique informatique.

1.1.1/ Les grandes étapes

L'ECD ou Extraction de Connaissances à partir des Données désigne tout le cycle de découverte de la connaissance à partir des données 1.1. Ce processus consiste à passer de données brutes à des connaissances. Fayyad définit ce concept comme "un processus non trivial qui construit un modèle valide, nouveau, potentiellement utile et au final compréhensible, à partir de données" [Fayyad et al., 1996]. Il décompose ce processus en un ensemble d'étapes complexes à savoir :

1. La sélection ou la création d'un ensemble de données à étudier ;
2. Le prétraitement qui permet d'éliminer le bruit et traiter les données manquantes ;
3. La transformation ou la définition des structures optimales de représentation des données ;
4. La fouille de données et la détermination de la tâche (classification, recherche de modèles, etc.) en définissant les paramètres appropriés ;
5. L'interprétation et l'évaluation durant laquelle les éléments extraits sont analysés pour aboutir à des connaissances stockées dans une base de connaissances.

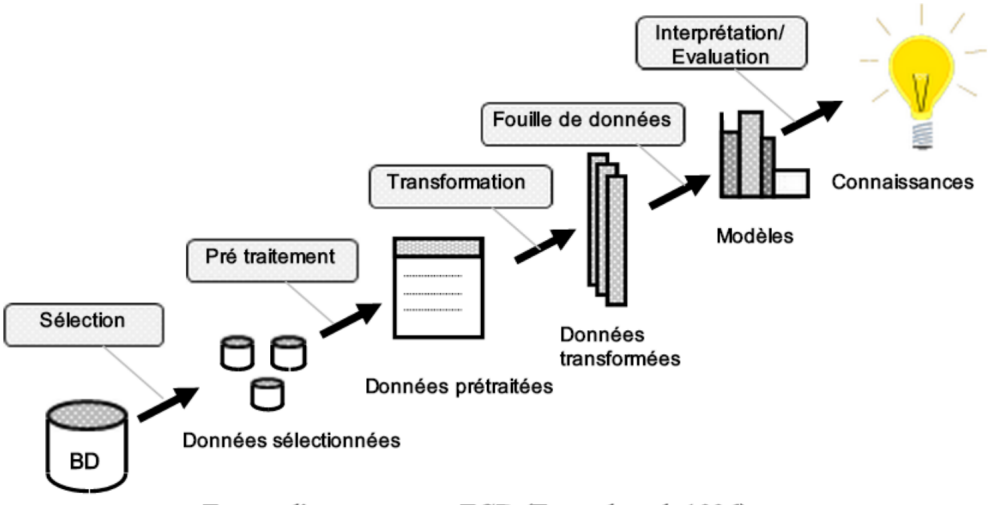

Figure 1.1 – Le processus de l'ECD

1.1.2/ LES FAMILLES DE DATA MINING

Pour extraire les connaissances, le data mining peut s'appuyer sur un expert à différents degrés dans le processus. C'est ce degré d'implication qui distingue les trois familles ci-dessous : supervisé, non supervisé et semi supervisé

1.1.3/ SUPERVISÉ (OU GUIDÉ PAR L'EXPERT)

Les méthodes de fouille supervisée implique un expert qui connaît les conclusions auxquelles ils souhaite aboutir (appelées classes), mais veut que l'ordinateur classe les données selon ses conclusions. Pour cela il va étiqueter un jeu de données test pour indiquer quelle donnée appartient à quelle classe. L'ordinateur va ensuite rechercher les corrélation entre les données d'une même classe pour déterminer les raisons de ce classement (classification en anglais). Après avoir tiré un certain nombre de règles de classement, il va traiter les données non étiquetées en leur appliquant ces règles de classement pour déterminer la classe à attribuer à la donnée.
Cette approche est très contraignante pour l'expert car le volume de données à étiqueter doit être assez conséquent pour que les règles soient assez précises. De plus tous les experts ne sont pas forcément d'accord entre eux quant aux conclusions auxquelles

aboutir... Par ailleurs, on peut tomber également dans le cas d'un apprentissage par cœur si on laisse l'ordinateur être trop précis dans les détermination de ces règles. Auquel cas le classement ne sera pas correct. Ces contraintes étant connues, si elles sont bien prises en compte, le résultat répond très bien aux besoins de l'utilisateur.

1.1.4/ NON SUPERVISÉ (OU GUIDÉ PAR LES DONNÉES)

Dans ce cadre appelé clustering en anglais, il n'y a aucune intervention d'expert. Il s'agit d'analyser les données les unes par rapport aux autres pour identifier leurs corrélations pour déterminer non seulement le nombre de classes (ou cluster en anglais) et les données qui les forment. L'ordinateur est totalement autonome dans ce cas. Cela permet de traiter de plus gros volumes de données et de gagner du temps et d'obtenir des résultats toujours similaires pour les mêmes données, car il n'y a pas de part humaine dans la démarche. Par contre, les résultats ne sont pas garantis pour répondre à une attente d'expert en sortie de processus, e.g. l'ordinateur peut déterminer l'existence de 3 classes alors que l'expert n'en voyait que 2.

1.1.5/ SEMI SUPERVISÉ

Cette approche permet de combiner à la fois l'objectivité de l'ordinateur et le traitement de gros volumes de données tout en ménageant l'implication fastidieuse de l'expert dans l'étiquetage des données, mais en lui permettant d'ajuster la classification. Par exemple dans le cas d'une analyse d'images cérébrales lors d'une IRM, l'ordinateur va analyser les données de tout le cerveau dans un premier temps, et indiquer ses conclusions. Le médecin qui suit l'IRM peut ainsi décider de vouloir recentrer l'IRM sur une zone particulière qui vient de lui être mise en exergue, et l'ordinateur reprendra ses calculs pour cette zone plus précise. Un dialogue s'instaure ainsi entre les deux.

Dans le cadre de l'Intuition Artificielle, les méthodes doivent être issues de l'analyse non supervisée des données. En effet, l'ordinateur doit se faire une opinion par lui-même à partir des données comme seule connaissance factuelle sans biais extérieur, et d'autre part les experts ne seront pas forcément disponibles pour tous les domaines que l'on souhaiterait traiter, et les délais de traitement sont largement allongés avec des interventions humaines.

1.2/ ANALYSE FORMELLE DE CONCEPTS

Ce domaine de recherche se situe dans le cadre de l'analyse de données non supervisée. Il s'appuie sur des théories mathématiques concernant les ensembles ordonnés et sur lesquelles s'appuie à la théorie des treillis. Rappelons ici les notions nécessaires à la compréhension du domaine.

Definition 1:

Relation binaire. Une relation binaire R entre deux ensembles O et T est un ensemble de couples (o,t) avec $o \in O$ et $t \in T$, i.e. un sous-ensemble de l'ensemble $O \times T$. On écrit oRt pour signifier $(o,t) \in R$, et R^{-1} est la relation inverse de R entre T et O avec $tR^{-1}o \Leftrightarrow oRt$.

Definition 2:

Relation d'ordre. Une relation binaire R sur $T \times T$ est appelée une relation d'ordre (ou simplement **ordre**) si elle satisfait les trois conditions suivantes pour tous les éléments $x, y, z \in T$:

On dit qu'un ordre est total si les éléments d'un ensemble sont tous comparables deux à deux :

$$\forall x, y \quad x \neq y \rightarrow x < y \; ou \; y < x$$

Un ordre qui n'est pas total est dit partiel.

Definition 3:

Ensemble ordonné. Un ensemble ordonné est un couple $(O, \leq)$ avec O un ensemble et $\leq$ une relation d'ordre sur O.

Definition 4:

Couverture. Soit O un ensemble ordonné et $x, y \in O$, on dit que x est couvert par y (ou que y couvre x) si $x < y$ et s'il n'existe pas de $z \in O$ tel que $x < z$ et $z < y$. On note alors $x \prec y$ et x est appelé prédécesseur de y (et respectivement y est appelé successeur de x).

1.2.1/ Diagramme de Hasse

Un diagramme de Hasse est une représentation visuelle d'un ordre fini. Cette représentation utilise des points pour représenter les éléments de O et des segments qui les inter-connectent pour indiquer les relations de couverture.

1.2.2/ Principe de dualité des ensembles ordonnés

Étant donné un ensemble ordonné $(O, \leq)$, on peut construire un nouvel ensemble ordonné $(O, \geq)$ (son dual) en définissant la relation inverse "$\geq$" de "$\leq$" comme étant aussi une relation d'ordre sur O qui sera alors appelée duale [Barbut et al., 1970].

1.2.3/ Treillis et propriétés des treillis

Les treillis ont été étudiés en parallèle aux Etats Unis, en France et en Allemagne. Il existe deux écoles, qui utilisent des vocabulaires différents. Nous définissons dans cette section les treillis de concepts selon l'approche allemande de Ganter et Wille [Ganter et al., 2012] et les treillis de fermés autrement appelés treillis de Galois selon l'approche

française de Barbut et Monjardet [Barbut et al., 1970]. Dans ce document, nous utiliserons de préférence la notation allemande.

Definition 5:

Infimum. Soient $(O, \leq)$ un ordre partiel, A un sous-ensemble de O, et m un élément quelconque de O. m est un infimum de A (noté $\bigwedge A$) si et seulement si :
$\forall a \in A, (m \leq a)$
et $\forall c \in O, (\forall a \in A,\ c \leq a) \rightarrow (c \leq m)$.
Synonymes d'infimum : plus grand commun diviseur, borne inférieure, greatest lower bound en anglais (glb).

Definition 6:

. **Supremum**. Soient $(O, \leq)$ un ordre partiel, A un sous-ensemble de O, et M un élément quelconque de O. M est un supremum de A (noté $\bigvee A$) si et seulement si :
$\forall a \in A, (a \leq M)$
et $\forall c \in O, (\forall a \in A,\ a \leq c) \rightarrow (M \leq c)$.
Synonymes de supremum : plus petit commun multiple, borne supérieure, lowest upper bound en anglais (lub).

Definition 7:

Treillis, treillis complet. Soit O un ensemble ordonné non vide.
O est un **treillis** si chaque couple (x, y) d'éléments de O possède un supremum et un infimum. O est un **treillis complet** si $\bigwedge S$ et $\bigvee S$ existent pour tout $S \subseteq O$.

1.2.4/ APPROCHE ALGÉBRIQUE

1.2.4.1/ PROPRIÉTÉS

Un treillis se définit également comme une structure algébrique $(O, \bigwedge, \bigvee, \leq)$. Il s'agit alors d'un ensemble O muni de deux lois de composition interne $\wedge$ et $\vee$ vérifiant :

- l'associativité : $\forall x, y, z \in O,\ (x \vee y) \vee z = x \vee (y \vee z)$ (resp. pour $\wedge$)
- la commutativité : $\forall x, y \in O,\ x \vee y = y \vee x$ (resp. pour $\wedge$)
- l'idempotence : $\forall x \in O,\ x \vee x = x\ et\ x \wedge x = x$
- l'absorption : $\forall x, y \in O,\ x \wedge (x \vee y) = x\ et\ x \vee (x \wedge y) = x$

Le principe de dualité des ensembles ordonnés s'applique aux treillis, et on obtient le dual d'un ordre en remplaçant chacun des symboles : $\leq, \frown, \smile, \bigvee, \bigwedge$... par $\geq, \smile, \frown, \bigwedge, \bigvee$...

1.2.4.2/ ELÉMENTS PARTICULIER

Dans un treillis, il existe un ensemble d'éléments par lesquels on peut représenter l'intégralité des éléments de ce treillis : les éléments inf et sup-irréductibles.

Definition 8:

Elément sup-irréductible. Considérons le treillis T et $x \in T$ un élément de ce treillis. On dit que x est un élément **sup-irréductible** (aussi noté $\bigvee$-*irréductible*) si et seulement s'il a un et un seul prédécesseur. Autrement dit le supremum de lui-même et de ce prédécesseur $x = x \wedge x_1$. On note l'ensemble de ces sup-irréductibles **J(T)**[a].

a. Pour **J**oin : terme anglais de l'union.

Definition 9:

Elément inf-irréductible. On définit de manière duale les éléments **inf-irréductibles** ($\bigwedge$-*irréductibles*) comme étant ceux n'ayant qu'un et un seul successeur. Un élément inf-irréductible est l'infimum de lui-même et de ce successeur : $x = x \vee x_1$. L'ensemble de ces éléments se note **M(T)**[a].

a. Pour **M**eet : terme anglais de l'intersection.

1.2.5/ TREILLIS DE GALOIS

Definition 10:

Fermeture [Barbut et al., 1970] Une fermeture dans un ensemble ordonné (O, $\geq$) est une isotonie $h : O \to O$ extensive et idempotente, c'est-à-dire, une application ayant les trois propriétés :

1. $\forall x, \forall y,\ x \geq y \Rightarrow h(x) \geq h(y)$ (monotone croissante)
2. $\forall x,\ h(x) \geq x$ (extensive)
3. $\forall x,\ h \circ h(x) = h(x)$ (idempotente)

Un élément x de O est dit fermé s'il est égal à sa fermeture : $x = h(x)$

Soient O et T deux ensembles ordonnés quelconques, et $\uparrow$: $O \to T$ et $\downarrow$: $T \to O$ deux applications monotones décroissantes :

$$\forall x, \forall y \in O, x \leq y \Rightarrow x^{\uparrow} \geq y^{\uparrow}$$
$$\forall x', \forall y' \in T, x' \leq y' \Rightarrow x'^{\downarrow} \geq y'^{\downarrow}$$

Considérons les applications composées $h = \downarrow\uparrow$ ($\uparrow$ suivie de $\downarrow$) de O dans O, et $h' = \uparrow\downarrow$ de T dans T. Ces deux applications sont monotones croissantes. Si en outre elles sont extensives alors elles sont idempotentes et sont donc des fermetures dans O et T respectivement.

Definition 11:

Connexion et fermeture de Galois. Le couple $(\uparrow, \downarrow)$ d'applications est une correspondance de Galois entre les ensembles ordonnés O et T. Les fermetures $h = \downarrow\uparrow$ dans O et $h' = \uparrow\downarrow$ dans T associées à une correspondance de Galois $(\uparrow, \downarrow)$ sont appelées fermetures de Galois [Barbut et al., 1970].

1.2.6/ Analyse de Concepts Formels

L'*Analyse de Concepts Formels* (FCA) dérive des mathématiques appliquées et se base sur les travaux de Birkhoff [Birkhoff, 1948] en 1940 (révisés en 1948 puis 1967). Sur ces travaux s'appuient ensuite ceux de Ganter et Wille en 1999 (repris en 2012) [Ganter et al., 2012] qui définissent les contextes et concepts formels.

1.2.6.1/ Contexte formel

Un contexte formel traduit une relation binaire entre un ensemble de propriétés et un ensemble d'objets.

Definition 12:

Un **contexte formel** $K := (O, T, I)$ est composé de deux ensembles O et T et d'une relation binaire I entre O et T. Les éléments de O sont appelés des objets et les éléments de T sont appelés des attributs. Pour exprimer qu'un objet o est en relation I avec un attribut t, on écrit oIt ou $(o, t) \in I$, ce qui se lit "l'objet $o \in O$ possède l'attribut $t \in T$".

Definition 13:

Concept formel. Soit la connexion de Galois suivante définie sur le contexte (O, T, I) :

$$\uparrow: 2^O \to 2^T$$
$$A \subseteq O \to A^\uparrow = \{t \in T | \forall o \in A, oIt\}$$

$$\downarrow: 2^T \to 2^O$$
$$B \subseteq T \to B^\downarrow = \{o \in O | \forall t \in B, oIt\}$$

Avec $A \subseteq O$ et $B \subseteq T$. Les couples (A, B) $\in 2^O \times 2^T$ d'ensembles vérifiant $A = B^\downarrow$ et $B = A^\uparrow$ sont appelés des concepts formels (ou simplement concepts). A est alors appelé l'extension du concept et B son intension. L'extension est l'ensemble de tous les objets partageant les attributs de l'intension. L'intension est l'ensemble de tous les attributs partagés par tous les objets de l'extension.
Le sous-ensemble A de O est l'extension du concept unique $(A, A^\uparrow)$ si et seulement si $A^{\uparrow\downarrow} = A$. On définit dualement les sous-ensembles B de T en tant qu'intensions. L'ensemble de tous les concepts du contexte (O, T, I) est noté par $\mathfrak{B}(O, T, I)$.

Par exemple, dans le contexte de la table 1.2, le concept ({Alien, 28 jours plus tard},{GB}) est un concept indiquant que ces deux films (et aucun autre) ont comme caractéristique(s) commune(s) d'être britanniques.

La table 1.2 permet d'introduire la notion de contexte, cependant classiquement en Analyse de Concepts Formels ce contexte sera plutôt présenté comme dans la table binarisée 1.3.

Film	Catégorie	Nationalité	Durée
La cité de la peur	Comédie	FR	1h30
Alien	Fant./SF	GB	1h30
Men in Black	Fantastique/SF	US	1h45
Stargate	Fantastique/SF	US	2h
Une nuit en enfer	Comédie	US	1h45
Scary movie	Comédie	US	1h30
28 jours plus tard	Horreur	GB	2h

Table 1.2 – Contexte filmographique.

	Catégorie			Nationalité			Durée		
Film	Horreur	Comédie	Fantast./SF	US	FR	GB	1h30	1h45	2h
La cité de la peur		×			×		×		
Alien			×			×	×		
Men in Black			×	×				×	
Stargate			×	×					×
Une nuit en enfer		×		×				×	
Scary movie		×		×			×		
28 jours plus tard	×					×			×

Table 1.3 – Contexte filmographique multivalué.

1.2.6.2/ ORDRE SUR LES CONCEPTS

Definition 14:

Ordre sur les concepts. Soit (O, T, I) un contexte. Soient $(A_1, B_1) \in \mathfrak{B}(O, T, I)$ et $(A_2, B_2) \in \mathfrak{B}(O, T, I)$, on écrit $(A_1, B_1) \leq (A_2, B_2)$, si et seulement si $A_1 \subseteq A_2$.
$A_1 \subseteq A_2$ implique $A_1^{\uparrow} \supseteq A_2^{\uparrow}$, et inversement parce que $A_1^{\uparrow\downarrow} = A_1$ et $A_2^{\uparrow\downarrow} = A_2$. Ainsi nous avons $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$.
(A_1, B_1) est alors appelé sous-concept de (A_2, B_2) qui, lui, est son super-concept.

Definition 15:

Treillis de concepts. L'ensemble $\mathfrak{B}(O, T, I)$ muni de l'ordre $\leq$ définit un treillis dit treillis de concepts et se note $< \mathfrak{B}(O, T, I); \leq >$.

1.2.6.3/ REPRÉSENTATION DES TREILLIS ET HÉRITAGE

Un treillis est représenté par un diagramme de Hasse où chaque nœud est un concept. Il peut être représenté de manière simplifiée, c'est-à-dire que les individus et propriétés indiqués sont ceux qui sont propres au concept et non pas ceux hérités. En effet, l'ordre permet également de définir un héritage entre les concepts. Ainsi chaque concept hérite des objets des concepts qui sont entre lui et $\perp$ (ses sous-concepts) et des attributs qui sont entre lui et $\top$ (ses super-concepts). C'est pour cela qu'un treillis est dit être une double hiérarchie de concepts. Nous appellerons donc cet ordre un ordre hiérarchique.

1.2.7/ TREILLIS SUR DES CONTEXTES NON TRIVIAUX

1.2.7.1/ CONTEXTE MULTIVALUÉ

La notion de contexte formel a été étendue par Ganter et Wille [Ganter et al., 1997] pour introduire le fait qu'un attribut puisse avoir plusieurs valeurs comme l'illustre l'exemple présenté dans la table 1.2. Dans le cadre de l'Intuition Artificielle il n'y a pas nécessité à préciser leur définition ni leur traitements.

1.2.7.2/ STRUCTURES DE PATRONS

Nous abordons ici d'une autre représentation des données, les structures de patrons, appelées en anglais : "pattern structures" et présentée par [Ganter et al., 2001]. Il s'agit d'une représentation générale pour les échelonnages conceptuels. Les objets possédent alors des descriptions ou "patrons" (*"patterns"* en anglais) qui sont structurées dans un demi-treillis et permettent des opérations latticielles.

Definition 16:

Structure de patrons. Soit G un ensemble, $(D, \sqcap)$ un inf-demi-treillis et $\delta : G \to D$ une application. Alors $(G, \underline{D}, \delta)$ avec $\underline{D} = (D, \sqcap)$ est appelé une structure de patrons impliquant que l'ensemble

$$\delta(G) := \{\delta(g) \mid g \in G\}$$

génère un sous-demi-treillis $(D_\delta, \sqcap)$ de $(D, \sqcap)$[a].

a. En considérant que chaque sous-ensemble X de $\delta(G)$ possède un infimum $\sqcap X$ de $(D, \sqcap)$ et que D_δ est l'ensemble de ces infima.

Les éléments de D sont appelés des patrons. L'ordre sur D, appelé ordre de subsomption est donné par :

$$c \sqsubseteq d :\Leftrightarrow c \sqcap d = c$$

Une structure de patrons s'interprète comme étant un ensemble d'objets avec des "descriptions" (patrons) acceptant entre elles une "opération de similarité" $\sqcap$, i.e. une opération telle que pour un ensemble quelconque d'objets, elle produise une "description" qui représente les similitudes entre les objets du sous-ensemble. La similarité est indépendante de l'ordre dans lequel les objets apparaissent, c'est une opération qui idempotente, commutative et associative.

Definition 17:

Opérateurs de fermeture. Si $(G, \underline{D}, \delta)$ est une structure de patrons, alors on définit les opérations de fermeture :

$$A^\square := \bigsqcap_{g \in A} \delta(g) \quad pour\ A \subseteq G$$

$$d^\square := \{g \in G \mid d \sqsubseteq \delta(g)\} \quad pour\ d \in D$$

Les opérateurs $(.)^\square$ forment une connexion de Galois entre l'ensemble des parties de G

et $(D, \sqsubseteq)$. Les couples (A, d) satisfaisant

$$A \subseteq G, \quad d \in D, \quad A^{\square} = d \quad et \quad A = d^{\square}$$

sont appelés des concepts patrons (en anglais : pattern concepts) de $(G, \underline{D}, \delta)$, avec A comme extension et d comme patron d'intension.

1.2.7.3/ CONTEXTE FLOU

La notion de contexte flou a été introduite dans [Zadeh, 1965] pour présenter un contexte où tout rattachement d'un attribut à un objet serait une question de degré d'appartenance. Elle a été étendue dans [Bělohlavek, 1999, Bělohlavek et al., 2005]. Des données floues permettent de représenter une relation floue entre un objet et un trait, ou, autrement dit : l'affinité d'un objet pour un trait.

Definition 18:

Contexte flou. Soit A un ensemble de degrés de vérité ou degrés d'appartenance. Un contexte flou est un triplet $K := (O, T, I)$ où O et T sont respectivement des ensembles d'objets et d'attributs, et $I : O \times T \rightarrow A$ est une relation floue entre O et T. Un degré $I(o, t) \in A$ est interprété comme un degré d'appartenance de l'attribut t pour l'objet o.

La table 1.4 présente un contexte flou de notre exemple. $I(Alien, Horreur) \in 40\%$ signifie que le film Alien est à 40% un film d'horreur[1].

Il est à noter que la définition d'un contexte formel est un cas particulier de la définition d'un contexte flou dans le sens où A est réduit à l'ensemble binaire {0,1}.

Film²	*Horreur*	*Comédie*	*Fant./SF*	*US*	*FR*	*GB*	*1h30*	*1h45*	*2h*
La cité de la peur	10%	90%			100%		40%	60%	
Alien	40%		60%	50%		50%	60%	40%	
Men in Black	5%	40%	55%	100%			47%	53%	
Stargate			100%	50%	50%				100%
Une nuit en enfer	30%	40%	30%	100%				87%	13%
Scary movie	40%	60%		100%			100%		
28 jours plus tard	90%		10%			100%		53%	47%

Table 1.4 – Contexte filmographique flou.

1.2.8/ CONNEXION DE GALOIS PAR SIMILARITÉ

Dans sa thèse [Messai, 2009], N. Messai s'intéresse à l'annotation sémantique. Le contexte sur lequel il travaille est multivalué et il propose une analyse par similarité. Dans une

1. Ces pourcentages n'ont qu'un caractère illustratif.

FCA standard deux objets sont regroupés s'ils ont un même attribut. En FCA par similarité, deux objets sont regroupés s'ils ont des valeurs similaires pour un même attribut.

La fonction de similarité entre les valeurs des attributs est calculée à partir de relations sémantiques (comme une distance dans une hiérarchie) ou par un paramètre dans la définition de partage d'attributs (par un seuil de similarité).

Les relations d'ordre partiel définies pour un tel treillis sont :

- Sur les objets : l'inclusion $\subseteq$
- Sur les attributs : l'inclusion par similarité $\subseteq_S$: Soient un contexte multivalué (O,T,M,I), t_1 *et* $t_2 \in T$, m_1 *et* $m_2 \in M$, B_1 *et* $B_2 \in \mathfrak{P}(T \times \mathfrak{P}(M))$.

  $$B_1 \subseteq_S B_2 \; ssi \; \forall t_{1|m_1} \in B_1, \; \exists \, t_{2|m_2} \in B_2 \; tel \; que \; t_1 = t_2 \; et \; M_1 \subseteq M_2.$$

  ($t_{|m}$ peut s'illustrer dans le contexte filmographique par *Nationalité*$_{|FR}$ et se lit "la nationalité française").

Les opérateurs de dérivation qui en découlent sont :

- Pour tout $A \subseteq O$ l'ensemble des éléments de $T \times \mathfrak{P}(M)$ partagés par A est :

  $$A^{\uparrow S} = \{t_{|\{t(o), o \in A\}} \in T \times \mathfrak{P}(M) \mid \forall o_i, o_j \in A, t(o_i) \simeq t(o_j)\}$$

- Pour tout $B \subseteq T \times \mathfrak{P}(M)$, l'ensemble des objets de O partageant B est :

  $$B^{\downarrow S} = \{o \in O \mid \forall t_{|M_t} \in B, t(o) \simeq m \; \forall m \in M_t\}$$

$\uparrow S$ et $\downarrow S$ forment une connexion de Galois appelée connexion de Galois par similarité entre $(O, \subseteq)$ et $(T \times \mathfrak{P}(M), \subseteq_S)$

Un concept multivalué (A, B) est tel que $A \subseteq O$, $B \subseteq T \times \mathfrak{P}(M)$, $A^{\uparrow S} = B$ et $B^{\downarrow S} = A$

1.3/ RÈGLES D'ASSOCIATION

L'extraction de ces règles est un problème important dans le domaine de l'extraction de connaissances et tire son origine de l'analyse de bases de données commerciales dans le but d'optimiser les ventes dans le cadre du "panier de la ménagère". Leur intérêt les a vite fait dépasser ce stade et maintenant les règles d'associations sont utilisées pour des applications très variées sortant largement du domaine commercial.

Dans [Agrawal et al., 1996], les auteurs définissent une règle d'association comme une relation d'implication entre deux ensembles d'articles X et Y. $X \rightarrow Y$ indique que les transactions contenant les articles de l'ensemble X ont tendance à contenir ceux de l'ensemble Y.
X est appelé condition ou prémisse, et Y résultat ou conclusion.

Afin de déterminer la qualité d'une règle, on utilise généralement différentes mesures :

- le support qui indique la probabilité que X et Y soient vrais en même temps. Il s'écrit formellement : $X \cup Y$. Par exemple le pourcentage de fois où de la bière et des bretzels ont été achetés ensemble. On appelle support d'un itemset X le rapport entre le nombre de transactions contenant X et le nombre total de transactions de D

- la confiance qui indique la probabilité que Y soit vrai sachant que X est vrai. Elle s'écrit formellement : $\frac{support(X \cup Y)}{support(X)}$. La confiance traduit par exemple, la probabilité qu'un client qui a acheté des bières achète aussi des bretzels.

- Le lift : cet indicateur s'appuie sur le principe qu'une règle, pour être intéressante, doit avoir une confiance supérieure à la probabilité absolue du résultat. Si un résultat a 80% de chances d'arriver dans l'absolu, une règle où le même résultat a 40% de chances de se produire (confiance de 40%) n'est pas intéressante. Un lift jugé intéressant est supérieur à 1.

- La progression brute peut être calculée en soustrayant la probabilité du résultat du lift à la confiance. Une règle pertinente a une progression brute supérieure à 0. Il est à noter que lorsqu'une règle est jugée inutile car elle a un lift et une progression brute trop faible, la règle inverse est nécessairement intéressante.

- La capacité de déploiement permet de donner la proportion des individus vérifiant les conditions mais ne satisfaisant pas encore le résultat.

Dans le cadre des treillis de Galois, nous pouvons extraire des itemsets fréquents, i.e. des ensembles d'attributs regroupés ensemble selon un support (nombre d'individus partageant ces attributs) supérieur à un seuil et des règles d'association avec une certaine confiance, soulignant les corrélations entre les ensemble d'attributs [Napoli, 2006]. Nous nous intéressons plus particulièrement aux implications [Duquenne et al., 1986] qui sont des règles d'association pour lesquelles la confiance est de 100%.

Une implication $A \rightarrow B$ concerne les attributs A et B d'un même concept. Le support d'une implication est la taille de l'extension du concept, c'est à dire le nombre d'objets partageant à la fois les attributs A et B.

Il est important de noter que le nombre de règles possibles dans un ensemble est très important. Pour un problème avec n variables, le nombre de règles possibles est donné par la formule $2 * n^{n-1} - 1$. Devant ce nombre de possibilités important, des moyens peuvent être mis en place pour limiter l'exploration à un certain type de règles. Par exemple, si la fréquence d'une variable (égale au nombre de fois où la variable vaut vraie ou 1) est trop faible, la variable ne sera pas utilisée comme condition d'une règle d'association. Un seuil de fréquence peut donc être fixé pour les variables utilisées comme conditions d'une règle.

De nombreux algorithmes existent pour traiter l'extrapolation des règles d'association à partir de jeux de données.
Le plus ancien et le plus populaire s'appelle l'algorithme APRIORI [Agrawal et al., 1994]. Le pseudo-code de l'algorithme APRIORI est donné dans 1.3 [Berka et al., 2010], où $minsup$ représente le seuil de prise en charge minimum, C en tant qu'éléments fréquents candidats, L_k est l'ensemble d'éléments avec k membres.

Algorithm 1 APRIORI

Require: Dataset{Entrée}
Ensure: Règles d'association{Sortie}
  Attribuer tous les éléments qui ont atteint le support de minsup dans L_1
  Let $k = 2$
  while $L_{K-1} \neq \emptyset$ **do**
    appeler la fonction apriori-gen pour créer un ensemble de candidats C_k à partir de L_{K-1}
    attribuer tous les itemsets de C_k ayant atteint le support de minsup à L_k
    augmenter k de 1
  end while

Algorithm 2 Fonction apriori-gen

Require: L_{K-1} {Entrée}
  for all itemsets C_b, C_q de L_{K-1} **do**
    if C_p et C_q partagent $k-2$ items **then**
      ajouter C_p et C_q to C_k
    end if
  end for
  for all itemsets C de C_k **do**
    if $\forall\ itemset \subset C \wedge |itemset| = k-1 \wedge itemset \notin L_{k-1}$ **then**
      retirer C de C_k
    end if
  end for

D'autres algorithmes plus récents tels que Eclat et Elimination récursive améliorent les algorithmes APRIORI d'origine [Rohit, 2016, Girotra et al., 2013].

2

Ma thèse : Treillis de Galois pour les contextes multivalués flous : application à l'étude des traits de vie en hydrobiologie

L'objet principal de ma thèse était de développer une approche pour la fouille de données complexes par l'analyse de concepts formels.

Cette thèse s'intéresse à l'étude de données particulières représentées par des *contextes multivalués flous* qui sont des contextes multivalués (dont les attributs sont divisés en modalités) pour lesquels la relation entre objets et modalités est floue, c'est à dire qu'un objet possède une modalité selon un certain pourcentage appelé *affinité*.

Deux démarches sont présentées pour traiter ces contextes. La première consiste à adapter les données aux traitements classiques par une binarisation. La seconde consiste à adapter les méthodes aux données.

Les méthodes sont ensuite confrontées au domaine hydrobiologique pour la sélection d'attributs écologiques et leur mise en correspondance avec des taxons et leurs attributs biologiques.

2.1/ Traitement binaire

Deux approches de binarisation ont été étudiées.

2.1.1/ Binarisation par disjonction totale

La première est une disjonction totale par un *échelonnage nominal*, nous fournissant un *contexte nominal* (dit "Lxy"). Ce contexte binaire permet alors d'utiliser des *règles d'implications*. Leur intérêt porte, d'une part, sur les relations qu'elles mettent en exergue entre les attributs, à travers les implications fréquentes et à fort support, mais également par les implications négatives ou de faible support. D'autre part, elles permettent d'ajuster le jeu de données, corriger les éventuelles erreurs, ajuster la plage des affinités, limiter la dispersion d'information au sein des modalités et les redondances d'informations entre les attributs et les modalités et ôter les attributs non pertinents. Par ailleurs, pour se

rapporter à des approches plus habituelles en biologie, des méthodes statistiques sont utilisées puisque le contexte est binaire et donc le permet. La comparaison et la combinaison de l'approche latticielle et de l'approche statistique, notamment par l'Analyse Factorielle Multiple (AFM) s'est montrée efficace car elle permet d'alléger l'approche latticielle très sensible au volume de données et améliore les conclusions que l'AFM est capable de fournir.

La combinaison des deux techniques se fait par une première étape de pré-sélection des données par l'AFM sur lesquelles est ensuite construit le treillis. Cela permet d'évaluer les apports réciproques de chaque méthode de laquelle se dégage un double intérêt. D'une part les treillis permettent de préciser les résultats de l'AFM. D'autre part l'AFM permet une pré-segmentation des données qui permet de construire des treillis plus petits et mieux focalisés. Différents niveaux de complexité peuvent être explorés, en fonction du nombre d'individus/propriétés retenus sur les axes de l'AFM ou en utilisant les informations de plusieurs axes. Ceci permet alors de sélectionner les données les plus pertinentes à prendre en compte et ainsi alléger les résultats à interpréter, ce qui est utile lorsque les contextes sont importants, et que les treillis deviennent rapidement volumineux et illisibles donc inexploitables. Pour palier ce type de contrainte, des travaux ont été menés, notamment par l'utilisation de calcul de χ^2 [**?**].

2.1.2/ BINARISATION PAR ÉCHELONNAGE HISTOGRAMME

La seconde méthode de binarisation est un échelonnage histogramme. Les notions d'histogrammes et d'échelonnage histogramme sont formellement définies dans un premier temps. Les histogrammes permettent de prendre en considération la répartition des affinités entre les différentes modalités de chaque attribut et qui forment ainsi la partie floue des données.

La comparaison des treillis nominal et histogramme pointe une lecture du treillis histogramme plus aisée, non seulement parce qu'il est moins volumineux, mais aussi parce qu'un concept histogramme regroupe l'équivalent de 3 ou 4 concepts Lxy. Pour un objet (ou groupe d'objets), un seul concept histogramme résume directement l'information sans avoir besoin de rechercher dans le treillis. L'inconvénient est qu'il faut une égalité des distributions d'affinités pour pouvoir construire les concepts.

2.1.3/ APPLICATION À LA HYDROBIOLOGIE

L'application de cette thèse à la biologie, présente la difficulté de traiter des données réelles d'un domaine avec lequel il faut se familiariser et qui est très complexe. Cependant, il offre également l'avantage de travailler sur des données et d'avoir des experts présents pour interpréter les résultats.

Par ailleurs, l'approche des treillis est novatrice dans ce domaine, où les biologistes se réfèrent préférablement à des méthodes statistiques. Etablir une comparaison et une combinaison des deux méthodes a permis de convaincre les biologistes de l'intérêt de l'approche latticielle. Les implications sont aussi très appréciées par les utilisateurs car elles se lisent facilement. Les informations qu'elles font ressortir sont intéressantes, cependant c'est lorsqu'on les combine qu'on découvre des relations particulièrement

digne d'intérêt, même si ce travail est rapidement complexe à cause des nombreuses combinaisons possibles et on peut facilement manquer certaines informations.

Concernant les conversions binaires, que l'on soit intéressé par le diagramme de Hasse ou uniquement par la liste des concepts, dans le cas de contextes multivalués flous, l'échelonnage histogramme est à préférer. D'une part parce que les attributs histogrammes sont moins nombreux que les attributs Lxy, et que le nombre de concepts générés par des connexions de Galois binaires et donc leur temps de calcul, sont fonction du nombre d'attributs. D'autre part, car ils permettent de conserver l'information de distribution des valeurs entre les modalités d'un trait.
les histogrammes sont moins nombreux et ils concernent donc plus d'objets contrairement aux attributs Lxy qui sont beaucoup trop spécifiques à chaque objets et les extensions (ensemble des objets concernés par les attributs) des concepts sont alors peu fournies.

2.2/ TRAITEMENT FLOU

La seconde approche pour le traitement des données multivaluées floues consiste à adapter les treillis aux données et passe par la définition de nouvelles connexions de Galois que appelées connexions min-max. Ces connexions manipulent des histogrammes et permettent de mettre en évidence les propriétés minimales et maximales communes à un groupe d'objets. Ainsi il est possible de comparer non seulement des distributions d'affinités sur les modalités mais également des "formes" de distribution (représentées par les histogrammes), ce qui est particulièrement significatif pour des modalités ordonnées ou temporelles. Par ailleurs, les fonctions sur lesquelles les connexions min-max sont établies sont croissantes monotones, et donc le nombre de concepts générés et les temps de calculs augmentent très rapidement avec le nombre d'objets et d'attributs. Pour contrer cet inconvénient, une notion de seuillage est introduite dans la fermeture de Galois qui ne conserve que les concepts dont l'écart entre les propriétés minimales et maximales est inférieur au seuil. Diverses options de seuillage ont été évaluées et la plus adaptée apparait être l'option d'un seuillage par trait et par quantile. Cette option permet non seulement de réduire le temps de calcul et le nombre de concepts générés, mais également de générer les concepts les plus pertinents.

Les histogrammes peuvent avoir des sens très différents en fonction des attributs et des modalités considérés. Ainsi avec des modalités qualitatives nous conseillons de conserver l'échelonnage histogramme actuel. En outre, pour des modalités quantitatives, même si on peut tout à fait conserver l'échelonnage histogramme proposé, il peut aussi être abordé différemment. Par exemple, considérons un attribut "distance" divisé en 3 modalités : courte (< 1km), moyenne (< 2km) et longue (> 2km). L'échelonnage considéré dans nos travaux estime qu'un individu parcourant 500 m parcours une petite distance. Cependant, un autre échelonnage pourra considéré que 0,5 < 1 mais également que 0,5 < 2. Auquel cas, 500 m représentent une petite et une moyenne distance.

2.3/ APPLICATION

Ces travaux ont été appliqués à l'hydrobiologie dans le cadre du projet Indices, mis en œuvre suite à la directive cadre européenne sur l'eau qui préconise l'évaluation de l'état écologique de toutes les masses d'eau d'ici 2015. Actuellement cette évaluation s'opère par des analyses physico-chimiques ou des prélèvements au sein de la faune et la flore des cours d'eau. La physico-chimie est précise mais ne conserve pas les traces des pressions antérieures aux prélèvements. Les paramètres biologiques quant à eux sont très compartimentés et difficilement comparables. De plus, les 5 indices normalisés qui en dépendent ne fonctionnent presque qu'en France, et cependant, pas de manière uniforme d'une hydro-éco-région à une autre. Ainsi un nouvel outil est nécessaire. Dans cette thèse, nous avons suivi l'idée de rechercher un outil basé sur les traits écologiques qui caractérisent le comportement des taxons au sein de leur environnement. Nous avons étudié les macrophytes et invertébrés de la plaine d'Alsace. Le chapitre 3 décrit une méthode de sélection des traits écologiques basée sur le traitement binaire du contexte histogramme des traits biologiques des macrophytes. En effet, il existe de nombreux traits écologiques mais nous recherchons les plus pertinents pour qualifier l'état écologique des cours d'eau en fonction des espèces et de leurs traits biologiques (morphologiques et physiologiques). Pour cela, les concepts sont analysés par un expert pour relever les relations entre les espèces des concepts sélectionnés et les traits écologiques (par rapport à leur environnement). Ensuite ces traits écologiques sont ajoutés au jeu de données et leurs affinités sont renseignées en accord avec la littérature et les connaissances de l'expert. Les concepts du treillis étendu sont ensuite examinés pour valider le processus et mettre en évidence les ensembles des traits biologiques et écologiques. Conformément à nos attentes, nous retrouvons les informations fournies par l'expert, et plus encore, le treillis nous fournit davantage d'informations et des informations plus précises.

Cette méthode a permis la sélection de 5 traits écologiques: stabilité de l'eau, tolérance à la sédimentation, tolérance aux variations d'humidité, tolérance aux matières organiques et tolérance à la trophie. Ces traits ont pu ensuite être associés à des groupes de plantes et de traits biologiques grâce aux connexions min-max. Ceci permet d'une part de pouvoir déterminer un profil écologique en rencontrant une espèce ou une autre sur un cours d'eau, car toutes les espèces d'un même groupe partagent les mêmes traits écologiques. D'autre part, cette analyse est utilisable dans d'autres hydro-éco-régions car si les plantes de ces groupes n'y vivent pas, d'autres espèces partageant les mêmes attributs biologiques peuvent les remplacer.

Enfin, les algorithmes ont été éprouvés également sur des données concernant des invertébrés de la plaine d'Alsace qui forment un contexte 16 fois plus volumineux que celui des macrophytes. Grâce au seuillage, les algorithmes aboutissent et nous fournissent des concepts plus intéressants pour l'analyse car en nombre plus réduit et plus pertinents.

Lors de la seconde phase de l'analyse des données qui associe les traits écologiques aux traits biologiques, nous nous sommes aperçues que certains traits écologiques qui avaient été sélectionnés par l'approche binaire, sont rarement mis en évidence. En effet, certains traits sont importants et récurrents tels que la tolérance à la trophie et à la sédimentation mais d'autres ne caractérisent jamais ou rarement un concept tels que la stabilité de l'eau ou la tolérance à la matière organique pour lesquelles les plantes sont équitablement réparties et ne manifestent donc pas de préférence. Cela implique que

ces traits sont informatifs à propos du milieu mais ne permettent pas de segmenter les espèces ou pointer un biais dans la méthode.

2.4/ DISCUSSION ET PERSPECTIVES

Au cours de ma thèse deux verrous principaux ont été levés, mais leur levée à souligner le besoin d'autres avancées.

D'un point de vue hydrobiologique, des analyses ont été débutées sur les invertébrés, cependant l'interprétation reste à terminer. En effet, la FCA fournit des concepts qui regroupent des éléments cohérents entre eux. Cependant la cohérence n'est pas explicite. Les attributs qui sont partagés traduisent une idée sous-jacente que l'expert doit identifier et interpréter. Une automatisation de cette tâche permettrait de simplifier le travail des experts. Cette piste a été suspendue depuis la fin de ma thèse mais en travaillant sur l'Intuition Artificielle j'ai pu revenir sur cette problématique. En effet, le besoin de qualifier la connaissance contenue dans les contextes est incontournable avec la volonté de contruire automatiquement une ontologie capable d'expliquer l'intuition de la machine en nommant les concepts qui deviendront les classes de l'ontologie. J'ai pu monter un projet de recherche industriel, me permettant le recrutement d'un post doctorant pour travailler sur cette partie (développée dans le chapitre 11), il a été appuyé par une doctorante de l'universite de Sfax m'ayant sollicitée pour l'encadrement de son stage de thèse à l'étranger.

Par ailleurs, il aurait été intéressant d'intégrer les paramètres physico-chimiques dans les concepts au même titre que les traits de vie (biologiques et écologiques). Les données physico-chimiques existent et sont, comme pour les espèces, associées aux stations où elles sont mesurées. Des travaux statistiques par des méthodes RLQ sont en cours à ce propos, et une étude par une Analyse Relationnelle de Concepts (ARC) a été esquissée également pour commencer à relier la physico-chimie aux traits biologiques. Précisons rapidement ce qu'est l'ARC et comment elle peut concerner nos données. L'ARC se base sur l'analyse de concepts formels et permet de prendre en compte plusieurs contextes formels et de regrouper des objets partageant les mêmes attributs qui peuvent être relationnels.

Le modèle de données de l'ARC est une famille de contextes relationnels (FCR) [**?**, **?**] définie ainsi :

Definition 19:

Famille de Contextes Relationnels. Une famille de contextes relationnels est un couple (K;R). K est un ensemble de contextes formels $K_i = (O_i;A_i; I_i)$, R est un ensemble de contextes relationnels $R_j = (O_k;O_l; I_j)$ (O_k et O_l sont les ensembles d'objets des contextes K_k et K_l de K).

Dans notre cas, nous disposons des contextes K_k = ({stations de prélèvement};{paramètres physico-chimiques}; I_k), K_l = ({taxons};{traits de vie}; I_l), et du contexte relationnel R_j = ({stations de prélèvement};{taxons}; I_j).

Déjà au moment de ma thèse le besoin d'intégrer d'autres dimensions étaient nécessaires. Là encore, j'ai du attendre mes travaux sur l'Intuition Artificielle pour reprendre cette problématique dans et lever ce verrou que j'explique dans le chapitre suivant.

3

EXTENSION DE LA FCA VERS LA MULTIDIMENSIONNALITÉ

Contribution : Alexandre Bazin recruté comme post doctorant sur le contrat Eurostars Predictive Smart Data Platform (détails dans la section 14.1.2)

La plupart des jeux de données réels simples prennent la forme de contextes formels et les modèles intéressants sont souvent des variations sur le thème des concepts formels, ce qui rend la FCA bien adaptée aux applications de tout domaine traitant des données [Carbonnel et al., 2017, Snelting, 2000, Kaytoue et al., 2011, Valtchev et al., 2004] tant que celles-ci restent simples. Cependant,concrètement, face à la complexité croissante des données, la FCA nécessite des extensions et des généralisations telles que des approches floues ou multidimensionnelles [Burusco Juandeaburre et al., 1994, Belohlavek, 2011, Lehmann et al., 1995, Voutsadakis, 2002].

Dans le cadre de l'Intuition Artificielle, il faut considérer deux aspects :

- la connaissance d'un domaine s'établit sur un ensemble de données venant d'horizons différents qui se complètent. Cette hétérogénéité (la Variété en Big Data) doit être prise en compte. C'est pourquoi je m'intéresse à l'extension de la FCA vers la multidimensionnalité.
- pour subodorer, la machine réfléchit dans un paradigme de graphes (partie 6). Les contributions de cette partie sont donc énoncées dans ce cadre.

Classiquement, en Analyse Foremelle de Concepts, les contextes formels sont bidimensionnels (i.e. des graphes bipartis). Il s'agit donc ici de généraliser les notions de contextes et de concepts à des graphes qui ne sont pas bipartis.

3.1/ CONSIDÉRATION DE LA FCA DANS LE PARADIGME DES GRAPHES

Les contextes formels sous leur forme de base sont des tables binaires - i.e. des graphes bipartis pour lesquels une bipartition en ensembles indépendants est donnée. L'une des

plus importantes généralisations de la FCA, l'Analyse de Concepts Polyadique (ACP) [Voutsadakis, 2002], traite des mêmes notions de contexte et de concept lorsque ce contexte est un hypergraphe n-uniform [1] n-partite [2] - modélisant la majorité des ensembles de données multidimensionnelles. Dans l'ACP, une partition n de l'hypergraphe est donnée. Quelleque soit la variante de la FCA, le nombre de dimensions est la taille des n-uplets des données.

Dans ce travail mené avec Alexandre Bazin[3], nous avons estimé qu'il serait intéressant, en définitive, de généraliser la FCA à des hypergraphes n-partis non n-uniformes afin de créer de nouvelles possibilités d'applications impliquant des données différentes, notamment celles inhérentes à l'Intuition Artificielle. Dans ce travail, comme première étape vers cet objectif, nous nous sommes concentrés sur le cas des graphes n-partitionnés (hypergraphes 2-uniformes) avec $n > 2$.

La première partie de cette section définit les "concepts" correspondants, étudie brièvement la complexité de leur énumération et montre qu'ils forment un n-treillis complet, ce qui implique que des algorithmes connus peuvent être utilisés pour les calculer.

3.1.1/ Analyse Formelle de Concepts Multidimensionnelle

Les notions de contextes et de concepts formels ont été largement étudiés et sont utilisés avec succès dans divers domaines tels que l'exploration et l'analyse de données, la récupération d'informations, la correction d'erreurs de code source, l'apprentissage automatique et la construction de taxonomies et d'ontologies [Škopljanac Mačina et al., 2014]. La généralisation multidimensionnelle de la FCA ou de l'ACP [Voutsadakis, 2002] a suscité relativement moins d'attention, mais constitue un domaine tout autant théorique qu'applicatif, dont voici les bases :

Definition 20:

n-contexte. Un n-contexte est un tuple $(S_1, \ldots, S_n, R)$ où S_i, $i \in \{1, \ldots, n\}$, est un ensemble appelé **dimension** et $R \subseteq \prod_{i \in \{1,\ldots,n\}} S_i$ est une **n-ary relation**.

Un n-contexte peut etre représenté par une table n-dimensionnelle.

	a	b	c	a	b	c	a	b	c
1	×	×		×			×		
2	×			×			×	×	
3	×			×		×			×
		α			β			γ	

Figure 3.1 – Un 3-contexte $(\{1,2,3\}, \{a,b,c\}, \{\alpha,\beta,\gamma\}, R)$

1. i.e. hypergraphe tel que toutes ses hyper-arêtes ont la taille n
2. i.e. l'ensemble de sommets de graphe est décomposé en n ensembles disjoints de sorte qu'aucun sommet de graphe du même ensemble ne soit adjacent
3. Post doctorant sur le projet Eurostars PSDP.

Definition 21:

n-concept Soit $C = (S_1, \ldots, S_n, R)$ un n-contexte. Un n-concept de C est un n-tuple $(T_1, \ldots, T_n)$ tel que $T_i \subseteq S_i$, $\prod_{i\in\{1,\ldots,n\}} T_i \subseteq R$ et tel qu'il n'existe pas $d \in \{1, \ldots, n\}$ et $k \in S_d \setminus T_d$ tel que $(T_1, \ldots, T_d \cup \{k\}, \ldots, T_n)$ respecte cette propriété.

Autrement dit, un n-concept est un rectangle n-dimensionnel maximal plein de croix dans C compte tenu de toutes les permutations à l'intérieur des dimensions.

Dans l'exemple de la figure 3.6, $(\{1,2,3\},\{a\},\{\alpha,\beta\})$ et $(\{2\},\{a,b\},\{\gamma\})$ sont des 3-concepts.

L'ensemble de tous les n-concepts dans un n-contexte et des n-quasi-ordres induits par une relation d'inclusion sur les sous-ensembles de chaque dimension, forme un n-treillis et tout n-treillis complet est isomorphe au treillis de concepts d'un n-contexte, comme mentionné dans l'analyse des concepts polyadiques [Voutsadakis, 2002].

3.1.2/ TRANSCRIPTION EN THÉORIE DES GRAPHES

Pour rappel, un graphe est une paire $G = (V, E)$ dans laquelle V est un ensemble d'éléments appelés arêtes et $E \subseteq V^2$ est un ensemble de nœuds.

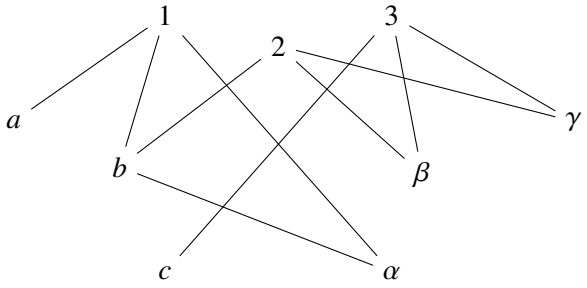

Figure 3.2 – Graphe servant d'exemple fil rouge.

Un ensemble $X \subseteq V$ de nœuds est une clique s'il existe une arête entre deux de ses éléments. Une clique est maximale si elle n'est pas contenue dans une autre clique. Un ensemble indépendant est un ensemble de sommets qui ne contient aucune arête. Un ensemble indépendant est maximal s'il n'est contenu dans aucun ensemble indépendant. Une couverture de nœuds est un ensemble de nœuds contenant au moins un sommet de chaque arête. Une couverture de nœuds est minimale si elle ne contient aucune couverture de nœuds. Un ensemble indépendant (maximal) dans un graphe G est une clique (maximale) dans le graphe complémentaire $\overline{G}$ et réciproquement. Le complément d'un ensemble indépendant (maximal) est une couverture de nœuds (minimale) et réciproquement.

Notons $\mathcal{M}(G)$ l'ensemble des cliques maximales dans un graphe G.

Un graphe $G = (V, E)$ est k-parti ssi V peut être partitionné en k ensembles indépendants.

Un graphe k-parti complet est un graphe k-parti tel qu'il existe une arête entre chaque paire de nœuds qui n'appartiennent pas au même ensemble indépendant.

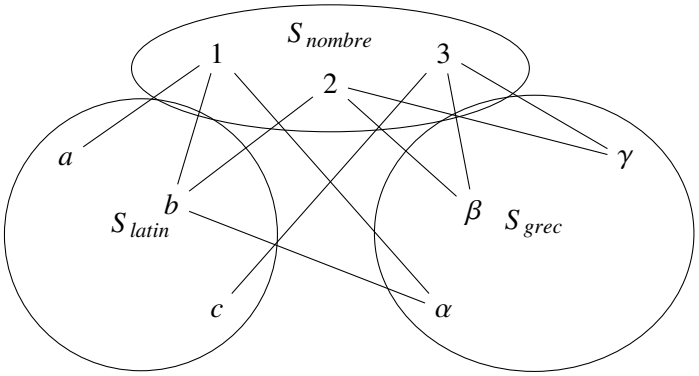

Figure 3.3 – Partition du graphe fil rouge en trois ensembles indépendants S_{nombre}, S_{latin} et S_{grec}.

Dans l'exemple fil rouge, les sous-graphes induits par les ensembles de nœuds $\{1, b, \alpha\}$ et $\{1, a, b\}$ sont, respectivement, triparti complet et biparti.

Les contextes formels bidimensionnels (S_1, S_2, R) sont des graphes bipartis $(S_1 \cup S_2, R)$ pour lesquels une bipartition est donnée. En terminologie des graphes, les 2-concepts sont ainsi des sous-graphes bipartis complets maximaux du contexte.

3.1.3/ CONSIDÉRATION DES GRAPHES K-PARTIS EN TANT QUE CONTEXTES

La FCA propose des outils pour rechercher et manipuler des motifs dans des graphes bipartis. Qu'advient-il de ces modèles et outils lorsque le graphe en entrée n'est pas biparti?

3.1.3.1/ DÉFINITION DES CONCEPTS

Commençons par définir les objets que nous recherchons. Les modèles centraux de la FCA sont des concepts : des sous-graphes bipartis complets maximaux du contexte. Lorsque le contexte est k-parti, une généralisation naturelle peut alors être exprimée comme suit.

Definition 22:

soit $G = (V, E)$ un graphe et $S = (S_1, \ldots, S_k)$ une partition de V en k ensembles indépendants. Soit $\{j_1, \ldots, j_m\} \subseteq \{1, \ldots, k\}$. Un m_2-concept de (S, E) est un tuple $C = (C_{j_1}, \ldots, C_{j_m})$, $C_{j_x} \neq \emptyset$, $C_{j_x} \subseteq S_{j_x}$, tel que $\bigcup_{x \in \{1,\ldots,m\}} C_{j_x}$ induit un sous-graphe maximal complet m-parti de G et qu'il n'existe pas $(C_{j_1}, \ldots, C_{j_m}, C_{j_{m+1}})$ ayant cette propriété.

Dans la désignation "m_2-concept", m signifie que l'on considère un graphe m-parti comme un "concept" (les m dimensions sont impliquées dans le motif) et $_2$ signifie que le motif se trouve dans un graphe 2-uniforme. Le choix de les définir comme des m-tuples au lieu de k-tuples avec $m \leq k$ évite de prendre en compte les éléments vides $m - k$ et la confusion avec les k-concepts de la PCA.

A partir de maintenant, supposons que notre exemple fil rouge est partitionné comme indiqué dans la figure 3.3. Dans ce cas, $(1, b, \alpha)$ est un 3_2-concept et $(1, ab)$ et $(23, \beta\gamma)$ sont des 2_2-concepts. Le tuple $(3, c, \beta\gamma)$ n'est pas un 3_2-concept parce que le sous-graphe induit est complet biparti et pas complet triparti[4]. Le tuple $(1, \alpha)$ n'est pas 2_2-concept parce que $(1, b, \alpha)$ est un 3_2-concept.

Lorsque le graphe est biparti et que la partition fournie est binaire, les 2_2-concepts sont des concepts formels d'intention et d'extension non vides. Il est important de noter que S_i, $i \in \{1, \ldots, k\}$, est un sous-graphe 1-parti complet - bien que (S_i) ne soit pas nécessairement un 1_2-concept.

Dénotons $\mathcal{T}((S, E))$ pour désigner l'ensemble des m_2-concepts, $1 < m \leq |S|$, d'un graphe k-parti (V, E) avec une partition S de V en k ensembles indépendants.

Soit (V, E) un graphe et $S = (S_1, \ldots, S_k)$ une partition de V en k ensembles independants.

$$\mathcal{T}((S, E)) = \mathcal{M}((V, E \cup X))$$

$$\text{avec } X = \bigcup_{i \in \{1,\ldots,k\}} \binom{S_i}{2}$$

Preuve : Dans $G = (V, E \bigcup_{i \in \{1,\ldots,k\}} \binom{S_i}{2})$, $\forall i \in \{1, \ldots, k\}$, S_i est une clique. Soit $C = (C_{j_1}, \ldots, C_{j_m})$ avec $C_{j_i} \subseteq S_{j_i}$ tel que $\bigcup_{i \in \{1,\ldots,m\}} C_{j_i}$ est une clique maximale dans G. Par définition, deux nœuds quelqu'ils soient $x \in C_{j_a}$ et $y \in C_{j_b}$, $a \neq b$ sont voisins dans G. En tant que tels, ils sont également voisins de (V, E). Clairement, cela fait de C un $m-$sous-graphe complet de (V, E). La propriété de maximalité est valable d'un graphe à l'autre, donc C est un m_2-concept de (V, E).

Soit $C = (C_{j_1}, \ldots, C_{j_m})$ un m_2-concept de (V, E). Par définition, deux nœuds quelqu'ils soient $x \in C_{j_a}$ et $y \in C_{j_b}$, $a \neq b$ sont voisins dans (V, E). Ainsi, ils sont voisins dans G. Comme, $\forall i \in \{1, \ldots, k\}$, S_i est une clique, $\bigcup_{i \in \{1,\ldots,m\}} C_{j_i}$ est une clique dans G. La propriété de maximalité tient encore d'un graphe à un autre donc $\bigcup_{i \in \{1,\ldots,m\}} C_{j_i}$ est une clique maximale dans G. □

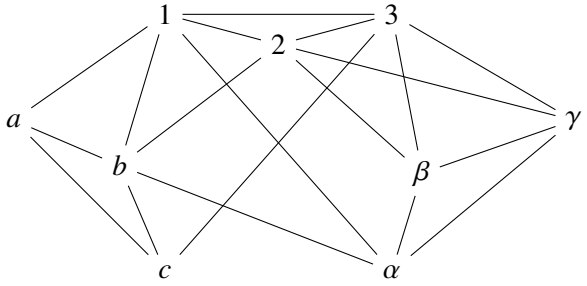

Figure 3.4 – Notre exemple fil rouge avec sa partition en cliques.

Cette proposition stipule que les m_2-concepts sont des cliques maximales dans un graphe qui peut être construit en temps polynomial à partir d'un contexte. Ceci implique que

4. Deux ensembles sont considérés : {3} et {c,$\beta\gamma$} sans relations entre {c} et {$\beta\gamma$}

$\mathcal{T}((S,E))$ peut être calculé à partir de (S,E) en temps polynomial en sortie [Tsukiyama et al., 1977].

3.1.3.2/ STRUCTURER LES CONCEPTS

Pour caractériser la structure de l'ensemble $\mathcal{T}((S,E))$, montrons qu'il forme un k-treillis lorsqu'il est mis en place avec les quasi-ordres appropriés. La meilleure façon de faire est de montrer que $\mathcal{T}((S,E))$ est isomorphe au concept k-treillis d'un k-contexte.

Soit $\mathcal{K}((S,E)) = (S_1 \cup \{s_1\}, \ldots, S_k \cup \{s_k\}, R)$ un k-contexte tel que $s_i \notin S_i$ et

$$(x_1, \ldots, x_k) \in R \Longleftrightarrow \forall x_i \neq s_i, x_j \neq s_j, \exists e \in E \text{ tel que } x_i, x_j \in e$$

Notons que, potentiellement, $x_i = x_j$. Dans le contexte $\mathcal{K}((S,E))$ chaque croix correspond à une clique du graphe (V,E), y compris les 1-elements, avec les éléments s_i representant le fait qu'une clique n'intersecte pas l'ensemble S_i. La figure 3.5 illustre le 3-contexte correspondant à notre exemple fil rouge.

Clairement, si $(X_1, \ldots, X_k)$ est un k-concept de $\mathcal{K}((S,E))$, alors $\forall i \in \{1, \ldots, m\}$, $s_i \in X_i$.

	a	b	c	s_2	a	b	c	s_2	a	b	c	s_2	a	b	c	s_2
1		×		×									×	×		×
2								×				×		×		×
3								×				×			×	×
s_1		×		×				×				×	×	×	×	×
		α				β				γ				s_3		

Figure 3.5 – Le 3-contexte $(\{1,2,3,s_1\}, \{a,b,c,s_2\}, \{\alpha,\beta,\gamma,s_3\}, R)$ correspondant à notre exemple fil rouge.

Soit (V,E) un graphe et S une k-partition de (V,E) en k ensembles indépendants. L'ensemble des m_2-concepts de (S,E), avec les k quasi-ordres induits par la relation d'inclusion sur chaque ensemble indépendant forme un k-treillis.

Preuve : Soit $(X_1, \ldots, X_k)$ un k-concept de $\mathcal{K}((S,E)) = (S_1 \cup \{s_1\}, \ldots, S_k \cup \{s_k\}, R)$. Par définition, $\prod_{i \in \{1,\ldots,k\}} (X_i \setminus \{s_i\}) \subseteq R$. A partir de la construction de $\mathcal{K}((S,E))$, nous obtenons $\forall x_i \in X_i \setminus \{s_i\}, x_j \in X_j \setminus \{s_j\}$, $\exists e \in E$ tel que $x_i, x_j \in e$. Cela signifie que le tuple $(X_{j_1} \setminus \{s_{j_1}\}, \ldots, X_{j_m} \setminus \{s_{j_m}\})$, tel que les differents $X_{j_i} \setminus \{s_{j_i}\}$ sont les composants non vides de $(X_1 \setminus \{s_1\}, \ldots, X_k \setminus \{s_k\})$, est un m_2-concept de (S,E).

Soit $(C_{j_1}, \ldots, C_{j_m})$ un m_2-concept de (S,E). Par définition, $\forall A \in \prod_{i \in \{1,\ldots,m\}} C_{j_i}$, $\forall x, y \in A$, $\exists e \in E$ tel que $x, y \in e$. Ainsi, le tuple $(X_1, \ldots, X_k)$ tel que

$$X_i = \begin{cases} C_i \cup \{s_i\} & \text{if } i \in \{j_1, \ldots, j_m\} \\ \{s_i\} & \text{sinon} \end{cases}$$

est un k-concept de $\mathcal{K}((S,E))$. □

Cela implique que les algorithmes [Cerf et al., 2008, Makhalova et al., 2017] pour calculer les n-concepts peuvent être utilisés pour calculer les m_2-concepts.

Dans la figure 3.5, les 3-concepts sont

$$\begin{array}{ll} (1s_1, bs_2, \alpha s_3) & (23s_1, s_2, \beta\gamma s_3) \\ (1s_1, abs_2, s_3) & (12s_1, bs_2, s_3) \\ (3s_1, cs_2, s_3) & (123s_1, s_2, s_3) \\ (s_1, abcs_2, s_3) & (s_1, s_2, \alpha\beta\gamma s_3) \end{array}$$

qui donne les m_2-concepts de notre exemple fil rouge une fois que les ensembles s_i et vides sont supprimés.

3.1.4/ CONCLUSION

Dans cette partie, les notions de contexte et concept formels ont été étendus aux graphes non bipartitionnés afin de permettre le traitement d'un type de données différent. Il a été montré qu'étant donné une k-partition du graphe en ensembles indépendants, l'ensemble de ses m_2-concepts forme un k-treillis. Cela permet l'utilisation de n'importe quel algorithme k-treillis pour calculer les m_2-concepts.

La prochaine étape consiste à généraliser la notion de n-concept aux hypergraphes qui ne sont pas n-partis n-uniform. Ceci, cependant, n'est pas aussi simple que les m_2-concepts. En effet, la structure du k-treillis des m_2-concepts provient du fait qu'une clique avec des n sommets peut être librement convertie en 2^n hyperarêtes (les sous-ensembles de sommets). Convertir une arête (a, b) en deux singletons (a) et (b) n'ajoute pas en complexité. Cependant, convertir une hyperarête (a, b, c) en un triangle (a, b), (b, c), (a, c) peut potentiellement créer de nouveaux triangles qui ne correspondent pas à hyperarêtes existantes de taille 3.

3.2/ REPRÉSENTATIONS CONDENSÉES DES RÈGLES D'ASSOCIATION DANS LES RELATIONS $n-aire$

L'extraction des règles d'association [Agrawal et al., 1993] est un problème qui a donné lieu à une littérature abondante, en particulier dans les données bidimensionnelles binaires classiques. En particulier, la représentation de l'ensemble de règles sans perte d'information est bien comprise. Ce n'est pas le cas dans les données binaires multidimensionnelles. Cette partie montre que la connaissance des n-ensembles fermés d'un tenseur booléen multidimensionnel est suffisante pour permettre de dériver la confiance de chaque règle d'association multidimensionnelle. Ceci généralise des résultats bien connus dans le cas bidimensionnel.

Depuis sa création, et afin de remédier au nombre excessif de schémas, l'extraction de règles d'association met l'accent sur la recherche du plus petit sous-ensemble de règles contenant des informations "intéressantes". Comme c'est généralement le cas lorsque deux critères doivent être optimisés - ici le nombre de règles et les informations qu'elles contiennent -, l'un d'eux a la priorité. Dans le domaine des règles d'association, le nombre de règles est souvent considéré comme primordial.

Le problème de la représentation de toutes les règles, dans le cas de données binaires bidimensionnelles, est résolu. Les règles peuvent être représentées de manière minimale par des règles entre les ensembles fermés. Ce résultat, basé sur le fait que les ensembles fermés sont des représentations uniques de leur classe d'équivalence w.r.t. de leur support, a produit de nombreuses combinaisons avec diverses mesures d'intérêt dans le but de réduire encore le nombre de règles, au détriment des informations qu'elles contiennent.

Ici, considérons le cas des données binaires multidimensionnelles. Ce cas est bien moins étudié que dans le cas bidimensionnel, cependant des moyens de réduire le nombre de règles en généralisant la mesure d'intérêt la plus connue, i.e. la fréquence, ont été proposés [Nguyen et al., 2011]. Cependant, il n'existe aucun résultat sur les représentations condensées de l'ensemble des règles. Pour y remédier, considérons les n-ensembles fermés d'un tenseur booléen pour obtenir le support de toute association et donc la confiance de toute règle.

Commençons par présenter les définitions et propriétés des tenseurs, ensembles fermés et règles en tenseurs bi- et n-dimensionnels; montrons que le support naturel de toute association peut être dérivé des supports des associations impliquant $n-1$ dimensions. Puis montrons qu'une simple transformation d'un tenseur permet de calculer le support de n'importe quelle association w.r.t. de n'importe quel nombre de dimensions.

3.2.1/ RAPPELS

Posons les notations et conventions suivantes : les lettres majuscules en police calligraphique, e.g. $\mathcal{A}$, dénotent des structures fixées ou ensembles. Les lettres majuscules normales, e.g. A, dénotent des ensembles variables, le plus souvent des sous-ensembles d'un autre ensemble, e.g. $A \subseteq \mathcal{A}$. Les lettres miniscules, e.g. a, dénotent des éléments d'ensembles. Pour l'écriture des ensembles les crochets sont souvent omis pour des questions de lisibilité.

Dans la manipulation des tuples, la notation . dénote la fusion de deux tuples. Par exemple, $(x_1, \ldots, x_i, \ldots, x_n) = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).(x_i)$.

3.2.2/ MATRICES, TENSEURS ET FERMETURES

Soit $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ un ensemble de dimensions, elles-mêmes ensembles d'éléments de même nature. Soit $\mathcal{R} \subseteq \prod_{\mathcal{D}_i \in \mathcal{D}} \mathcal{D}_i$ une relation $n-aire$. Le tenseur booléen $\mathcal{T} = (\mathcal{D}, \mathcal{R})$ est une table binaire, n-dimensionnelle qui représente les données. $\mathcal{T}$ est parfois appelé contexte n-dimensionnel ou n-contexte. Le tenseur de la figure 3.6 sert d'exemple fil rouge.

	p_1	p_2	p_3	p_1	p_2	p_3	p_1	p_2	p_3
c_1	×		×				×		
c_2		×		×	×	×	×		×
c_3			×		×			×	
		s_1			s_2			s_3	

Figure 3.6 – Un tenseur 3-dimensionnel représentant des consommateurs (c_1,c_2,c_3) achetant des produits (p_1,p_2,p_3) dans différents magasins (s_1,s_2,s_3).

Definition 23:

Soit $\mathcal{T}$ un tenseur booléen n-dimensionnel. Un n-ensemble fermé, dans $\mathcal{T}$, est un tuple $(X_1,\ldots,X_n)$ avec $X_i \subseteq \mathcal{D}_i$ tel que

$$\prod_{i\in\{1,\ldots,n\}} X_i \subseteq \mathcal{R}$$

et chaque composant est maximal pour cette propriété.

Dans le cas classique bidimentionnel, les 2-ensembles fermés sont alors des paires (A,B) pour lesquelles A et B sont maximaux tels que $\forall a \in A$, $\forall b \in B$, $(a,b) \in \mathcal{R}$. Dans ce cas, A et B sont tous deux appelés ensembles fermés. Si X est un sous-ensemble d'une des deux dimensions, alors la notation $c(X)$ dénote le plus petit ensemble fermé contenant X.

Notre exemple fil rouge contient les 3-ensembles fermés suivants :

$$\begin{array}{ccc}
(c_1,p_1p_3,s_1) & (c_1c_3,p_3,s_1) & (c_2,p_2,s_1s_2) \\
(c_1,p_1,s_1s_3) & (c_2,p_1p_2p_3,s_2) & (c_2c_3,p_2,s_2) \\
(c_2,p_1p_3,s_2s_3) & (c_1c_2,p_1,s_3) & (c_3,p_2,s_2s_3) \\
(\emptyset,p_1p_2p_3,s_1s_2s_3) & (c_1c_2c_3,\emptyset,s_1s_2s_3) & (c_1c_2c_3,p_1p_2p_3,\emptyset)
\end{array}$$

L'ensemble des 2-ensembles fermés ordonnés par une relation d'inclusion d'ensembles sur l'un ou l'autre de leurs composants forment un treillis complet [Ganter et al., 2012]. La généralisation de ce résultat montre que l'ensemble des n-ensembles fermés, avec des n quasi-ordres induits par la relation d'inclusion d'ensembles sur leurs n composants forment un n-treillis complet [Voutsadakis, 2002]. Les n-ensembles fermés sont aisément calculables par DATA-PEELER [Cerf et al., 2008].

3.2.3/ RÈGLES D'ASSOCIATION DANS LE CAS BIDIMENSIONNEL

Sans restrictions supplémentaires, il existe $2^{2|\mathcal{D}_2|}$ règles d'association possibles. Si l'on veut utiliser ces règles dans le monde réel, il est donc nécessaire de réduire leur nombre en ne considérant que certaines d'entre elles, idéalement les plus intéressantes. À cette fin, un certain nombre de mesures d'intérêt ont été proposées [Zhang et al., 2009]. La plus utilisée, est la fréquence. Une règle $A \rightarrow B$ est fréquente, w.r.t. un seuil $t \in [0,1]$, ssi $|S(A \cup B)| \geq t \times |\mathcal{D}_1|$. Les seuils de fréquence sont souvent combinés avec des seuils de confiance pour réduire davantage le nombre de règles.

Cependant, la fréquence, comme la plupart des mesures intéressantes, provoque la perte des informations contenues dans des règles non fréquentes. Chaque fois que la taille du

jeu de règles doit être réduite sans perte d'information, les propriétés suivantes traditionnelles [Luxenburger, 1991] peuvent être utilisées:

- $conf(A \rightarrow B) = conf(A \rightarrow A \cup B)$
- $conf(A \rightarrow B) = conf(c(A) \rightarrow c(B))$

Le premier énonce que les règles importantes sont celles dont le conséquent reprend l'antécédent, alors que le second indique qu'il suffit de ne prendre en compte que les règles entre ensembles fermés. À partir de l'ensemble de ces règles, la confiance de toute autre règle entre n'importe quel A et B peut être déduite.

Definition 24:

Base. Un ensemble de règles d'association permettant de déduire la confiance de toute autre règle est appelé une base.

Cependant, la construction de bases de règles d'association utilisant toutes les règles entre des ensembles fermés comparables n'est toujours pas assez efficace, car il peut y avoir jusqu'à $2^{|\mathcal{D}_2|}$ ensembles fermés. Pour réduire davantage la taille de la base, on peut se limiter aux règles de la forme $A \rightarrow B$ telles que $A = c(A)$, $B = c(B)$, $A \subset B$ et il n'existe pas d'ensemble fermé X tel que $A \subset X \subset B$. Ceci correspond à une règle par arête dans le diagramme de Hasse du treillis de 2-ensembles fermés. La confiance de toute règle peut ensuite être calculée en trouvant un chemin entre son antécédent et son conséquent dans le diagramme et en multipliant les confiances des règles trouvées le long du chemin.

[Luxenburger, 1991] a montré que des bases encore plus petites peuvent être obtenues en considérant uniquement un arbre couvrant du diagramme de Hasse. Cependant, l'utilisation d'une telle base pour obtenir la confiance d'autres règles implique la résolution de problèmes d'optimisation linéaire et prend trop de temps pour la plupart des applications pratiques.

Comme, par définition, les ensembles ont le même support que leur fermeture, les ensembles fréquents ont des fermetures fréquentes. Les deux approches permettant de réduire le nombre de règles peuvent ensuite être combinées en ne calculant que les règles entre les ensembles fermés fréquents voisins [Pasquier et al., 1999a,Pasquier et al., 1999b]. Cela réduit encore le nombre de règles au détriment de l'information.

3.2.4/ RÈGLES D'ASSOCIATION DANS LE CAS MULTIDIMENSIONNEL

Les règles pour les tenseurs booléens de plus de deux dimensions n'ont pas été étudiées de manière aussi approfondie. Différentes généralisations des règles d'association pour les relations n-aires ont été discutées. Dans [Nguyen et al., 2011], les auteurs ont proposé ce qui est, à notre connaissance, le plus général et le plus inclusif. Il est présenté ici et utilisé par la suite.

Soit $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ un ensemble de n dimensions et $\mathcal{R} \subseteq \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ une relation n-aire entre elles. L'objectif est d'extraire les "règles d'association" du tenseur $(\mathcal{D}, \mathcal{R})$. Cependant, contrairement au cas bidimensionnel, les patterns qui composent une règle peuvent concerner plusieurs dimensions.

Definition 25:

Association et Domaine. Soit $D \subseteq \mathcal{D}$ un ensemble de dimensions. soit $X_d \subseteq \mathcal{D}_d$, $\mathcal{D}_d \in D$, un ensemble non vide d'éléments de la dimension $\mathcal{D}_d$. L'ensemble $\prod_{\mathcal{D}_d \in D} X_d$ est appelé une association sur D et D est appelé son domaine

Dans notre exemple fil rouge, p_1 et $p_3 \times s_1$ sont des associations sur, respectivement, les domaines $\{\mathcal{D}_{produits}\}$ et $\{\mathcal{D}_{produits}, \mathcal{D}_{magasins}\}$.

Notons $dom(X)$ pour qualifier le domaine d'une association X. Dans le reste de ce travail et pour un ensemble D de dimensions, $\overline{D}$ désigne l'ensemble $\mathcal{D} \setminus D$.

Definition 26:

Projection. Soit $\mathcal{D}_i$ une dimension et $X = \prod_{\mathcal{D}_d \in Dom(X)} X_d$ une association. La projection $\pi_{\mathcal{D}_i}(X)$ de X sur $\mathcal{D}_i$ est X_i si $\mathcal{D}_i \in Dom(X)$ ou $\emptyset$ sinon.

Dans notre exemple fil rouge, $\pi_{Products}(p_3 \times s_1) = p_3$, $\pi_{Customers}(p_3 \times s_1) = \emptyset$ et $\pi_{Shops}(p_3 \times s_1) = s_1$.

Dans le cas bidimensionnel, le support d'une association sur une dimension trouve ses racines dans l'autre dimension. De même, dans le cas multidimensionnel, le support d'une association est basé sur toutes les dimensions qui ne sont pas dans son domaine.

Definition 27:

Support multidimensionnel. Le support d'une association X est l'ensemble

$$s(X) = \{t \in \prod_{\mathcal{D}_i \in \overline{dom(X)}} \mathcal{D}_i || forall x \in X, x.t \in \mathcal{R}\}$$

de tuples du produit cartésien des dimensions n'étant pas le domaine de X et qui forment une entrée de $\mathcal{R}$ avec un tuple contenu dans X.

Dans notre exemple fil rouge, nous avons $s(p_1) = \{(c_1, s_1), (c_2, s_2), (c_1, s_3), (c_2, s_3)\}$ et $s(p_3 \times s_1) = \{c_1, c_3\}$.

Definition 28:

L'union de deux associations X et Y est une association $X \sqcup Y$ telle que, pour toute dimension $\mathcal{D}_i$, $\pi_{\mathcal{D}_i}(X \sqcup Y) = \pi_{\mathcal{D}_i}(X) \cup \pi_{\mathcal{D}_i}(Y)$.

Par exemple, $p_1 \sqcup p_2 \times s_2 = p_1 p_2 \times s_1$.

Definition 29:

Une règle d'association multidimensionnelle est une règle $X \rightarrow Y$ entre deux associations X et Y. Le domaine de la règle est le domaine de $X \sqcup Y$.

La règle d'association $p_1 \rightarrow p_3 \times s_1$ est sur le domaine $\{Produits, Magasins\}$.

Soit $X \rightarrow Y$ une règle sur $dom(X \sqcup Y)$. Si nous avions été dans le cas bidimensionnel, la confiance de la règle pourrait être trouvée en comparant les supports de X et de $X \sqcup$

Y. Toutefois, dans le cas multidimensionnel, il est possible que $s(X)$ et $s(X \sqcup Y)$ soient définis sur des ensembles de dimensions différents. Comparer directement leurs tailles n'aurait aucun sens. Pour résoudre ce problème, nous utilisons une définition différente du support.

Definition 30:

Support multidimensionnel. Le support d'une association X par rapport à un domaine $D \supseteq dom(X)$ est l'ensemble

$$s_D(X) = \{t \in \prod_{\mathcal{D}_d \in \overline{D}} \mathcal{D}_d || existsu \in \prod_{\mathcal{D}_i \in D \setminus dom(X)} \mathcal{D}_i$$

$$\text{tel que } \forall x \in X, x.u.t \in \mathcal{R}\}$$

Clairement, le support $s(X)$ d'une association est égal au support relatif $s_{dom(X)}(X)$ de l'association par rapport à son domaine. Avec ce support relatif, la confiance des règles vient naturellement.

Definition 31:

Confiance multidimensionnelle. La confiance d'une règle d'association $X \rightarrow Y$ est

$$conf(X \rightarrow Y) = \frac{|s(X \sqcup Y)|}{|s_{dom(X \sqcup Y)}(X)|}$$

Dans notre exemple fil rouge, $s_{\{Produits, Magains\}}(p_1) = \{c_1, c_2\}$ et $s(p_1 p_3 \times s_1) = \{c_1\}$. En tant que telle, la confiance de $p_1 \rightarrow p_3 \times s_1$ est

$$\frac{|\{c_1\}|}{|\{c_1, c_2\}|} = \frac{1}{2}$$

ce qui signifie que la moitié des clients qui ont acheté le produit p_1 quelque part, a acheté les deux produits p_1 et p_3 au magasin s_1. Notons que la sémantique de la règle est contrainte par l'union de l'antécédent et du conséquent. Par exemple, il est impossible d'exprimer la relation entre l'achat d'un premier produit quelque part et l'achat d'un second produit dans un magasin spécifique avec une seule règle. Ces règles d'association multidimensionnelles respectent toujours la propriété

$$conf(X \rightarrow Y) = conf(X \rightarrow X \sqcup Y)$$

Le nombre de règles est encore bien plus élevé que dans le cas bidimensionnel. Dans [Nguyen et al., 2011], les auteurs utilisent la fréquence et la confiance pour le réduire. Il est certain que d'autres mesures intéressantes pourraient également être généralisées. Cependant, pourquoi garder toutes les informations? Il semblerait naturel d'essayer de mimer le cas bidimensionnel et de représenter également l'ensemble des règles d'association n-dimensionnelles avec des n-ensembles fermés. Cependant, aujourd'hui, il n'existe aucun résultat garantissant que cela soit réalisable. C'est ce que nous proposons donc ici.

3.2.5/ TRANSFORMATIONS DE TENSEURS

Cette partie contient les définitions des transformations de tenseurs utilisées dans les preuves.

Soit $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ un ensemble de dimensions, $\mathcal{R} \subseteq \mathcal{D}_1 \times \cdots \times \mathcal{D}_n$ une relation n-aire et $\mathcal{T} = (\mathcal{D}, \mathcal{R})$ un tenseur. Soit $D \subseteq \mathcal{D}$ un sous-ensembles de dimensions et $\mathcal{D}_d \in D$ une dimension. Deux operations peuvent être utilisées pour transformer le tenseur :

- "Fixer" les éléments d'une dimension
- Combiner les dimensions

La première opération consiste à limiter le tenseur à un sous-ensemble de l'une de ses dimensions. Notons $X_d \subseteq \mathcal{D}_d$ le sous-ensemble d'une dimension $\mathcal{D}_d$ et notons $X_D = \{X_{j_1}, \ldots, X_{j_{|D|}}\}$ une collection de sous-ensembles des dimensions dans D. Le tenseur $\mathcal{T}_{X_d} = (\mathcal{D} \setminus \mathcal{D}_d, \mathcal{R}_{X_d})$ avec

$$\mathcal{R}_{X_d} = \{(x_1, \ldots, x_{d-1}, x_{d+1}, \ldots, x_n) || forall x_d \in X_d, (x_1, \ldots, x_d, \ldots, x_n) \in \mathcal{R}\}$$

est construit par l'intersection des "tranches" de $\mathcal{T}$ correspondant aux éléments de X_d, résultant d'un tenseur $(n-1)$-dimensionnel. Quand des dimensions multiples sont restreintes simultanément, on écrit $\mathcal{T}_{X_D} = (((\mathcal{T}_{X_{j_1}})_{X_{j_2}}) \ldots)_{X_{j_{|D|}}}$.

	p_1	p_2	p_3
c_1	×		×
c_2		×	
c_3			×

p_1	p_2	p_3
×		×

	s_1	s_2	s_3
c_1	×		
c_2		×	×
c_3			

Figure 3.7 – les transformations $\mathcal{T}_{s_1}$, $\mathcal{T}_{s_1,c_1}$ et $\mathcal{T}_{p_1,p_3}$ de notre exemple fil rouge $\mathcal{T}$

La deuxième opération consiste à remplacer des ensembles de dimensions par leurs produits cartésiens. Soit $\Omega = (\omega_1, \ldots, \omega_m)$ une partition de $\mathcal{D}$ en m ensembles et

$$\mathcal{D}^\Omega = \{\prod_{\mathcal{D}_i \in \omega_1} \mathcal{D}_i, \ldots, \prod_{\mathcal{D}_j \in \omega_m} \mathcal{D}_j\}$$

le nouvel ensemble de dimensions. Le nouveau tenseur est alors $\mathcal{T}^\Omega = (\mathcal{D}^\Omega, \mathcal{R}^\Omega)$ avec

$$\mathcal{R}^\Omega = \{(x_1, \ldots, x_m) ||_1 . x_2 . \ldots . x_m \in \mathcal{R}\}$$

3.2.6/ DÉRIVER LES CONFIANCES DES RÈGLES

L'objectif est d'identifier un petit ensemble de règles d'association multidimensionnelles permettant de dériver les confiances de toutes les autres règles potentiellement intéressantes d'un tenseur booléen. Pour ce faire, commençons par montrer que la taille du

	(p_1,s_1)	(p_2,s_1)	(p_3,s_1)	(p_1,s_2)	(p_2,s_2)	(p_3,s_2)	(p_1,s_3)	(p_2,s_3)	(p_3,s_3)
c_1	×		×				×		
c_2		×		×	×	×	×		×
c_3			×		×			×	

Figure 3.8 – La transformation $\mathcal{T}^{(\{Customers\},\{Products,Shops\})}$ de notre exemple fil rouge $\mathcal{T}$.

support de toute association dans un tenseur peut être déduite de la taille des composants de n-ensembles fermés. La seule hypothèse est qu'il y a une dimension qui n'est jamais dans le domaine d'une telle association. Nous pensons que cette hypothèse est raisonnable dans la mesure où, en pratique, une dimension contient généralement les "objets" ou les "transactions" et que ses éléments n'apparaissent pas dans les règles. Sans perte de généralité, nous supposerons que cette dimension est $\mathcal{D}_1$.

Comme l'illustre la figure 3.9, le tenseur n-dimensionnel $\mathcal{T}$ peut être vu comme l'empilement de tenseurs $(n-1)$ dimensionnels. Par conséquent, la taille du support d'une association X est la somme des tailles de ses supports dans les tranches qui composent le tenseur.

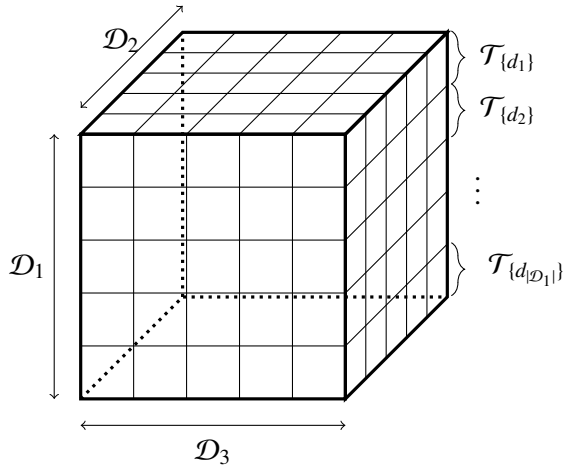

Figure 3.9 – Tenseurs n-dimensionnels empilés ou empilement de tenseurs $(n-1)$-dimensionnels

Proposition [1] : Soit X une association. soit

$$D = \prod_{\mathcal{D}_i \in \mathcal{D} \setminus (dom(X) \cup \mathcal{D}_1)} \mathcal{D}_i$$

le produit cartésien de toutes les dimensions de X exceptées $\mathcal{D}_1$. Nous avons

$$|s(X)| = \sum_{d \in D} |s_d|$$

où s_d est le support de X dans $\mathcal{T}_d$.

Preuve : Par définition, chaque $r \in \mathcal{R}_d$ est telle que $r.d \in \mathcal{R}$. Ainsi, le support de X dans $\mathcal{T}_d$ est le sous-ensemble O de $\mathcal{D}_1$ tel que, $\forall o \in O$, $o.d$ est dans le support de X dans $\mathcal{T}$. Comme les tuples $d \in D$ sont distincts deux à deux, la taille du support de X dans $\mathcal{T}$ est donc la somme des tailles des supports de X dans les divers $\mathcal{T}_d$. □

La proposition 1 indique que la taille des supports des associations X et Y dans un tenseur $\mathcal{T}$ peut être déduite des tailles des supports de X et Y dans les divers tenseurs obtenus en fixant les éléments du produit cartésien des supports de toutes les dimensions, à l'exception de $\mathcal{D}_1$. Dans ces tenseurs, à la fois $s(X)$ et $s(Y)$ sont des sous-ensembles de $\mathcal{D}_1$.

Supposons que nous recherchions les tailles des supports de p_1 et p_1p_3 dans notre exemple fil rouge $\mathcal{T}$. Le produit cartésien de la seule dimension de support autre que $\mathcal{D}_1$ est $\{s_1, s_2, s_3\}$. Les supports de p_1 dans $\mathcal{T}_{\{s_1\}}$, $\mathcal{T}_{\{s_2\}}$ et $\mathcal{T}_{\{s_3\}}$ ont, respectivement, les tailles 1, 1 et 2 ainsi $|s(p_1)| = 4$ dans $\mathcal{T}$. Les tailles des supports de p_1p_3 dans les mêmes tenseurs sont 1, 1 et 1 ainsi $|s(p_1p_3)| = 3$ dans $\mathcal{T}$.

Proposition [2] : Soit X une association dans $\mathcal{T}$ et Z un élément du produit cartésien d'un sous-ensemble des dimensions support de X. Le support de X dans $\mathcal{T}_Z$ est le support de $X \sqcup Z$ dans $\mathcal{T}$.

Preuve : $\Rightarrow$. Soit S le support de X dans $\mathcal{T}_Z$. Cela signifie, par définition, que $\forall s \in S$, $\forall x \in X$, $s.x \in \mathcal{R}_Z$. Comme $\mathcal{T}_Z$ est construit à partir de $\mathcal{T}$ par l'intersection des tranches correspondant uniquement à Z, alors $s.x \in \mathcal{R}_Z$ implique $s.Z.x \in \mathcal{R}$. Par conséquent, S est un sous-ensemble du support de $X \sqcup Z$ dans $\mathcal{T}$.

$\Leftarrow$. Soit S' le support de $X \sqcup Z$ dans $\mathcal{T}$. Celà signifie que, par définition, $\forall s \in S$, $\forall x \in X$, $s.Z.x \in \mathcal{R}$. De la construction de $\mathcal{T}_Z$, nous savons que $s.Z.x \in \mathcal{R}$ implique $s.x \in \mathcal{R}_Z$. Par conséquent, S' est un sous-ensemble du support de X dans $\mathcal{T}_Z$ tel que, $S = S'$. □

Dans notre exemple fil rouge, le support de $p_1.s_3$ dans $\mathcal{T}$ est $\{c_1, c_2\}$. Le support de s_3 dans $\mathcal{T}_{p_1}$ est aussi $\{c_1, c_2\}$.

La proposition 2 implique que les supports des associations dans les tenseurs $\mathcal{T}_d$ mentionnés dans la proposition 1 peuvent être dérivés des supports des associations sur $\overline{\mathcal{D}_1}$ dans $\mathcal{T}$. Des propositions 1 et 2, nous déduisons que la taille du support de toute association dans $\mathcal{T}$ peut être dérivée uniquement des supports des associations sur $\mathcal{D} \setminus \{\mathcal{D}_1\}$.

3.2.7/ RÈGLES ENTRE ASSOCIATIONS DE DIFFÉRENTS DOMAINES

Dans la partie précédente, il a été montré que la taille du support de toute association peut être déduite de la taille des supports des associations sur $\overline{\mathcal{D}_1}$. Ceci est utile lorsque l'on essaye de calculer la confiance d'une règle entre deux associations sur le même domaine. Cependant, les règles de la forme $X \rightarrow Y$ avec $dom(X) \subset dom(X \sqcup Y)$ impliquent le support $s_{dom(X \sqcup Y)}(X)$ de X en ce qui concerne $dom(X \sqcup Y)$. Dans cette partie, montrons que le tenseur peut être transformé pour unifier les domaines des antécédents et les conséquents de telle sorte que le support de toute association à l'égard de tout domaine puisse être dérivé d'associations sur $\overline{\mathcal{D}_1}$.

Soit $\mathcal{T} = (\mathcal{D}_1, \ldots, \mathcal{D}_n, \mathcal{R})$ un tenseur. Soit $\mathcal{T}^{\uparrow} = (\mathcal{D}_1, \mathcal{D}_2 \cup \{\vee_2\}, \ldots, \mathcal{D}_n \cup \{\vee_n\}, \mathcal{R}^{\uparrow})$ avec

$$\mathcal{R}^{\uparrow} = \mathcal{R} \cup \{(x_1, \ldots, x_n) \| forall x_i = \vee_i, \exists x'_i \neq x_i \text{ such that } (x_1, \ldots, x'_i, \ldots, x_n) \in \mathcal{R}^{\uparrow}\}$$

En d'autres mots, le tenseur $\mathcal{T}^{\uparrow}$ est construit à partir de $\mathcal{T}$ en ajoutant un élément à chaque dimension (excepté $\mathcal{D}_1$) et projetant les croix sur ces éléments jusqu'à saturation,

comme illustré par la figure 3.10. Chaque nouvel élément représente une disjonction sur sa dimension, c'est-à-dire une sorte de "joker". Dans l'exemple de la figure 3.10, la croix dans $(c_3, \vee_p, s_2)$ signifie que le client c_3 a acheté **un produit** dans la boutique s_2.

	p_1	p_2	p_3	$\vee_p$	p_1	p_2	p_3	$\vee_p$	p_1	p_2	p_3	$\vee_p$	p_1	p_2	p_3	$\vee_p$
c_1	×		×	×					×			×	×		×	×
c_2		×		×	×	×	×	×	×		×	×	×	×	×	×
c_3			×	×		×		×		×		×		×	×	×
	s_1				s_2				s_3				$\vee_s$			

Figure 3.10 – Le tenseur $\mathcal{T}^{\uparrow}$ correspondant à l'exemple $\mathcal{T}$ de la figure 3.6. La dimension $Clients$ joue le rôle de $\mathcal{D}_1$.

Definition 32:

Soit X une association dans $\mathcal{T}$ et $D \supseteq dom(X)$ un domaine. $X^{D\uparrow}$ est une association dans $\mathcal{T}^{\uparrow}$ telle que

$$\pi_{\mathcal{D}_i}(X^{D\uparrow}) = \begin{cases} \pi_{\mathcal{D}_i}(X) \cup \{\vee_i\} & \text{if } \mathcal{D}_i \in D \\ \emptyset & \text{otherwise} \end{cases}$$

Dans notre exemple fil rouge, si $D = \{Produits, Magasins\}$, alors $p_3^{D\uparrow} = p_3 \vee_p . \vee_s$ et $p_2.s_1^{D\uparrow} = p_2 \vee_p . s_1 \vee_s$.

Proposition [3] : Soit X une association dans $\mathcal{T}$ et $D \supseteq dom(X)$ un domaine. Alors, $s_D(X)$ dans $\mathcal{T}$ est égal à $s(X^{D\uparrow})$ dans $\mathcal{T}^{\uparrow}$.

Preuve : Soit t dans le support de $X^{D\uparrow}$ de $\mathcal{T}^{\uparrow}$. A partir de la construction de $\mathcal{T}^{\uparrow}$, nous savons que, $\forall \mathcal{D}_i \in \overline{\mathcal{D}_1}$, $(x_1, \ldots, \vee_i, \ldots, x_n) \in \mathcal{R}^{\uparrow}$ implique que $(x_1, \ldots, x_i, \ldots, x_n) \in \mathcal{R}^{\uparrow}$ avec $x_i \neq \vee_i$. En suivant ce raisonnement de manière récursive sur les dimensions dans D, on obtient que, pour tout $x \in X$, il existe un tuple $a \in \prod_{\mathcal{D}_i \in D \setminus dom(X)} \mathcal{D}_i$ tel que $t.a.x \in \mathcal{R}$. En tant que tel, par définition, $t \in s_D(X)$ in $\mathcal{T}$. Par conséquent, $s(X^{D\uparrow})$ dans $\mathcal{T}^{\uparrow}$ est un sous ensemble de $s_D(X)$ in $\mathcal{T}$.

Soit t' dans $s_D(X)$ dans $\mathcal{T}$. Si t' n'est pas dans le support de $X^{D\uparrow}$, alors cela signifie qu'il n'y a pas de tuple $a \in \prod_{\mathcal{D}_i \in D \setminus dom(X)} \mathcal{D}_i$ tel que $t'.a.x \in \mathcal{R}$. Cela contredit notre affirmation. Par conséquent, $s_D(X)$ in $\mathcal{T}$ est un sous-ensemble de $s(X^{D\uparrow})$ in $\mathcal{T}^{\uparrow}$ et ils sont alors égaux. □

La proposition 3 déclare que le soutien de X en ce qui concerne un domaine D dans $\mathcal{T}$ est le support de $X^{D\uparrow}$ dans $\mathcal{T}^{\uparrow}$. A partir de cela et des propositions 1 et 2, on en déduit que la taille du support de toute association X en ce qui concerne n'importe quel domaine dans $\mathcal{T}$ peut être dérivée de la taille des supports des associations sur $\overline{\mathcal{D}_1}$ dans $\mathcal{T}^{\uparrow}$.

Dans l'exemple $\mathcal{T}$ de la figure 3.6, le support de s_3 en ce qui concerne le domaine $D = \{Produits, Magasins\}$ est $s_{Clients}(s_3) = \{c_1, c_2, c_3\}$. L'association $s_3^{D\uparrow} = \vee_p.s_3 \vee_s$ a aussi $\{c_1, c_2, c_3\}$ pour support $\mathcal{T}^{\uparrow}$ dans la figure. 3.10. Similairement, le support des deux $p_1.s_3$ dans $\mathcal{T}$ et $p_1 \vee_p . s_3 \vee_s$ dans $\mathcal{T}^{\uparrow}$ est $\{c_1, c_2\}$.

3.2.8/ BASES DES RÈGLES D'ASSOCIATION

Dans les parties 3.2.6 et 3.2.7, il a été montré que le support de toute association en ce qui concerne n'importe quel domaine peut être dérivé uniquement à partir des supports des associations sur $\overline{\mathcal{D}_1}$ dans $\mathcal{T}^{\uparrow}$. Montrons maintenant que l'ensemble des n-ensembles fermés est suffisant pour récupérer ces supports, puis identifions deux ensembles de règles d'association qui représentent toutes les autres sans perte d'information.

Soit X une association sur $\overline{\mathcal{D}_1}$ dans $\mathcal{T}^{\uparrow}$. A partir de la définition d'une association et de son support, nous savons que $s(X) \times X \subseteq \mathcal{R}$. En d'autres mots, c'est une boîte de croix n-dimensionnelles dans le tenseur. Elle n'est pas nécessairement maximale sur toutes les dimensions mais le support lui-même l'est. Ainsi, il doit y avoir au moins un n-ensemble fermé $(s(X), C_2, \ldots, C_n)$ avec $\pi_{\mathcal{D}_i}(X) \subseteq C_i$, $\forall i \in \{2, \ldots, n\}$. Ce qui implique que $s(X) = s(\prod_{i\in\{2,\ldots,n\}} C_i)$. Dans le cas bidimensionnel, il n'existe qu'un seul 2-ensemble fermé pour X. Quand $n \geq 3$, il peut y en avoir beaucoup. Abusons légèrement des notations en utilisant $c(X)$ pour désigner l'association résultant du produit cartésien des $n-1$ derniers composants d'un tel n-ensemble fermé pour l'association X.

Dans notre exemple, on peut dire que $c(\vee_p.s_1) = \vee_p.s_1s_3\vee_s$ car $s(\vee_p.s_1) = \{c_1c_2c_3\}$ et le triplet $(c_1c_2c_3, \vee_p, s_1s_3\vee_s)$ est un 3-ensemble fermé dans le tenseur $\mathcal{T}^{\uparrow}$ décrit dans la figure 3.10.

Comme X et $c(X)$ ont le même support, les règles d'association $X \rightarrow Y$ et $c(X) \rightarrow c(Y)$ ont la même confiance. Cela signifie que, comme dans le cas bidimensionnel, la connaissance des règles entre les n-ensembles fermés est suffisante pour déduire la confiance de toutes les autres règles sur $\mathcal{D}\backslash\{\mathcal{D}_1\}$. A partir de cette propriété, deux bases différentes de règles d'association peuvent être identifiées.

Soit $\emptyset$ une association vide et X une association sur $\overline{\mathcal{D}_1}$. Le support de $\emptyset$ concernant $\overline{\mathcal{D}_1}$ est $\mathcal{D}_1$. Ainsi, $s(c(\emptyset)) = \mathcal{D}_1$. La confiance de $\emptyset \rightarrow X$ est donc $conf(\emptyset \rightarrow X) = \frac{|s(X)|}{|\mathcal{D}_1|}$. Par conséquent, $|s(X)| = conf(\emptyset \rightarrow X) \times |\mathcal{D}_1|$. L'expression de la proposition 1 peut être réécrite ainsi :

$$|s(X)| = \sum_{Z} conf(\emptyset \rightarrow X \sqcup Z) \times |\mathcal{D}_1|$$

avec Z un élément du produit cartésien de $\overline{dom(X)} \setminus \mathcal{D}_1$.

Proposition [4] *:* Soit $\mathcal{T} = (\mathcal{D}_1, \ldots, \mathcal{D}_n, \mathcal{R})$ un tenseur booléen multidimensionnel. L'ensemble des règles de la forme $c(\emptyset) \rightarrow c(X)$ où X est une association sur le domaine $\overline{\mathcal{D}_1}$ est une base de règle d'association pour $\mathcal{T}$.

Ainsi, l'ensemble des règles d'association de la forme $c(\emptyset) \rightarrow c(X)$ permet de dériver la confiance de toutes les autres règles du tenseur. Cependant, même si ces règles représentent l'ensemble des règles d'association, elles sont difficiles à interpréter pour un analyste humain. Une base composée de règles plus expressives serait utile.

Soit $c(X)$, $c(Y)$ et $c(Z)$ trois associations correspondant aux $n-1$ derniers composants des n-ensembles fermés tels que $c(X) \subseteq c(Y) \subseteq c(Z)$. Puisque $s(Z) \subseteq s(Y) \subseteq s(X)$, nous avons $conf(X \rightarrow Z) = conf(X \rightarrow Y) \times conf(Y \rightarrow Z)$. En tant que tel, l'ensemble des règles d'association de la forme $c(X) \rightarrow c(Z)$ tel qu'il n'existe pas d'association $c(Y)$ avec $c(X) \subseteq c(Y) \subseteq c(Z)$ est suffisant pour dériver la confiance de toutes les autres règles sur

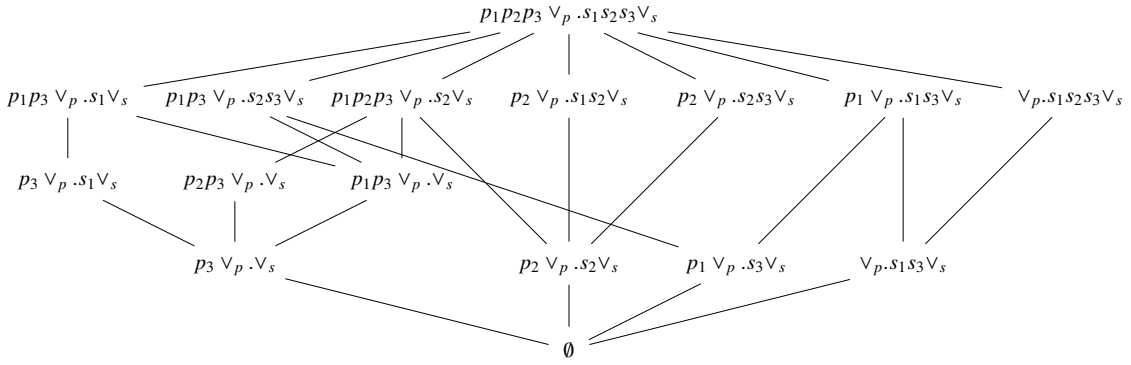

Figure 3.11 – Diagramme de Hasse des associations construit à partir des second et troisième composants des 3-ensembles fermés dans l'exemple du tenseur $\mathcal{T}^{\uparrow}$ de la figure 3.10 ordonné par la relation d'inclusion.

$\overline{\mathcal{D}_1}$. Plus important encore, il peut être utilisé pour obtenir la confiance des règles de la forme $c(\emptyset) \rightarrow c(X)$ et ainsi de la confiance de toute règle.

Proposition [5] *:* Soit $\mathcal{T} = (\mathcal{D}_1, \ldots, \mathcal{D}_n, \mathcal{R})$ un tenseur booléen multidimentionnel. L'ensemble des règles de la forme $c(X) \rightarrow c(Z)$ tel qu'il n'existe pas d'association $c(Y)$ avec $c(X) \subseteq c(Y) \subseteq c(Z)$ est une base de règles pour $\mathcal{T}$.

Lorsque la première base de règles d'association est fondamentalement l'ensemble des n-ensembles fermés, la seconde correspond aux arêtes dans le diagramme de Hasse de l'ensemble des associations de $\overline{\mathcal{D}_1}$ ordonnées par la relation d'inclusion, comme représenté sur la figure 3.11. La deuxième base contient nécessairement plus de règles, mais elles sont beaucoup plus compréhensibles du point de vue humain.

Cependant, malgré le fait que la deuxième base généralise la base traditionnelle du cas bidimensionnel, une différence cruciale apparaît quant à son interaction avec la fréquence. En effet, bien que les ensembles fermés fréquent soient suffisants pour obtenir le support de tous les ensembles fréquents lorsque $n = 2$, ce n'est pas le cas ici. Par exemple, supposons que le seuil de fréquence dans notre exemple de la figure 3.10 est 0.5. L'association p_1 a pour support $\{(c_1, s_1), (c_2, s_2), (c_1, s_3), (c_2, s_3), (c_1, \vee_s), (c_2, \vee_s)\}$ et donc une fréquence d'exactement $0,5$. Une des associations utilisées par la proposition 1 pour calculer la taille du support de p_1 est $p_1.s_1$, ce qui a une fréquence de $0,33$. Par conséquent, déterminer la taille des supports d'associations fréquentes nécessite la connaissance d'associations non fréquentes.

Puisque les deux notions ne peuvent pas être combinées, une question se pose naturellement : comment comparer les nombres des n-ensembles fermés et les associations fréquentes? Une première réponse expérimentale est présentée dans un article en soumission à TKDE et les algorithmes sont présentés ci-après.

3.2.9/ ALGORITHMES

Dans cette partie, construisons des algorithmes explicites pour calculer les deux bases de règles d'association définies dans la partie 3.2.6, ainsi que pour obtenir la confiance de toute règle.

$$
\begin{array}{rcl}
\emptyset & \xrightarrow{2/3} & p_2 \vee_p . s_2 \vee_s \\
\emptyset & \xrightarrow{1} & p_3 \vee_p . \vee_s \\
p_2 \vee_p . s_2 \vee_s & \xrightarrow{1/2} & p_2 \vee_p . s_2 s_3 \vee_s \\
p_3 \vee_p . \vee_s & \xrightarrow{2/3} & p_2 p_3 \vee_p . \vee_s \\
p_2 \vee_p . s_2 s_3 \vee_s & \xrightarrow{0} & p_1 p_2 p_3 \vee_p . s_1 s_2 s_3 \vee_s \\
p_2 p_3 \vee_p . \vee_s & \xrightarrow{1/2} & p_1 p_2 p_3 \vee_p . s_2 \vee_s
\end{array}
$$

Figure 3.12 – Quelques règles extraites du diagramme de la figure 3.11.

L'algorithme 3 calcule l'ensemble des règles entre les n-ensembles fermés qui sont voisins au regard de la relation d'inclusion sur leurs derniers $n-1$ composants. Il transforme le tenseur d'entrée $\mathcal{T}$ en $\mathcal{T}^{\uparrow}$, puis calcule les n-ensembles fermés. Les règles sont ensuite construites soit en utilisant les n-ensembles fermés tels quels, soit en calculant leur relation voisine au regard de l'inclusion sur leurs derniers $n-1$ composants.

L'algorithme 4 calcule la confiance d'une règle entre deux associations A et B en utilisant un ensemble de règles d'association de la forme $\emptyset \rightarrow c(X)$. Il commence par transformer $A \rightarrow B$ en $A^{dom(A \sqcup B)\uparrow} \rightarrow (A \sqcup B)^{dom(A \sqcup B)\uparrow}$. Ensuite, en utilisant les propriétés introduites dans les propositions 1 et 2, il calcule la confiance. Comme les confiances de règles de la forme $\emptyset \rightarrow c(X)$ sont facilement disponibles dans le jeu de règles, elles sont simplement extraites de là.

L'algorithme 5 calcule la confiance d'une règle entre deux associations A et B en utilisant un ensemble de règles d'association de la forme $c(X) \rightarrow c(Y)$ telles que $c(X)$ et $c(Y)$ sont voisins. Il débute par transformer $A \rightarrow B$ en $A^{dom(A \sqcup B)\uparrow} \rightarrow (A \sqcup B)^{dom(A \sqcup B)\uparrow}$. Ensuite, en utilisant les propriétés introduites dans les propositions 1 et 2, il calcule la confiance. Contrairement à l'algorithme 4, les confiances de règles de la forme $\emptyset \rightarrow c(X)$ doivent être calculées à partir du jeu de règles. Cela se fait en trouvant un chemin entre $\emptyset$ et $c(X)$ dans le graphe orienté induit par les règles et en multipliant les confiances le long du chemin.

3.2.10/ DISCUSSION

Nous avons montré que des représentations condensées sans perte de règles d'association dans des tenseurs booléens multidimensionnels peuvent être construites à partir des n-ensembles fermés par simples transformations de ces tenseurs. Ceci généralise les résultats connus dans le cas bidimensionnel.

Nous avons également discuté de l'impossibilité de combiner cette représentation avec la mesure d'intérêt la plus populaire, la fréquence. Les résultats expérimentaux montrent que le seuil pour lequel il existe des associations plus fréquentes que les n-ensembles augmente à la fois avec la densité de données et le nombre de dimensions.

Cependant, cela ne signifie pas que nous ne pourrions pas construire des ensembles de règles entre des n-ensembles fermés résumant les informations repérées par d'autres mesures d'intérêt. Il serait alors intéressant d'étudier quelles mesures ou catégories de mesures sont compatibles avec ces résultats.

Par ailleurs, le verrou du volume de données en sortie (nombre de règles) reste présent. En effet, même si cette base est minimaliste, elle reste très volumineuse. L'intégration des concepts et des règles aussi volumineuses dans une ontologie pose problème car les ontologies supportent très mal la charge. Pour répondre à ce verrou j'ai monté un projet de recherche FUI permettant le recrutement d'une post doctorante (Amira Mouakher) avec qui nous travaillons actuellement pour l'intégration de l'algorithme Quality Cover à MARM (donc en multidimensionnel) permettant d'extraire un minimum de concepts tout en maximisant la couverture de la connaissance contenue dans le treillis. Pour cela nous étudions aussi les mesures sémantiques permettant de calculer la couverture dans le cas multidimensionnel.

Algorithm 3 COMPUTERULES($\mathcal{T}$)

Input : Un tenseur $\mathcal{T}$

Output : Une base pour les règles d'association dont le domaine ne contient pas $\mathcal{D}_1$

$R \leftarrow \emptyset$ $\mathcal{T} \leftarrow \mathcal{T}^{\uparrow}$ $C \leftarrow ClosedNSets(\mathcal{T})$ $R \leftarrow BuildRules(C)$

return R

Algorithm 4 INFER($A, B, \mathcal{T}^{\uparrow}$)

Input : Deux associations A et B dans le tenseur $\mathcal{T}^{\uparrow}$

Output : La confiance de $A \rightarrow B$

$B \leftarrow (A \sqcup B)^{dom(A \sqcup B)\uparrow}$ $A \leftarrow A^{dom(B)\uparrow}$

$R_A \leftarrow 0$ $R_B \leftarrow 0$

foreach $Z \in \prod_{\mathcal{D}_i \in \overline{dom(B)} \setminus \{\mathcal{D}_1\}} \mathcal{D}_i$ **do**

  $C \leftarrow c(A \sqcup Z)$ $D \leftarrow c(B \sqcup Z)$ $R_A \leftarrow R_A + conf(\emptyset \rightarrow C)$ $R_B \leftarrow R_B + conf(\emptyset \rightarrow D)$

return $\frac{R_B}{R_A}$

Algorithm 5 INFER2(A, B)

Input : Deux associations A et B

Output : La confiance de $A \rightarrow B$

$B \leftarrow (A \sqcup B)^{dom(A \sqcup B)\uparrow}$ $A \leftarrow A^{dom(B)\uparrow}$

$R_A \leftarrow 0$ $R_B \leftarrow 0$

foreach $Z \in \prod_{\mathcal{D}_i \in \overline{dom(B)} \setminus \{\mathcal{D}_1\}} \mathcal{D}_i$ **do**

  $C \leftarrow c(A \sqcup Z)$ $D \leftarrow c(B \sqcup Z)$ $P \leftarrow Path(\emptyset, C)$ $R \leftarrow 1$ **foreach** $(X, Y) \in P$ **do**

    $R \leftarrow R \times conf(X \rightarrow Y)$

  $R_A \leftarrow R_A + R$ $P \leftarrow Path(\emptyset, D)$ **foreach** $(X, Y) \in P$ **do**

    $R \leftarrow R \times conf(X \rightarrow Y)$

  $R_B \leftarrow R_B + R$

return $\frac{R_B}{R_A}$

4

CONCLUSION

Depuis mes débuts en fouille de données au sein de l'équipe AFD du laboratoire lE2I de Strasbourg en master recherche première année, j'ai rapidement trouvé ma spécialité dans ce domaine : l'Analyse Formelle de Concepts. La confrontation de ce domaine avec des données réelles et la rencontre avec l'Intuition Artificielle ont ciblé un ensemble de verrous à lever. Dans cette partie j'ai présenté ce qu'est la fouille de données et plus précisément en quoi consiste l'Analyse Formelle de Concepts et les règles d'associations qui lui sont liées. J'ai présenté ma thèse et les verrous essentiels levés par ces travaux : l'adaptation de la FCA à la prise en charge de données multi-valuées floues. Puis j'ai présenté les travaux menés lors du projet PSDP (cf section 14.1.2) avec Alexandre Bazin pour lever le verrou de la multidimensionnalité des données (l'outil MARM) capital pour l'Intuition Artificielle. Toujours dans le cadre des verrous bloquant l'Intuition Artificielle, se pose le problème du volume de concepts générés. Ce travail en cours d'élaboration par le portage en multidimensionnel de l'algorithme Quality Cover, son intégration à MARM et la redéfinition de distances en multidimensionnalité pour permettre à Quality Cover de calculer au mieux la couverture concernée. Pour cela j'ai monté le projet FUI WineCloud (cf section 14.1) pour le recrutement d'Amira Mouakher (conceptrice de Quality Cover) comme post doctorante. Enfin il reste un verrou qui n'est pas encore levé : le flou au niveau multidimensionnel. En effet, les données binaires ne représentent qu'une part de la réalité des données. Pour lever ce verrou, un projet qui démarrera en 2020 va permettre le recrutement de deux post doctorants afin de lever ce verrou et l'étendre par le développement d'un moteur d'inférence floue multidimensionnelle pour permettre d'une part à la FCA d'intégrer le flou et d'autre part, aux ontologies d'inférer sur ces problématiques, ce dont elles sont pour l'instant incapables.

II

NK-Correlated Sample Graph au coeur de l'Intuition Artificielle

NK-CORRELATED SAMPLE GRAPH AU COEUR DE L'INTUITION ARTIFICIELLE

Dans le cadre de mes travaux, l'Intuition Artificielle est la faculté que possède la machine à se rendre compte que quelque n'est pas cohérent en fonction des connaissances qu'elle possède. Cela se traduit concrètement par sa capacité à identifier grâce à un socle de connaissances (construit à l'étape précédente du processus) conséquent qu'il y a des incohérences, des manques...

Cette faculté à subodorer est décrite dans ce chapitre dans un paradigme d'analyse de graphes présentant une approche d'identification de relations directes ou indirectes (explicites ou implicites) qui font défaut et qui traduisent l'Intuition Artificielle.

Dans un premier temps, afin de contextualiser ces travaux, je commence par donner les éléments nécessaires à la compréhension du paradigme des graphes utilisé pour créer l'outil que je présente en section partie : nk-CSG (nk-Correlated Sample Graph).

5

L'ANALYSE DE GRAPHES

Dans le cadre de mon post doctorat au LIG, je travaillais sur l'approximation de sous-graphes et à une mesure capable de caractériser leur distance au(x) graphes de départ. Autrement dit, si les arêtes extraites existent dans le graphe le matching est exact, si elles existent à condition de transiter par un noeud intermédiaire le matching est approché où le nombre de noeuds de transition augmente la distance de matching.

En parallèle, en supervisant un doctorant (Seyed Hamid Mirisaee) qui travaillait sur une problématique de factorisation de matrices non négatives pour optimiser la recherche d'itemsets fréquents. La combinaison de ces recherches m'a mené vers l'idée de travailler sur les graphes en considérant leur matrice d'adjacence. Ainsi est né le travail présenté dans le chapitre suivant.
Cette section rappelle les définitions de graphe nécessaires à la compréhension de la définition de matrice polynomiale permettant de calculer k-relation.

5.1/ LES GRAPHES

Definition 33:

Graphe. Soit G un graphe défini par $G = (V, E)$, avec V un ensemble de sommets (ou nœuds) et $E \subseteq V \times V$ un ensemble d'arêtes. Chaque $e \in E$ est un couple de sommets (v_i, v_j)

Definition 34:

Graphe orienté. Soit Go un graphe orienté défini par $Go = (V, A)$, avec V un ensemble de sommets et $A \subseteq V \times V$ un ensemble d'arcs. Chaque $a \in A$ est constitué de quelques sommets (v_i, v_j) avec $v_i, v_j \in V$ orientés de v_i à v_j.

Definition 35:

Sous-graphe. Soit SG un sous-graphe tel que SG = (V ', E') iif $V' \subseteq V$ et $E' \subseteq E$

Definition 36:

Sample graph ou graphe échantillon. Soit S un échantillon de graphe tel que $S = (V', E')$ iif $E' \subseteq E$

Definition 37:

Sous-graphe étiqueté. Un sous-graphe étiqueté est un tuple $S' = (V, E, \varphi)$ où $\varphi : V \rightarrow L$ est une fonction d'étiquetage avec un ensemble d'étiquettes L.

Definition 38:

Chemin.
Un chemin est une séquence $(v_0, v_1, ..., v_{n-1}, v_n)$ de sommets de G tels que deux sommets consécutifs v_i et v_{i+1} sont reliés par un bord de G: $\forall i, 0 \leq i \leq n-2, (v_i, v_{i+1}) \in E$. Les sommets v_0 et v_n sont respectivement origine et fin du chemin.

Definition 39:

Longueur d'un chemin. La longueur d'un chemin contenant n + 1 sommets et n arcs est n.

Definition 40:

Matrice d'adjacence. Une matrice d'adjacence est la représentation d'un graphe par une table. Les noeuds sont donnés en colonnes et en lignes, et s'il existe une relation entre deux noeuds qui forme un chemin de longueur 1, la valeur 1 est donnée à l'intersection de la colonne et de la ligne considérées.
Dans le cas d'un graphe non orienté, la matrice est alors symétrique.

Definition 41:

Matrice à la puissance. Soient : G un graphe, adj(G) sa table d'adjacence, $(adj(G))^i$ sa table d'adjacence élevée à la puissance i.

$$V^k = \sum_{i=1}^{k} [(adj(G)]^i$$

Avec *i* : longueur de chemin entre chaque paire de noeuds.
k : longueur maximale de chemin entre chaque paire de noeuds.
Note : cette matrice est normalisée (i.e. chaque cellule de V^k est divisée par la valeur maximale présente dans V^k) pour pouvoir en comparer plusieurs entre elles.

V^k stocke le nombre (normalisé) de chemins ayant au maximum la longueur k qui relient chaque paire de noeuds.

La pertinence d'une relation est liée au nombre de relations directes ou indirectes entre deux nœuds.

5.2/ TRAVAUX CONNEXES EN ÉTUDE DE GRAPHES POUVANT INTÉRESSER L'INTUITION ARTIFICIELLE

Les travaux menés dans le cadre de l'Intuition Artificielle m'ont amené, comme je le présente dans le chapitre suivant à définir les nk-Correlated Sample Graphs (nk-CSG). Chacun de ces termes à un sens bien précis issus de travaux connexes que je présente ci-dessous. Préalablement je fixe certaines notions dans le cadre de ces travaux :

- le jeu de données peut être sous forme d'un ou plusieurs graphes
- "relation principale" signifie majoritairement relié (sans notion de pondération d'arêtes)
- "relié" signifie explicite (i.e. directement connecté) ou implicite (i.e. indirectement connecté)
- les relations peuvent être orientées ou non

Le travail effectué est novateur mais est en connexion avec d'autres travaux présentés ci-dessous.

5.2.1/ NOTION DE "CONNEXION"

5.2.1.1/ CONNECTION DE SOUS-GRAPHES

[Faloutsos et al., 2004] a introduit un domaine pour l'exploration de graphes : connection sub-graphs, qu'il définit comme "un petit sous-graphe d'un grand graphe capturant au mieux la relation entre deux nœuds", ce qui est exactement la signification de la connaissance à extraire pour l'Intuition Artificielle. Avant cet article, il n'existait "aucun travail qui aborde directement ce problème dans la littérature publiée", pas même dans des domaines tels que "PageRank [Page et al., 1998], [...], clustering de graphes, partitionnement et réorganisation de matrices [Brandes et al., 2003, Dhillon et al., 2003, Karypis et al., 1996, Satuluri et al., 2009, Virtanen, 2003], [...] propagation d'influence [Kempe et al., 2003], "customer value" d'un noeud ou d'autres graphes clairsemés [Albert et al., 1999, Dorogovtsev et al., 2002, Faloutsos et al., 1999, Newman, 2003] ", qui sont également des domaines intéressant l'Intuition Artificielle.

Après cette publication, aucun autre travail plus proche de la théorie de connexion de sous-graphes n'a été effectué.

L'approche proposée dans [Faloutsos et al., 2004] concerne les graphes non orientés avec arêtes pondérées, avec un centre d'intérêt ciblé sur deux nœuds. Au contraire, l'Intuition Artificielle considère les caractéristiques suivantes :

- des graphes non pondérés
- des graphes orientés comme non-orienté
- capable de traiter avec un noeud d'intérêt comme avec un graphe entier ou une collection de graphes

5.2.1.2/ LINK MINING OU ÉTUDE DES CONNEXIONS

Parler de relation conduit à considérer le domaine de l'étude des connexions qui est défini par [Getoor et al., 2005] comme "un domaine de recherche émergent qui se situe à l'intersection du travail en analyse de connexion et [...] d'extraction de graphes" et est composé de 8 tâches. Notre recherche correspond à la tâche LBR (Link-Based Object Ranking), qui vise à "exploiter la structure de connexions d'un graphe pour ordonner ou hiérarchiser l'ensemble d'objets dans le graphe". Mon travail complète cette approche en envisageant de hiérarchiser les relations des graphes.

5.2.2/ COMMUNAUTÉ DE GRAPHES VERSUS ÉCHANTILLON DE GRAPHE

[Leskovec et al., 2010] définit la communauté de graphes comme "un groupe de nœuds avec plus et/ou de meilleures interactions entre ses membres qu'entre ses membres et le reste du graphe".
La notion de communauté implique que les nœuds sont connectés entre eux. Mais il n'y a aucune raison pour que les relations les plus fortes concernent le même groupe de nœuds, et le résultat de l'extraction peut être un graphe déconnecté.
C'est pourquoi il est plus pertinent de considérer les échantillons de graphes qui considèrent les propriétés "similaires par rapport au graphe original" [Leskovec et al., 2006].
En effet, l'extraction de sous-graphes, qui considère que le sous-graphe extrait doit être une sous-partie exacte du graphe original, alors qu'un échantillon de graphe, permet d'introduire la notion de similarité, et peut avoir des relations qui ne sont pas présentes dans le graphe original (relations indirectes). Cette dernière est plus pertinente.

5.2.3/ DISTANCE K

5.2.3.1/ PROXIMITÉ DE RÉSEAU

Dans l'objectif d'extraire des relations fortes entre les nœuds, considérés comme étant les nœuds ayant le plus grand nombre de relations directes ou indirectes. Ceci conduit à considérer la notion de proximité telle que présentée par [Koren et al., 2007] comme "une notion subtile dont la définition peut dépendre d'une application spécifique" et "peut aider à découvrir des communautés inattendues dans n'importe quel graphe". Ces descriptions correspondent à notre compréhension de ce que nous recherchons, mais les mesures de proximité définies sont basées sur des arêtes pondérées, ce qui est hors sujet. De plus, [Faloutsos et al., 2004] considère que cette approche "échoue à capturer la notion de *meilleur chemin* dans un réseau social".

5.2.3.2/ INFÉRENCE CONFIANTE OU TRUST INFERENCE

Pour calculer le nombre de chemins de longueur variables de chaque paire de nœuds dans un graphe, je distingue mon approche de l'inférence confiante qui vise à déduire un score de fiabilité entre deux éléments. L'inférence confiante intéresse principalement deux domaines: l'inférence basée sur le chemin et l'inférence basée sur les composants, mais les deux ne sont pas mesurables et supposent l'existence d'un sous-graphe qui est toujours une question ouverte depuis que [Yao et al., 2012] en a fait mention.

Pour l'Intuition Artificielle, il faut un algorithme évolutif et considérer un échantillon de graphe au lieu d'un sous-graphe.

5.2.4/ NOMBRE N DE RELATIONS À EXTRAIRE

Dans le cadre de ces travaux, le nombre de relations à extraire correspond au n dans les $nk - CSG$ et peut être lié à deux sources :

5.2.4.1/ N-CSDP

Le premier est la découverte de sous-graphes n-connectés (n-CSDP) présentée par [Chen et al., 2006] comme une approche permettant de "trouver un sous-graphe de petite taille pouvant capturer correctement la relation entre les n nœuds donnés dans un grand graphe". Dans le cadre de l'Intuition Artificielle, il s'agit d'extraire un graphe avec des dimensions restreintes et contrôlées : contenant n arêtes capturant les meilleures relations du graphe. Mais dans le cadre des nk-CSG, n détermine le nombre d'informations extraites et non le nombre d'informations saisies.

5.2.4.2/ K-NN OU K PLUS PROCHES VOISINS

La seconde approche est une méthode non paramétrique pour la classification de patterns, connue depuis sous l'appellation des k plus proches voisins [kNN, 1951]. Cette approche recherche les nœuds les plus proches d'une source et k représente le nombre de ces nœuds les plus proches. Pour l'Intuition Artificielle, k signifie la longueur maximale du chemin reposant sur deux nœuds. n étant le nombre de relations extraites.

K-NN une approche pertinente, sauf que nous extrayons n relations ayant le plus grand nombre de chemins de longueur 1 à k, dans l'ensemble du jeu de données, et non les n nœuds les plus proches d'un nœud donné.

6

NK-CORRELATED SAMPLE GRAPH POUR IDENTIFIER LES RELATIONS MANQUANTES

Intuition. Cette faculté prodigieuse à saisir les indices les plus subtils, ceux que personne n'aperçoit.

Jean-Claude Lalanne-Cassou

La recherche dans les domaines touchant les réseaux complexes et immenses tels que les réseaux sociaux ou les réseaux de gènes se sont bien développés tant leur utilisation est devenue incontournable. En raison de leurs structures, ces réseaux sont généralement représentés par des graphes. Des méthodes d'exploration de graphes sont donc nécessaires pour explorer ces données. Il existe de nombreux domaines de recherche sur l'extraction de graphes, tels que ceux listés par [Jiang et al., 2013], tous cherchant à extraire des connaissances contenues dans des données. Dans le cas des graphes, trouver l'information principale consiste généralement à extraire les relations les plus pertinentes. Dans le cadre de l'Intuition Artificielle, l'intérêt se porte sur l'identification des relations principales du ou des graphes même si le jeu de données n'exprime pas explicitement cette information.

[Faloutsos et al., 2004] introduit la définition de connexion de sous-graphes comme "un petit sous-graphe d'un grand graphe qui capture le mieux la relation entre deux nœuds". Cette définition correspond à la compréhension de la relation entre les éléments au sens de l'Intuition Artificielle. Mais à ma connaissance, il n'y a pas d'autre travail prenant en compte cette notion de relation dans la littérature.
Plus précisément, mon intérêt est de mettre en évidence, à partir d'un ou plusieurs graphes, les relations explicites et / ou implicites les plus pertinentes en considérant que les relations les plus pertinentes à rechercher sont celles pour lesquelles deux nœuds sont le plus (directement ou indirectement) en corrélation. Pour qualifier au mieux cette idée nous parlerons de nk-correlated sample graph (nk-CSG) avec k représentant une distance telle que le nombre de poignées de main (dans la théorie du petit monde) séparant deux personnes. Nous parlons de nœuds en corrélation (et non de noeuds connectés) car dans le cas de l'Intuition Artificielle il faut considérer les relations indirectes aussi bien que les relations directes. Ensuite, l'idée principale est de rechercher l'échantillon des nœuds les n plus corrélés ayant une distance inférieure à k.

Commençons par rappeler quelques définitions inhérentes aux graphes et aux matrices pour introduire la notion de matrice polynomiale permettant de qualifier le degré de relation entre deux noeuds. Puis nous verrons comment extraire les nk-correlated sample graph et comment traiter les gros volumes de données, que ce soit dans un graphe unique ou une collection de graphes. Pour cela je présente un algorithme naïf de complexité $\theta(n^3)$ appelé multiplication de matrices à intersection non nulle et utilisé pour des expérimentations qui ne sont pas présentées dans ce manuscrit car non encore publiés d'une part, mais également parce qu'ils ne sont pas pertinent à l'introduction du domaine de l'Intuition Artificielle.

6.1/ APPROCHE DES MATRICES POLYNOMIALES POUR EXTRAIRE LES NK-CORRELATED SAMPLE GRAPHS

L'utilisation des matrices polynomiales pour extraire les nk-CSG porte un double intérêt. Premièrement, il s'agit d'identifier et d'extraire les relations explicites ou implicites les plus solides et, deuxièmement, gérer un volume de données conséquent. Il y a donc nécessité à trouver une méthode efficace pour résoudre ces problèmes.

L'exploration de graphes permet de travailler avec des matrices d'adjacence qui sont des matrices binaires et sont donc faciles à manipuler. De plus, élever une matrice binaire à une puissance est un calcul léger qui nous donne un bon indicateur de la force d'une relation entre deux nœuds. En effet, plus deux nœuds sont liés par plusieurs chemins de longueur déterminée (inférieure à une distance k), plus ils sont pertinents à extraire. Dans cette partie, je présente l'approche menée pour ces travaux, par la définition de matrice polynomiale et la méthode de construction de cette matrice à partir de laquelle extraire les n relations les plus fortes.

6.2/ MATRICE POLYNOMIALE

Les matrices d'adjacence sont simple à manipuler et possèdent des propriétés permettant, en plus d'être évolutive, de répondre notamment aux contraintes de volume. En effet, ce sont des tables binaires stockable facilement en mémoire machine et le calcul matriciel est léger et efficace. Je m'intéresse à une méthode permettant de traiter un jeu de données très volumineux. Dans le cadre de l'exploration de graphes, cela signifie qu'il faut traiter avec un gros graphe ou une collection de graphes. Pour cela, j'introduis la définition de matrice polynomiale pour prendre en compte ces deux considérations.

Les matrices polynomiales sont établies en additionnant les élévations successives à la puissance de la matrice d'adjacence de la puissance 1 à la puissance k.

Notons que le fait d'élever une matrice à une puissance maintient la relation dans la table et peut être exploitée avec les graphes dirigés comme non dirigés.

Definition 42:

Elévation d'une matrice d'adjacence à la puissance.
Soit G un graphe, et $adj(G)$ sa matrice d'adjacence. $(adj(G))^i$ est alors cette matrice élevée à puissance i.

La propriété qui m'intéresse particulièrement est le fait que dans le cas d'une matrice élevée à une puissance i, chaque valeur qu'elle contient correspond au nombre de chemins de longueur i entre chaque paire de nœuds.
Un i spécifique ne caractérise qu'une seule longueur de chemin. Ainsi, pour considérer tous les chemins possibles de longueur 1 à k les matrices polynomiales sont utilisées.

Definition 43:

Matrice polynomiale pour un graphe unique
Une matrice polynomiale contient le nombre total de tous les chemins des longueurs $l \in [1,k]$ entre chaque paire de nœuds. Elle correspond donc à la somme de chaque matrice d'adjacence élevée successivement aux puissances 1 à k.

$$V_G^k = \sum_{i=1}^{k} [(adj(G)]^i \; with \; k \geq 3$$

Pour fixer k il est préférable de respecter $k \geq 3$ afin d'éviter les problèmes de dépendance de parité [van Dongen, 2000], et d'avoir k inférieur au diamètre du graphe car un graphe complet est inutile pour la considération sémantique des nk-CSG. De manière générale, les expérimentations montrent que k = 3 est un bon choix.

6.3/ MATRICE POLYNOMIALE POUR UNE COLLECTION DE GRAPHES

Construire cette matrice combinée est similaire à la construction pour un grand graphe. La différence réside dans le fait que la matrice polynomiale d'un graphe a la même taille pour chaque puissance. Dans le cas d'une collection de graphes, les nœuds ne sont pas nécessairement les mêmes, alors l'ensemble des nœuds de la matrice polynomiale correspond à l'union de l'ensemble des nœuds de tous les graphes. Notons que, dans l'objectif d'extraire les nœuds qui sont les plus corrélés, un ensemble de graphes sans relations entre eux perd son intérêt.

$$V^k = \sum_{g=1}^{|S|} \sum_{i=1}^{k} [(adj(G_g)]^i \; with \; k \geq 3$$

avec $k \in [3; diametre[$[1] et S l'ensemble de tous les graphes G_g.

6.4/ PROCESSUS D'EXTRACTION DU NK-CORRELATED SAMPLE GRAPH

Le processus d'extraction du nk-correlated sample graph se déroule en deux étapes présentées ci-après. La première consiste à construire la matrice polynomiale et la seconde à extraire l'échantillon du graphe.

1. Le diamètre d'un graphe est la plus grande distance possible qui puisse exister entre deux de ses sommets ; la distance entre deux sommets étant définie par la longueur d'un plus court chemin entre ces deux sommets. En d'autres termes, le diamètre est l'excentricité maximale de ses sommets.[wikipedia]

Algorithm 6 $build\_V^k(adj(G), k)$

Require: k: longueur de chemin maximale entre les noeuds
adj(G): matrice d'adjacence du graphe G
Ensure: V_G^k {*Retourne la matrice polynomiale* V_G^k *de G*}
1: $V_G^k = adj(G)$;
2: $i = 2$;
3: **for** $i \leq k$ **do**
4: $V_G^k = V_G^k + ajd(G)^i$;
5: **end for**
6: **return** V_G^k

6.4.1/ CONSTRUIRE LA MATRICE POLYNOMIALE

L'algorithme 6 décrit comment construire la matrice polynomiale. Il prend en paramètre la table d'adjacence d'un graphe et la puissance k à laquelle l'élever.

Cet algorithme 6 est appelé par l'algorithme principal 7 pour chaque graphe de la collection et construit la matrice polynomiale de la collection.
Notons que dans le cas d'un graphe unique, au lieu d'une collection, l'algorithme n'est appelé qu'une seule fois et le processus se déroule de la même manière.

6.4.2/ EXTRACTION DU NK-CORRELATED SAMPLE GRAPH

Le nk-correlated sample graph extrait (algorithme 7) est composé des n arêtes ayant les valeurs les plus élevées de la matrice polynomiale, c'est-à-dire celles ayant le plus grand nombre de relations à leur voisinage proche (distance inférieure à k), autrement dit : si les nœuds étaient des personnes, ce serait les plus populaires.
Plusieurs autres critères d'extraction sont possibles, selon la sémantique souhaitée pour le graphe. Par exemple, n peut correspondre au pourcentage de représentativité du graphe échantillon : extraire toutes les meilleures arêtes jusqu'à ce que la somme des valeurs de ces arêtes extraites corresponde à un pourcentage de la valeur totale de la matrice polynomiale. Ou un pourcentage du nombre total d'arêtes afin d'élaguer le graphe... Ici, la connectivité du sous-graphe n'est pas prise en compte, mais cela pourrait également être un critère d'extraction.

J'ai développé une méthode de multiplication de matrice appellée Non-Zero Intersection Matrix Multiplication pour optimiser l'efficacité en terme de temps d'exécution et les espaces mémoires.

6.5/ MÉTHODE DE MULTIPLICATION DE MATRICES D'INTERSECTION NON NULLE

Afin d'évaluer mon approche, j'ai développé un algorithme, car l'efficacité dépend directement de cette méthode. La méthode s'appelle Multiplication de matrices d'intersection non nulle car elle ne calcule que les résultats non nuls. Cela vise à économiser du temps de calcul et le volume de mémoire nécessaire. Cette méthode est donc particulièrement utile pour les matrices creuses.

Algorithm 7 Extraction du k-correlated sample graph

Require: k: longueur maximale du chemin entre deux noeuds
S: ensemble des graphes G
n: taille du sous graphe extrait en terme de nombre d'arêtes
Ensure: nkCSG : nk-correlated sample graph
1: Matrice Polynomiale P; {*// Somme des* V_G^k *de chaque* G}
2: **for all** $G \in S$ **do**
3: P=P$\oplus build_V^k(adj(G), k)$;
4: **end for**
5: kCSG ←extraire_meilleures_aretes(P,n); {*//Rechercher dans P pour les n aretes ayants les plus grandes valeurs.*}
6: **return** kCSG

M	A	B	C	D
A	0	1	1	0
B	0	0	1	1
C	0	1	0	1
D	0	0	0	0

Table 6.1 – Matrice d'adjacence M du graphe orienté de la figure 6.1

Considérons le graphe orienté présenté dans la figure 6.1. Sa matrice d'adjacence M correspondante est donnée dans le tableau 6.1.
En regardant ce graphe, chaque humain conclut immédiatement qu'il se passe quelque chose entre A et D. Mais ce graphe ne montre aucune relation entre ces deux nœuds. Tout l'intérêt de ce travail est de présenter une méthode permettant à la machine de souligner cette évidence.

Présentons ici les différentes étapes d'une méthode pour multiplier deux matrices M1 et M2. Pour simplifier l'explication, considérons M1 = M2 = M.

6.5.1/ CRÉER LA TABLE "LIGNES"

La première étape consiste à créer une table Lignes. Cette table contient les indices des lignes de M1 pour lesquelles M1 n'est pas nulle telle que présentée dans la table gauche de la table 6.2.

Notons que Lignes[3] est vide, nous ne tiendrons donc pas compte de cette ligne dans les étapes suivantes.

6.5.2/ CRÉER LA TABLE "COLONNES"

La seconde étape créé la table Colonnes. Cette table présente dans table 6.2 (à droite) les indices des colonnes de M2 pour lesquelles M2 n'est pas nulle.

Notons que Colonnes[0] est vide, nous ne considérons donc pas cette colonne.

Figure 6.1 – Petit exemple de graphe orienté.

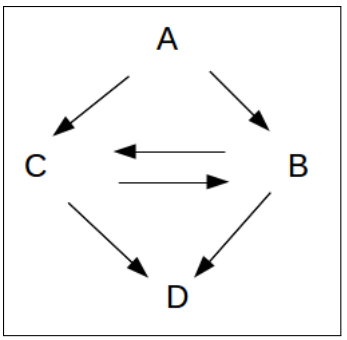

	Table Lignes	
A	1	2
B	2	3
C	1	3
D	vide	

	Table Colonnes	
A	vide	
B	0	2
C	0	1
D	1	2

Table 6.2 – Tables Lignes et Colonnes

6.5.3/ CALCULER LA MATRICE D'INTERSECTION

La table Intersection contient les seuls éléments nécessaires au calcul de la matrice résultante M_{finale} de la multiplication entre les matrices M1 et M2 (algorithme 8).

Les résultats obtenus sont affichés dans le tableau 6.3.
Notons que si M1 et M2 sont des matrices binaires telles que des matrices d'adjacence, alors $M_{finale}[i, j] = |Intersection[i, j]|$, ce qui est le cas dans notre exemple.

6.6/ ALGORITHME D'INTERSECTION NON NULLE ET COMPLEXITÉ

6.6.1/ ALGORITHME D'INTERSECTION NON NULLE

Considérons que les deux tables Lignes et Colonnes sont construites. L'algorithme présenté dans la figure 9 calcule la matrice M_{finale} = M1 * M2, qui consiste à intersecter la table Lignes de M1 avec la table Colonnes de M2.

Dans cet algorithme, nbElt correspond au nombre d'éléments dans le jeu de données, Lignes[i].nbIndex correspond au nombre d'indices pour un élément dont la valeur est non nulle dans la ligne correspondante de la matrice (nombre d'éléments de la ligne i dans le tableau Lignes) et respectivement pour les colonnes[j].nbIndex.

6.6.2/ CAS PARTICULIERS POUR DES OPTIMISATIONS DE TEMPS DE CALCUL

Il existe trois types de multiplications possibles : multiplier deux matrices de degré 1, une matrice de degré 1 et une autre de degré quelconque, et deux matrices de degré quelconque. Les deux premiers cas permettent une optimisation de l'algorithme 9 visant à gagner du temps même s'il n'y a pas de changement de complexité de l'algorithme.

Algorithm 8 Fonction pour calculer l'intersection de matrices

```
1:  for (int i=0 ; i < nbLignes ; i++) do
2:    for (int j=0 ; j < nbColonnes ; j++) do
3:      if (Intersection[i,j] != ∅) then
4:        int sum=0;
5:        for  each   k ∈ Intersection[i,j] do
6:          sum += M1[i,k] * M2[k,j];
7:        end for
8:        M[i,j] = sum;
9:      end if
10:   end for
11: end for
```

M_{finale}	A	B	C	D
A	0	1	1	2
B	0	1	0	1
C	0	0	1	1
D	0	0	0	0

Table 6.3 – Matrice finale $M_{finale} = M1 \times M2$.

- Cas de deux matrices du degré 1. La valeur à définir dans M[i,j] est égale au nombre d'éléments de l'intersection des lignes[i] et des colonnes[j]. En effet, le calcul est la somme des multiplications entre deux éléments valant 1, il n'y a donc aucun intérêt à calculer 1 * 1, il suffit de savoir combien de fois cela est fait.
- Cas d'une seule matrice de degré 1. Le calcul est basé sur la multiplication des valeurs de la matrice de degré 1 (M1) (qui ne peut être que 1 ou 0 mais on ne considère que l'indice pour une valeur non nulle), et des valeurs v de l'autre matrice M2. Ainsi $1 * v = v$, la ligne 9 de l'algorithme doit donc être modifiée par $valeur+ = M2[Lignes[i][l]][j]$;.
- Cas sans matrice de degré 1. Dans ce cas, il n'y a pas d'optimisation et le calcul doit être entièrement effectué. La ligne 9 doit donc être modifiée par $sum+ = M1[i][Lignes[i][l]] * M2[Lignes[i][l]][j]$; (parce que Lignes[i][l] = Colonnes[j][c] en cas d'intersection).

6.6.3/ COMPLEXITÉ DE L'ALGORITHME DE MULTIPLICATION D'INTERSECTION NON NUL

Considérons n comme le nombre d'éléments dans l'ensemble de données. L'algorithme 9 qui calcule la multiplication de deux matrices est tel que :

- ligne 1: $\theta(n)$
- ligne 2: $\theta(n)$
- ligne 5 à 17: $\theta(2n)$ car les tables sont ordonnées

Alors la complexité est $\theta(n * n * 2n) = \theta(2n^3) \equiv \theta(n^3)$.

6.6.4/ COMPLEXITÉ DE LA MÉTHODE DU POLYNÔME DE MATRICES

- Calcul des lignes: $\theta(n^2)$
- Calcul des colonnes: $\theta(n^2)$
- Multiplication M2 = M1 * M1: $\theta(n^3)$
- Multiplication M3 = M1 * M2: $\theta(n^3)$
- Somme M1 + M2 + M3: $\theta(n)$

Alors la complexité est $\theta(2n^2 + 2n^3 + n) \equiv \theta(n^3)$

6.6.5/ SÉMANTIQUE DE LA PONDÉRATION DANS LE CADRE DES RELATIONS DIRECTES OU INDIRECTES

Selon le type de relation (directe ou indirecte) que l'on souhaite prendre en compte, il est possible de choisir d'ajouter du poids aux matrices. Dans le cas d'une relation réelle, c'est-à-dire celles indiquées dans le graphe du jeu de données, i.e. une relation directe, qui signifie que l'on s'intéresse simplement à distinguer ces relations existantes entre elles. Par contre, si l'on recherche une relation principale (cachée) indirecte, il ne faut pas pondérer la matrice. En effet, plus la puissance d'une matrice est grande, plus ses valeurs sont élevées.

Le principe de la matrice pondérée est de considérer que plus la matrice est puissante, plus les chemins sont longs. Donc, plus les nœuds sont éloignés, moins ils sont concernés. Il est donc proposé de diviser chaque valeur de la matrice par sa puissance avant de les cumuler.

Ainsi si l'on est intéressé par des liens existants réellement dans les données, un calcul pondéré est la solution.
Par ailleurs, si l'on souhaite découvrir ce qui est caché derrière les données, la solution non pondérée est la meilleure. En montrant les résultats, la machine nous dit ce que nous soupçonnions au début : la relation la plus forte indique $A \rightarrow D$.

Cette approche soutend l'Intuition Artificielle.

6.7/ EXPÉRIMENTATIONS

Le projet ANR HYBRIDE mené par le LORIA, laboratoire pour lequel j'ai effectué un service d'ATER à la suite de mon post doctorat, m'a permis de confronter ce sujet à un cas réel. Ce projet visait à développer de nouvelles méthodes et de nouveaux outils pour guider la découverte de connaissances à partir de textes issues des méthodes de traitement du langage naturel (NLP) et d'analyse de données (KDD). L'idée principale était de

concevoir un processus menant à une interaction entre la NLP et le KDD où les méthodes de NLP guident les méthodes d'exploration de données et réciproquement, pour analyser et exploiter les documents textuels en fonction de leur contenu. Les méthodes KDD utilisées incluaient l'Analyse Formelle de Concepts (FCA) qui permet de regrouper les propriétés des objets en partage. Ma contribution a consisté à inclure les polynômes de matrices d'adjacence (nk-CSG), et ont permis de mettre en évidence les relations indirectes les plus intéressantes. En travaillant sur le jeu de données d'Orphadata, en particulier avec les maladies rares liées aux symptômes et les maladies rares liées aux gènes. Mon approche a permis d'identifier des relations indirectes: entre gènes, symptômes et maladies non mentionnées dans l'ensemble de données et significatives d'erreurs ou d'oublis.

6.8/ CONCLUSION ET AMÉLIORATIONS

Les humains sont capables d'avoir l'intuition mais ne peuvent pas l'appliquer à une énorme quantité de données. Au contraire, l'ordinateur peut gérer un tel volume mais ne sait pas comment "ressentir" les données. Donc, cette approche s'intéresse à mélanger ces deux forces en permettant à l'ordinateur de se doter d'une l'intuition.Cette approche est basée sur la théorie des graphes car elle permet (i) de représenter facilement les relations entre éléments (relations dirigées ou non dirigées), (ii) de gérer leur matrice d'adjacence qui est facile à stocker et à manipuler pour le calcul, (iii) permet exploiter la théorie consistant à dire que le fait d'élever une matrice à une puissance x donne le nombre de chemins distincts de longueur x entre chaque paire de nœuds. J'introduis la notion de *matrice polynomiale* pour exprimer la corrélation entre les noeuds visant à extraire le graphe échantillon *nk-correlated* caractérisant les n relations les plus fortes dans le graphe ou dans une collection de graphes. La relation la plus forte étant celle ayant le plus grand nombre de chemins allant de la longueur 1 (relation directe) à k (relation indirecte).
De plus, cette approche est capable de distinguer les liens existants (informations explicites) ou de mettre en évidence des relations fortes inconnues (informations implicites). Enfin, cette approche est adaptable à différents types d'extraction des n relations permettant de répondre aux besoins de l'utilisateur.

Les différents algorithmes présentés se veulent intéressants, mais n'ont pas été étudiés pour être compétitifs. Ils permettent de répondre aux expérimentations menées et permettent au lecteur d'appréhender facilement le processus. Mais plusieurs optimisations peuvent être facilement réalisées. En effet, la méthode de multiplication *matricielle d'intersection non nulle* a une complexité de $\theta(n^3)$. Cet algorithme peut être modifié par des algorithmes optimisés tels que Strassen ($\theta(n^{2.807})$) ou Coppersmith-Winograd ($\theta(n^{2.376})$).

Par ailleurs, comme la méthode est basée sur des multiplications de matrices, elles pourraient être facilement calculées en parallèle.

Algorithm 9 Fonction pour calculer M_{finale} = M1 * M2

Require: Table Lignes de M1 (chaque ligne est ordonnée par ordre croissant des index)
Table Colonnes de M2 (chaque colonne est ordonnée par ordre croissant des index)
Ensure: M

```
1: for (int i=0 ; i<Lignes.nbElt ; i++) do
2:   for (int j=0; j< Colonnes.nbElt ; j++) do
3:     int valeur=0; {Valeur pour M[i][j]}
4:     int restart=0; {Index pour arrêter et reprendre la recherche}
5:     for (int l=0 ; l<Lignes[i].nbIndex ; l++) do
6:       bool trouver=faux; {vrai s'il y a une intersection}
7:       for (int c=restart ; c< Colonnes[j].nbIndex AND !find AND
         Lignes[i][l]<Colonnes[j][c] ; c++) do
8:         if (Lignes[i][l] == Colonnes[j][c]) then
9:           valeur ++; {pour la multiplication de deux matrices de degré 1}
10:          restart = c+1;
11:          trouver = vrai;
12:        end if
13:        if (Lignes[i][l] > Colonnes[j][c]) then
14:          restart = c+1;
15:        end if
16:      end for
17:    end for
18:    if (valeur != 0) then
19:      M[i][j]=valeur;
20:    end if
21:  end for
22: end for
23: return M
```

III

LES DONNÉES

LES DONNÉES

Nous avons vu l'étape de création de l'Intuition Artificielle qui s'appuie sur un socle de connaissances. Ce socle est construit à partir de données collectées en fonction du centre d'intérêt dans lequel on souhaite voir l'Intuition Artificielle opérer. Ce chapitre aborde la question de la gestion de ces données. En effet, elles doivent répondre à certaines contraintes afin d'être pertinentes pour ces travaux :

- Volume et Pertinence. Tout comme dans le cas de l'intuition humaine, le socle de connaissances de l'Intuition Artificielle doit être suffisamment large pour que l'intuition soit pertinente, tout en s'assurant que les données utilisées soient cohérentes de façon à ce que le socle soit large mais focalisé. Il y a deux intérêts à la Pertinence, celui des données en entrée et des résultats en sortie du système. Dans le cas des données en entrée, c'est au moment de la collecte des données que cette question se pose. Je la prends donc en considération en même temps que le Volume concerné lors de la collecte des données notamment. Pour ce qui est de la Pertinence en sortie, le système est paramétrable pour répondre aux sollicitations de l'utilisateur quant à la précision de la recherche et son nombre de résultats. Puis il est transmis à une ontologie en charge d'expliquer ce résultat et c'est l'utilisateur qui décide s'il l'intéresse ou non.
- Hétérogénéité. Autrement appelée Variété dans le contexte du Big Data. Les sources de données peuvent être multiples et chacune présenter des aspects cohérents avec le domaine étudié. Il s'agit donc de pouvoir toutes les prendre en charge.

En analysant cette liste, on trouve une analogie avec les 5 V du Big Data : Volume, Variété, Vélocité, Véracité et Valeur. Dans mes travaux, la Vélocité n'est pas prises en compte en tant que contraintes de l'Intuition Artificielle. En effet, la Vélocité, qui correspond à la vitesse à laquelle les données sont générées et/ou modifiées n'influent pas sur la constitution d'un bloc de connaissances pour l'Intuition Artificielle. Cependant dans le cadre d'évolutions futures où ce système serait rendu dynamique avec des facultés d'adaptation, cette notion devra être prise en considération.

La Véracité est spécifiquement traitée dans le dernier chapitre.

Ces contraintes sont aussi des verrous dans la mesure où les outils mis en oeuvre pour l'Intuition Artificielle ne résolvent pas ces contraintes. En effet, le Volume et l'Hétérogénéité sont des verrous pour la FCA et les ontologies, alors que la Pertinence l'est uniquement pour l'ontologie seule.
Dans le cadre de la FCA, l'Hétérogénéité à été levée par MARM et le Volume est en cours de résolution par l'intégration de Quality Cover à MARM.
Cependant il reste le domaine ontologique. C'est pourquoi j'ai co-encadré 3 thèses

pour lever ces 3 verrous dans ce domaine. Ma contribution à ces travaux, issue de mon expérience établie par la FCA, passent notamment par toute la partie formelle que j'y ai placée. Bien sur ces thèses ont abordé d'autres questions également, mais dans le cadre de ce manuscrit je n'aborde, que les travaux menés pour lever ces 3 verrous de l'Intuition Artificielle.

Tout d'abord je présente les travaux de Thomas Hassan sur les contraintes simultanées du Volume et de la Pertinence. Puis je présente les travaux de David Werner qui ont précédé ceux de Thomas Hassan et qui s'intéressent à l'Hétérogénéité des données.

7

LE VOLUME ET LA PERTINENCE DES DONNÉES

Contributions :

Thomas Hassan dans le cadre de sa thèse sur crawling web pour la recherche d'informations (détails en section 15.3).

La première thèse menée par David Werner, dans le cadre d'une architecture d'une classification sémantique de l'information appliquée à la classification d'articles de presse, montre que les raisonneurs sémantiques basés sur la logique de description (DL) ne sont pas adaptés pour la montée en charge, et sont donc inadaptés au contexte du Big Data. Des raisonneurs à l'échelle du web [Urbani, 2013] utilisent un raisonnement basé sur des règles OWL Horst, et permettent un passage à l'échelle par parallélisation et distribution de la charge de calcul sur des clusters de machines. La montée en charge peut ainsi s'effectuer via des techniques telles que Map-Reduce. Ce type d'approche peut permettre de palier cette limitation. L'architecture de classification hiérarchique multi-label (ou SHMC) développée dans la thèse de Thomas Hassan dans la continuité de la thèse de David Werner, se base sur ces méthodes et l'approche de David Werner pour palier les limitations de l'état de l'art. En particulier la construction d'un modèle de classification dans un contexte Big Data, et l'application du mécanisme d'inférence pour effectuer la classification selon un modèle décrit par une base de connaissances. L'approche consiste à utiliser le raisonnement sur des règles de Horn, dont l'expressivité est limitée mais permet un passage à l'échelle dans un contexte Big Data, tout en conservant des performances élevées pour la classification.

Dans ce chapitre je précise tout d'abord ce qu'est la classification hiérarchique multi-étiquette puis je présente le travail en 3 volets effectué pour concevoir (i) un outil d'apprentissage non supervisé d'une ontologie, (ii) la classification qu'elle peut réaliser afin d'être utilisée évaluer la Pertinence des données que récupère qu'un (iii) crawler web de notre conception.

7.1/ CLASSIFICATION HIÉRARCHIQUE MULTI-ÉTIQUETTE

Cette partie précise la notion de classification qui consiste à associer un ou plusieurs attributs de classe, ou labels, à un item définit par un ensemble d'attributs ou caractéristiques (features). Lorsque les attributs de classe sont indépendants, la classification est dite "plate". Dans la plupart des cas réels de classification, les labels sont dépendent les uns des autres et ne sont pas traités séparément : ils peuvent être organisés en groupes, et notamment agencés en fonction de leur généralité/spécificité. Le résultat de cet agencement est un structure hiérarchique [Bi et al., 2011, Cerri et al., 2014, Tsoumakas et al., 2007, Silla et al., 2011]. La classification hiérarchique multi-label (hierarchical multi-label classification) est la combinaison de la classification multi-label et de la classification hiérarchique [Bi et al., 2011, Santos et al., 2009, Cerri et al., 2014]. Par définition, la classification hiérarchique multi-étiquette bénéficie des avantages de la classification hiérarchique et de la classification multi-étiquette. Dans cette approche, plusieurs chemins de la hiérarchie de classes peuvent être attribués à chaque item, qui peut donc appartenir à différentes classes d'un même niveau. On différencie deux types de structures hiérarchiques pour l'organisation des labels : les arbres et les graphes acycliques [Cerri et al., 2014]. La figure 7.1 décrit les trois types de structures et leur correspondance avec les méthodes existantes de classification.

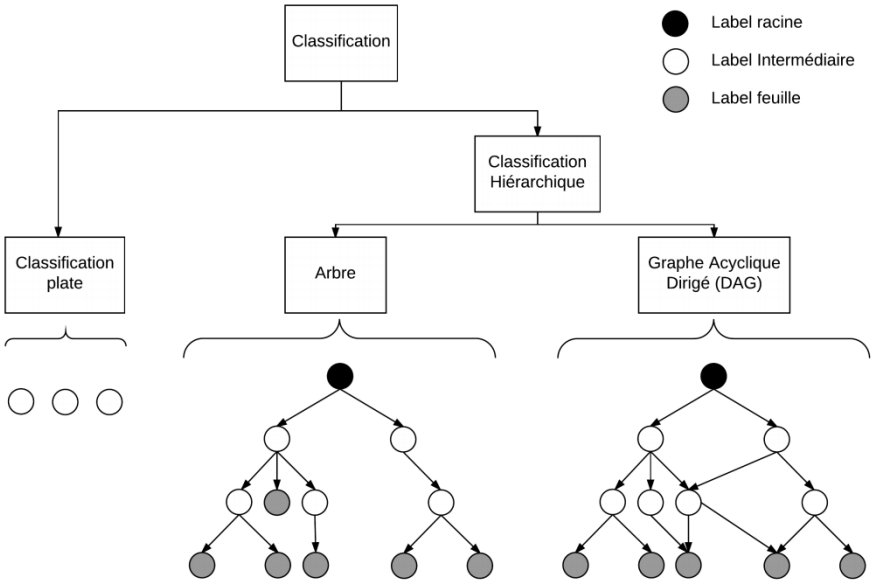

Figure 7.1 – Types de structure pour la classification

L'appartenance (classification) d'une instance à une classe induit l'appartenance aux classes parentes, i.e. assigner une classe à un item revient à assigner le chemin de la classe jusqu'au nœud racine. On considère généralement 2 types d'approches pour résoudre le problème de classification hiérarchique, la classification locale (par des algorithmes de classification conventionnels et une fonction de prédiction) et globale (uniquement par une fonction de prédiction).

L'état de l'art de la classification multi-étiquette (hiérarchique) est éprouvé. Les principales approches, sont basées sur des modèles probabilistes, notamment bayésiens, et des

variantes de ceux-ci. Ces approches bien que performantes sont purement statistiques, et ne tirent pas partie des métadonnées ou de la sémantique éventuellement associée aux données. Nous retenons le dernier aspect de la classification, par l'intégration de d'ontologies dans le processus de classification. L'ontologie [Studer et al., 1998] joue un rôle déterminant dans la définition des termes utilisés pour représenter la connaissance. L'exploitation de bases de connaissances dans un processus de classification est un moyen de rapprocher le processus de classification et les connaissances de l'utilisateur humain. La plupart des travaux de la littérature se concentrent sur l'amélioration du processus de classification en utilisant les ontologies, ce qui permet d'améliorer la description des items, et la performance de la classification. En revanche, ils ne tirent pas avantage des capacités des raisonneurs sémantiques pour classer automatiquement des items [Peixoto et al., 2016]). Au delà de l'aspect descriptif de haut niveau sémantique que permet l'usage d'une ontologie, l'utilisation de raisonneurs pour la tâche de classification peut améliorer les performances du processus de classification [Moeller et al., 2009]). L'architecture de classification hiérarchique multi-label, ou SHMC, se base sur ces méthodes et l'approche de la thèse de David Werner et al. pour palier les limitations de l'état de l'art, en particulier la construction d'un modèle de classification dans un contexte Big Data, et l'application du mécanisme d'inférence pour effectuer la classification selon un modèle décrit par une base de connaissances. Notre approche consiste à utiliser le raisonnement sur des règles de Horn, dont l'expressivité est limitée mais permet un passage à l'échelle dans un contexte Big Data, tout en conservant des performances élevées pour la classification.

7.2/ APPRENTISSAGE NON SUPERVISÉ D'UNE ONTOLOGIE

Le premier outil pour traiter les problématiques de Volume et de Pertinence des données est un processus de classification automatique de données Big Data (en 5 étapes) qui s'appuie sur les technologies du web sémantique pour représenter l'information et effectuer la classification. L'approche permet de classer les nouvelles données via un moteur inférence en respectant les spécificités sémantiques du modèle, tout en palliant les problématiques de passage à l'échelle définies par la thèse de David Werner.

D

Le processus nommé HMC 1 Sémantique, ou SHMC, est composé de 5 étapes (Figure 7.2) :

1. L'étape d'indexation extrait les termes à partir des données d'entraînement (a), i.e. les items non classés, et crée un index inversé (b) des items.
2. L'étape de vectorisation calcule les vecteurs de fréquence des items indexés, et une matrice de cooccurrence (c) à partir des vecteurs (b). Cette matrice recense l'ensemble des termes contenus dans l'index, et l'ensemble des cooccurrences entre chaque paire de termes, i.e. le nombre d'items où deux termes apparaissent simultanément.
3. L'étape de hiérarchisation créé la hiérarchie de labels (d) à partir de la matrice de cooccurrence (c).
4. L'étape de résolution créé les règles de classification (e) à partir de la matrice de cooccurrence (c).

5. L'étape de réalisation peuple l'ontologie avec les nouveaux item (f). Un moteur de raisonnement permet d'inférer les labels les plus spécifiques pour chaque item à partir des règles de classification et la hiérarchie de labels. Le résultat de la classification est composé de l'ensemble des relations liant un item aux labels inférés (g).

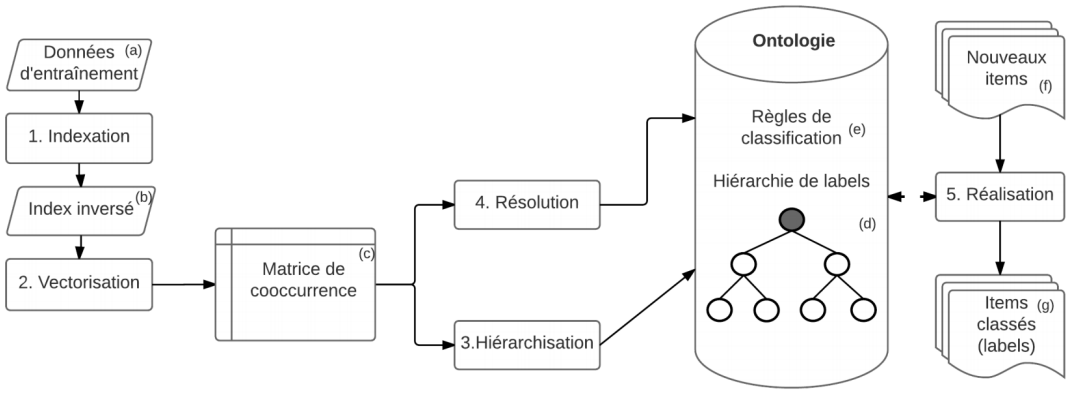

Figure 7.2 – Le processus HMC Sémantique

Tout d'abord,je présente l'apprentissage automatique non-supervisé d'un modèle de classification à partir de grands volumes de données, qui est traité et présenté par les 3 premières étapes. Puis, l'utilisation de ce modèle de classification dans un contexte Big Data, qui est traité et présenté par les 2 dernières étapes..

7.2.1/ INDEXATION

L'indexation est une étape obligatoire qui permet d'une part de palier l'Analyse de Contenu Limitée, et d'autre part de gérer une plus grande Variété de données. L'analyse de Contenu Limitée répond à la difficulté d'extraire automatiquement des informations fiables à partir de données non structurées (texte, image, vidéo...), notamment pour les systèmes de recherche d'information et de recommandation [Lops et al., 2011, Bobadilla et al., 2013]. Le type des données et le cas d'utilisation engendrent de nombreuses contraintes additionnelles qui doivent être prises en compte lors de l'indexation.

Dans le cadre de l'apprentissage automatique, l'indexation est également la composante principale de la réduction de dimension des données initiales. Les principaux objectifs de la réduction de dimension sont les suivantes [Guérif, 2006, Ricci et al., 2011] :

- faciliter la visualisation et la compréhension des données
- réduire l'espace de stockage nécessaire
- réduire le temps d'apprentissage et d'utilisation
- identifier les caractéristique pertinentes

L'indexation doit notamment réduire le bruit et la redondance des données d'apprentissage, et ainsi affiner la sélection des caractéristiques. Cette tâche est d'autant plus importante dans un contexte non supervisé, où les classes peuvent être directement

affectées par les imperfections des données d'apprentissage. Les méthodes employées lors de l'étape d'indexation peuvent être séparées entre la sélection de variables, qui consiste à choisir des caractéristiques dans l'espace de mesure, et l'extraction de caractéristiques, sélectionnées dans un espace transformé [Guérif, 2006].

L'indexation extrait les caractéristiques depuis une collection d'items et génère un index inversé des items. Chaque item est représenté par un vecteur v_{item_i} des termes extraits dans celui-ci, tel que :

$$v_{item_i} < term_1, k_1, term_2, k_2 ..., term_m, k_m >, \forall i \in C$$

où pour tout i, $item_i \in Item$ dans l'ontologie noyau, $k_1 ... k_m$ est le nombre d'occurrences du terme dans l'item, et m est le nombre de termes distincts contenus dans l'item. La création de l'index inversé permet de passer d'un vecteur de termes pour chaque item à un vecteur d'items v_{item_i} pour chaque terme tel que :

$$v_{term_j} < item_1, item_2, ..., item_n >$$

où pour tout j, $v_{term_j} \in Term$ dans l'ontologie noyau, $< item_1, item_2, ..., item_n >$ sont les items où $term_j$ apparaît au moins une fois. Le résultat de l'indexation est un index inversé des items (à chaque terme, sa liste d'items). De la même façon que dans un base de données relationnelle indexée, l'index inversé permet un accès direct et rapide à tout item de la collection à partir d'un ou plusieurs attributs (termes). L'index est utilisé comme base de construction de la matrice de cooccurrence des termes dans l'étape suivante du processus : la Vectorisation.

7.2.2/ VECTORISATION

L'étape de vectorisation a pour objectif de construire une matrice de coocurrence des termes extraits lors de l'étape d'indexation. La construction d'une matrice de coocurrence est une phase commune en apprentissage automatique, et est souvent le point de départ nécessaire à l'utilisation d'autres algorithmes d'apprentissage statistique, i.e. la plupart des mesures d'association basés sur l'étude de la corrélation et de la dépendance de variables aléatoires. De fait, ces méthodes sont utilisées pour des types très variés de données. Pour des données textuelles, les variables aléatoires peuvent être des termes ou des expressions plus complexes comme des collocations ou des entités nommées. [Lin et al., 2010] donne une définition formelle d'une matrice de termes : la matrice de coocurrence d'un corpus de documents est une matrice carrée de taille n × n, où n est le nombre de termes uniques dans le corpus. Une cellule $m_{i,j}$ contient le nombre de cooccurrences entre deux termes $term_i$ et $term_j$, i.e. le nombre d'apparitions simultanées des deux termes dans un contexte donné. Le nombre de cooccurrences est grandement affecté par le choix du contexte, et permet d'influer sur le nombre total de cooccurrences détectées,et par conséquent sur le coût calculatoire de la matrice.

La Vectorisation crée deux types de vecteurs de fréquence de termes pour les items indexés [Salton et al., 1988] : un vecteur de fréquence des termes pour chaque item, ou term-frequency noté TF (1), et un vecteur de fréquence des termes dans l'ensemble du corpus, ou document-frequency noté DF (2). Une matrice de cooccurrence de termes est construite à partir de ces vecteurs. La matrice permet d'analyser les corrélations entre les termes extraits de la collection d'items, en l'occurrence leur proportion conditionnelle, ou fréquence relative. Les étapes suivantes du processus, hiérarchisation et résolution,

exploitent la matrice de cooccurrence afin de générer le modèle prédictif. Les deux types de vecteurs sont définis comme suit :

1. Term-frequency (pour chaque item), $V^i_{tfidf} = \{t, tfidf_{i,t,C}\}$, où t est le terme, i est l'item, C est la collection de documents et $tfidf_{i,t,C}$ est la valeur TF-IDF Salton et al.(1988) du terme t dans l'item i.
2. Document frequency (collection de documents), $V_{df} = \{t, df_{t,C}\}$ où t est le terme, C est la collection d'items, et $df_{t,C}$ est le nombre d'items de la collection C où t apparaît.

La matrice de cooccurrence, notée cfm, représente la cooccurrence entre toute paire de termes $(term_i, term_j)$ dans la collection de documents C, et est définie comme suit :

$$cfm(term_i, term_j) = |\{item_{i,j} \in C | item_{i,j} \in vector_{term_i} \wedge item_{i,j} \in vector_{term_j}$$

où $item_{i,j}$ est un item, $vector_{term_i}$ est le vecteur issu de l'index inversé pour le terme $term_i$, et $vector_{term_j}$ est le vecteur correspondant pour le terme $term_j$

La matrice de cooccurrence cfm est une matrice symétrique de taille n × n où n est le nombre de termes différent dans l'index inversé. La diagonale de la matrice $cfm(term_i, term_i)$ représente le nombre d'occurrences total (document frequency) du terme $term_i$ dans la collection C. Par conséquent $cfm(term_i, term_i)$ est la valeur maximale de la ligne de la matrice correspondant au terme $term_i$.

7.2.3/ HIÉRARCHISATION

La hiérarchisation sélectionne les termes pertinents à partir des vecteurs de termes et définit ainsi les concepts qui appartiennent à la hiérarchie. Deux méthodes pour construire automatiquement une hiérarchie de classes (ou taxonomie) sont utilisées dans la littérature [De Knijff et al., 2013, Meijer et al., 2014] :

- La méthode de subsomption qui construit les relations de généralisation - spécialisation en se basant sur la cooccurrence des concepts
- Le regroupement hiérarchique qui consiste à regrouper les concepts les plus proches les uns des autres

La méthode de subsomption est en l'occurrence adaptée à notre solution, de par sa performance. Les vecteurs générés lors de la vectorisation sont utilisés en entrée du processus de hiérarchisation. La sortie du processus est la hiérarchie de labels, composée des relations de subsomptions entre les labels.

La hiérarchisation détermine les labels (classes) et les relations hiérarchiques entre les labels. Les labels sont un sous-ensemble des termes : les termes les plus pertinents sont désignés comme labels, selon l'approche décrite dans [Feldman et al., 1998]. Cette méthode exploite se base sur les vecteurs de fréquence calculés lors de la vectorisation. Pour chaque terme $term_j$, la proportion $PC(term_j)$ d'items de C où apparaît $term_j$ est définie par :

$$PC(term_j) = \frac{dfC(term_j)}{|C|}$$

où $dfC(term_j)$ est la fréquence de $term_j$ dans l'ensemble du corpus C, et $|C|$ est la taille du corpus C. Pour l'ensemble des termes dont la proportion $PC(term_j)$ est supérieure à un seuil (lT), i.e. $PC(term_j) \geq lT$, où $term_j \in Term$, l'ensemble ω_{lT} est défini tel que :

$$\omega_{lT} = \{term_j \in Term | PC(term_j) \geq lTo\}$$

Les termes appartenant à cet ensemble sont classés comme Label dans l'ontologie, tel que $Label \sqsubseteq Term$. Pour construire les relations hiérarchiques entre les labels, nous adoptons la méthode de subsomption pour sa performance.La matrice de cooccurrence est utilisée pour construire les relations hiérarchiques, en suivant l'approche de [Sanderson et al., 1999]. Soit deux labels x et y, x subsume y, i.e. $x \prec y$ si :

$$(PC(x|y) = 1) \wedge (PC(y|x) < 1)$$

où :

- $PC(x|y)$ est la proportion conditionnelle, i.e. le nombre d'items de C communs à x et y, en fonction du nombre d'items où y apparaît tel que :

$$PC(x|y) = \frac{x \cap y}{y} = \frac{cfm(x,y)}{df_y}$$

- PC(y|x) est la proportion conditionnelle inverse, i.e. le nombre d'items de C communs à x et y, en fonction du nombre d'items où x apparaît :

$$PC(y|x) = \frac{x \cap y}{x} = \frac{cfm(x,y)}{df_x}$$

Dans Sanderson et al. (1999) les auteurs remarquent que beaucoup de labels ne sont pas inclus par les conditions définies ci-dessus, et proposent d'assouplir la première condition $PC(x|y) = 1$ pour les inclure. La subsomption est alors redéfinie par :

$$(PC(x|y) \geq st1 \wedge (PC(y|x) < 1)$$

où $st1 \in [0, 1]$ est le seuil de subsomption pour la première condition. De Knijff et al.(2013) propose également d'assouplir la seconde condition $PC(y|x) < 1$.

La définition de la subsomption est alors :

$$x \prec y = (PC(x|y) \geq st_1) \wedge (PC(y|x) < st_2)$$

où $st_1 \in [0, 1]$ est le seuil pour la première condition, et $st_2 \in [0, 1]$ le seuil pour la seconde condition tel que $st1 \geq st2$. De façon verbeuse, si x apparaît avec une proportion supérieure à st_1 dans les documents où y apparaît, et y apparaît avec une proportion supérieure à st_2 dans les documents où y apparaît, alors x subsume y, i.e. $x \prec y$. L'application de cette méthode à l'ensemble des labels définie un graphe acyclique dirigé, i.e. la hiérarchie de labels qui est ensuite intégrée à l'ontologie noyau.

7.2.4/ CONCLUSION

Cette section a décrit l'approche d'apprentissage d'une taxonomie dans un contexte Big Data, et son intégration au sein d'une ontologie pour la construction d'un modèle de classification hiérarchique multi-label. Les résultats ont démontré la faisabilité de l'approche, et les capacités d'extraction de caractéristiques pertinentes à partir de grands volumes de données. L'évaluation des temps de calcul a montré que le point critique du processus est la construction de la matrice de cooccurrence, dont le coût calculatoire est élevé. La montée en charge horizontale, i.e. l'augmentation des ressources matérielles par ajout du nombre de machines, permet cependant de palier cette limitation, comme indiqué dans la littérature [Lin, 2013]. En outre, des optimisations ont par la suite été apportées au processus (notamment l'ajout d'un combiner), et d'autres solutions ont été proposées pour optimiser cette étape, i.e. la réduction de la fenêtre de voisinage pour la détection des cooccurrences. Les limitations des méthodes d'extraction de labels et de construction de relations hiérarchiques pertinentes ont été soulignées.

7.3/ CLASSIFICATION BASÉE SUR UN MOTEUR DE RAISONNEMENT

Cette section présente les deux dernières étapes du processus SHMC : d'une part la méthode d'apprentissage de règles de classification (résolution), et d'autre part la méthode de classification par raisonnement sémantique (réalisation) du processus SHMC.

7.3.1/ RÉSOLUTION

La résolution crée les règles de l'ontologie utilisées pour la classification des items par rapport à la taxonomie, à partir de la matrice de cooccurrence de termes issue des données d'apprentissage. Le processus de création des règles utilise une méthode de seuils pour sélectionner les termes les plus pertinents pour chaque concept de la hiérarchie, et inclue ces termes dans la définition de la règle. Les règles sont transcrites au format SWRL (Semantic Web Rule Language), plutôt que de les traduire en tant que contraintes logiques de l'ontologie. Le principal intérêt des règles SWRL est de réduire la charge de calcul du raisonneur, donc d'améliorer la performance du système. Un ensemble conséquent de règles SWRL simples (i.e. courtes et peu complexes) est généré lors de cette phase afin de classer les items.

Definition 44:

Raisonneur sémantique. Un raisonneur sémantique, ou raisonneur, est un logiciel qui implémente des algorithmes de raisonnement déductif, selon une logique de description donnée. Les raisonneurs et les moteurs d'inférence répondent aux mêmes principes. Les raisonneurs fournissent en revanche un plus grand nombre de mécanismes. Ils font partie intégrante des technologies du web sémantique, et respectent les standard définis par le W3C[a]. Les raisonneurs sont notamment intégrés dans différents logiciels de gestion de bases de données de triplets (triple-stores), et permettent d'inférer de nouvelles connaissances à partir d'ontologies existantes contenues dans la base.

a. Le World Wide Web Consortium, ou W3C, est un organisme de standardisation à but non lucratif, chargé de promouvoir la compatibilité des technologies du World Wide Web. https://www.w3.org/

Definition 45:

Triple Store. Un triple store est une base de données conçue pour le stockage et la récupération de données RDF[a]. Un triplestore ne stocke des triplets de type <Sujet,Prédicat,Objet>, tels que définis dans une logique de description. La base de données représente ainsi un graphe RDF formé de l'ensemble des triplets. Le triple store permet par la suite de requêter le graphe RDF de la même manière qu'une base de données relationnelle. Le langage standard de requête de données RDF est SPARQL[b].

a. Ressource Description Framework, https://www.w3.org/RDF/
b. https://www.w3.org/TR/rdf-sparql-query/

La résolution génère en partie le modèle prédictif, en déterminant pour chaque label un ensemble de règles de classification, qui constituent les conditions pour qu'un label $label_1$ soit attribué à un item $item_1$. Les règles sont générées à partir de la matrice de cooccurrence définie lors de la vectorisation. Chaque règle est composée d'un ensemble de termes nécessaires pour qu'un item soit classé avec le label correspondant. Des relations hasAlpha et hasBeta définissent les règles de classification pour chaque label. Deux seuils sont définis pour créer ces relations :

- Un seuil alpha (α) tel que $\alpha < P_C(term_i|term_j)$, où $term_i \in Label$ et $term_j \in Term$.
- Un seuil beta (β) tel que $\beta \leq P_C(term_i|term_j) \leq \alpha$, où $term_i \in Label$ et $term_j \in Term$.

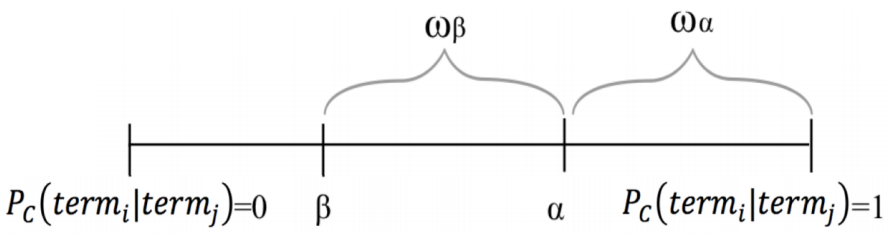

Figure 7.3 – Ensembles Alpha et Beta

Ces deux seuils sont définis par l'utilisateur dans l'intervalle [0, 1] avant la création des règles. En fonction des seuils, deux ensembles de termes sont identifiés (figure 7.3) : —

L'ensemble Alpha ($\omega_{\alpha}^{(term_i)}$) est défini pour chaque label par :

$$\omega_{\alpha}^{(term_i)} = \{term_j | \forall term_j \in Term : P_C(term_i|term_j) > \alpha\}$$

i.e. l'ensemble des termes $term_j$ qui co-occurrent avec $term_i \in Label$, et dont la proportion conditionelle est supérieure à α. — L'ensemble Beta ($\omega_{\beta}^{(term_i)}$) est défini pour chaque label par :

$$\omega_{\beta}^{(term_i)} = term_j | \forall term_j \in Term : \beta \leq P_C(term_i|term_j) \leq \alpha$$

i.e. l'ensemble des termes $term_j$ qui co-occurrent avec $term_i \in Label$ et dont la proportion conditionelle est comprise entre les seuils β et α. Pour chaque terme $term_j$ compris dans $\omega_{\beta}^{(term_i)}$, une relation hasAlpha est créée entre $term_i$ et $term_j$. De la même façon, pour chaque terme $term_j$ compris dans $\omega_{\beta}^{(term_i)}$, une relation hasBeta entre $term_i$ et $term_j$ est créée. En fonction de la cardinalité des ensembles Alpha et Beta déterminés pour chaque item, quatre cas sont possibles pour catégoriser un item :

1. Beta Vide : $|\omega_{\alpha}^{(term_i)}| > 0 \wedge |\omega_{\beta}^{(term_i)}| = 0$
2. Alpha vide : $|\omega_{\alpha}^{(term_i)}| = 0 \wedge |\omega_{\beta}^{(Label_i)}| > 0$
3. Alpha et Beta non vides : $|\omega_{\alpha}^{(term_i)}| > 0 \wedge |\omega_{\beta}^{(term_i)}| > 0$
4. Alpha et Beta vides : $|\omega_{\alpha}^{(term_i)}| = 0 \wedge |\omega_{\beta}^{(term_i)}| = 0$

Pour les trois premières catégories, la création des règles est définie ci-dessous (la quatrième catégorie (4) ne génère aucune règle).

Dans le cas Beta Vide (1), seul l'ensemble ω_{α} est utilisé pour la création de règles. La condition de classification est alors :

$$\forall label \forall item \exists term : hasTerm(item, term) \wedge term \in \omega_{\alpha}^{(label)} \rightarrow isClassified(item, label)$$

i.e. si l'item possède au moins un terme de $\omega_{\alpha}^{(term_i)}$ il est classé avec le terme $term_i$, où $term_i \in Label$. Une règle est créée pour chaque label unique respectant cette condition, puis sérialisée au format SWRL (voir section suivante).

Dans le cas Alpha vide (2), seul l'ensemble ω_{β} est utilisé. Les items sont classés avec un label via la condition suivante :

$$\forall label \forall item : |\{\forall term : hasTerm(item, term) \wedge term \in \omega_{\beta}^{label}\}| \delta \rightarrow isClassified(item, label)$$

i.e. si l'item possède un nombre de termes de $\omega_{\beta}^{(term_i)}$ supérieur ou égal à δ, il est classé avec $term_i$, où $term_i \in Label$. L'ensemble des combinaisons de termes possibles est alors considéré pour la création des règles SWRL : une règle est créée pour chaque combinaison de termes $term_j \in \omega_{\beta}^{(term_i)}$ où le nombre de termes combinés est $\delta = \lceil |\omega_{\beta}^{(term_i)}| p \rceil$, et $0 \leq p \leq 0.5$. L'ensemble des règles Beta est la combinaison C_n^m de m termes d'un ensemble de n éléments. Dans notre approche, n est le nombre de termes possibles $|\omega_{\beta}^{(term_i)}|$, et m le nombre minimum de termes δ qui composent chaque règle (par exemple, $C_{20}^{10} = 184756$). Pour limiter le nombre total de règles générées, on fixe $n \leq 10$. Les termes sont sélectionnés par classement dans l'ensemble $\omega_{\beta}^{(term_i)}$, en utilisant la proportion conditionnelle $P_C(term_j|term_i)$ comme fonction de classement. On note que les règles décrites

par un nombre de termes supérieur à δ ne nécessitent pas d'être intégrées au modèle, car redondantes avec la combinaison de δ termes qui suffit à effectuer la classification. Dans le cas Alpha et Beta non vides (3), les deux ensembles de termes sont considérés pour la création de règles.Les règles alpha et beta alors déduites des deux ensembles sont définis de la même façon que dans les cas (1) et (2). Une modification est cependant faite dans le calcul des combinaisons de termes, car les règles beta sont par définition moins pertinentes que les règles alpha. On définit alors $q = p2$, tel que $\delta = \lceil |\omega_{\beta}^{(term_i)}| q \rceil$, avec $0 \leq q \leq 1$ et $q = p2$.

Enfin, le cas Alpha et Beta vide (4) où la cardinalité des ensembles est zéro n'induit pas de création de règles. L'ensemble des règles définies pour l'ensemble des labels sont sérialisées en SWRL et intégrées à la base de connaissances. La section 4.2 décrit comment les règles sont utilisées pour effectuer la classification.

7.3.2/ RÉALISATION

La réalisation effectue la classification de nouveaux items selon le modèle prédictif, généré à partir de toutes les étapes précédentes du processus SHMC. Le modèle final est composé de la hiérarchie de labels issue de la hiérarchisation, et des règles de classification issues de la résolution. Une ontologie regroupe l'ontologie noyau, et le modèle prédictif. Cette ontologie est intégrée dans un triple store pour effectuer la classification de nouveaux items au sein de l'ontologie.

La réalisation peuple l'ontologie avec les nouveaux items, et effectue leur classification en accord avec la taxonomie et les règles de classification. L'ontologie est d'abord peuplée avec les nouveaux items au niveau Abox (assertion). Les règles SWRL générées lors de la résolution sont utilisées afin d'associer des étiquettes aux items. Le moteur d'inférence utilise ensuite les règles et la hiérarchie pour inférer les concepts correspondants aux caractéristiques des items pour les classer. La classification des items est ainsi multi-étiquette et hiérarchique.

La réalisation est divisée en deux sous-étapes : le peuplement de l'ontologie en tant qu'instances de la classe Item (Abox), puis leur classification en fonction de leurs attributs (termes) et du modèle prédictif. Chaque item est d'abord décrit par l'ensemble des termes extraits de celui-ci lors de son indexation. Le processus d'indexation des nouveaux items est similaire à l'indexation des items des données d'entraînement. Chaque item est décrit par un ensemble de termes pertinents $\omega_{\gamma}^{(item_i)}$ tel que :

$$\omega_{\gamma}^{(item_i)} = \{term_j | \forall term_j \in Term \gamma < tfidf(item_i, term_j, C)\}$$

où γ est le seuil de pertinence,

$$\gamma < tfidf(item_i, term_j, C), term_j \in Term, item_i \in Item$$

La valeur de tfidf est calculée de la même façon que lors de la vectorisation. L'étape de classification attribue à chaque item les labels appropriées en fonction du modèle prédictif décrit par la hiérarchie et les règles de classification. Des implémentations de raisonneurs sémantiques basés sur des algorithmes d'inférence (tableau, résolution) tels que Pellet [Sirin et al., 2007], FaCT++ [Tsarkov et al., 2006], ou Hermit [Motik et al., 2009] permettent d'effectuer une inférence complète des axiomes implicites pour des

ontologies ayant une expressivité élevée, telle que OWL2 SROIQ(D). Les travaux de David Werner ont montré que dans un contexte réel ou le nombre de règles est élevé, ces moteurs ne permettent pas de passer l'échelle et de gérer un volume important de données. Notre approche consiste à utiliser le raisonnement sur des règles de Horn, dont l'expressivité est limitée mais permet un passage à l'échelle dans un contexte Big Data. Le raisonnement sur des règles applique de façon exhaustive un ensemble de règles à un ensemble de triplets, pour inférer les axiomes implicites [Urbani et al., 2011]. Dans le cas de la tâche de classification, ces axiomes sont les classifications (labels) des items représentés par un ensemble d'attributs (triplets). Le moteur d'inférence infère les labels les plus spécifiques de la hiérarchie de labels pour chaque nouvel item. Les règles de classification et la hiérarchie de subsomption sont utilisées par le raisonneur pour cette tâche. Le résultat est une classification hiérarchique multi-label des items. Pour inférer l'intégralité des labels en fonction de la hiérarchie de labels, la règle SWRL suivante est ajoutée à l'ontologie :

$$Item(?item), Label(?labelA), Label(?labelB), broader(?labelA, ?labelB),$$

$$isClassified(?item, ?labelA) \rightarrow isClassified(?item, ?labelB)$$

Les règles de classification peuvent ensuite être appliquées selon deux types d'inférence, i.e. le chaînage avant (matérialisation) ou chaînage arrière. A ces deux types d'inférence correspondent deux stratégies de classification : la classification avant la requête et la classification au moment de la requête. La classification avant la requête est effectuée par un moteur d'inférence qui détermine par chaînage avant l'inférence de toutes les assertions implicites[1] de la base de connaissance (matérialisation). Les règles d'inférence sont appliquées sur la totalité de l'ontologie, jusqu'à ce que toutes les assertions soient matérialisées dans la base. Une fois la matérialisation terminée, les requêtes effectuées sur la base permettent de retrouver les classifications rapidement. Cette méthode est peu adaptée dans un système où l'ontologie change fréquemment (le raisonnement doit alors être ré-exécuté entièrement). La classification au moment de la requête est effectuée par le moteur d'inférence par chaînage arrière, qui applique les règles uniquement sur les données concernées par la requête. Contrairement au chaînage avant, le raisonneur est activé par la requête. Le moteur applique les règles uniquement sur les axiomes nécessaires pour répondre à la requête et déterminer de la classification. L'inconvénient est que les requêtes demanderont un temps d'exécution relativement long, indépendamment de leur fréquence ou de leur complexité. Appliquer les règles uniquement sur une partie des données, i.e. les items à classer, représente un avantage dans un contexte Big Data, en particulier gérer la vélocité des données. En revanche, le désavantage est la nécessité d'activer le moteur d'inférence à chaque requête, qui est impacté par le volume et l'expressivité des règles. Ces deux stratégies d'inférence sont compatibles avec le processus SHMC, cependant il faut adopter l'une ou l'autre en fonction du cas d'utilisation.

7.3.3/ CONCLUSION

Cette section a présenté une approche d'apprentissage de règles de classification à partir dans un contexte Big Data, l'intégration de ces règles dans un modèle prédictif au sein d'une ontologie, et l'utilisation du modèle prédictif pour la tâche de classification hiérarchique multi-étiquette. Le processus est décrit de façon exhaustive, et a été évalué

1. i.e. le calcul de la fermeture de l'ontologie, ou "ontology closure"

lors de la thèse pour la tâche de classification. Une première évaluation a montré la faisabilité de l'approche, qui pallie notamment les limitations techniques des raisonneurs DL identifiées par la thèse de David Werner, et permet la classification d'importants volumes de données en temps raisonnable. La performance du modèle prédictif généré par l'approche a ensuite été évaluée à la fois en contexte non-supervisé et en contexte supervisé, et comparé à des approches de la littérature dans le second cas. L'évaluation qualitative en contexte non supervisé basée sur la méthode de validation croisée a montré des performances encourageantes de l'approche proposée. L'évaluation permet en outre d'optimiser le paramétrage du processus en fonction de la performance du modèle généré, ce qui est essentiel pour l'optimisation du processus et l'amélioration continue du modèle. Suivant les critères sélectionnés, l'évaluation qualitative en contexte supervisé montre que la performance l'approche SHMC est comparable aux autres approches de l'état de l'art pour la tâche de classification multi-étiquette. Elle est plus performante que les autres approches selon certains critères d'évaluation pour le jeu de données Delicious. L'orientation de traitement massif de données de l'approche SHMC conserve des performances proches de l'état de l'art pour la tâche de classification supervisée, ce qui montre les avantages de l'approche.

7.4/ ARCHITECTURE SEMXDM POUR CRAWLER ET CLASSIFIER

L'architecture SEMXDM (Semantic Cross-Referencing Data Mining) comprend deux modules principaux : un crawler ciblé, ou module de recherche, qui permet le parcours du web et la recherche d'items, et un classifieur, ou module de classification, basé sur l'architecture SHMC, qui permet de déterminer la pertinence des items. Le modèle prédictif est décrit par une ontologie d'expressivité $\mathcal{ALCI}$ dont les axiomes en logique de description sont présentés dans le tableau 7.1 (cette ontologie est identique à l'ontologie de l'architecture SHMC). Le modèle prédictif effectue une classification hiérarchique multi-étiquette des items, en fonction des attributs (termes) extraits des items. Le modèle est composé d'une hiérarchie de labels, i.e. les concepts pertinents décrivant les items, et de règles de classification qui permettent d'associer les labels aux items. Les items du web indexés par le crawler sont intégrés dans l'ontologie en tant qu'individus (ABox), et la classification est effectuée par un raisonneur qui applique de façon exhaustive les règles de classifications du modèle prédictif aux items.

La figure 7.4 présente l'architecture globale et le rôle de chaque module. Les parties suivantes décrivent chaque module séparément.

1. **Le module de recommandation** répond aux requêtes de l'utilisateur, notamment les requêtes de croisement d'information sur le web, à partir desquelles une tâche de recherche est lancée.

2. La combinaison du **module de recherche**

3. et du **module de classification** est un crawler ciblé basé sur le modèle prédictif issu de l'architecture SHMC, qui effectue une classification sémantique des items. La matrice de cooccurrence et l'ontologie, décrite dans le tableau 7.1 sont des éléments déjà abordés dans le processus SHMC. L'architecture SemXDM comprend trois modules complémentaires, qui définissent la stratégie de recherche du crawler ciblé, et permettent d'adapter cette stratégie au cours du temps.

Concepts DL	Description
$Item \vee \exists hasTerm.Term$	associe un item à ses attributs (termes)
$Term \vee asString.String$	Termes extraits des items
$Label \vee Term$	Termes pertinents pour classer les items
$Label \vee .Label$	Relation de généralisation
$Label \vee .Label$	Relation de spécialisation
$broader \equiv narrower$	les relations Broader et Narrower sont inverses
$Item \cup Term \equiv \perp$	Items et Termes sont disjoints
$Label \vee \forall hasAlpha.Term$	Termes composant une règle Alpha
$Label \vee \forall hasBeta.Term$	Termes composant une règle Beta
$Item \vee \exists isClassified.Label$	Représente la classification d'un item

Table 7.1 – Concepts du modèle prédictif

4. le **module de maintenance** permet l'adaptation du modèle prédictif, à partir des nouvelles données récoltées. L'objectif de ce module est de répondre à la problématique de classification d'items issus du web, dont les propriétés évoluent au cours du temps.

5. Un **module de priorité** calcule la pertinence des items du web, en fonction de leur similarité à la requête de l'utilisateur. Le calcul de pertinence exploite le résultat de la classification des items d'une part, et d'autre part le graphe de liens issu du module de recherche.

7.4.1/ MODULE DE RECOMMANDATION

Le module de recommandation est le point d'interaction entre l'utilisateur (expert du domaine) et le système de recherche. Le module répond aux requêtes des utilisateurs, et renvoie les items les plus pertinents en fonction de la requête. Un système de recherche d'information peut considérer différents types de requête utilisateur. Tous les types de requête ne sont pas traitées par l'architecture SEMXDM. Nous considérons essentiellement deux types de requêtes, pour effectuer la tâche de croisement d'information : les requêtes de type "mot-clé" (Item queries), et les requêtes "item" ("full-document queries"). Deux types de requête considérées :

- Requête par termes (requête mot clé) : ce type de requête prend en entrée un ou plusieurs mots clés définis par l'utilisateur, et recherche dans l'index les items correspondant à la requête. Les mots clés sont normalisés par une chaîne de traitement syntaxique, afin de faire correspondre les mots clés aux termes de l'index sur lesquels les items sont indexés. La recherche est effectuée hors-ligne, i.e. les items retournés sont déjà intégrés dans l'index et la base de connaissances. La requête est alors composée d'un ensemble de termes $\omega_{term_i} = (term_1, term_2, ..., term_n)$. La réponse à la requête est un ensemble d'items $\omega_{term_j} = (item_1, item_2, ..., item_n)$ où $\forall j, \exists item_j hasTerm(term_i)$, et $term_i \in \omega_{term_j}$. Un classement des items de ω_{term_j} retournés est effectué par ordre décroissant selon la similarité cosinus entre chaque vecteur de termes et la requête initiale.

- Requête-item : ce type de requête prend en entrée un item déjà présent dans le système de recherche d'information, et classé avec des labels du modèle prédictif.

La requête prend alors la forme d'un vecteur de termes correspondant à l'item initial. La requête est envoyée au module de recherche (crawler), qui a pour objectif de chercher des items similaires sur le web. Le calcul de similarité entre les nouveaux items crawlés est similaire au calcul hors ligne. La similarité cosinus entre les vecteurs de termes de chaque item est définie comme score de pertinence. Le calcul du score potentiel des liens sortants de chaque item est décrit dans la section 7.4.5

7.4.2/ MODULE DE RECHERCHE

Ce module consiste en un crawler web, qui permet une recherche performante d'un grand volume d'items du web. Le module de recherche est le module central de l'architecture, qui connecte l'ensemble des modules. Une recherche d'items est déclenchée lorsqu'une requête utilisateur de croisement d'information est reçue depuis le module de recommandation. La recherche est divisée en cycles. A chaque cycle, un nombre fixe d'items de la frontière du crawl est fetché. La recherche s'arrête après un nombre pré-déterminé de cycles, ou après un nombre total d'items fetchés Les items fetchés sont indexés en fonction de leur contenu (termes). Un index inversé des termes est construit à partir des items. Dans le cas du crawler ciblé, l'index est mis à jour avec les nouveaux items crawlés. L'indexation des items est effectuée après chaque cycle du crawl. L'index permet d'accéder aux fréquences de chaque terme, nécessaires notamment pour l'évolution du modèle prédictif effectuée par le module de maintenance. L'index inversé est composé d'un ensemble de vecteurs d'items, i.e. un vecteur pour chaque terme $term_i \in Term$, de la forme : $\omega_{term} < item_1, item_2, ..., item_n >$. Les termes sont extraits de chaque item. Pour chaque item, un processus d'extraction des termes discrimine les termes non pertinents (bruits). Le résultat de l'extraction est un vecteur de termes $\omega_{item_i} < term_1, term_2, ..., term_n >$.

7.4.3/ MODULE DE CLASSIFICATION

L'objectif du module de classification est de déterminer la pertinence des items crawlés, en s'appuyant sur la méthode de classification hiérarchique multi-étiquette développée dans les chapitres précédents pour définir la pertinence des items selon les termes les plus pertinents du modèle de classification (labels). Dans cette approche, une matrice de cooccurrence de termes est créée à partir d'un corpus comprenant un grand nombre d'items. Les termes les plus pertinents sont d'abord identifiés en temps que Label dans l'ontologie. Puis, à partir de la matrice de cooccurrence, des règles de classification pour chaque label sont générées, ainsi que des relations hiérarchiques entre les labels. Le module de classification reprend le principe de la dernière étape du processus SHMC, la réalisation, qui effectue la classification des items dans l'ontologie par un raisonneur. Le résultat de la classification est un ensemble de relations de classifications entre un item et des labels, i.e. $isClassified(item, label)$. Dans l'approche SEMXDM, un modèle de classification pré-établi est utilisé à l'initialisation du processus. L'architecture SHMC permet de générer ce modèle et de l'intégrer à une base de connaissances.

La méthode de réalisation effectue la classification par un moteur d'inférence selon le modèle, pour classer les nouveaux items fetchés. Elle est composée de deux sous-étapes :

7.4.3.1/ PEUPLEMENT

Lors du peuplement, chaque item est d'abord décrit par l'ensemble des termes extraits de celui-ci lors de son indexation. L'ontologie est peuplée avec les nouveaux items en tant qu'individus (Abox), selon les axiomes du modèle (tableau 7.1). Chaque item est décrit par un ensemble de termes pertinents $\omega_{\gamma}^{item_i}$ tel que : $\omega_{\gamma}^{item_i} = term_j | \forall term_j \in Term \wedge \gamma < tfidf(item_i, term_j, C)$ où ω est le seuil de pertinence,$\omega < tfidf(item_i, term_j, C), term_j \in Term, item_i \in Item$. La valeur de $tfidf$ est la fréquence inverse des termes dans l'ensemble des items fetchés, et est calculée de la même façon que lors de la vectorisation.

7.4.3.2/ CLASSIFICATION

Un moteur d'inférence est ensuite utilisé pour effectuer la classification hiérarchique multiétiquette des items, en se basant sur le modèle prédictif i.e. la hiérarchie de labels et les règles de classification. L'étape de classification attribue à chaque item les labels appropriées en fonction du modèle prédictif décrit par la hiérarchie et les règles de classification. Pour inférer les labels à partir d'un ensemble de termes, le processus de création de règles (Résolution) de l'approche SHMC est utilisé. Les règles sont générées à partir d'un jeu de données d'entraînement antérieurement au crawl. Ces règles définissent les conditions minimales pour qu'un individu $item_i \in Item$ de l'ontologie soit classé avec un individu $label_j \in Label$. Deux types de règles sont intégrées dans l'ontologie : Les règles de type Alpha sont basées sur l'appariement entre un terme unique et un label, tel que l'apparition du terme suffise pour déduire une classification. Les règles de type Beta sont basées sur une combinaison de plusieurs termes pertinents pour effectuer la classification. Pour inférer l'intégralité des labels en fonction de la hiérarchie de labels, la règle SWRL suivante est ajoutée à l'ontologie afin de prendre en compte les relations hiérarchiques entre les labels :

$$Item(?item), Label(?labelA), Label(?labelB), broader(?labelA, ?labelB),$$

$$isClassified(?item, ?labelA) \rightarrow isClassified(?item, ?labelB)$$

Comme décrit précédemment, les règles de classification peuvent être appliquées par chaînage avant ou chaînage arrière, ce qui induit respectivement une classification avant la requête ou au moment de la requête. Dans le cas de l'architecture SEMXDM, la classification au moment de la requête est adaptée car la classification dépend de l'évolution du modèle au cours du temps, et n'est donc pas permanente. Le moteur d'inférence applique donc les règles par chaînage arrière lors du crawl.

Le résultat de la classification est un vecteur $v_{item_i}^{label}$, où chaque dimension est un label (classe Label de l'ontologie noyau). Le vecteur de labels est utilisé par le module priorité pour déterminer le potentiel de pertinence des liens extraits de l'item. Après qu'un item soit classé avec succès, sa similarité avec la requête initiale peut être calculée. Dans la littérature, les crawlers ciblés standards ne parviennent pas à s'adapter à un environnement non contrôlé tel que le web au cours du temps [Dong et al., 2014]. D'après [Dong et al., 2014], le modèle prédictif sur lequel s'appuie le crawler ciblé doit pouvoir s'adapter aux nouvelles données rencontrées, et aux changements induits dans les caractéristiques des items. Aussi, les crawlers devraient apprendre par expérience comment optimiser les temps de recherche des items pertinents, en se basant sur les chemins (liens hypertexte) connus menant à des items pertinents, [Diligenti et al., 2000]. Pour répondre à

ces problématiques, l'architecture proposée se base sur deux modules supplémentaires décrits dans les sections suivantes. Un premier module de Maintenance a pour objectif l'adaptation du modèle prédictif en fonction de nouvelles données rencontrées. Le second module est le module de priorité, qui a pour objectif de déduire à partir des items déjà fetchés les chemins menant potentiellement vers des items pertinents. Ce dernier modèle calcule dans un premier temps la pertinence des items fetchés par rapport à la requête initiale. Dans un second temps, le module réordonne la frontière du crawl, en adaptant la priorité des items de la frontière en fonction de la pertinence de leurs parents dans le graphe web.

7.4.4/ Module de Maintenance

Dans un environnement web peu structuré et non contrôlé, le contenu est dynamique et en constante évolution. Les données sont considérées comme non stationnaires, les propriétés des données issues du web évoluent au cours du temps. Ces changements représentent une difficulté pour produire un modèle prédictif représentatif des données, car ils induisent des changements radicaux dans les classes cibles (targets concepts). Cette problématique est nommée dérive conceptuelle (concept drift) [Widmer et al., 1996]. Dans ce cas de figure, les performances d'un modèle prédictif sont impactées négativement, car les données d'entraînement utilisées pour le construire ne correspondent plus aux données réelles. La classification des données réelles est alors dégradée. Nous proposons de palier ces difficultés par l'intégration du module d'évolution dans l'architecture. Ce module est basé sur l'approche d'apprentissage adaptatif [Peixoto et al., 2015], qui apporte constamment des modifications au modèle prédictif de classification représenté par une ontologie. Le processus permet d'adapter le modèle prédictif en fonction d'un flux non stationnaire de données non structurées. [Peixoto et al., 2015] a montré que les performances pour la tâche de classification de données inconnues sont améliorées lorsque le modèle prédictif est mis à jour par le processus d'apprentissage adaptatif. Les données émises par le crawler sont utilisées comme entrée du module d'évolution. Le processus d'apprentissage adaptatif se focalise sur l'adaptation du modèle prédictif décrit par l'ontologie, en fonction du flux de données textuelles issu du module de crawling. Trois caractéristiques de gestion de flux de données continues sont considérées par le processus :

- Évolution des attributs (feature-evolution), qui correspond à l'apparition de nouveaux termes (features), i.e. des termes jusqu'alors absents de l'index inversé des items. L'évolution des attributs concerne aussi la disparition de termes qui ne sont plus représentatifs des données, lorsque leur fréquence diminue.
- Évolution des concepts (concept-evolution), qui correspond à l'addition ou la suppression des attributs de classes, i.e. es labels dans notre approche.
- Dérive des concepts (concept-drift) : le flux constant de données apporte des changements dans les propriétés statistiques relatives des concepts et des attributs, ce qui induit des modifications du modèle prédictif (règles de classification et hiérarchie de labels).

L'approche d'apprentissage adaptatif de [Peixoto et al., 2015] répercute l'impact des trois caractéristiques ci-dessus dans le modèle prédictif. Le processus est basé sur une adaptation incrémentale du modèle de classification, [Hulten et al., 2001, Aggarwal et al., 2006]

Requête de modification	Description
AddTerm	Ajoute un Term au modèle de classification
AddLabel	Ajoute un Label au modèle de classification
DeleteLabel	Supprime un Label du modèle de classification
AddHRelation	Ajoute une relation hiérarchique entre deux labels
DeleteHRelation	Supprime une relation hiérarchique entre deux labels
AddAlphaTerm	Ajoute une relation Alpha entre un label et un terme
DeleteAlphaTerm	Supprime une relation Alpha entre un label et un terme
AddBetaTerm	Ajoute une relation Beta entre un label et un terme
DeleteBetaTerm	Supprime une relation Beta entre un label et un terme

Table 7.2 – Types de requêtes de modification du modèle

selon les spécificités de l'ontologie noyau. L'architecture de module de maintenance n'est présentée que de façon superficielle dans ce manuscrit. Une description exhaustive de l'approche peut être trouvée dans [Peixoto et al., 2015]. Le module de maintenance. Le processus est divisé en 4 étapes :

1. détection des changements depuis le flux de données,
2. dérivation des requêtes de modification à partir des changements détectés,
3. traduction des requêtes de modifications en modifications appliqués au modèle, et
4. application des modification (évolution du modèle)

La dernière étape considère la résolution des inconsistances résultant de l'application des modifications selon 4 sous-étapes : application des changements, vérification de la consistance, résolution des inconsistances et application définitive (commit) des modifications.

Les changements de propriétés statistiques, sont nommées Input Changes dans le processus d'apprentissage. Les différents types de changements sont capturés par des capteurs (change sensors) dédiés. Le tableau 7.2 décrit chaque type de requête de modification qui affectent le modèle de classification. Les changements détectés et validés en entrée, ou "Input Changes", génèrent des "Requêtes de Modification" au cours du crawl pour adapter le modèle de classification. Chaque entrée de cooccurrence implique une mise à jour de la fréquence document du terme, ce qui impacte les proportions relatives aux termes cooccurrents. Les requêtes sont propagées en tant que "Modifications du modèle", i.e. les mises à jour de l'ontologie. Les mises à jour du modèle peuvent être effectuées après chaque cycle de crawl ou en parallèle du crawl, suivant le cas d'utilisation. Dans le second cas, le résultat de la classification d'un item est impacté après chaque changement du modèle de classification, issu d'autres items. Comme décrit dans la section précédente, l'architecture SemXDM effectue la classification des nouveaux items au moment de la requête. L'évolution de modèle est donc prise en compte de façon transparente, sans interruption du processus.

7.4.5/ MODULE DE PRIORITÉ

L'objectif du module de priorité est de combiner les informations issues du graphe web et le contenu des items pour améliorer les décisions du crawler, i.e. pour redéfinir les

priorités des items de la frontière de crawl. En opposition au module de classification qui calcule le score des items fetchés, le module de priorité calcule un score potentiel des items de la frontière (items non fetchés). Pour chaque item $item_i$, le module de priorité peuple l'ontologie avec les liens extraits des items en tant qu'instances (Abox).

Les liens sont mis à jour dans la base de connaissances après chaque cycle de crawl. Les liens sortants (outlinks) sont mis à jour en fonction des liens extraits des items lors du crawl. Les liens entrants (inlinks) sont la relation opposé des liens entrants et sont mis à jour en conséquence. L'ajout de liens entrants ne fait pas appel à un processus externe tel qu'une requête à un moteur de recherche [2], par conséquent les liens entrants sont incomplets. Le résultat du peuplement de l'ontologie est l'ensemble de propriétés $hasInlink : \omega_{in}^{item_i} = |item_1, item_2, ..., item_n|$ et $hasOutlink : \omega_{out}^{item_i} = |item_1, item_2, ..., item_m|$, où n est le nombre total de liens entrants, et m le nombre total de liens sortants extraits des items. Pour identifier les chemins menant potentiellement vers des items pertinents, l'approche graphe contextuel de [Diligenti et al., 2000] est utilisée. Dans cette approche, les items pertinents sont d'abord identifiés. A partir de ces items, les liens entrants de ces items sont extraits de façon récursive. L'ensemble des liens forme un sous-graphe, découpé en couches. Chaque couche correspond à une distance à un item pertinent. Après chaque cycle de crawl, les couches du graphe contextuel sont recalculées à partir de l'ensemble des items pertinents à jour. Un seuil de pertinence défini par l'utilisateur permet de discriminer les pages non pertinentes. L'ensemble des items pertinents $\omega_{rel}^{item_i}$ est défini par : $\omega_{rel}^{item_i} = item_i | \forall item_i \in Item : Relevance_i > \theta$ où $Relevance_i$ est le score de pertinence de l'item. Lors du premier cycle de crawl, l'ensemble d'items pertinents est composé des seed (pages initiales) fournies par l'utilisateur. A partir de l'ensemble $\omega_{rel}^{item_i}$, chaque couche L_j du graphe contextuel est définie par l'ensemble des items à une distance de j liens entrants d'un $item_i \in \omega_{rel}^{item_i} : L_j = \omega^{item_k} = \{itemk | \forall item_k \in Item : \exists path(item_k, item_i)\}$ où $path(item_k, item_i)$ est le plus court chemin dans la base de connaissances entre $item_i$ et $item_k$, composé de j relations $hasInlink(|path| = j)$.

Chaque couche du graphe contextuel est mise à jour avec l'ensemble des termes contenus dans les items qui la composent : L_j est définie par un couple $(term_k, valeur_k)$ tel que :

- $\forall j, item_j \in L_j$
- $\forall k, term_k \in Term$
- $\forall_k, \exists item_j hasTerm(label_k)$
- $\forall j, valeur_k = |item_j|, item_j hasTerm(term_k)$

Un vecteur pondéré v_i^{term} est créé pour chaque couche L_i, où chaque dimension du vecteur est un terme. Chaque vecteur v_i^{term} est mis à jour après chaque phase de crawl avec les nouveaux termes et liens découverts. Un second vecteur $v_{item_i}^{label}$ est généré, où chaque dimension est un Label dans l'ontologie. La probabilité qu'un item appartienne à une couche L_i est ensuite approximée par calcul de la distance entre les vecteurs de termes/labels d'une couche du graphe et du nouvel item.

$$cos(v_{L_i}^{term}, v_{item_j}^{term}) = \frac{v_{L_i}^{term} \cdot v_{item_j}^{term}}{||v_{L_i}^{term}|| \cdot ||v_{item_j}^{term}||}$$

2. Certains moteurs de recherche peuvent être requêtés pour retrouver l'ensemble des liens entrant d'une page

Le calcul de similarité entre les deux types de vecteurs est identique. Les deux types de vecteurs sont ensuite utilisés dans le calcul du degré d'appartenance de chaque couche du graphe. On définit le degré d'appartenance d_i pour chaque couche du graphe comme la moyenne arithmétique deux deux vecteurs :

$$d_i = \frac{cos(v_{L_i}^{term} v_{item_j}^{term}) + cos(v_{L_i}^{label} v_{item_j}^{label})}{2}$$

Le score de pertinence final est définit ci-dessous. Une heuristique supplémentaire est employée pour palier les limitations du modèle de classification, qui ne peut permettre de décrire avec précision l'ensemble des items dans un environnement web[3] :

$$r_{item_j} = max(cos(v_{L_i}^{term} v_{item_j}^{term}), d_i)$$

où $r_{item_j} \in [0, 1]$.

Cette heuristique permet de conserver une priorité importante pour les items dont seule la distance des vecteurs de termes est élevée. Le score de pertinence résultant, compris dans l'intervalle [0, 1] est utilisé pour estimer la distance, en nombre de liens, entre un nouvel item et un item pertinent (couche L0). Les priorités de la frontière de crawl sont mises à jour après chaque crawl en fonction de cette distance. Plus une page est proche de la couche L_0, plus sa priorité est élevée (une priorité fixe est attribuée à chaque couche du graphe). Les liens qui ne correspondent à aucune couche ont une priorité minimale dans la frontière. Après chaque cycle de crawl, le graphe contextuel est mis à jour : la première couche du graphe est composée de l'ensemble des items dont la pertinence $\omega_{L_0}^{rel}$ dépasse un seuil Θ définit par l'utilisateur, i.e. tous les items dont le score de pertinence avec la première couche du graphe est supérieure au seuil : $r_{item_i} > \Theta$. Pour le premier cycle, la couche L_0 du graphe est uniquement composée du vecteur de termes issu de la requête initiale. Les autres couches du graphe sont modifiées de façon dynamique, en fonction du graphe de liens mis à jour.

7.4.6/ CONCLUSION

Ce chapitre a décrit une architecture de recherche d'information basée sur un crawler ciblé, et tire partie de l'exploitation d'un modèle de classification sémantique pour décrire les items du web, dont la construction est basée sur l'approche SHMC. Afin de répondre aux limitations de l'état de l'art, l'approche considère premièrement l'adaptation du modèle de classification dans le temps; et deuxièmement, une nouvelle méthode de définition de la priorité dans la frontière du crawl basée sur une approche de graphe contextuelle. Une évaluation qualitative montre que notre approche semble être performante pour discréditer rapidement les informations pertinentes au départ de la recherche, mais ne parvient pas à maintenir une similarité moyenne élevée pendant toute la recherche contrairement à l'approche Best-N-First ayant servi à la comparaison. Afin de maximiser la performance de l'approche durant toute la recherche, modifier la stratégie du crawler durant le crawl semble être pertinent. Le changement de stratégie permet de maximiser la performance du crawler tout au long de la recherche.

3. Cette problématique nommée "incomplete label assignment", est définie par le coût important de création d'un jeu de données complètement étiquetté, i.e. le jeu d'entraînement ne comprend pas tous les labels possibles. Cette problématique est d'autant plus importante dans un contexte Big Data, où le nombre d'items à étiqueter est important, rendant la création d'un modèle parfaitement représentatif des données quasiment impossible [Yu et al., 2014].

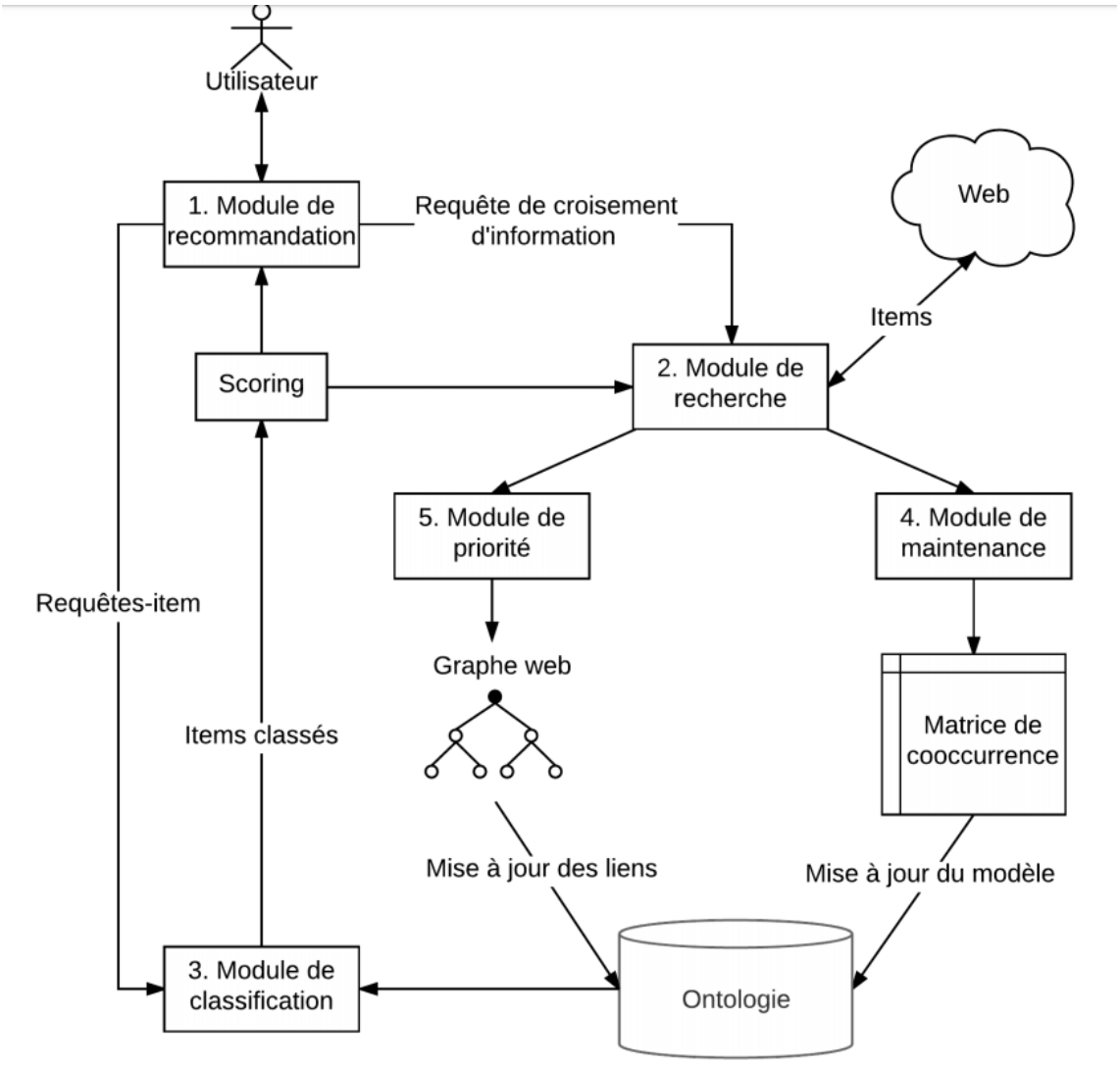

Figure 7.4 – Architecture SemXDM

8

L'HÉTÉROGÉNÉITÉ DES DONNÉES

Contributions :

David Werner dans le cadre de sa thèse sur la recommandation d'articles pour la veille économique (détails en section 15.5).

La variété autrement appelée hétérogénéité implique la prise en considération de données issues de sources variées et de forme différente.
Ce verrou a été levé par la thèse de David Werner. L'architecture développée est adaptable à différents contextes métiers et domaines afin de fournir un système de recommandation sémantique, efficace et simple d'utilisation. Son adaptation au domaine de l'économie dans le contexte métier de l'entreprise partenaire, First ECO, a permis la commercialisation d'un nouveau produit hautement compétitif, FirstECO Pro'fil. Un des trois principaux éléments de l'architecture concerne la base de connaissances. L'ensemble du fonctionnement du système repose sur cette base répondant aux différentes contraintes d'intégration des données hétérogènes. Elle contient le vocabulaire d'indexation et forme une modélisation du domaine traité correspondant au contexte métier d'utilisation du système.

La modélisation de la base de connaissance doit en effet permettre la modélisation de domaines riches variés et au plus près de la vision métier des experts. Le système doit proposer un vocabulaire modélisant le domaine qui soit aisément accessible et utilisable par la machine comme par l'humain. Il doit permettre la maintenance ainsi que l'adaptation rapide et simple du vocabulaire. C'est pourquoi le modèle propose la description des items à l'aide de facettes de descriptions. Ce qui a pour avantage de permettre la gestion de domaines complexes tout en conservant des descriptions aisément accessibles pour un humain. Chaque facette étant composée d'une ressource terminologique de type langage documentaire (i.e. liste, taxonomie ou thésaurus).

Une seconde contribution consiste à proposer une approche pour l'automatisation du processus d'indexation. L'approche considère l'indexation comme une tâche d'indexation multi-label (ou multi-label hiérarchique dans le cas de vocabulaires d'indexation organisés de façon hiérarchique).

La troisième contribution de David Werber consiste à proposer un algorithme de

comparaison qui exploite pleinement la richesse des descriptions d'items permise par le modèle sur lequel repose la base de connaissances. En effet, cette modélisation permet de décrire différents aspects (i.e. facettes) d'un item. Les vocabulaires utilisés pour la définition des facettes peuvent être organisés sous la forme de simples listes, mais aussi structurés sous la forme de taxonomies ou de thésaurus. Les items peuvent ainsi être décrits sur chacune des facettes à l'aide de termes plus ou moins précis. Les algorithmes de comparaisons classiques déduisent directement la pertinence de la précision et ne peuvent donc pas prendre en compte la différence de précision qu'il peut y avoir entre l'expression du besoin de celle de l'offre d'information. Notre algorithme comble un manque de l'état de l'art en ce qui concerne la prise en compte du degré de précision de la description des items. Bien que l'architecture ait été mise en place dans le contexte de la recommandation d'articles de synthèse d'informations économiques régionales, elle a été pensée pour être évolutive et adaptable à d'autres domaines.

Les hétérogénéités structurelles, syntaxiques et sémantiques sont liées à l'existence de différents types de ressources terminologiques prenant en compte différentes relations sémantiques entre les termes qui les composent. Ces différents types répondent à différentes normes ou standards définissant le format de représentation de l'information (i.e. sa syntaxe), ainsi que son organisation (i.e. sa structure) et son interprétation (i.e. sémantique associée à sa syntaxe).

Nous distinguons trois types de structures concernant les ressources terminologiques à destination des humains : les listes plates, les taxonomies (i.e. classifications et nomenclatures) et les thésaurus. Par rapport à la présentation faite dans les travaux de thèse de Thomas Hassan, je complète la définition des vocabulaires contrôlés.

Comme le présente la figure 8.1, les différents types de vocabulaires contrôlés sont donc une combinaison de termes avec ou sans relations hiérarchiques, associatives ou synonymiques. Dans cette figure, les relations associatives y sont représentées par des flèches en pointillés, les relations hiérarchiques par des flèches classiques, les termes principaux (i.e. termes choisis parmi leur groupe de synonymes pour représenter une notion) sont représentés par des ronds blancs, les termes synonymes sont représentés par des ronds pleins. Les relations de synonymie entre termes principaux et les termes synonymes sont représentées par des traits noirs.

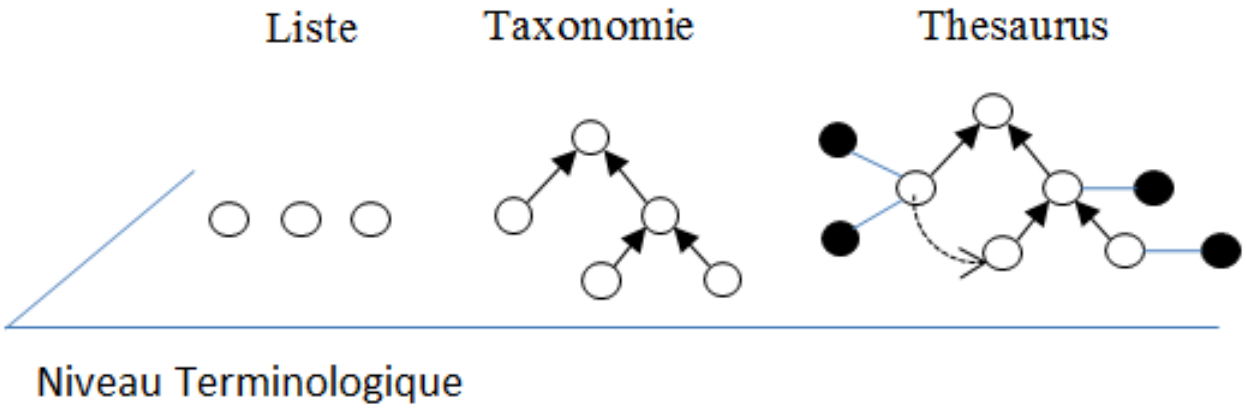

Figure 8.1 – Présentation des différents types de ressources et de leur structuration

8.1/ MODÈLES “*unificateur*” ET “*intégrateur*”

Les différentes formes d'hétérogénéités existantes entre les types de ressources terminologiques ont été montrées tout en soulignant leurs similitudes structurelles. Néanmoins, l'hétérogénéité des ressources terminologiques reste un problème lors de l'utilisation simultanée de plusieurs ressources pour la description d'un item. Bien que ces ressources soient hétérogènes, nous pouvons mettre en avant l'existence d'une structure sous-jacente commune, basée sur une combinaison de termes et de relations.

Nous présentons dans cette section, un modèle mathématique unificateur et formel basé sur des ensembles, pour la manipulation des ressources terminologiques en vue de leur intégration dans une base de connaissances. Ensuite nous présentons le modèle intégrateur de la base de connaissances, permettant la prise en compte de ressources terminologiques pour l'indexation d'items.

L'objectif du modèle unificateur est de traduire en fonction de leur type les ressources terminologiques disponibles de façon à ce qu'elles soient intégrables dans le modèle intégrateur. De cette manière, nous supprimons l'hétérogénéité syntaxique et structurelle. Le modèle intégrateur s'attache a résoudre l'hétérogénéité sémantique.

8.1.1/ MODÈLE “*unificateur*”

Le modèle unificateur proposé est un langage artificiel, basé sur des mathématiques ensemblistes qui possèdent une sémantique formelle. Ce modèle permet la description de l'ensemble des types de ressources terminologiques.

Par conséquent cette section présente une modélisation mathématique pour chaque type de ressources distinguées afin de proposer une modélisation unique correspondante pour n'importe quel type de ressource terminologique. Nous distinguons trois types : les listes, les taxonomies (i.e. classification, nomenclatures) et les thésaurus.

Definition 46:

Listes plates.

$$l \in L = \{ t \in T \mid l \, R_L \, t \}$$

$$R_L \subseteq L \times T \subseteq \mathcal{P}(T)$$

Soit L l'ensemble des listes plates de termes. Soit l une liste plate composée d'un ensemble de termes t appartenant à l'ensemble de tous les termes T. Soit R_L la relation entre les termes t et la liste l.

Definition 47:

Taxonomie.

$$t_x \in T_x = \{ t \in T \mid t_x \, R_x \, t \, , \leq_{T_x} \}$$

$$R_x \subseteq T_x \times T \subseteq \mathcal{P}(T)$$

Soit t_x une taxonomie appartenant à l'ensemble des taxonomies T_x. La taxonomie t_x est composée d'un ensemble de termes t appartenant à l'ensemble de tous les termes T organisés selon une hiérarchie $\leq_{T_x}$. Soit R_x la relation entre les termes t du sous ensemble de termes et la taxonomie T_x. Soit $\leq_{T_x}$ une relation d'ordre partiel entre ces termes telle

que :

- $t_1 \leq_{T_x} t_1$ (réflexivité)
- $t_1 \leq_{T_x} t_2, t_2 \leq_{T_x} t_3 \Rightarrow t_1 \leq_{T_x} t_3$ (transitivité)
- $t_1 \leq_{T_x} t_2, t_2 \leq_{T_x} t_1 \Rightarrow t_1 = t_2$ (antisymétrie)

Exemple : Si n_1 = Automobile et n_2 = Véhicule, alors $n_1 \leq_{T_x} n_2$ signifie que le terme Automobile spécialise le terme Véhicule.

Definition 48:

Thésaurus.

$$n \in N = \langle t_p \in T, \{ t_s \in T \mid t_p R_s t_s \} \rangle$$

$$R_s \subseteq T \times T$$

Soit n un nœud appartenant à l'ensemble des nœuds N. Un nœud est un couple composé d'un terme principal t_p appartenant a l'ensemble des termes T ainsi que d'un ensemble de termes synonymes t_s appartenant à l'ensemble de termes T. Soit R_s la relation de synonymie entre un terme principal t_p et un terme synonyme t_s.

$$t_h \in T_h = \langle N_h = \{ n \in N \mid n R_h t_h, \leq_{T_h} \}, g \in G \mid V_g \subseteq N_h \rangle$$

$$A \subseteq V \times V \subseteq N \times N$$

$$R_h \subseteq N \times T_h$$

Soit t_h un thésaurus appartenant à l'ensemble des thésaurus T_h. Un thésaurus t_h est un couple composé d'un ensemble de nœuds n appartenant à l'ensemble de tous les nœuds N, organisés selon une hiérarchie $\leq_{T_h}$ ainsi que d'un graphe g appartenant à l'ensemble des graphes G. Soit R_h la relation qui lie l'ensemble des les nœuds n au thésaurus t_h. Soit $g \in G$ un graphe orienté défini par $g = (V_g, A_g)$ tel que $V_g \subseteq N_h$ et A_g est l'ensemble des relations non hiérarchiques entre certains nœuds du thésaurus. Nous rappelons la définition d'un graphe orienté : $G = (V, A)$, avec V un ensemble de nœuds et A un ensemble d'arcs. Chaque $a \in A_g$ est un couple de nœuds (v_i, v_j) avec $v_i, v_j \in V_g$ orienté de v_i vers v_j. Soit $\leq_{T_h}$ une relation d'ordre partiel telle que :

- $n_1 \leq_{T_h} n_1$ (réflexivité)
- $n_1 \leq_{T_h} n_2, n_2 \leq_{T_x} n_3 \Rightarrow n_1 \leq_{T_h} n_3$ (transitivité)
- $n_1 \leq_{T_h} n_2, n_2 \leq_{T_h} n_1 \Rightarrow n_1 = n_2$ (antisymétrie)

Définition mathématique du modèle unifié des ressources terminologiques Afin de n'utiliser qu'une seule modélisation nous définissons une ressource terminologique de la façon suivante:

1) La définition mathématique que nous avons formulée d'un thésaurus est suffisamment expressive pour permettre la gestion des thésaurus comme des vocabulaires plus simples.

Definition 49:

Ressource terminologique.

$$r_t \in R_T = \langle N_h = \{ n \in N \mid n\, R_h\, t_h, \leq_{T_h} \}, g \in G \mid V_g \subseteq N_h \rangle$$

2) Si g est un graphe vide $V_g = \varnothing$, $A_g = \varnothing$ et que chacun des nœuds n contient une liste de termes synonymes vide, alors, la ressource est une taxonomie.

3) Si g est un graphe vide $V_g = \varnothing$, $A_g = \varnothing$, que chacun des nœuds n contient une liste de termes synonymes vide et qu'aucun ordre partiel $\leq_{T_h}$ n'est défini entre ces nœuds, alors, la ressource est une simple liste.

8.1.2/ MODÈLE “*intégrateur*”

Nous venons de présenter le modèle mathématique unificateur permettant la manipulation des différents types des ressources terminologiques. Ce modèle est nécessaire pour le processus d'intégration des ressources terminologiques dans notre modèle intégrateur. Le modèle intégrateur permettant l'intégration des ressources terminologiques unifiées sera exploitable par le processus d'indexation des documents. Il structure de cette manière notre base de connaissances.

Les connaissances contenues dans la base de connaissances sont les suivantes : (i) description du domaine traité définie par des facettes, (ii) terminologies associées aux facettes et (iii) items indexés sur la base de ce vocabulaire. La sous-section suivante présente la notion de facettes liées aux ressources terminologiques qui seront exploitées dans une seconde section décrivant le modèle intégrateur.

8.1.2.1/ LES FACETTES DES RESSOURCES CONCEPTUELLES

Une facette constitue une dimension descriptive d'un item. Cette description se matérialise par l'usage d'une ressource terminologique telle qu'une taxonomie. Le contexte de description d'un item dépend du domaine d'application pour lequel cette description est nécessaire. Toutefois, il est difficile de prendre en compte l'ensemble des connaissances d'un domaine nécessaire à une application avec une seule ressource terminologie [Aussenac-Gilles, 2005] [Vandenbussche et al., 2009]. Raganathan [Ranganathan, 1933] a défini la notion de facette, afin de répondre aux problèmes liés à la rigidité des classifications taxonomiques de type “Classification Décimal de Dewey” et “Classification Décimal Universel”. Cette difficulté à couvrir l'ensemble d'un domaine, nommée rigidité, n'est pas propre aux taxonomies. L'ensemble des types de ressources terminologiques utilise un nombre limité de relations sémantiques avec lesquelles il est difficile, voir presque impossible, de couvrir l'ensemble des dimensions descriptives d'un item. C'est pourquoi nous proposons l'utilisation d'un ensemble de ressources terminologiques, chacune couvrant une dimension descriptive. Nous proposons donc l'utilisation de plusieurs facettes pour la description d'un item. La définition des facettes ainsi que la structuration et la définition des terminologies qui sont associées dépendent de la vision métier sur le domaine traité. Par conséquent, les facettes résolvent les problèmes liés de manière générale à la rigidité des ressources terminologiques et permettent une prise en compte de la vision métier. L'ensemble des connaissances métiers du domaine modélisé à l'aide de facettes constitue la fondation de notre base de connaissances nécessaire au pro-

cessus d'indexation des items. Toutefois, il est à noter que certaines ressources terminologiques intègrent dans leur définition la notion de facette afin de décrire différentes dimensions d'un domaine, c'est le cas de thésaurus à facettes.

Une des propriétés des facettes est d'être manipulable intuitivement par les humains. Leurs modélisations à l'aide du modèle intégrateur, formel permettent de les rendre manipulables tout aussi aisément par la machine, c'est-à-dire aux processus automatiques comme un processus de recommandation. Il est nécessaire de préciser que ces deux aspects sont importants pour le système décrit dans ce mémoire, car celui-ci sera utilisé à la fois par des processus informatiques et par des utilisateurs allant du simple utilisateur à l'expert. Les facettes sont très populaires sur les sites de e-commerce, car chacune des facettes permet d'appréhender une des dimensions décrivant les items du site. C'est pour cela que les utilisateurs sont aujourd'hui familiers avec leur utilisation [Hearst et al., 1998].

Par conséquent, nous proposons un modèle intégrateur utilisant la notion de facettes prenant la forme de ressources terminologiques.Tout comme le modèle unificateur, l'ensemble des ressources terminologiques intégrées à notre modèle intégrateur sont unifiées dans leur structure, leur syntaxe et leur sémantique. Ce modèle repose sur un langage artificiel, permettant la définition d'une ontologie dont la sémantique formelle repose sur la logique de description. Afin de prendre en compte les dernières avancées concernant les langages documentaires (normes ISO 25964 et BS 8723), nous distinguons la notion de concept et celle de terme généralement confondues dans les ressources existantes. Ainsi lors de l'intégration de ressources terminologiques au modèle intégrateur, une abstraction conceptuelle est réalisée. Les termes et les concepts auxquels ils renvoient sont distingués. L'objectif étant lors de la modélisation de distinguer le concept, c'est a dire la signification d'un terme, du ou des termes pouvant y faire référence.

La section suivante propose une formalisation de notre modèle suivi d'une explication et d'un exemple concernant l'intégration de ressources terminologiques au modèle. Puis nous finissons en présentant l'utilisation de ce modèle pour l'indexation d'items.

8.1.2.2/ FORMALISATION DU MODÈLE INTÉGRATEUR

Le modèle intégrateur illustré par la figure 8.2, prend la forme d'une ontologie formelle. Dans cette figure (f) signifie fonctionnelle, (t) signifie transitive, (r) signifie réflexive et $\neq$ signifie disjonction entre les concepts. Plus d'informations sur G-OWL dans la publication suivante : [Héon et al., 2013]. Nous la modélisons à l'aide d'un méta-modèle inspiré du meta-modèle de Karlsruhe utilisé notamment par Ehrig [Ehrig et al., 2004] et étendu aux bases de connaissances d'expressivité logique $\mathcal{SHOIN}(\mathcal{D})$ par Pittet [Pittet, 2014]. Une base de connaissances est un ensemble d'axiomes [Haase et al., 2005] qui peut être définie comme une structure mathématique. Nous définissons une structure [Marker, 2002] comme un n-uplet $\mathcal{S} = (\Omega, \Sigma, \Phi, E)$:

- Soit Ω un ensemble appelé l'ensemble sous-jacent de $\mathcal{S}$.
- Soit Σ une collection d'axiomes de signature $\{\sigma_i : i \in I_1\}$ ou $\sigma_i \subseteq \Omega^{m_i}$ pour $m_i \geq 1$.
- Soit Φ une collection d'axiomes de fonction $\{\varphi_i : i \in I_0\}$ ou $\varphi_i : \Omega^{n_i} \to \Omega$ pour $n_i \geq 1$.
- Soit E une collection d'éléments distincts $\{\varepsilon_i : i \in I_2\} \subseteq \Omega$.

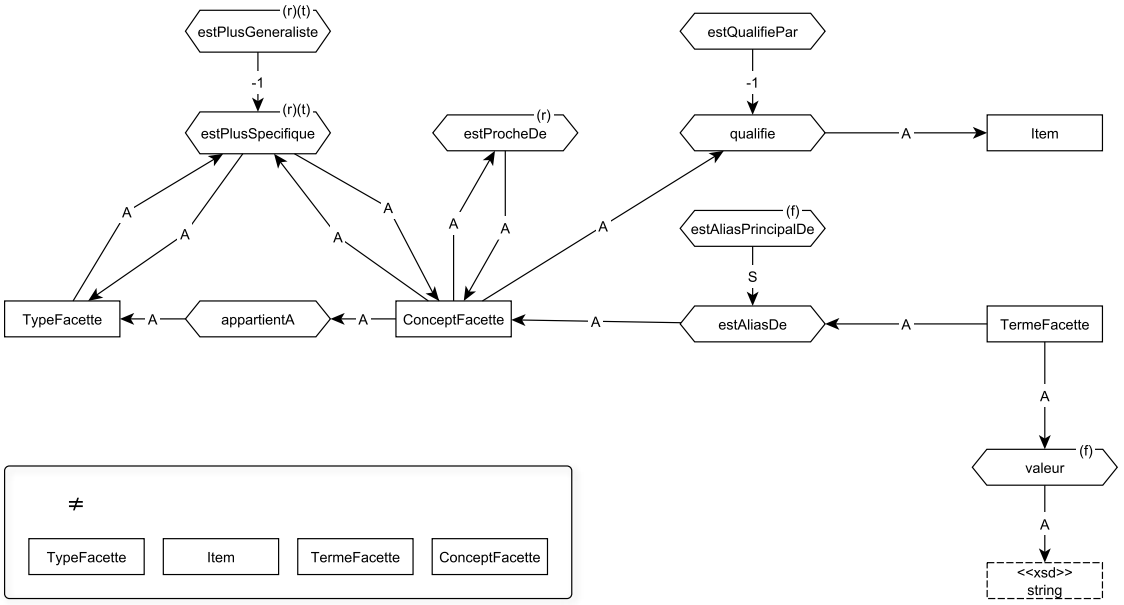

Figure 8.2 – Présentation du schéma de l’ontologie au format G-OWL

Les ensembles I_0 I_1 et I_2 peuvent être vides. n_i et m_i sont les arités respectives de φ_i et σ_i.

Notre schéma $\mathcal{S}_{integration}$:

Définition de $\Omega_{integration} = \{(sC, \leq_C), (sT, \leq_T), (sR, \leq_R), (sA, \leq_A), sI, sV, sK_R, sK_A\}$:

- sC = {TopConcept, TypeFacette, ConceptFacette, Item, TermeFacette},
- $\leq_C$ = {(TopConcept, TypeFacette), (TopConcept, TypeFacette), (TopConcept, ConceptFacette), (TopConcept, Item), (TopConcept, TermeFacette)},

sC et $\leq_C$ définissent respectivement l’ensemble des concepts et l’ensemble de définition des relations de subsomption entre ces concepts. C’est-à-dire leur organisation selon une hiérarchie de subsomption. Le **TopConcept** est le concept le plus général, celui qui subsume tous les autres. Nous avons défini deux concepts particulièrement important pour la modélisation des facettes. Le **ConceptFacette** et le **TermeFacette** permettant la distinction entre concepts et termes lors de l’intégration des ressources terminologiques.

- sT = {TopDataType, xsd:string},
- $\leq_C$ = {(TopDataType, xsd:string)},

sT et $\leq_C$ définissent respectivement l’ensemble des types de données, ainsi que l’ensemble de définition des relations de subsomption entre ces types de donnée. Le **TopDataType** est le type de donnée le plus général, celui qui subsume tous les autres. Nous avons besoin ici du type **xsd:string**[1] afin de gérer les chaînes de caractères correspondant à la valeur des termes des ressources terminologiques.

1. xsd est le préfixe de l’espace de nom http://www.w3.org/2001/XMLSchema

- sR = {TopRelation, estPlusSpecifiqueQue, estPlusGeneralisteQue, appartientA, estAliasDe, estAliasPrincipalDe, qualifie, estQualifiePar},
- $\leq_R$ = {(TopRelation, estPlusSpecifiqueQue), (TopRelation, estPlusGeneralisteQue), (TopRelation, appartientA), (TopRelation, estAliasDe),(TopRelation, estAliasPrincipalDe), (TopRelation, qualifie), (TopRelation, estQualifiePar), (estAliasDe, estAliasPrincipalDe)},

sR et $\leq_R$ définissent respectivement l'ensemble des rôles et l'ensemble de définition des relations de subsomption entre ces rôles. Le rôle **topRelation** est le rôle le plus général, celui qui subsume tous les autres. Les relations **estPlusSpecifiqueQue** et **estPlusGeneralisteQue** permettent la prise en compte des relations hiérarchiques provenant des ressources terminologiques, notamment des taxonomies et des thésaurus. Les relations **estAliasDe** et **estAliasPrincipalDe** permettent la prise en compte des relations de synonymie provenant des ressources terminologiques, notamment des thésaurus. La relation **estAliasPrincipalDe** permet la gestion du terme principal (i.e terme utilisé de préférence) alors que la relation **estAliasDe** permet la gestion des non-termes (i.e. termes non principaux). Les relations **qualifie** et **estQualifiePar** permettent elles comme nous le verrons ci-dessous la qualification des articles.

- sA = {TopAttribute, possedeValeur},
- $\leq_A$ = {(TopAttribute, possedeValeur)},

sA et $\leq_A$ définissent respectivement l'ensemble des attributs et l'ensemble de définition des relations de subsomption entre ces attributs. L'attribut **topAttribute** est l'attribut le plus général, celui qui subsume tous les autres. L'attribut **possedeValeur** permet comme nous le voyons ci-dessous de donner une valeur au terme.

- $sI = \emptyset$,
- $sV = \emptyset$,

Nous ne présentons ici que le schéma de l'ontologie, c'est-à-dire, l'ensemble des éléments permettant de définir la TBox de la base de connaissances. Les ensembles d'instanciation sI et sV correspondant à la ABox sont donc vides car le modèle n'est pas, pour le moment, peuplé par des instances.

- sk_R = {Functional, Inverse Functional, Transitive},

L'ensemble sk_R permet de définir l'ensemble des caractéristiques de rôle. C'est-à-dire l'ensemble des contraintes logiques qui peuvent être appliquées à une relation définie dans l'ensemble des rôles sR.

- sK_A = {Functional},

L'ensemble sk_A permet de définir l'ensemble des caractéristiques d'attributs. C'est-à-dire l'ensemble des contraintes logiques qui peuvent être appliquées à un attribut défini dans l'ensemble des attributs sA.

Définition de $\Sigma_{integration} = \{\sigma_R, \sigma_A\}$:

- σ_R = {(estPlusSpecifiqueQue, (TypeFacette, TypeFacette)), (estPlusSpecifiqueQue, (ConceptFacette, ConceptFacette)), (estPlusGeneralisteQue, (TypeFacette, TypeFacette)), (estPlusGeneralisteQue, (ConceptFacette, ConceptFacette)), (qualifie, (ConceptFacette, Item)), (estQualifiePar, (Item, ConceptFacette)), (estAliasDe, (TermeFacette, ConceptFacette)), (estAliasPrincipalDe, (TermeFacette, ConceptFacette)), (appartientA, (ConceptFacette, TypeFacette))},

- σ_A = {(possedeValeur,(TermeFacette, xsd:string))},

σ_R et σ_A définissent respectivement la signature des rôles définis dans l'ensemble des rôles sR et la signature des attributs définis dans l'ensemble des attributs sA. La signature consiste en la définition du **range** et du **domain** qui permettent de définir des contraintes sur les relations et les attributs. Dans le cas de relations, ou rôles, ils permettent de définir pour une relation donnée, quels sont les concepts dont les instances peuvent instancier cette relation (i.e. le domaine), et quels sont les concepts dont les instances peuvent être cibles de l'instanciation de cette relation (i.e. le range). Ces relations sont orientées. Dans le cas d'attributs, ils permettent pour un attribut donné de définir quels sont les concepts dont les instances peuvent utiliser cet attribut et quels sont les types de données de l'attribut.

Notre modèle permet à une instance du concept **TermeFacette** d'avoir un attribut **vpossedeValeur** qui a pour type une chaîne de caractères. Cet attribut permet au modèle de gérer la valeur des termes.

Les relations **estPlusSpecifiqueQue** et **estPlusGeneralisteQue** permettent la prise en compte par les facettes des relations de hiérarchie existantes entre les termes provenant de ressources terminologiques de type taxonomie ou thésaurus. Elles s'appliquent ici notamment aux **ConceptFacette**, car notre modèle permet une abstraction conceptuelle. L'organisation hiérarchique s'applique sur les notions (i.e. **ConceptsFacette**) auxquelles font référence les termes (i.e. **TermeFacette**) et non plus sur les termes principaux comme c'est le cas dans la plupart des thésaurus, ou directement sur les termes comme c'est le cas avec des ressources terminologiques de type taxonomie.

La relation **estQualifiePar** permet la description des items. C'est son instanciation qui permet l'indexation. Elle s'applique entre une instance d'**Item** et une instance de **ConceptFacette**, elle a donc lieu au niveau conceptuel et non au niveau terminologique comme c'est généralement le cas lors de l'utilisation de ressources terminologiques (i.e. langages documentaires) pour l'indexation de documents.

Les relations **estAliasDe** et **estAliasPrincipalDe** permettent la prise en compte des relations de synonymie, dans notre modèle. Elle permettent de faire le lien entre le niveau terminologique et le niveau conceptuel. Toutes les instances du concept **TermeFacette** en relation avec une même instance de **ConceptFacette** étant des synonymes.

Definition 50:

$\Phi_{integration}$ = $\{iC, iT, iR, iA, K_R, K_A, \varepsilon_C, \varepsilon_R, \varepsilon_A, \varepsilon_I, \delta_C, \delta_I, -_C, -_R, maxCard_R, minCard_R, \sqcap_C, \sqcup_C, \sqcup_I, \sqcup_V, \rho_{\exists R}, \rho_{\forall R}, \rho_R, \rho_{\exists_A}, \rho_{\forall_A}, \rho_A\}$

- $iC = \emptyset$,

- $iT = \emptyset$,

- $iR = \emptyset$,
- $iA = \emptyset$,

iC, iT, iR et iA sont les ensembles d'instanciation respectivement de concepts, types de données, rôles et attributs. Ils sont vides ici, car nous définissons dans cette section notre modèle, c'est-à-dire le schéma ontologique de la base de connaissances ou TBox.

- K_R = {(estPlusSpecifiqueQue, Transitive), (estPlusGeneralisteQue, Transitive), (estAliasPrincipal, Functional)},
- K_A = {(possedeValeur, Functional)},

K_R et K_A sont les ensembles de caractérisation respectivement des rôles et des attributs. K_R permet de définir les contraintes logiques s'appliquant sur les rôles de l'ensemble de rôles s_R en fonction de contraintes de rôles définies dans l'ensemble de contraintes de rôles sK_R. K_A permet de définir les contraintes logiques s'appliquant sur les attributs de l'ensemble d'attributs s_A en fonction des contraintes d'attributs définies dans l'ensemble des contraintes d'attributs sK_A.

Les relations **estPlusSpecifiqueQue** et **estPlusGeneralisteQue** sont formellement définies comme transitives. La relation **estAliasPrincipalDe** est défini comme formel ce qui permet de contraindre la relation. Ainsi un **ConceptFacette** ne peut avoir qu'un et un seul **TermeFacette** défini comme étant son terme principal (i.e. le terme préféré).

L'attribut **vpossedeValeur** est lui aussi défini comme fonctionnel, ainsi un **TermeFacette** ne peut avoir qu'une et une seule chaîne de caractères définissant la valeur du terme.

- $\varepsilon_C = \emptyset$,
- $\varepsilon_R = \emptyset$,
- $\varepsilon_A = \emptyset$,
- $\varepsilon_I = \emptyset$,

ε_C, ε_R, ε_A et ε_I permettent de définir respectivement l'ensemble des concepts équivalents, rôles équivalents, attributs équivalents et instances équivalentes. L'ensemble des instances équivalentes ε_I est vide, car nous ne nous intéressons pas ici à la ABox, nous définissions un modèle (i.e. un schéma d'ontologie). Les autres ensembles sont vides aussi car notre modèle ne contient aucun concepts, rôles ou attributs équivalents.

- $-_C = \emptyset$,
- $-_R$ = {estPlusGenerique, estPlusSpecifique), (estQualifiePar, qualifie)},

$-_C$ et $-_R$ permettent respectivement de définir l'ensemble des concepts complémentaires deux à deux ainsi que l'ensemble des relations inverses. Notre modèle ne présente aucun concept comme étant le complément d'un autre concept. Notre modèle définit, par contre, que les relations **estPlusGenerique** et **estPlusSpecifique** sont des relations inverses, de même les relations **qualifie** et **estQualifiePar** sont elles aussi des relations inverses.

- δ_C = {TypeFacette, Item, TermeFacette, ConceptFacette},
- $\delta_I = \emptyset$,

δ_C et δ_I définissent respectivement les ensembles de disjonction de concepts ainsi que de différentiation d'instances. Nous n'avons pas d'instances ici car nous définissons le modèle, l'ensemble δ_I est donc vide. Nous distinguons par contre les concepts **TypeFacette**, **Item**, **TermeFacette** et **ConceptFacette** comme définissant des ensembles disjoints d'instances.

- $maxCard_R = \emptyset$,
- $minCard_R = \emptyset$,

$maxCard_R$ et $minCard_R$ sont des ensemble permettant de définir des contraintes de cardinalité sur les rôles.

- $\sqcap_C = \emptyset$,
- $\sqcup_C = \emptyset$,
- $\sqcup_I = \emptyset$,
- $\sqcup_V = \emptyset$,

$\sqcap_C$, $\sqcup_C$, $\sqcup_I$ et $\sqcup_V$ permettent de définir des ensembles respectivement, d'intersection de concepts, d'union de concepts, d'énumération d'instances et d'énumération de valeurs de données.

- $\rho_{\exists R} = \emptyset$,
- $\rho_{\forall R} = \emptyset$,
- $\rho_R = \emptyset$,
- $\rho_{\exists_A} = \emptyset$,
- $\rho_{\forall_A} = \emptyset$,
- $\rho_A = \emptyset$,

$\rho_{\exists R}$, $\rho_{\forall R}$, ρ_R, $\rho_{\exists_A}$, $\rho_{\forall_A}$ et ρ_A permettent de définir des ensembles de restrictions sur les rôles et les attributs. Ces ensembles correspondent respectivement aux restrictions existentielles, universelles et de valeurs de rôles ainsi que d'attributs.

Définition de l'ensemble des éléments distincts de $E_{integration}$:

- **TopConcept** Concept spécial, subsumant tous les concepts,
- **BottomConcept** Concept spécial, subsumé par tous les concepts,
- **TopAttribute** Attribut spécial, subsumant tous les attributs,

- **BottomAttribute** Attribut spécial, subsumé par tous les attributs,
- **TopRole** Rôle spécial, subsumant tous les rôles,
- **BottomRole** Rôle spécial, subsumé par tous les rôles,
- **TopDataType** Type de données spécial, subsumant tous les types de données,
- **BottomDataType** Type de donnés spécial, subsumé par tous les type de donnés,

$E_{integration}$ = {TopConcept, BottomConcept, TopAttribute, BottomAttribute, TopRole, BottomRole, TopDataType, BottomDataType},

Nous venons de présenter les modèles unificateur et intégrateur, la section suivante illustre leur utilisation.

8.2/ PROCESSUS D'INTÉGRATION

Les modèles unificateur et intégrateur permettent respectivement, la manipulation unifiée des ressources terminologiques et leurs utilisations en tant que facettes de description pour l'indexation d'items. Nous abordons ci-dessous l'intégration des ressources terminologiques au modèle intégrateur par la présentation du processus, illustré par différents algorithmes, ainsi qu'un exemple de résultat.

8.2.1/ PROCESSUS D'ALIMENTATION DU MODÈLE INTÉGRATEUR

Le modèle intégrateur étant une ontologie formelle, le processus d'intégration peut être qualifié de processus de peuplement de l'ontologie. Ainsi, nous présentons ce processus via la création d'une facette à l'aide d'un algorithme, présenté dans les figures 10, et 12 dont le résultat sera illustré par un exemple. Ce processus est composé de trois phases : (i) peuplement de l'ontologie par la création de la facette ainsi que du niveau conceptuel du vocabulaire associé, (ii) peuplement de l'ontologie par la création du niveau terminologique de la facette et sa mise en relation avec le niveau conceptuel, (iii) peuplement de l'ontologie par l'organisation des relations au sein du niveau conceptuel de la facette.

L'algorithme prend en entrée la ressource terminologique à intégrer, le nom de la facette qui va être créée à l'aide de cette ressource, ainsi que la base de connaissances dans laquelle cette facette sera créée.

Phase (i) : Peuplement de l'ontologie par la création de la facette, ainsi que du niveau conceptuel du vocabulaire associé.
Cette partie de l'algorithme présente la création de la facette ainsi que du niveau conceptuel du vocabulaire associé à celle-ci. C'est sur ce niveau conceptuel que le niveau terminologique viendra se greffer. Pour faciliter la compréhension c'est le nom du terme principal auquel la chaîne de caractères "_c" est concaténée qui est utilisée afin de nommer le concept représentant un ensemble de synonymes.

Phase (ii) : Peuplement de l'ontologie par la création du niveau terminologique de la facette et sa mise en relation avec le niveau conceptuel.
Cette partie de l'algorithme présente la création du niveau terminologique et sa mise en relation avec le niveau conceptuel.

Algorithm 10 Algorithme de création d'une facette dans la base de connaissances à partir d'une ressource terminologique - phase 1

Require: RT une ressource terminologique, Tel Que $RT = \{T_x,\ g\ \in\ G\ |v \in V_g, v \in T_h, v \in N\ \}$.
NA le nom de la facette correspondant à ce nouveau vocabulaire.
$\mathcal{S}$ la base de connaissances, telle que $\mathcal{S} = (\Omega, \Sigma, \Phi, E)$

1: $sI = sI.add(NA)$;
2: $iC = iC.add((\text{"}TypeFacette\text{"}, NA))$; *//création de la facette*
3: **for** $n \in RT$ **do**
4: $termePrefere = n.getTermePref()$; *//récupération du terme principal*
5: $sI = sI.add(termePrefere + \text{"}\_c\text{"}$;
6: $iC = iC.add((\text{"}ConceptFacette\text{"}, termePrefere + \text{"}\_c\text{"}))$; *//création d'une instance de ConceptFacette avec le suffixe "_c" permettant de distinguer l'instance de ConceptFacette et celle de TermeFacette pour le terme principal.*
7: $iR = iR.add((\text{"}appartientA\text{"}, (NA, termePrefere + \text{"}\_c\text{"})))$; *//mise en relation des instances du niveau conceptuel avec la facette à laquelle elles appartiennent*
8: **end for**

Phase (iii) : Peuplement de l'ontologie par l'organisation des relations au sein du niveau conceptuel de la facette.
Cette partie de l'algorithme présente la création des relations hiérarchiques au niveau conceptuel du vocabulaire définissant une facette.

8.2.2/ INDEXATION DES ITEMS

La base de connaissances du système repose sur une ontologie formelle : le modèle intégrateur. Cette base de connaissances est composée de modules comme l'illustre la figure 8.3. Chacun de ces modules distinguent différents types de connaissances. La base de connaissances repose sur le modèle intégrateur. Il est donc utilisé comme une ontologie de haut niveau à partir duquel la base de connaissances peut être enrichie en fonction du domaine d'application.

Les différents modules liés à l'ontologie de haut niveau, nommé module de référence, permettent (i) la gestion des concepts généraux sous lesquels viennent se positionner les concepts plus spécifiques de chaque sous-module, (ii) la gestion de concepts transcendants différents types de connaissances. La base de connaissances comporte ainsi les modules suivants:

1. Un module contenant les connaissances générales, c'est-à-dire les connaissances associées aux facettes de description qui ne dépendent pas d'une vision métier sur un domaine précis. Cette partie de la connaissance n'a généralement pas à être adaptée, elle contient les connaissances pour la gestion de l'espace de façon administrative ou du temps afin d'indexer les items en fonction de la géolocalisation et de la temporalité.

2. Un module contenant les connaissances spécifiques au domaine, c'est-à-dire les connaissances associées aux facettes de description qui dépendent du domaine d'application ainsi que de la vision métier sur le domaine.

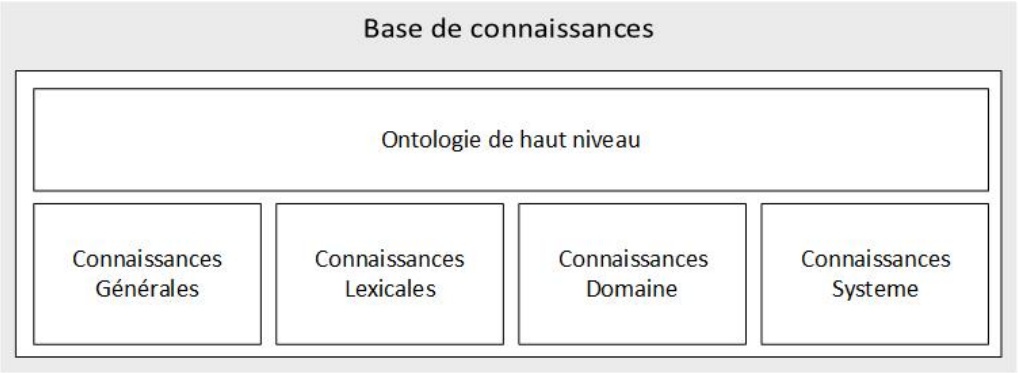

Figure 8.3 – Présentation des modules de notre base de connaissances

Ces deux modules contiennent le niveau conceptuel des ressources terminologiques provenant du processus d'intégration.

3. Un module contenant les connaissances nécessaires au système, c'est-à-dire l'indexation des items en tant que tels, et éventuellement des spécialisations des concepts ou relations nécessaires à la description des items en fonction des besoins de l'application.

4. Un module contenant les connaissances lexicales. Ces connaissances correspondent au niveau terminologique des facettes définies par les ressources intégrées.

La figure 8.4 illustre un exemple d'indexation sur des ressources terminologiques intégrées à notre modèle. Elle illustre le fait que l'indexation des items se fait au niveau conceptuel (cf. flèches rouges) et non au niveau terminologique, comme cela est généralement le cas. Cette approche, est avantageuse car les termes et langues utilisés pour désigner une notion peuvent évoluer indépendamment de la structuration conceptuelle.

8.3/ CONCLUSION

Cette partie présente les différents modèles pour la représentation des connaissances nécessaires à la recommandation d'items. Ce point est central au système, car il permet de gérer les descriptions d'items de façon à ce que celles-ci soient manipulables par la machine et par l'humain. Tous deux intervenant dans un système automatique de recommandation.

Face à l'hétérogénéité des ressources et de leurs types, nous présentons un modèle mathématique unificateur permettant de manipuler les ressources lors de la phase d'intégration de ressources dans notre modèle intégrateur. Celui-ci permettant la description des items en fonction de facettes. L'intégration des ressources terminologiques à la base de connaissances en respectant le modèle défini possède divers avantages :

1. Il permet, et c'est là l'objectif principal, la description d'items à l'aide de différentes facettes, chacune associée à une terminologie. Ces descriptions sont aisément accessibles pour une compréhension humaine, car elles sont organisées sous la forme de facettes associées à des ressources terminologiques (i.e. listes plates, taxonomies, thésaurus) pensées en premier lieu pour une utilisation humaine.

Les facettes sont une façon simple de modéliser un domaine même complexe. Leur utilisation est aisée pour des humains qui y sont habitués, notamment sur les plate-formes de e-commerce.
Les descriptions d'items sont de plus directement manipulables par la machine, car basées sur un modèle formel unifiant tous les types de ressources.
L'utilisation de facettes permet de gérer la complexité et la richesse de la modélisation d'un domaine tout en conservant une modélisation compréhensible et manipulable pour l'humain.

2. Notre modèle étant une ontologie formelle basée sur les logiques de description, des contraintes et propriétés logiques telles que les rôles inverses, les relations transitives et les relations formelles ont pu être ajoutées aux ressources. Ces contraintes et propriétés sont exploitables par des processus d'inférence.

3. Lors de l'intégration de ressources conceptuelles au modèle, nous réalisons une abstraction conceptuelle. Le niveau conceptuel et le niveau terminologique sont distingués contrairement aux ressources terminologiques qui ne font pas cette distinction. L'indexation des items a lieu sur le plan conceptuel, indépendamment de l'évolution de la terminologie permettant l'adaptation à une nouvelle langue par exemple. De même qu'une évolution de la conceptualisation peut être réalisée sans impacter la terminologie.

4. Notre modèle permet d'augmenter le niveau d'expressivité d'une ressource. Par exemple, il permet d'organiser hiérarchiquement les éléments d'une liste afin d'en faire une taxonomie.

La base de connaissances résultant du modèle proposé et permettant la gestion de l'hétérogénéité des données contient ainsi les connaissances suivantes :

- La modélisation du domaine d'application sous la forme de facettes et le vocabulaire associé.
- La description de chacun des items sur la base de ces facettes.

Algorithm 11 Algorithme de création d'une facette dans la base de connaissances à partir d'une ressource terminologique - phase 2

for $n \in RT$ **do**
  $termePrefere = n.getTermePref()$; *//récupération du terme principal*
  $listeSynonymes = n.getListSyno()$; *//récupération de la liste des termes synonymes*
  $sI = sI.add(termePrefere)$; *//ajout du terme principal à l'ensemble des instances*
  $iC = iC.add("TermeFacette", termePrefere)$; *//déclaration du terme principal en tant qu'instance de TermeFacette*
  $iR = iR.add(("hasMainAlias", (termePrefere, termePrefere + "_c")))$; *//création d'une relation hasMainAlias entre l'instance de termePéféré de ConceptFacette et celle de TermeFacette*
  $sV = sV.add(termePrefere)$;
  $iT = iT.add((xsd : string, termePrefere))$;
  $iA = iA.add((valeur, (termePrefere, termePrefere)))$; *//instantiation de la valeur de l'attribut valeur de chaque instance de TermeFacette alias principal*
  for $termeSynonyme$ de $listeSynonymes$ **do**
    $sI = sI.add(termeSynonyme)$; *//ajout du terme synonyme a l'ensemble des instances*
    $iC = iC.add(("TermeFacette", termeSynonyme))$; *//déclaration du synonyme en tant qu'instance de TermeFacette*
    $iR = iR.add((hasAlias, (termeSynonyme, termePrefere + "_c")))$; *//création d'une relation hasAlias entre l'instance de termePéféré de ConceptFacette et l'instance de termeSynonyme de TermeFacette*
    $sV = sV.add(termeSynonyme)$;
    $iT = iT.add((xsd : string, termeSynonyme))$;
    $iA = iA.add((valeur, (termeSynonyme, termeSynonyme)))$; *//instantiation de la valeur de l'attribut valeur de chaque instance de TermeFacette alias non-principal*
  end for
end for

Algorithm 12 Algorithme de création d'une facette dans la base de connaissances à partir d'une ressource terminologique - phase 3

for $n \in RT$ **do**
  $listePeres = n.getPeres()$; *//récupération de la liste des nœuds pères*
  for $npere \in listePeres$ **do**
    $iR = iR.add((\text{"}estPlusGeneralisteQue\text{"}, (npere.getTermePref() + \text{"}\_c\text{"}, n.getTermePref() + \text{"}\_c\text{"})))$; *//création des relations hiérarchiques au niveau conceptuel, c'est à dire entre les instances de ConceptsFacette*
    $iR = iR.add((\text{"}estPlusSpecifiqueQue\text{"}, (n.getTermePref() + \text{"}\_c\text{"}, npere.getTermePref() + \text{"}\_c\text{"})))$; *//cette relation n'est pas créée mais déduite d'un raisonnement, estPlusSpecifiqueQue étant la relation inverse de estPlusGeneralisteQue*
  end for
end for
for $n \in RT$ **do**
  $listeProches = n.getProches()$; *//récupération de la liste des nœuds proches*
  for $nproche \in listeProches$ **do**
    $iR = iR.add((estProcheDe, (n.getTermePref() + \text{"}\_c\text{"}, nproche.getTermePref() + \text{"}\_c\text{"})))$; *//instantiation des relations estProcheDe au niveau conceptuel c'est à dire entre les instances de ConceptFacette*
  end for
end for

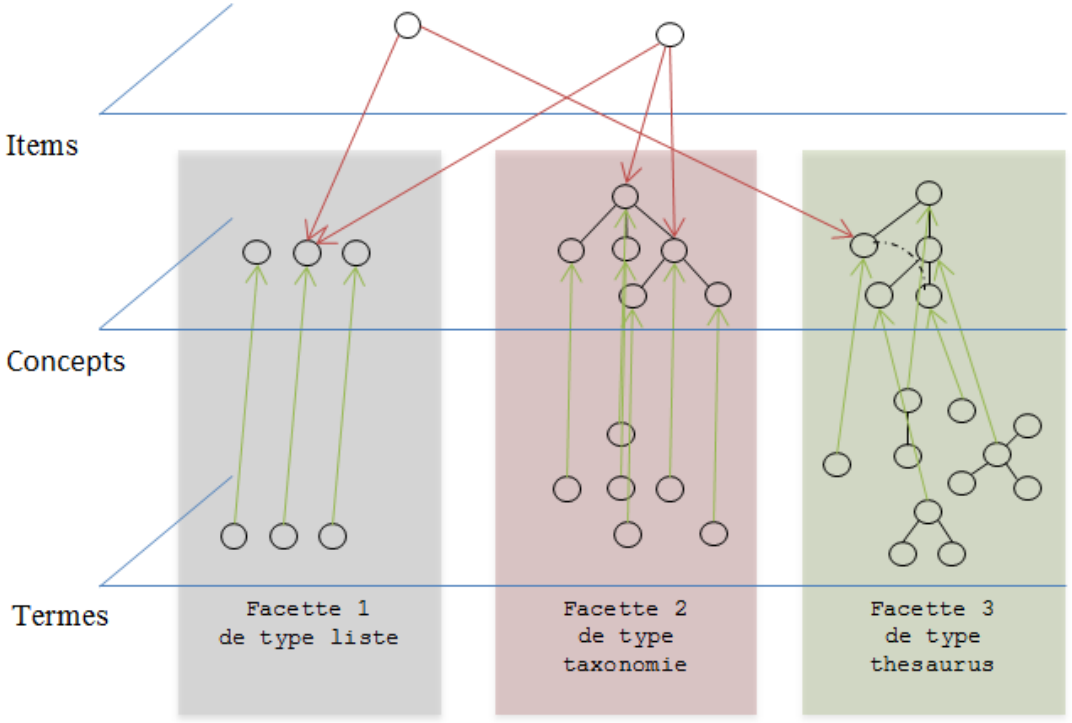

Figure 8.4 – Indexation de deux items

9

CONCLUSION

Ce chapitre présente les travaux menés dans le but de solutionner les problèmes que posent les données dans le processus d'Intuition Artificielle. Plus précisément il s'agit des problématiques de collecte des données qui concernent un sujet d'intérêt et pour laquelle il convient donc de pouvoir évaluer la pertinence des données et leur volume. Le dernier verrou solutionné est la gestion de l'hétérogénéité des données qui sont collectées auprès de sources très variées.

Les deux thèses ont été menées dans le cadre ontologique car dans la suite du processus d'Intuition Artificielle ces verrous concernent la dernière étape de l'approche, celle de la compréhension et l'expression de l'intuition (chapitre IV) traitées par le biais d'ontologies, qui, malheureusement, ne sont naturellement pas adaptées pour gérer ces contraintes sur les données.

Les travaux menés démontrent mathématiquement et fournissent les algorithmes nécessaires à la lever de ces verrous.

IV

La sémantique pour comprendre et s'exprimer

LA SÉMANTIQUE POUR COMPRENDRE ET S'EXPRIMER

Les étapes préalables d'élaboration de l'Intuition Artificielle ont permis de collecter massivement des données, d'en tirer des connaissances et d'identifier dans ces connaissances les éléments qui manquaient.
En complément, le système faisant d'intuition, il faut qu'il puisse l'exprimer, et surtout l'exprimer de telle sorte que l'humain puisse comprendre les tenants et aboutissants de cette intuition pour qu'il décide ce qu'il souhaite en faire.

Pour cela l'accès à la sémantique est primordial. Mon choix pour y accéder est d'utiliser les ontologies et leurs moteurs d'inférence. Cependant, cela pose plusieurs problèmes. Tout d'abord, une ontologie est construite généralement par le truchement d'experts qui fournissent leurs connaissances. Or, faire appel à un expert est difficile, comme en fouille de données supervisée, car il faut trouver un expert disponible, et fastidieux pour lui, car la tâche est ardue. De plus, dans notre cas il s'agit de pouvoir construire une ontologie adaptée au domaine de l'intuition concernée, et donc qui est fonction des données et des connaissances qui pourront être collectées pour enrichir ce domaine.
L'idée est donc de construire l'ontologie automatiquement à partir des connaissances du système. Cette opération en plus d'être économe en temps et en énergie, permet aussi d'avoir une ontologie possédant exactement les mêmes connaissances que l'Intuition Artificielle qui va la solliciter.

Une autre contrainte des ontologies réside dans le fait que les moteurs d'inférence actuels n'ont pas la faculté de traiter notre environnement : trop de données, trop de changements, trop d'incertitude...

Pour répondre à ces deux axes, un ensemble de verrous sont à lever : la qualification des connaissances, l'inférence multidimensionnelle floue,l'automatisation de la construction d'ontologies et la véracité des conclusions émises par le système.

Je présente quatre parties dans ce chapitre. Dans un premier temps, j'expose un état de l'art des mesures de similarité sémantique ciblé au domaine des graphes puisque c'est le paradigme de l'Intuition Artificielle. La seconde partie s'intéresse au verrou de la qualification des connaissances par le nommage des concepts de la FCA afin de participer au processus d'automatisation du processus de création non supervisée d'ontologies à partir des données présenté dans le partie suivante. Enfin la dernière partie démontre que l'on peut prouver la véracité des conclusions émises.

10

MESURES DE SIMILARITÉ SÉMANTIQUE BASÉES SUR LES GRAPHES

Contribution : Naeimeh Laleh recrutée comme post doctorante dans le cadre du projet Eurostars Predictive Smart Data Platform.

Il existe une grande diversité de mesures sémantiques pour la comparaison de termes, de concepts, de mots, de phrases et de graphes de connaissances. Le principal objectif décrit ici est de mesurer la similarité sémantique entre concepts en considérant les graphes sémantiques de concepts reposant sur des graphes de connaissances puisque l'Intuition Artificielle repose sur les relations manquantes dans les graphes de connaissances. Les mesures présentées ici, prennent donc uniquement en compte le paramètre relationnel pour définir les relations sémantiques entre les concepts dans le graphe de connaissances.

La similarité sémantique s'établit sur des mesures permettant d'évaluer la valeur, l'importance ou la force des interactions sémantiques (distance / similarité / relation) entre deux éléments ciblés [Harispe et al., 2013].

Les mesures de similarité sémantiques (MSS) sont utilisés dans de nombreuses applications, notamment les systèmes de recommandation, mais aussi plus précisément pour des systèmes de recommandation basés sur des graphes [de Gemmis et al., 2015, Gori et al., 2007, Abbassi et al., 2007, Noulas et al., 2012], dans lesquels les éléments sont reliés entre eux en fonction de leur similarité. Une autre application des mesures sémantiques est l'intégration de représentations de la connaissance, qui sont des modèles informatiques ou des expressions formelles de la connaissance sous une forme compréhensible par la machine. Le but de ces modèles est de définir un ensemble de concepts, leurs relations et leurs axiomes en considérant la sémantique des concepts et leurs relations. Les MSS sont utilisées pour trouver des concepts similaires définis notamment dans différentes ontologies [Euzenat et al., 2007]. Par conséquent, les MSS sont utiles dans la conception de systèmes de récupération d'informations basés sur des ontologies [Hliaoutakis et al., 2006, Hliaoutakis, 2005, Varelas et al., 2005, Baziz et al., 2007, Saruladha et al., 2010, Sy et al., 2012].

Il existe plusieurs approches pour mesurer la similarité sémantique. Les plus importantes sont les approches distributionnelles et basées sur les connaissances.

Les distributions peuvent être: géométriques, basées sur les ensembles, probabilistes et capturer des co-occurrences plus profondes [Harispe et al., 2013]. Toutes ces approches sont basées sur des distributions de mots en comptant les occurrences de mots dans de grandes collections de textes pour obtenir des MSS et ajouter une sémantique en prenant en compte la relation entre les mots. Certaines de ces approches représentent les significations conceptuelles dans des vecteurs de grande dimension telles que l'Analyse Sémantique Explicite [Gabrilovich et al., 2007] et l'Analyse Sémantique Latente [Landauer et al., 1997] et certains travaux récents sont basés sur des modèles informatiques tels que GLOVE [Pennington et al., 2014] ou Word2Vec [Mikolov et al., 2013] qui représentent les mots ou les concepts avec des vecteurs de faible dimension. Ces approches mesurent généralement la relation sémantique plutôt que la similarité sémantique prenant en compte les relations hiérarchiques [Turney et al., 2010]. La relation entre deux concepts comme "voiture" et "essence" est forte alors que leur similarité est faible.

Les approches basées sur la connaissance [Mihalcea et al., 2006] mesurent la similarité sémantique entre les concepts basés sur une analyse de graphe. Certaines de ces mesures sémantiques utilisent la structure du graphe en analysant les relations entre les concepts et en effectuant des marches aléatoires ou en calculant les chemins les plus courts [**?**,Olsson et al., 2011,Chebotarev et al., 2006a,Jeh et al., 2002]. Certaines d'entre elles sont basées sur le modèle d'entités défini par Tversky [Tversky, 1977] et prennent en compte la liste des propriétés spécifiques de chaque élément et comparent ces entités pour estimer la similarité sémantique. Il existe d'autres mesures sémantiques basées sur la quantité d'informations partagée par les concepts. Ces approches utilisent le contenu informationnel des concepts ou des éléments du graphe [SáNchez et al., 2013, Sánchez et al., 2012a].

Dans le cadre de l'Intuition Artificielle, seules les approches basées sur la connaissance sont explorées ici.

10.1/ VUE D'ENSEMBLE

Les approches basées sur la connaissance [Mihalcea et al., 2006] mesurent la similarité sémantique entre des concepts en considérant des graphes de concepts tels que des graphes sémantiques. Un graphe sémantique est une représentation de la connaissance qui, au lieu de s'appuyer sur des constructions logiques comprenant: conjonction, disjonction, négation, consiste en un ensemble d'instructions comprenant des triplets: sujet, prédicat, objet [Harispe et al., 2013]. Cette de représentation s'appelle un graphe de connaissances et correspond à un graphe construit à partir d'une vaste collection (pouvant représenter des milliards) d'énoncés sous forme de triplets <sujet-prédicat-objet> utilisés pour décrire formellement toute connaissance. Cette connaissance peut être modélisée sous la forme d'un graphe RDF (Resource Description Framework), dans lequel les sommets représentent des sujets et des objets, et les arêtes étiquetées correspondent à des prédicats. Il existe plusieurs exemples de ces grands référentiels de connaissances RDF, tels que WordNet, DBpedia, Freebasei, Wikidata, Yago. En outre, il existe des bases de données de graphes, notamment Neo4J, Tita et OrientDB. Ces bases de données sont basées sur une structure de graphe permettant de décrire des informations en NoSQL: mode [Webber et al., 2013].

Les mesures sémantiques basées sur les graphes de connaissances analysent le graphe

en prenant en compte le paramètre relationnel pour définir les relations sémantiques entre les concepts du graphe. Elles utilisent la transitivité de la relation taxonomique dans la conception des mesures sémantiques. De plus, dans certaines de ces approches, la taxonomie des types de relations sémantiques (prédicats) prend en compte le calcul de la similarité sémantique. Certaines de ces mesures sémantiques sont basées sur la structure du graphe en analysant les relations entre les concepts d'un graphe, tels que la marche aléatoire, le chemin le plus court (pondéré ou non) ou l'estimation de la proximité pour calculer l'interconnexion entre les deux concepts à l'aide du théorème matrix-forest [Chebotarev et al., 2006b, Olsson et al., 2011, Chebotarev et al., 2006a, Jeh et al., 2002]...

La similarité sémantique et la relation sémantique [Harispe et al., 2013] sont deux notions distinctes. La similarité sémantique est un cas particulier de parenté lié à la ressemblance des concepts [Jiang et al., 2015]. Par exemple, le mot "augmentation" est associé au mot "diminution", mais ils ne sont pas similaires [Yazdani et al., 2013].

Plus en détail, certaines approches utilisent le graphe pour extraire les caractéristiques des éléments, puis pour estimer la similarité entre deux éléments ciblés en analysant ces caractéristiques [Gamallo, 2017, Taieb et al., 2014]. La fonction de similarité consiste à calculer la somme pondérée des caractéristiques communes et distinctes des deux éléments comparés. Outre la description du concept, les auteurs considèrent les informations taxonomiques et non taxonomiques modélisées dans les ontologies [Ngan et al., 2006, Meng et al., 2013, Rodríguez et al., 2003]. Les mesures sémantiques s'appuient sur différentes approches basées sur :

- l'indice de Jaccard, proposée dans [Maedche et al., 2001, Stojanovic et al., 2001],
- les ensembles [Bulskov et al., 2002],
- la distance taxonomique entre deux classes en fonction de caractéristiques partagées et distinctes [Sánchez et al., 2012b],
- des vecteurs et la représentation vectorielle, construitent en fonction de l'ensemble des instances des classes [Bodenreider et al., 2005, Sun et al., 2016c] et sont basées sur le modèle d'espace vectoriel (VMS),
- la mesure des concepts, des mots ou des termes tirés de Wikipedia [Jiang et al., 2015]. Plus précisément, l'approche présente d'abord une représentation formelle des concepts de Wikipédia, puis plusieurs approches basées sur les caractéristiques pour les mesures de similarité sémantique calculées,
- nombre de descendants de la LCA des concepts comparés [Jain et al., 2010, Ranwez et al., 2006],
- un journal négatif [Batet et al., 2010, Batet et al., 2011]. Ainsi, dans cette approche, les propriétés de distance ont été prouvées : inégalité triangulaire, positivité et symétrie sont démontrées,
- la quantité d'informations partagée et distinctes entre les concepts. Ces approches ont utilisé le contenu informationnel de concepts ou d'éléments du graphe [SáNchez et al., 2013, Sánchez et al., 2012a],
- contenu informationnel de leur ancêtre commun le plus informatif (MICA) [Resnik, 1995],

- un méta-graphique pour réduire l'ontologie originale en un groupe de concepts connexes. Ensuite, une formulation spécifique d'une mesure basée sur un ensemble considère les classes comme leurs ensembles d'ancêtres [Mazandu et al., 2011],

- le contenu informationnel du MICA des classes comparées divisé par le contenu informationnel maximal des classes comparées [Mazandu et al., 2013],

- les spécificités des deux concepts comparés [Lin et al., 1998, Jiang et al., 1997a, Mazandu et al., 2013, Pirró et al., 2008, Pirró, 2009, Pirró et al., 2010] pour contourner l'inconvénient que les concepts avec le même MICA ont le même score de similarité sémantique,

- la spécificité des concepts MICA [Schlicker et al., 2006, Li et al., 2010, Mazandu et al., 2011, Cross et al., 2011],

- l'expression théorique de l'information de l'indice de Jaccard [Pirró et al., 2010],

- le chemin le plus court contraint par le LCA des deux classes comparées. Les arêtes entre deux concepts sont pondérées en fonction de la différence de contenu informationnel des classes auxquelles elles sont liées [Jiang et al., 1997a, Couto et al., 2003, Othman et al., 2008],

- des combinaisons hybrides entre le contenu informationnel des concepts et les autres approches structurelles citées [Singh et al., 2013, Paul et al., 2012].

Le schéma de la figure 10.2 ordonne les différentes approches de mesures de similarité sémantique détaillées dans la partie ci-après.

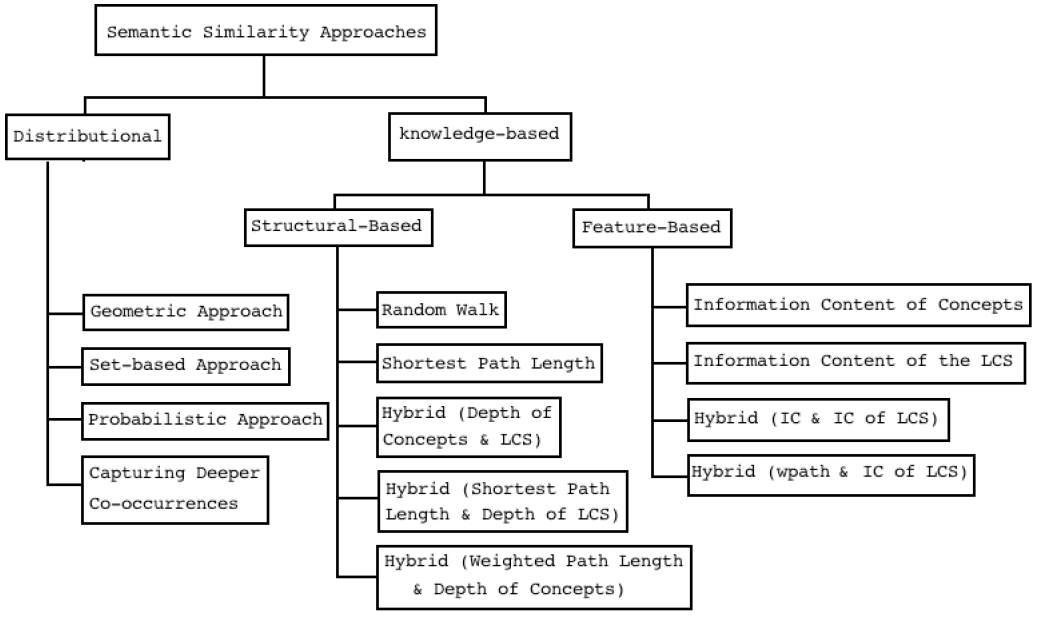

Figure 10.1 – Approches de mesures de similarités sémantiques

10.2/ MESURES SÉMANTIQUES BASÉES SUR LA CONNAISSANCE

Le but des mesures basées sur des graphes est d'analyser les interactions des nœuds dans le graphe de connaissances pour estimer la relation des paires de nœuds. Par conséquent, plus les interactions entre les deux nœuds sont nombreuses, plus ils sont considérées comme liés.

10.2.1/ MESURES STRUCTURELLES

MSS basées sur la marche aléatoire La mesure de la marche aléatoire est une mesure de traversée du graphe pour estimer la parenté des paires de nœuds basée sur le modèle des chaînes de Markov [Harispe et al., 2013].

Il existe plusieurs mesures de marche aléatoire [Fouss et al., 2007] : l'approche basée sur le nombre moyen d'étapes entre deux noeuds appelé "temps moyen de premier passage" ou l'ECT (Euclidean Commute Time [1]) pour un graphe pondéré qui présente la propriété intéressante de décroître lorsque le nombre de chemins reliant deux nœuds augmente ou lorsque la "longueur" d'un chemin diminue, ce qui la rend bien adaptée aux tâches de clustering "[Hughes et al., 2007, Fouss et al., 2007, Ramage et al., 2009, Alvarez et al., 2011b, Alvarez et al., 2011a, Garla et al., 2012]. Par conséquent, plus le nombre de chemins reliant le nœud A au nœud B dans le graphe de connaissances est faible, plus leur distance est courte [Sarkar et al., 2008, Von Luxburg et al., 2010].

MSS basée sur la longueur du plus court chemin. Cette approche calcule le poids du chemin le plus court qui relie deux nœuds en fonction de leur relation [Harispe et al., 2013] en tant que distance sémantique. Ensuite, ils inversent la fonction de distance pour calculer la similarité sémantique. L'inconvénient de ces approches est que la similarité sémantique entre les deux concepts qui ont la même longueur de chemin sera la même. De cette manière, ils ne considèrent pas le sens ni la sémantique des relations. Par conséquent, certains travaux de recherche examinent les relations sémantiques avec plusieurs prédicats en attribuant différents scores à différents prédicats [Hirst et al., 1998, Bulskov et al., 2002]. En outre, la méthode LCH [Leacock et al., 1998a] tente d'améliorer la mesure de similarité en proposant une mesure basées sur la longueur du chemin le plus court avec une fonction non linéaire, en prenant également en compte la profondeur maximale de la taxonomie du concept dans le graphe (i.e. le nombre maximum de super classes) [Harispe et al., 2013]. Une autre approche est également basée sur la profondeur maximale de la taxonomie [Resnik, 1995].

MSS hybride basée sur la profondeur des concepts et leurs LCS La profondeur de l'information dans la taxonomie des concepts permet de définir les concepts de niveau supérieur comme étant plus généraux et donc moins similaire.

[Wu et al., 1994a] considère à la fois la profondeur de chaque paire de concepts et la profondeur du LCS (Least Common Subsumer correspondant au LCA) de ces concepts. Bien que ces approches améliorent les performances de similarité sémantique en utilisant la profondeur de l'information par rapport aux approches de longueur de chemin pure, de

1. Une "commute distance" est le temps qu'il faut pour aller d'un point à un autre et en revenir par un déplacement aléatoire.

nombreux concepts présentent la même similarité en raison de la même profondeur dans le graphe de connaissances hiérarchique.

Dans d'autres approches, les auteurs considèrent la spécificité des concepts comparés [Mao et al., 2002] ou la profondeur du LCA des deux concepts, [Wu et al., 1994a, Pekar et al., 2002, Zhong et al., 2002, Wang et al., 2012].

MSS hybride basée sur la longueur du chemin le plus court & profondeur du LCS Il existe une approche hybride [Li et al., 2003] qui combine la longueur de chemin la plus courte et la profondeur du LCS de deux concepts ciblés pour améliorer les performances de l'approche de Wup [Wu et al., 1994a]. Il mesure la similarité sémantique à l'aide d'une fonction non linéaire. Une autre approche est également une combinaison de deux fonctions, la première fonction étant le chemin le plus court entre les deux mots de Wordnet et la seconde fonction représentant la profondeur du LCS à partir de la racine du graphe [Li et al., 2006, Li et al., 2003, Li et al., 2006].

[Ganesan et al., 2012, Shet et al., 2012] ont proposé une mesure en considérant également à la fois la profondeur de l'information et le poids du plus court chemin calculé mais en pénalisant les chemins avec de multiples changements de type de relations .

MSS hybride basée sur la longueur pondérée du chemin & la profondeur des concepts Considérer la longueur de chemin ne suffit pas pour avoir une distance sémantique efficace. Par conséquent, certains travaux de recherche associent des poids aux arêtes du graphe de connaissances [Rada et al., 1989, Sussna, 1993]. L'approche [Sussna, 1993], utilise deux axes. Le premier consiste à considérer le type d'arête pour y attribuer un poids inversement proportionnel au nombre d'arêtes du même type. Le deuxième consiste à examiner la profondeur des concepts pour améliorer l'approche de la longueur du trajet. En utilisant la profondeur, la valeur de similarité avec les concepts de niveau inférieur est plus élevée, car ils sont situés plus en profondeur dans la taxonomie.

10.2.2/ MESURES BASÉES SUR LES CARACTÉRISTIQUES DES GRAPHES

MSS basée sur le contenu informationnel des concepts (IC) Considérer le contenu informationnel des concepts [Resnik, 1995] dans la mesure de la similarité sémantique est utile pour améliorer les performances des approches basées sur le chemin ou sur la profondeur, lorsque les deux concepts ciblés ont la même longueur de chemin et la même profondeur d'information [Jiang et al., 1997a, Lin et al., 1998, Resnik et al., 1999]. La mesure est basée sur l'idée que des concepts plus spécifiques (plus profond dans le graphe) ont une valeur plus élevée dans le contenu de l'information. Par conséquent, si les deux concepts partagent des concepts plus spécifiques, ils sont plus similaires car ils partagent davantage d'informations.

MSS basée sur le contenu informationnel du LCS (IC-LCS) [Resnik, 1993, Resnik, 1995, Resnik et al., 1999] utilisent les informations contenues dans un corpus et les relations de subsomption entre les concepts du graphe de connaissances. L'auteur a déclaré que la mesure de similarité entre deux concepts est liée à la quantité d'informations qu'ils partagent. Par conséquent, plus les concepts qu'ils partagent sont nombreux, plus la similarité entre eux est forte, la quantité d'informations partagées étant liée à la position du

dernier parent commun. Ils utilisent le contenu de l'information pour évaluer la spécificité d'un concept en calculant la probabilité de le rencontrer dans le graphe hiérarchique. La mesure de spécificité est utile pour le calcul du LCS des deux concepts. Cependant, cette approche présente un inconvénient : en attribuant une mesure de similarité égale pour les couples de concepts qui partagent un parent commun. Par conséquent, les paires de concepts spécifiques avec un nombre différent de liens entre elles dans les approches basées sur les graphes auront la même mesure de similarité sémantique dans cette approche [Budanitsky et al., 2006].

Par conséquent, certaines autres approches ont tenté d'améliorer l'inconvénient en prenant en compte le contenu informationnel des concepts ainsi que le contenu informationnel de leur LCS.

MSS hybride basé sur IC & IC de LCS (Hybrid-IC-LCS) Il existe certaines approches hybrides pour calculer la similarité sémantique en prenant en compte le contenu informationnel des deux concepts ciblés et le contenu informationnel de leur LSC. À titre d'exemple, l'approche [Lin et al., 1998] calcule la similarité sémantique entre les concepts en considérant le rapport entre le contenu informationnel des concepts et leur LCS. En outre, l'approche JCN [Jiang et al., 1997a] mesure la différence entre les concepts en soustrayant la somme du contenu informationnel de chaque concept du contenu informationnel de leur LCS en tant que mesure de distance. Ensuite, ils calculent l'inverse de la distance pour la mesure de similarité.

Cependant, considérer uniquement le contenu informationnel de l'information pour calculer la différence entre les concepts ne considère pas l'information de distance précieuse basée sur la longueur du chemin qui les sépare. Par conséquent, ces approches échouent lorsque le LCS des deux concepts est l'entité de concept racine dans le graphe d'information. De plus, les mesures hybrides basées sur le contenu informationnel des concepts et leur LCS ne prennent pas en compte les informations de niveau hiérarchique. Par conséquent, les concepts généraux qui semblent être moins similaires ont donné un score de similarité plus élevé basé sur ces deux approches.

MSS hybride basé sur wpath & IC de LCS [Gao et al., 2015] combine le comptage des arêtes et le contenu informationnel. Dans cette approche, les auteurs appliquent la longueur de chemin la plus courte pour les deux concepts en attribuant une pondération à la longueur des chemins. Ils ont utilisé le contenu informationnel du LCS des paires de concepts pour attribuer le poids à la longueur.

[Zhu et al., 2017] propose une longueur de chemin pondérée (wpath).Cette approche considère différents poids pour la longueur de chemin la plus courte parmi les paires de concepts en fonction des informations partagées. Par conséquent, lorsque la longueur de chemin entre les concepts est élevée, la similarité sémantique entre les paires de concepts est plus petite. Les auteurs ont déclaré que cette approche ne prend pas en compte la profondeur de l'information, mais que le contenu de l'information de LCS est similaire à la profondeur de concept, ce qui indique que le niveau plus profond des concepts de la taxonomie est plus spécifique, donc plus similaire. [Zhu et al., 2017]. Parce que LCS définit le niveau hiérarchique de la taxonomie et que le contenu de l'information exploite l'occurrence statistique de l'information des concepts. D'autre part, le contenu de l'information inclut la fréquence des concepts et a donc plus de valeur que la profondeur de l'information. Par conséquent, cette approche tente d'améliorer les performances des

approches basées sur le contenu de l'information qui ne tiennent pas compte des niveaux hiérarchiques et de la distance conceptuelle des concepts. De cette manière, cette approche donne un score de similarité plus élevé aux concepts plus spécifiques ayant la même longueur de chemin, mais donne également un score de similarité plus élevé aux concepts qui partagent le même contenu d'information et se situent plus près dans la taxonomie.

10.3/ MSS BASÉE SUR PLUSIEURS RELATIONS CONCEPTUELLES

Considérer l'intégration de multiples relations conceptuelles dans les ontologies est un aspect important pour proposer des mesures de similarité sémantique. Il existe plusieurs types de relations comme is-a, has-a, etc. dans les graphes de connaissances et ces approches prennent en compte la distance sémantique relationnelle entre les concepts comme ceux de WordNet ou d'autres ontologies.

Dans le cadre de vecteurs de termes, un terme peut recouvrir trois types de sens : les noms, les verbes et les adjectifs / adverbes. Dans ce type d'approche, la similarité des noms et des verbes et la similarité des adjectifs sont définies différemment en fonction de la distance sémantique. De manière plus détaillée, dans WordNet, les mesures de similarité pour les concepts de noms et de verbes organisés selon une structure hiérarchique et les concepts d'adjectifs et d'adverbes sont regroupés dans des structures bipolaires. Les relations sémantiques formées entre les noms et les verbes couvrent is-a, has-a et l'antinomie[2], alors que les relations sémantiques d'adjectifs et d'adverbes incluent similaire-à et l'antinomie.

Pour mesurer la similarité sémantique ou la distance sémantique des concepts de noms et de verbes, la méthode basée sur les chemins est la longueur du plus court chemin reliant les deux concepts [Wenyin et al., 2010, Rada et al., 1989]. [Chen et al., 2017] utilise un chemin pondéré pour la force du lien en tant que distance sémantique en prenant en compte le type de relation sémantique, la profondeur et la densité locale (nombre d'enfants selon un concept parent). De manière plus détaillée, pour les concepts parent-enfant avec une relation is-a ou has-a, les concepts d'un niveau inférieur ont une sémantique plus spécifique que les concepts d'un niveau supérieur. Par conséquent, la distance sémantique entre les concepts de niveau inférieur est inférieure à celle de niveau supérieur.

Pour estimer la similarité sémantique ou la distance sémantique des concepts d'adjectifs et d'adverbes, les auteurs ont utilisé une méthode basée sur le gloss des adjectifs fournis dans WordNet. Le gloss d'un adjectif est une courte définition textuelle qui décrit brièvement le sens d'un concept. Notez que deux adjectifs dans un synset partagent le même gloss. Plus précisément, pour mesurer la distance sémantique entre les adjectifs, on utilise un Vector Space Model (VSM) [Billhardt et al., 2002, Chen et al., 2012].

10.4/ MSS POUR LA COMPARAISON DE GRAPHES RDF

Mesurer la similarité des graphes est une tâche ardue. Plusieurs travaux de recherche ont proposé de mesurer la similarité structurelle basée sur leur topologie. Il existe deux

2. Contradiction, opposition totale entre deux idées, concepts, principes.

types de distance de structure pour mesurer leur similarité : la distance basée sur les caractéristiques et la distance basée sur les coûts [Sanfeliu et al., 1983]. Dans le premier cas, l'ensemble des caractéristiques est extrait de la structure en tant que vecteur, puis la distance euclidienne est appliquée pour calculer la distance. Dans le second cas, on prend en compte les opérations permettant de transformer le premier graphe en le second.

Cependant, il existe plusieurs graphes sémantiques tels que les graphes sémantiques, les graphes liés à des événements et les graphes de liaisons sémantiques. Par exemple, les graphes de connaissances RDF sont sans schéma. De plus, le même type d'informations est parfois représenté dans divers graphes sémantiques et cela présente un gros problème pour les unifier et appliquer des requêtes SPARQL complexes sur plusieurs structures de graphes. Dans ce type de graphes sémantiques, la mesure de la similarité structurelle basée sur la sémantique est un problème clé et ces mesures de similarité structurelle basées sur la topologie ne peuvent pas être utiles pour détecter les similarités.

Par conséquent, certains travaux de recherche ont proposé certaines approches pour accéder à ces graphes RDF et comparer leurs similarités en proposant des mesures malgré une connaissance incomplète du schéma sous-jacent. [Zheng et al., 2016] propose une mesure de similarité appelée ”distance d'édition du graphe sémantique“ pour mesurer la similarité entre les graphes RDF en considérant leur graphe de synthèse sémantique qui les résume. Plus précisément, la mesure de ”distance d'édition du graphe sémantique“ uniforme, est une intégration de la similarité structurelle des graphes, de la similarité au niveau des concepts et de divers modèles de structure d'équivalence sémantique.

[Sun et al., 2016a] propose une mesure permettant de calculer la similarité structurelle en plaçant les relations sémantiques au coeur du calcul. Plus en détail, l'auteur a utilisé la théorie de l'espace pour les relations sémantiques [Sun et al., 2016b] afin de calculer la similarité structurelle sémantique afin d'obtenir un résultat efficace comparable à la similarité de Jaccard ou de Dice. Ce Calcul s'appuie sur les bases des relations sémantiques entre les paires de ressources et leur orthogonalisation.

11

LA QUALIFICATION DES CONNAISSANCES PAR LE NOMMAGE DE CONCEPTS

Contributions :
Quention Brabant recruté comme post doctorant dans le cadre du projet QAPE.
Wided Selmi en stage de fin de doctorat de l'université de Sfax.

La FCA est la source de fourniture des connaissances nécessaires à l'Intuition Artificielle. Cependant la connaissance générée l'est sous forme de concepts formels qui peuvent être vu comme des clusters d'objets, avec une définition intensionnelle fournie par son ensemble d'attributs.

Dans l'intention d'exporter ces concepts vers des classes pour construire l'ontologie et de bâtir la table d'adjacence nécessaire à l'identification de relations manquantes dans les connaissances fournies, il est nécessaire de qualifier, i.e. de nommer simplement ces concepts par un mot ou une expression simple.

Pour lever ce verrou, j'ai accepté en encadrement de stage de thèse une doctorante (Wided Selmi) de l'université de Sfax, et j'ai recruté un post doctorant sur un projet industriel (QAPE dont la description se trouve en section **??**) : Quentin Brabant.

Concrètement, si l'on considère le contexte formel de la table 11.1, certains des concepts peuvent être identifiés comme "mammifères" ou "oiseaux". L'objectif est de nommer automatiquement les concepts d'un contexte donné. Bien que cette tâche ne nécessite pas de surmonter de grandes difficultés théoriques, elle présente un intérêt pratique et n'est pas anodin. Étonnamment, il semble que très peu d'ouvrages sur ce sujet aient encore été publiés. A notre connaissance, ce problème n'est explicitement abordé que dans [Benaicha, 2017].

Les objets et les attributs sont généralement identifiés et référencés via des "noms" se présentant sous la forme d'une chaîne de caractères. Nous désignons ce nom par (x), pour tout objet ($\in G$) ou attribut ($\in M$) x. Dans notre approche, nous nous appuyons sur WordNet et sur l'hypothèse suivante :

H_1: pour chaque objet $a \in G$ (a) est un mot dans WordNet,

H_2: pour chaque attribut $b \in M$, (b) est un mot dans WordNet.

Comme nous l'expliquerons, il est possible de gérer les cas où H_1 n'est pas vérifié pour certains objets et où H_2 n'est vérifié pour aucun attribut. Cependant, il est nécessaire que H_1 soit vérifié pour la plupart des objets.

	poils	plumes	oeufs	lait	vol	aquatique	prédateur	dents	colonne vertébrale	respiration	venimeux
ours	×			×			×	×	×	×	
poisson chat			×			×	×	×	×		
dauphin				×		×	×	×	×	×	
canard		×	×		×	×			×	×	
éléphant	×			×				×	×	×	
grenouille			×			×	×	×	×	×	
mouette		×	×		×	×	×		×	×	
abeille	×		×		×					×	×
guêpe	×		×		×					×	×

Table 11.1 – Exemple de contexte formel. Cette table représente un sous-ensemble du jeu de données "zoo"[1]. Ici, on peut voir que les lignes abeille et guêpe et les colonnes poils, oeufs, vol, respiration et venimeux forment un rectangle maximal, et donc $(\{abeille, gupe\}, \{poils, oeufs, vol, respiration, venimeux\})$ est un concept.

11.1/ L'APPROCHE ET SES VERROUS

Le dictionnaire choisi pour cette approche est WordNet qui est une base de données lexicale développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton depuis une vingtaine d'années. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise. La composante atomique sur laquelle repose le système entier est le synset (synonym set), un groupe de mots interchangeables, dénotant un sens ou un usage particulier. Il offre aussi les relations d'hyperonymie, i.e. la relation sémantique hiérarchique d'une unité lexicale à une autre selon laquelle l'extension du premier terme, plus général, englobe l'extension du second, plus spécifique. [Wikipedia].

Puisque la relation hypernyme peut être interprétée comme une relation is-a, une solution simple et naturelle à notre problème consisterait à associer à chaque concept (A, B) l'hypernyme le moins commun (LCH : least common hypernym) des éléments de A. Cependant, cette approche simple pose plusieurs difficultés :

1. **L'ambiguïté des mots.** Dans WordNet, chaque mot est lié à plusieurs synsets. Chacun de ces synsets correspond à un sens du mot. De plus, seuls les synsets

(et non les mots) sont liés via la relation hypernyme. Ainsi, afin de mettre en œuvre l'idée intuitive de LCH sur les mots, nous devons décider comment gérer le fait que chaque mot peut correspondre à plusieurs synsets.

2. **Plusieurs LCH.** La relation hypernyme est un ordre avec un élément maximal. Par conséquent, tous les synsets $S_1, \ldots, S_n$ ont un ensemble non vide d'hypernymes communs. Cependant, cet ensemble n'a pas toujours un élément le plus petit qui soit unique. Dans ces cas, l'un des plus petits éléments doit être choisi.

3. **Information manquante dans WordNet.** Le jeu de données "zoo" contient un objet "fille". Techniquement, cette fille peut être considérée comme une femme et donc comme un mammifère. Cependant, "mammifère" ou même "animal" ne sont pas des hypernymes de "fille" dans WordNet. En ce sens, nous pouvons dire que des informations sont manquantes dans WordNet. En conséquence, l'hypernyme le moins commun de $\{fille, ours, dauphin, elephant\}$ est "organisme", qui est moins spécifique que "mammifère". Notre méthode devrait pouvoir détecter lorsqu'un ensemble de noms d'objet contient une valeur aberrante, telle que "fille" dans notre exemple.

4. **Même nom pour plusieurs concepts.** Deux extensions de deux concepts différents peuvent donner le même LCH. Notre algorithme doit garantir que le nom de chaque concept est unique.

5. **Noms composés.** L'hypothèse H_1 est en réalité très restrictive. En effet, certains objets peuvent être décrits par un groupe nominal au lieu d'un seul mot. Par exemple, un contexte plus précis pourrait contenir un objet "Grand lion à crinière rouge". Bien que le nom de cet objet ne soit pas un mot de WordNet, il s'agit d'un groupe nominal valide, à partir duquel nous devrions pouvoir obtenir des informations.

Pour contrer ces difficultés, voici les propositions envisagées et/ou mises en oeuvre :

11.1.1/ AMBIGUÏTÉ DES MOTS

Les mots ont souvent plus d'une signification, c'est-à-dire qu'ils sont liés à plus d'un synset dans WordNet. Il convient cependant de sélectionner un seul synset par mot : le plus pertinent. Pour cela nous nous appuyons sur le contexte disponible (autres mots dans l'extension et dans l'intension).

L'identification de la signification correcte d'un mot dans un contexte particulier s'appelle *Word Sense Disambiguation* (WSD). Les informations de contexte sont très importantes pour identifier le sens d'un mot ambigu et peuvent donner des indices sur le WSD. Pour un aperçu complet des approches en matière de WSD, nous orientons les lecteurs vers l'enquête de [Navigli, 2009], qui suggère principalement trois catégories : approches supervisées, basées sur la connaissance [Chaplot et al., 2018, Lesk, 1986, Rais et al., 2016] et non supervisées.

Notre approche relève de la catégorie des méthodes basées sur la connaissance, qui s'appuient sur des ressources de connaissances (sémantiques), telles que des dictionnaires, des thésaurus et des ontologies. L'objectif principal est de récupérer le sens adéquat d'un mot polysémique à l'aide de mesures de connexité sémantique / similarité. Ces mesures calculent la similarité entre des concepts, ayant la même signification ou

des informations connexes, contenus dans une ressource donnée. Il existe différentes mesures telles que LCH [Leacock et al., 1998b], WUP [Wu et al., 1994b], JCN [Jiang et al., 1997b] ou LIN [Lin, 1998] telles que présentées dans l'état de l'art.

Notre méthode utilise WordNet en tant que ressource de connaissances pour WSD et les mesures de parenté. Elle lève l'ambiguïté d'un mot w à partir d'un contexte (G, M, I) (intensions et extensions) fourni en entrée. Ensuite, les étapes suivantes sont effectuées :

1. Soit l'ensemble U de tous les synsets qui correspondent à un objet ou à un attribut avec un nom non ambigu. Un mot w est sans ambiguïté s'il appartient à un seul synset. Dans ce cas, on note S_w le synset de w. On définit U ainsi :

$$U = \left\{ S_{(w)} \;\middle|\; x \in G \cup M \text{ et } |S_{(w)}| = 1 \right\}.$$

2. Pour chaque synset S auquel w appartient, on calcule un score pour le synset S comme la ressemblance moyenne de S avec les synsets de U:

$$\text{score} = \frac{1}{|U|} \sum_{w \in U} \text{sim}(S, \boldsymbol{S}((w))).$$

   L'idée ici est de calculer le degré de parenté sémantique entre tous les sens du mot cible, tout sens d'un mot appartenant à son contexte. Notons que la fonction sim() peut être n'importe quelle fonction de similarité.

3. Renvoyer le synset avec le score le plus élevé.

11.1.2/ PLUSIEURS LCHS

Il y aurait plusieurs manières de départager les LCH potentiels (prendre le plus proche selon une certaine distance sémantique, au hasard, etc), mais nous n'avons pas eu à faire ce choix car la bibliothèque Java WordNet que nous utilisons sélectionne déjà un unique LCH, même lorsqu'il en existe plusieurs.

11.1.3/ INFORMATIONS MANQUANTES DANS WORDNET

Pour la détection d'anomalies, le problème qui se pose est celui-ci : lorsqu'on a un ensemble {chat, chien, otarie, fille}, on s'attend à ce que Fille descende de Humain et de Mammifère, et donc que le LCH de l'ensemble soit "Mammifère". Le problème c'est que Fille, dans wordnet n'a pas de lien avec Humain ou Mammifère. Elle est sur une branche sans lien avec les notion de classification des animaux. Ce qui fait que le LCH de l'ensemble donne "Entité Vivante". L'objectif serait donc d'identifier que Fille est très éloignée (dans la hiérarchie de wordnet) des autres éléments de l'ensemble. Grâce aux mesures de distance de wordnet, on peut créer une matrice de distance pour Chat, Chien, Otarie et Fille. Les méthodes de détection d'anomalies telles que DBSCAN interviennent alors pour traiter cette matrice de distances et détecter que Fille est un outlier, qui doit être ignoré pendant le calcul du LCH.

11.1.4/ MÊME NOM POUR PLUSIEURS CONCEPTS

Soit $(A_1, B_1), \ldots (A_n, B_n)$ des concepts auxquels on donne le même nom w. Pour donner un nom unique w_i à chaque concept (A_i, B_i), nous définissons w_i comme le mot w concaténé avec les noms des attributs distincts de l'ensemble

$$B_i \setminus \left(\bigcap_{j=1}^{n} B_j \right),$$

où \ indique l'opération de différence définie.

Cette approche garantit que chaque concept a un nom différent. Cependant, les noms ainsi obtenus peuvent parfois contenir de nombreux attributs. Notons également que si $\{(A_1, B_1), \ldots (A_n, B_n)\}$ contient le plus grand concept, alors l'intension de ce concept est : $\bigcap_{j=1}^{n} B_j$ et ainsi aucun attribut n'est ajouté à son nom.

11.1.5/ OBJETS AVEC UN NOM COMPOSÉ

Dans le cas où un objet possède un nom composé tel que la "Grand lion à crinière rouge", l'approche ne peut pas aller chercher le ou les synsets correspondant car il lui faut un nom unique. Pour contrer cette problématique nous utilisons un analyseur pour extraire le nom principal du groupe nominal, "Lion" dans cet exemple .

11.2/ APPROCHES ALTERNATIVES : EXTRACTION DE NOUVELLES DESCRIPTIONS

Redescription mining [Galbrun et al., 2017] est une tâche de data mining dont le but est de trouver des équivalences entre des définitions alternatives des mêmes sous-ensembles d'objets. L'exploration de redescription commence généralement par un *contexte de redescription*, qui est similaire à un contexte formel de FCA : on considère un ensemble G d'objets, un ensemble M d'attributs et une relation $I \subseteq G \times M$ indiquant quel objet a quels attributs. De plus, nous avons deux ensembles d'attributs disjoints, appelés vues et désignés par V_1 et V_2. Le processus de *Redescription mining* consiste à rechercher des sous-ensembles d'objets pouvant être définis à la fois en utilisant uniquement les attributs de V_1 et en utilisant uniquement l'attribut de V_2. Le résultat du processus est un ensemble de relations d'équivalence de la forme

$$\Phi_1 \quad \Leftrightarrow \quad \Phi_2$$

où Φ_1 et Φ_2 sont des formules logiques dont les variables sont, respectivement, les attributs de V_1 et les attributs de V_2.

Plusieurs outils de *Redescription mining*, tels que [Galbrun et al., 2012, Leeuwen et al., 2015], sont disponibles gratuitement et peuvent être configurés de manière à limiter le nombre de variables ou à interdire des opérateurs logiques autres que la conjonction. Ainsi, un moyen simple d'affecter des noms aux concepts formels d'un contexte pourrait être de créer un contexte de redescription, où V_1 contient les attributs distincts et V_2 serait défini par

$$V_2 = \bigcup_{a \in G} H((a)),$$

$H((a))$ étant l'ensemble des hypernymes de (a), et de configurer l'un des outils disponibles pour générer des règles d'équivalence de la forme :

$$b_1 \wedge b_2 \wedge \cdots \wedge b_n \quad \Leftrightarrow \quad w,$$

où $b_1, b_2, \ldots, b_n \in V_1$ et $w \in V_2$. Ensuite, à chacune de ces règles d'équivalence, le nom w est attribué au concept avec la plus petite intension contenant $b_1, b_2, \ldots, b_n$.

11.3/ CONCLUSION

Cette approche permet de pouvoir nommer les concepts formels de la FCA de façon à leur donner une sémantique d'une part, et d'autre part pour permettre de transformer ces concepts en classes ontologiques permettant le raisonnement et en matrice d'adjacence pour l'Intuition Artificielle.

12

CONSTRUCTION AUTOMATIQUE D'ONTOLOGIE À PARTIR DES DONNÉES

Contribution : Marwan Batrouni dans le cadre de sa thèse sur l'analyse de scénarios (détails en section 15.2).

La dernière thèse que j'ai encadrée est celle de Marwan Batrouni. Cette thèse part du postulat qu'en exploitant les technologies et les outils Big Data, il est possible d'améliorer considérablement le domaine de l'analyse de scénarios. Pour esquisser une méthode possible pour une telle utilisation, une méthode dont l'élément clé consiste à créer une ontologie et des règles dites "FZS.Bayes" à partir d'un ou de plusieurs ensembles de données a été mise au point.
La figure 12.1 illustre les principales étapes de ce processus. La raison principale d'un tel objectif est double. D'une part, l'ontologie et les règles constituent l'une des matières premières essentielles à la transformation en un modèle dynamique, tel qu'un système markovien, dont le système d'analyse de scénario dépend. D'autre part, l'ontologie et les règles générées à partir des données peuvent augmenter et compléter une ontologie plus large issue de différentes sources, telles que l'opinion d'experts ou à partir d'ontologies et de règles préexistantes et réutilisées.
Différents travaux se sont intéressés à la création d'ontologies à partir de données, ils sont présentés en première partie. Puis vient la question des techniques d'exploration de données permettant de générer des règles d'association à partir de données, et sur la façon de mapper ces règles dans des règles FZS.Bayes.

12.1/ ETAT DE L'ART SUR LA CONSTRUCTION D'ONTOLOGIES À PARTIR DES DONNÉES

Les approches et méthodologies de création d'ontologies constituent un immense champ d'investigation. Dans cette partie, un bref aperçu des différentes approches et méthodologies est passé en revue. Cependant, le focus est essentiellement mis sur le domaine de l'ontologie tirée de l'apprentissage des données.

Globalement, pour la création d'ontologies, différentes sources sont exploitées :

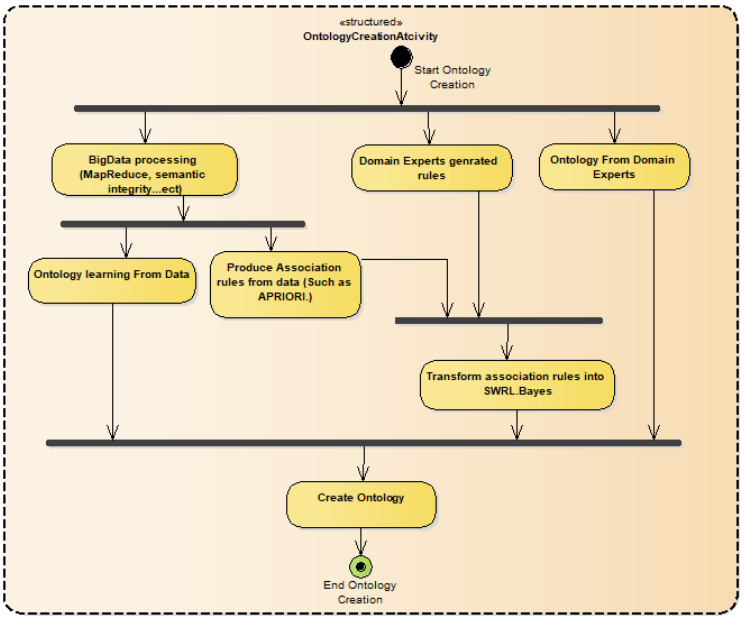

Figure 12.1 – Ontologie et processus de création de règles

- Sollicitation des experts du domaine
- La réutilisation d'ontologie existantes
- Ontology learning: à partir de texte ou de données
- Différentes combinaisons des approches ci-dessus

Construction par experts. La construction d'ontologies par experts de domaine est une activité coûteuse, car elle implique l'expertise humaine en modélisation et d'experts de domaine travaillant de concert, ce qui laisse place à des problèmes potentiels de communication et de conceptualisation. Les principales étapes de l'ingénierie ontologique peuvent être résumées par [Sure et al., 2009, Cimiano et al., 2006] en ce qu'on appelle *Knowledge Meta Process*, qui est divisé en cinq étapes principales : l'étude de faisabilité, la création, le raffinement, l'évaluation et l'application / évolution.

Construction par réutilisation. La réutilisation des ontologies est un moyen très efficace et recommandé de commencer la phase de modélisation.
[d'Aquin et al., 2012] répertorie certaines des sources les plus populaires, y compris des domaines spécifiques tels que BioPortal [NF et al., 2009] pour le biomédical et génériques tels que Ontology Design Patterns (ODP) [1]. Le principal défi de la réutilisation des ontologies est principalement d'assurer la qualité de l'ontologie réutilisée. Il existe plusieurs tentatives pour créer un repère standard. Cependant, cela reste un domaine de recherche actif [Fernández et al., 2009].

Construction par Ontology learning. L'apprentissage ontologique est un vaste sujet; il peut être principalement subdivisé en apprentissage à partir de texte ou à partir de

1. http://ontologydesignpatterns.org.

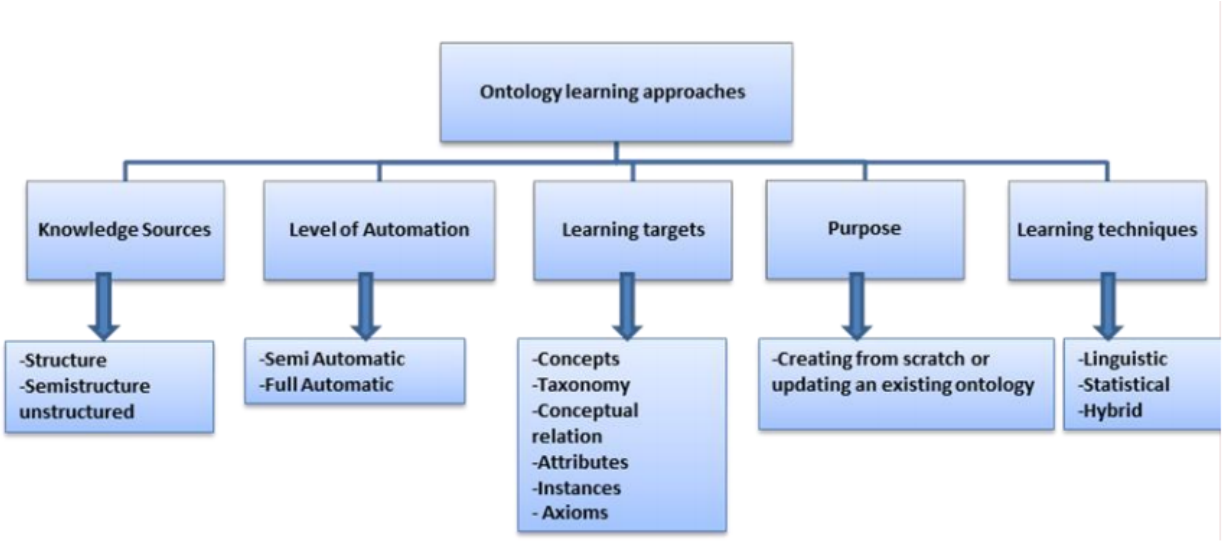

Figure 12.2 – Classification des Approches d'Ontology Learning (Source [Abeer Al-Arfaj, 2015])

données numériques.
En ce qui concerne l'apprentissage d'une ontologie à partir d'un texte, [Abeer Al-Arfaj, 2015] classe l'apprentissage de l'ontologie en plusieurs segments (figure 12.2).

Dans l'apprentissage à partir de l'espace texte, il existe plusieurs méthodologies et outils, les principales approches d'apprentissage étant centrées sur les approches linguistique, statistique et hybride.
Des outils tels que DODDLe2 [Yamaguchi, 2001], Text2Onto [Cimiano et al., 2005] sont des outils populaires utilisés.

L'apprentissage de l'ontologie à partir de données numériques est un domaine en plein essor qui prend de l'ampleur avec l'avènement des technologies et des outils Big Data.

La section suivante présente de plus près la principale méthodologie utilisée aujourd'hui pour l'extraction d'ontologies à partir de données, à savoir l'Analyse de Formelle de Concepts (FCA).

12.2/ CONSTRUCTION D'UNE ONTOLOGIE BASÉE SUR LES DONNÉES

12.2.1/ LA STRUCTURE ONTOLOGIQUE

Les stratégies orientées données pour l'apprentissage d'ontologies ont des approches différentes selon le degré de structure des données.
Le nettoyage des données et les techniques ETL, ainsi que l'utilisation de MapReduce pour rendre la taille des données gérable, peuvent constituer un préalable indispensable à toute tentative d'exploitation des données pour la construction d'ontologies. L'enquête [Hazman et al., 2011] décrit l'approche d'exploration de données utilisée pour traiter les

données semi-structurées.
Cependant, la méthodologie principale pour l'apprentissage des ontologies à partir de données non structurées et semi-structurées est l'Analyse Formelle de Concepts [Hazman et al., 2011, Priya et al., 2015].

Il existe plusieurs méthodologies pour transformer les concepts d'un treillis en une ontologie, comme résumé dans la table 12.1 [Priya et al., 2015], ces méthodes traduisant généralement les concepts en classes d'ontologies où les sous-concepts deviennent des sous-classes et des super-concepts des classes parentes. Une fois la correspondance établie entre les concepts et l'ontologie, les données peuvent également être exploitées pour générer des règles d'association entre les concepts.

12.2.2/ LE RAISONNEMENT ONTOLOGIQUE.

Les règles d'association ont largement été abordées dans ce manuscrit. Cependant, l'un des avantages de ce type de règles dont il convient de parler est leur similarité avec les règles SWRL, qui ont également la forme $Antecedent \Rightarrow consequent$. Un autre avantage est la possibilité de limiter le niveau de confiance et de support permettant ainsi de contrôler le nombre et la qualité des règles obtenues. Cependant, générer les règles d'association à partir de données ne suffit pas sous leur forme brute, il faut transformer ces règles en règles SWRL. En effet, pour pouvoir utiliser ces règles ultérieurement, il faut les intégrer dans le contexte général de l'ontologie et dans le cadre de la thèse de Marwan Batrouni : les rendre compatibles avec l'analyse de scénarios. Pour accomplir la transformation des règles d'association, voici le processus suivi :

1. Préparer les données pour l'analyse par la FCA et générer un treillis de concepts.
2. Transformer le treillis de concepts en une ontologie.
3. Générer des règles d'association à l'aide d'algorithmes dédiés.
4. Mapper les attributs des antécédents et des conséquents sur les classes de l'ontologie de l'étape 2, générant dans le processus les règles SWRL de la forme $C_0 \Rightarrow C_1$ Où C_0, C_1 sont des classes d'ontologie.
5. Transformer les règles de l'étape 4 en transférant le conséquent avec l'antécédent et en utilisant la confiance comme conséquent. Par exemple, si la confiance pour les règles que nous avons à l'étape 4 est de 75 %, alors:
   $C_0 \wedge C_1 \Rightarrow 0.75$
6. Trouvez l'intervalle de probabilité qui correspond au niveau de confiance concerné, ce qui nous donne par exemple que 0.75 correspond à la plage de valeurs *Probable*, puis remplacez cette valeur par sa sémantique, ainsi: $C_0 \wedge C_1 \Rightarrow Probable$

En utilisant les règles transformées et la position des classes dans la hiérarchie ontologique, on peut déduire les classes concernées par une règle. La figure 12.3 présente

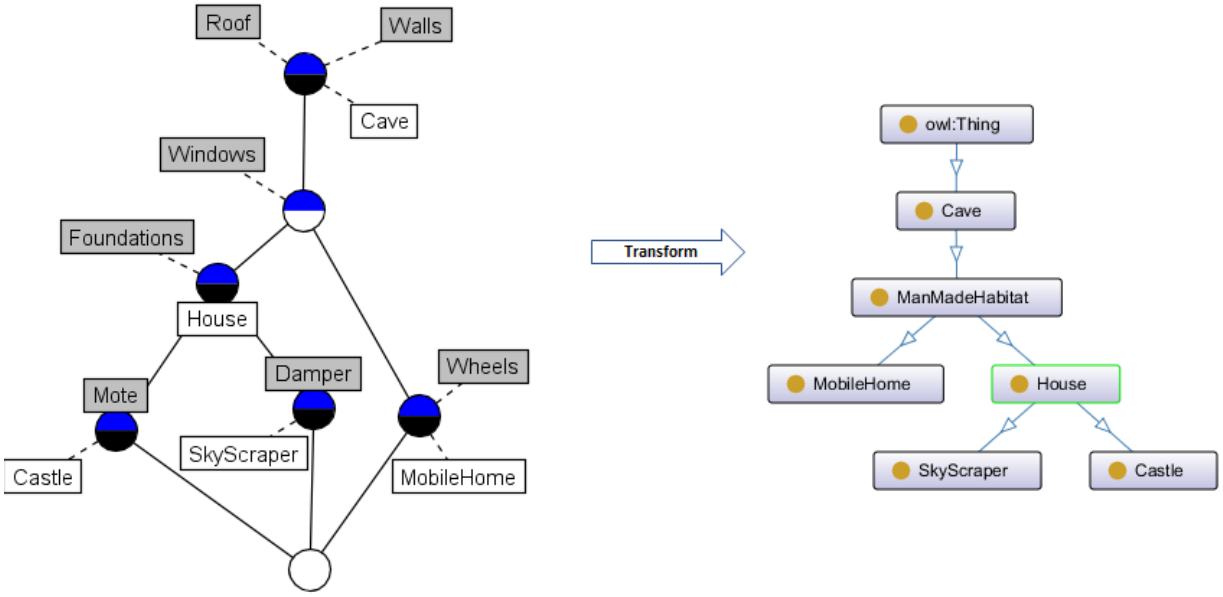

Figure 12.3 – Exemple de conversion d'un treillis en ontologie

un exemple de treillis d'habitat et sa structure ontologique résultante.
Supposons maintenant que nous avons une règle d'association telle que:

$$Walls \wedge Roof \wedge Windows \Rightarrow Foundations$$

avec un niveau de confiance de 75 %. Cela correspond aux classes d'ontologie en tant que telles :

$$Habitat \Rightarrow House$$

Enfin, la transformation de la règle produit :

$$Habitat \wedge House \Rightarrow Probable$$

Une fois que le mapping des règles d'association est effectué, il faut générer les tables de probabilités conditionnelles (CPT). L'algorithme examine d'abord quelle est la relation entre *House* et *Habitat*, puisqu'il s'agit d'une relation d'héritage, il est alors interprété comme une dépendance bayésienne et se lit comme la probabilité conditionnelle de $p(House|Habitat) = Probable$.

L'un des principaux avantages de la transformation du treillis et ses règles d'association en une ontologie et ses règles SWRL est la possibilité d'utiliser des raisonneurs d'ontologie pour vérifier la cohérence des règles. L'ontologie et les règles générées sont des préliminaires et doivent être vérifiées pour leur pertinence.

12.3/ DISCUSSION

Pour cette thèse qui s'intéresse à l'analyse de scénarios dans un objectif de prédiction et prescription, nous avons exploré la voie de la création d'ontologie à partir des données

pour répondre à ce paradigme en orientant les règles vers une réponse de probabilité à la survenue d'un événement scénaristique.
Je souhaite étendre ce travail pour qu'il soit compatible à n'importe quel domaine outre que l'analyse de scénarios, cependant ce travail a déjà permis d'établir un état de l'art sur le sujet et un prototype, en cours de réalisation montre sa faisabilité.

De ce travail nous avons appris que la construction automatique d'ontologie à partir des données nécessite 3 éléments :

- La création des classes de l'ontologie et leur subsumption. Pour cela la FCA est toute indiquée.
- Le raisonnement de l'ontologie. Des règles d'inférence peuvent être composées à partir de règles d'association qui prennent la même forme que des clauses de Horn : $(r_1 \wedge r_2 \wedge \ldots \wedge r_n) \Rightarrow h$, avec $n \in \mathbb{N}$
- Les caractéristiques relationnelles (fonctionnelle, inverse, transitive, symétrique, réflexive) et les restrictions existentielle (some) et universelle (only) doivent aussi pouvoir être comprises et transcrites automatiquement.

Ce travail est préliminaire et évolue pour trouver sa place au sein de l'Intuition Artificielle afin de pouvoir expliquer l'intuition qui émerge.

Méthode / Auteur	Résumé
Guoqian Jiang et al [Jiang et al., 2003]	Méthode semi-supervisée utilisant la FCA et un module de traitement du langage naturel (NLP) pour la terminologie médicale, le treillis est d'abord proposé dans une interface Protégé destinée à être mappée aux classes d'ontologie.
Hele-Mai Haav [Haav, 2004]	Combine un langage basé sur des règles avec la FCA pour construire une ontologie de domaine semi-automatique. Dans cette approche, initialement, les contextes formels pour le domaine donné peuvent être construits à partir des données du domaine extraites par des techniques de traitement du langage naturel.
Marek Obitko et al [M et al., 2004]	Processus manuel : Les concepts sont reconnus selon leurs propriétés (intension). Les propriétés spécifient la hiérarchie des concepts. Lorsque les propriétés des différents concepts sont identiques, les concepts sont identiques.
Xin Peng et al [Peng X, 2007]	Processus collaboratif avec les fournisseurs d'éléments, implique la collecte des descriptions basiques des entrées / sorties des éléments et concepts à valeur hiérarchique à partir des descriptions des éléments d'origine. Le but est de construire une hiérarchie ontologique pour chaque élément pouvant être rempli automatiquement.
Suqin Tang et al [Tang S, 2010]	Une méthode pour construire une ontologie touristique appelée (TOCM) en utilisant la FCA, cette méthode permet de mapper des concepts du treillis en classes d'ontologie.
Liu Ning et al [Ning L, 2010]	Pour créer une ontologie maritime à l'aide de la FCA. Le processus de construction d'une ontologie maritime comporte quatre étapes. 1-Calculez l'ontologie initiale à partir du thésaurus avec l'aide d'experts du domaine maritime. 2-Effectuer la création d'ontologies maritimes sur la base de la FCA. En utilisant la technologie NLP pour identifier les objets et les attributs du domaine, et construire ensuite le treillis de concepts. 3-Mappez les nouveaux concepts entre l'ontologie initiale et la nouvelle ontologie générée. 4- Ontologie résultante est décrite formellement en utilisant Protégé
Chien Duy et al [Ta et al., 2012]	Génère une ontologie à l'aide d'une valeur de seuil, d'un gain d'information et d'une Entropie. L'algorithme mesure le gain d'information pour chaque attribut à ajouter à une classe; si le gain d'information est supérieur au seuil, l'attribut sera ajouté.

Table 12.1 – Quelques méthodes de construction d'ontologies à l'aide de la FCA

13

La Véracité des conclusions

Contributions :

Yoan Chabot dans le cadre de sa thèse sur la criminalistique informatique (détails en section 15.4).

La thèse de Yoan CHabot portant sur la criminalistique s'alliait très bien avec mes besoins pour l'Intuition Artificielle. En effet, les résultats de cette thèse pouvant être exploités en tribunal, non seulement la véracité des données et conclusions devaient être prise en compte mais également prouvées et reproductibles.

Pour satisfaire aux exigences légales inhérentes au domaine, la plateforme SADFC (Semantic Analysis ofDigital Forensic Cases) créée s'appuie d'abord sur des fondements théoriques [Chabot et al., 2014] par des définitions formelles d'une scène de crime et des entités qui la composent afin de lever l'ambiguïté de ces concepts. Sur la base de cette formalisation, les opérateurs permettent de réaliser les différentes tâches composant la reconstruction d'événements, de l'extraction des empreintes à l'analyse de la chronologie. La nature formelle de ces opérateurs contribue à satisfaire le besoin de crédibilité en rendant leur opération explicite. Pour interfacer les opérateurs de reconstruction d'événements dans le processus plus large de l'enquête, ils sont intégrés dans un processus d'investigation. Ce modèle de processus décrit toutes les étapes composant l'investigation et répond ainsi au besoin de reproductibilité via une description explicite du processus d'investigation. De plus, l'ontologie proposée dans ces travaux intègre une couche dédiée à la représentation des activités au cours de l'enquête. Le stockage de la nature de chaque tâche, en plus des informations utilisées en entrée et en sortie, aide à comprendre comment chaque résultat est produit. Les enquêteurs peuvent utiliser cette fonctionnalité pour étayer leurs résultats avec des explications claires. Les principes et les outils développés dans les fondements théoriques de SADFC sont mis en œuvre sous la forme d'une architecture décrite dans [Chabot et al., 2015].

Voici les grandes lignes de la formalisation proposée dont les détails sont fournis dans les publications mentionnées et additionnés d'exemples illustratifs.

13.1/ DÉFINITION FORMELLE D'UNE SCÈNE DE CRIME.

Une scène de crime est un espace où un ensemble d'événements $E = \{e_1, e_2, ..., e_i\}$ a lieu. Un événement est une action unique se produisant à un moment donné et pour une certaine durée. Un événement peut être la rédaction d'un document, la lecture d'une page Web ou une conversation via un logiciel de messagerie instantanée.

Definition 51:

Sujet
Un sujet $x \in S$ correspond à une entité impliquée dans un ou plusieurs événements $e \in E$ et est défini par $x = \{a \in A_s | \ x\alpha_s \ a\}$ où:

- A_s est un ensemble contenant tous les attributs pouvant être utilisés pour décrire un sujet. Un attribut de sujet peut être le prénom et le nom de famille d'une personne, l'identifiant d'une session Web, le nom d'une session Windows, etc.
- α_s est la relation utilisée pour lier un sujet aux attributs de A_s qui le décrivent.

Definition 52:

Objet
Au cours de son cycle de vie, un événement peut également interagir avec des *objets*. Un objet peut être une page Web, un fichier ou une clé de registre par exemple. Un objet $x \in O$ est défini par $x = \{a \in A_o | \ x\alpha_o \ a\}$ où:

- A_o est un ensemble contenant tous les attributs pouvant être utilisés pour décrire un objet. Un attribut d'objet peut être par exemple un nom de fichier, l'emplacement de l'objet sur un disque dur, etc.
- α_o est la relation utilisée pour lier un objet aux attributs de A_o le décrivant.

Definition 53:

Evénement
Chaque événement se déroule dans un intervalle de temps défini par une heure de début et une heure de fin nécessitant l'introduction d'une algèbre spécifique. Dans nos travaux, l'algèbre d'Allen est utilisée. Un*événement* $e \in E$ est défini par $e = \{t_{dbut}, t_{fin}, l, S_e, O_e, E_e\}$ où:

- t_{dbut} est la date à laquelle l'événement commence.
- t_{fin} est la date à laquelle l'événement se termine.
- l est le lieu où l'événement se produit. Ce lieu peut être l'adresse IP ou l'adresse MAC de la machine sur laquelle l'événement a été initié.
- S_e est un ensemble contenant tous les sujets impliqués dans l'événement. $S_e = \{s \in S |\ e \in E,\ s\sigma_s\ e\}$ où σ_s est une relation composée utilisée pour lier un événement $e \in E$ à un sujet $s \in S$. La relation σ_s est définie ci-dessous.
- O_e est un ensemble contenant tous les objets liés à l'événement e. $O_e = \{o \in O |\ e \in E,\ e\sigma_o\ o\}$ où σ_o est une relation composée utilisée pour lier un événement $e \in E$ à un objet $o \in O$. La relation σ_o est définie ci-dessous.
- E_e est l'ensemble contenant tous les événements avec lesquels l'événement est corrélé. $E_e = \{x \in E |\ e \in E,\ e\sigma_e\ x\}$ où σ_e est une relation composée utilisée pour lier un événement $e \in E$ à un événement $x \in E$. La relation σ_e est définie ci-dessous.

13.2/ RELATIONS

Pour relier les entités présentées précédemment, quatre relations composées qui ont été introduites dans la partie précédente sont détaillées ici.

Definition 54:

Relations Sujets
σ_s est composée de deux types de relations pour lier un événement $e \in E$ à un sujet $s \in S$ et peut être définie de la manière suivante: $\sigma_S = s\ estImplique\ e \vee s\ subit\ e$

- *Relation de participation*: $s\ estImplique\ e$ signifie que s a été initié ou a été impliqué dans e.
- *Relation de répercussion*: $s\ subit\ e$ signifie que s est affecté par l'exécution de e.

Definition 55:

Relations d'objets
σ_o est composé de quatre types de relations pour lier un événement $e \in E$ à un objet $o \in O$ et peut être définie de la manière suivante: $\sigma_o = e\ cre\ o \vee supprime\ o \vee e\ o \vee e\ utilise\ o$:

- *Relation de création*: $e\ cree\ o$ signifie que o n'existe pas avant l'exécution de e et que o est créé par e.
- *Relation de suppression*: $e\ supprime\ o$ signifie que o n'existe plus après l'exécution de e et que o est supprimé par e.
- *Relation de modification*: $e\ modifie\ o$ signifie qu'un ou plusieurs attributs de o sont modifiés lors de l'exécution de e.
- *Relation d'utilisation*: $e\ utilise\ o$ signifie qu'un ou plusieurs attributs de o sont utilisés par e pour s'acquitter de sa tâche.

Definition 56:

Relations événementielles
σ_e est composée des relations utilisées pour lier deux événements $x, e \in E$ et peut être définie de la manière suivante: $\sigma_e = x\ compose\ e \vee e\ compose\ x \vee x\ cause\ e \vee e\ cause\ x$. Dans nos travaux, $x\ estCorrele\ e$ signifie que x est lié à e sur la base de critères multiples. On distingue deux cas particuliers de la relation de corrélation :

- *Relation de composition* : $x\ compose\ e$ signifie que x est un événement composant e. Soit $x = \{t_{x_{dbut}}, t_{x_{fin}}, S_x, O_x, E_x\}$ un événement composant $e = \{t_{e_{dbut}}, t_{e_{fin}}, S_e, O_e, E_e\}$, la relation de composition implique un ensemble de contraintes : e.g. des sous-événements ont lieu pendant l'événement parent. En utilisant les relations Allen, si $x\ compose\ e$ alors $egal(x, e)$ ou $pendant(x, e)$ ou $commence(x, e)$ ou $termine(x, e)$. Les sous-événements ont également des contraintes sur les sujets, objets et événements avec lesquels il est corrélé. Si $x\ compose\ e$ alors $S_x \subseteq S_e$ et $O_x \subseteq O_e$. Ainsi, $x\ compose\ e = [egal(x, e) \vee pendant(x, e) \vee commence(x, e) \vee termine(x, e)] \wedge (S_x \subseteq S_e) \wedge (O_x \subseteq O_e)$.
- *Relation de causalité* : $x\ cause\ e$ signifie que x doit arriver pour permettre à e de se produire. Un événement peut avoir plusieurs causes et peut être la cause de plusieurs événements. Soit $e = \{t_{e_{dbut}}, t_{e_{fin}}, S_e, O_e, E_e\}$ un événement provoqué par $x = \{t_{x_{dbut}}, t_{x_{fin}}, S_x, O_x, E_x\}$, la relation de causalité implique une contrainte temporelle imposant que la cause soit antérieure à la conséquence . En utilisant l'algèbre d'Allen, $x\ cause\ e = [avantque(x, e) \vee rencontre(x, e) \vee sechevauche(x, e) \vee commence(x, e)] \vee (S_x \cap S_e) \vee (O_x \cap O_e)$.

Definition 57:

Relation entre les empreintes.
σ_f est une relation utilisée pour relier une empreinte $f \in F$ à une entité $en \in \{E \times O \times S\}$. Cette relation s'appelle *Relation de support*: f $supporte$ en signifie que f est utilisé pour déduire un ou plusieurs attributs de en. Nous définissons une fonction $support$ qui peut être utilisée pour connaître les empreintes utilisées pour déduire une entité donnée: $support(en \in \{E \times O \times S\}) = \{f \in F | \ f\sigma_f \ en\}$. Par exemple, une entrée d'un historique Web peut être utilisée pour reconstruire un événement représentant la visite d'une page Web par un utilisateur.

Definition 58:

Scène de Crime
Dans nos travaux, nous définissons une *scène de crime* CS comme un environnement dans lequel un incident se produit et par $CS = \{PCS, DCS\}$ où:

- PCS est un ensemble contenant les *scènes de crime physique*. Au début d'une enquête, PCS est initialisé avec la localisation où l'incident a lieu. Toutefois, dans une enquête, la scène du crime ne se limite pas à un seul bâtiment.
- DCS est un ensemble contenant les *scènes de crimes numériques*.

Definition 59:

Actions
Soit A l'ensemble des actions telles que $A = \{I \cup L\}$ avec I l'ensemble de toutes les actions considérées comme des infractions par les lois et $L = A \setminus I$. Une scène de crime est également liée à des événements qui s'y déroulent $E_{CS} = \{E_{i_{CS}} \cup E_{c_{CS}} \cup E_{n_{CS}}\}$. Pour une notation plus facile, nous écrivons ici $E = \{E_i \cup E_c \cup E_n\}$ où:

- E_i est un ensemble contenant des événements *illicites* tels que $E_i = \{e \in E | \ e\sigma_i \ i, i \in I\}$ avec I un ensemble contenant toutes les actions considérées comme des infractions par les lois et σ_i la relation liant un événement à une action. Par exemple, un événement représentant le téléchargement de documents diffamatoires sur un site Web est un événement de E_i.
- E_c est un ensemble contenant les événements *corrélés* sous la forme $E_c = \{e \in E | \ e\sigma_i \ l, l \in L, e\sigma_e \ x, x \in E_i\}$. Cet ensemble contient tous les événements juridiques liés à un ensemble d'événements illicites x.
- $E_n = \{E \setminus (E_i \cup E_c)\}$ est un ensemble contenant les événements qui ne sont pas pertinents pour l'enquête.

13.3/ OPÉRATEURS FORMELS POUR LA RECONSTRUCTION D'ÉVÉNEMENTS

La reconstruction d'événements a pour but de passer d'une scène de crime statique contenant les traces d'événements passés à une chronologie décrivant les événements survenus dans le passé. L'objectif final de la reconstruction de l'événement est de tirer des conclusions en utilisant les informations contenues dans la chronologie produite.

Chaque événement est créé pour réaliser une action (par exemple, supprimer un fichier, modifier une clé de registre, charger une page Web, etc.). Les événements se déroulant sur une scène de crime numérique peuvent être classés comme suit: $E_{CS} = \{E_{i_{CS}} \cup E_{c_{CS}} \cup E_{n_{CS}}\}$. Pour une notation plus facile, nous écrivons ici $E = \{E_i \cup E_c \cup E_n\}$ où:

- E_i est un ensemble contenant des événements *Illicites*. Cet ensemble contient toutes les actions considérées comme des infractions par les lois.
- E_c est un ensemble contenant les événements *en corrélation* avec l'incident sous forme de $E_c = \{e \in E | e\sigma_E\ x, xE_i\}$. Cet ensemble contient tous les événements juridiques liés à un ensemble d'événements illicites x.
- $E_n = \{E \backslash (E_i \cup E_c)\}$ est un ensemble contenant les événements qui sont *Non pertinent* pour l'enquête.

Definition 60:

Empreinte.
Une empreinte est le signe d'une activité passée et une information permettant de reconstruire des événements passés. Une empreinte peut être une entrée de journal ou un historique Web. Soit F un ensemble contenant toutes les empreintes liées à un cas; une empreinte $x \in F$ est définie par $x = \{c \in C_F |\ x\gamma_F\ c\}$ où:

- C_F est un ensemble contenant toutes les informations effectuées par une empreinte. Par exemple, une entrée de journal d'un navigateur Web sur la visite de la page Web peut contenir le titre et l'URL de la page Web en plus de la date de la visite et du numéro de session de l'utilisateur.
- γ_F est la relation utilisée pour lier une empreinte avec les attributs de C_f utilisés pour la décrire.

Les relations σ_F permettent de lier une empreinte $f \in F$ à une entité $en \in \{E \times O \times S\}$. Cette relation est appelée relation de support: $f\ supporte\ en$ signifie que f est utilisé pour déduire un ou plusieurs attributs de en. Nous définissons une fonction $support$ qui peut être utilisée pour connaître les empreintes utilisées pour déduire une entité donnée:

$$support(en \in \{E \times O \times S\}) = \{f \in F |\ f\sigma_f\ en\}$$

Les empreintes sont les seules informations disponibles pour définir les événements passés et peuvent être utilisées par les enquêteurs pour reconstruire les événements survenus lors d'un incident. Cependant, la nature imparfaite et incomplète des empreintes peut conduire à des résultats approximatifs. Il n'est donc pas toujours possible de déterminer quel événement est associé à une empreinte donnée. De plus, il n'est pas toujours

possible de reconstruire complètement un événement à partir d'une empreinte. Ainsi, une empreinte peut être utilisée pour identifier une ou plusieurs fonctionnalités :

- Les caractéristiques temporelles ou l'emplacement d'un événement. Par exemple, chaque entrée d'un journal (pouvant être considérée comme une empreinte) fournit des informations temporelles permettant d'établir l'heure à laquelle l'action s'est produite. Les entrées peuvent également fournir des informations sur l'emplacement d'un événement, telles que le nom de la machine ou une adresse IP.
- Une relation entre un événement et un objet. Par exemple, une entrée d'un journal écrit par le système de fichiers peut donner des informations sur la modification d'un fichier. Par conséquent, un lien peut être établi avant l'événement de modification et le fichier.
- Une relation entre un événement et un sujet. Par exemple, toutes les empreintes générées par le navigateur Web Firefox sont stockées dans un dossier nommé avec le nom de profil de l'utilisateur. Cela permet de lier chaque événement Firefox à l'utilisateur désigné par ce nom.
- Caractéristiques d'un objet. Par exemple, une empreinte extraite des journaux des navigateurs Web peut être utilisée pour déterminer l'URL et le titre d'une page Web.
- Caractéristiques d'un sujet. Par exemple, les journaux du système d'exploitation peuvent être utilisés pour déterminer l'identifiant de session d'un utilisateur.

13.3.1/ OPÉRATEURS D'ANALYSE

Des opérateurs de mapping créent des entités (événements, objets et sujets) associées aux empreintes extraites. Ces opérateurs se présentent sous la forme de règles de mapping permettant de relier les attributs extraits d'empreintes à des attributs d'événements, d'objets et de sujets. Une grande partie des caractéristiques d'un événement peut être déterminée par les opérateurs d'extraction à partir des empreintes recueillies sur la scène du crime. Toutefois, l'identification de certains types de caractéristiques nécessite l'utilisation de techniques avancées telles que l'inférence.

Les opérateurs d'inférence permettent de déduire de nouvelles connaissances sur les entités à partir des connaissances existantes. Contrairement aux opérateurs d'extraction qui utilisent la connaissance des empreintes, les opérateurs d'inférence utilisent la connaissance des événements, des objets et des sujets (connaissances générées par les opérateurs de mapping). L'objectif de ces opérateurs est d'améliorer la connaissance que nous avons du passé afin d'améliorer l'efficacité de l'analyse ultérieure. En effet, plus les enquêteurs en savent, plus ils sont en mesure de tirer des conclusions précises et vraies.

13.3.2/ ANALYSE DE LA CHRONOLOGIE ET EXTRACTION DU SCÉNARIO D'INCIDENT

Après la spécification des événements, l'analyse de la chronologie vise à identifier le scénario de l'incident. Identifier un incident signifie identifier tous les événements $E_{inc} = E_i \cup E_c$ en utilisant les empreintes de E où E_inc est un ensemble contenant tous les

événements directement *liés à l' INCident*. Les événements classés chronologiquement décrivant un incident sont appelés le scénario de l'incident. Les opérateurs d'analyse sont utilisés pour aider les enquêteurs lors de l'interprétation de la chronologie et de la reconstruction du scénario de l'incident. Ces opérateurs sont utilisés pour identifier les relations entre les événements et pour mettre en évidence les informations pertinentes de la chronologie. Nous proposons un seul outil d'analyse permettant d'identifier les corrélations entre les événements. L'identification des événements illicites est un autre aspect important de l'analyse. Cependant, il n'a pas été possible d'introduire un tel outil dans le temps alloué à cette thèse. Ainsi, cette étape doit être effectuée manuellement par l'utilisateur.

Le premier outil d'analyse proposé dans notre approche est un processus permettant de détecter une corrélation entre deux événements. La corrélation est admise comme une relation avec une vaste sémantique qui couvre les relations de causalité et d'autres liens sémantiques. L'identification de telles relations revêt une importance particulière pour les enquêteurs car elle leur permet d'avoir rapidement une vue d'ensemble des événements qui composent un incident et des liens qui les unissent.

La corrélation entre deux événements e, xE est mesurée par la fonction suivante:

Definition 61:

Corrélation entre événements.

$$Corr(e,x) = \begin{cases} \textbf{If } Corr_{KBR}(e,x) = 1 \textbf{ then } 1 \\ \textbf{else } \frac{\alpha * Corr_T(e,x) + \beta * Corr_S(e,x) + \gamma * Corr_O(e,x)}{3} \end{cases}$$

$Corr_T(e,x)$, $Corr_S(e,x)$ et $Corr_O(e,x)$ sont normalisées pour obtenir une valeur comprise entre 0 et 1. $Corr_T(e,x)$, $Corr_S(e,x)$ et $Corr_O(e,x)$ peuvent être pondérées pour permettre de donner plus d'importance à l'une des fonctions de corrélation en utilisant les facteurs α, β et γ.

Si $Corr_{KBR}(e,x) = 1$, ce qui signifie qu'une règle basée sur les connaissances est satisfaite, alors $Corr(e,x) = 1$. Contrairement aux autres facteurs de corrélation, $Corr_{KBR}(e,x)$ est basé sur des connaissances définies par des experts. Cette connaissance étant fiable, lorsqu'une règle de l'ensemble est satisfaite, on peut considérer que les deux événements sont corrélés. $Corr(e,x)$ peut également être ordonné et doté d'un seuil pour traiter les contraintes de volume de données en sélectionnant les corrélations les plus significatives. Ces quatre corrélations sont décrites de la manière suivante:

Corrélation temporelle, $Corr_T(e,x)$ Tout d'abord, un ensemble d'hypothèses concernant l'aspect temporel est défini (selon l'algèbre d'Allen) :

- Plus la différence relative entre les deux événements $avant(e,x)$ est grande, plus la relation temporelle est basse et réciproquement.

- La relation temporelle est maximale (égale à 1) pour les fonctions $rpond(e,x)$, $sechevauche(e,x)$, $pendant(e,x)$, $termine(e,x)$, $commence(e,x)$ et $estegal(e,x)$.

Definition 62:

Corrélation temporelle.

$$Corr_T(e,x) = commence(e,x) + estegal(e,x) + rencontre(e,x) + sechevauche(e,x) + pendant(e,x) + termine(e,x) + avant(e,x) \quad (13.1)$$

où $commence(e,x)$, $estegal(e,x)$, $rencontre(e,x)$, $sechevauche(e,x)$, $pendant(e,x)$, $termine(e,x)$ sont des fonctions binaires et $avant(e,x) = \frac{1}{(x_{t_{dbut}} - e_{t_{fin}})}$.

Les hypothèses précédentes indiquent que plus deux événements sont proches dans le temps, plus il est probable que ces événements soient corrélés. En raison des granularités temporelles et des ordinateurs multitâches, si deux événements commencent en même temps, la relation est maximale.

Corrélation de sujets, $Corr_S(e,x)$ Ce score quantifie la corrélation entre deux événements concernant les sujets impliqués dans chaque événement. Les hypothèses suivantes sont définies en fonction de l'idée de base du domaine de l'exploration de données. Par exemple, l'Analyse Formelle de Concepts [**?**] regroupe des objets en fonction des attributs qu'ils partagent. De la même manière, en statistique, l'analyse en composantes principales regroupe les observations mesurant leur étendue (variance). La corrélation entre e et x augmente proportionnellement au nombre de sujets communs qu'ils partagent concernant les relations de *participation* et de *répercussion*.

Definition 63:

Corrélation de sujets.
$Corr_S(e,x) = \frac{|S_e \cap S_x|}{max(|S_e|,|S_x|)}$

Corrélation d'objets, $Correlation_O(e,x)$ Ce score quantifie la corrélation entre deux événements concernant les objets utilisés, générés, modifiés ou supprimés par les événements. L'hypothèse suivante est basée sur la même idée de base que la corrélation de sujets. La corrélation entre e et x augmente proportionnellement au nombre d'objets qu'ils partagent en ce qui concerne les relations de *création*, *suppression*, *modification* et *utilisation*. Nous ne considérons pas uniquement les objets communs, mais également les objets ayant des emplacements virtuels similaires. Par conséquent, le score de corrélation des objets augmente si deux événements interagissent avec un même objet ou si deux événements interagissent avec deux objets différents situés à proximité. Par exemple, si deux événements interagissent avec deux objets différents situés tous deux dans `C:\ProgramFiles(x86)\Adobe`, nous pouvons admettre que les deux événements sont liés car ils interagissent tous les deux avec des objets liés à *Adobe* outils. Étant donné deux objets $o_1, o_2 \in O$ situés virtuellement dans $l_{o1}, l_{o2} \in L$, la distance entre ces objets est donnée par la formule suivante:

$$d(o_1,o_2) = \frac{|l_{o1}| + |l_{o2}| - 2 * |l_{o1} \cap_H l_{o2}|}{|l_{o1}| + |l_{o2}|} \quad (13.2)$$

L'intersection $l_1 \cap_H l_2$ entre deux emplacements virtuels $l_1, l_2 \in L$ est définie comme suit:

$$l_1 \cap_H l_2 =< h_{l_1 r} \in H \mid \leq_H r \in [0;\ max_i(h_{l_1 i}, h_{l_2 i}) \mid \nexists\, h_{l_1 j} \neq h_{l_2 j},\ j \leq i] > \tag{13.3}$$

Definition 64:

Corrélation d'objets.

$$Corr_O(e,x) = \begin{cases} \textbf{If } \sum_{i=0}^{n}\sum_{j=0}^{m} d(O_{ei}, O_{xj}) = 0 \textbf{ then } 1 \\ \textbf{else } \frac{n+m}{\sum_{i=0}^{n}\sum_{j=0}^{m} d(O_{ei}, O_{xj})} \end{cases}$$

avec $n = |O_e|$, le nombre d'objets interagissant avec l'événement e et $m = |O_x|$ le nombre d'objets interagissant avec l'événement x.

Corrélation basée sur les règles, $Corr_{KBR}(e,x)$ En plus des facteurs précédents (heure, sujet, objet), des règles basées sur des connaissances expertes peuvent être utilisées pour corréler des événements.

Definition 65:

Corrélations sur les règles.

$Corr_{KBR}(e,x) = 1\ si\ \sum_{r=1}^{n} regle_r(e,x) > 0$

avec $regle_r(e,x) = 1$ si la règle est satisfaite et 0 sinon.

13.4/ CONCLUSION

La crédibilité des résultats et leur reproductibilité sont deux critères importants pour garantir l'admissibilité des preuves au tribunal. Dans le but de créer des outils qui répondent à ces exigences légales inhérentes au domaine de la criminalistique numérique, ce chapitre introduit les fondements théoriques de l'approche SADFC. Premièrement, la scène du crime et les entités qui le composent ont été définies afin de lever l'ambiguïté des notions impliquées dans ce travail. Deuxièmement, il propose une formalisation du problème de reconstruction d'événements. Sur la base de cette formalisation, des opérateurs nécessaires à la reconstruction ont été définis. L'introduction d'une telle formalisation vise à accroître la crédibilité des résultats en expliquant mathématiquement comment ils sont produits.

BILAN DE LA PARTIE SÉMANTIQUE DE L'INTUITION ARTIFICIELLE

Dans ce chapitre, je montre qu'il est possible de construire automatiquement une ontologie à partir de connaissances formées par la FCA, ce qui permet d'assurer une compréhension et une expression des conclusions de l'Intuition Artificielle émises à partir du même jeu de connaissances.

Pour réaliser cette automatisation, le nommage des classes de l'ontologie est nécessaire et je montre comment qualifier les concepts de la FCA pour les traduire en classes.

Le processus de construction automatique des ontologies implique trois verrous : la génération des classes et leur hiérarchie, la création des règles d'inférence et la généra-

tion des caractéristiques relationnelles et des restrictions (i.e. le verrou ”contraintes“). Les deux premiers verrous ont été levés et le troisième sera traité dans le cadre d'un projet ultérieur.

Ainsi pour synthétiser l'avancée des verrous identifiés pour cette partie, à savoir : la qualification, l'automatisation, l'inférence, et la véracité; la qualification et la véracité sont résolues, l'automatisation l'est au 2/3 car il manque le verrou des contraintes. Et enfin, l'inférence sera prise en charge d'un projet industriel qui va débuter en 2020 dans le but de créer un moteur d'inférence multidimensionnelle flou.

CONCLUSION ET TRAVAUX À VENIR

C'est avec la logique que nous prouvons et avec l'intuition que nous trouvons.
Henri Poincaré - *(1854 - 1912)*

Les propos d'Henri Poincaré soulignent un positionnement largement partagé de l'intuition humaine. Mais un peu plus d'un siècle plus tard, avec l'informatique, nous pouvons concilier les deux comme je le montre dans ce manuscrit : doter la machine d'une faculté à subodorer qui s'appuie sur la logique et dont les conclusions peuvent être expliquées et prouvées : l'Intuition Artificielle.

Pour cela j'ai défini une approche en quatre étapes en précisant les verrous pour chacune d'elle comme le présente la figure 13.1. Certains de ces verrous sont levés (mentionnés en vert sur la figure), d'autres sont en cours de traitement (en orange) et les derniers sont envisagés pour des projets qui débuteront ultérieurement, notamment en 2020.

Ce travail de recherche a trouvé son origine dans des travaux personnels que j'ai eu nécessité à étendre. Pour se faire, j'ai monté différents travaux recherches académiques et industriels pour lever progressivement ces verrous.

Ainsi sur la première étape de l'approche, concernant les **données**, les verrous de volume et de pertinence ont pu être traités dans le cadre de la thèse de Thomas Hassan par le développement d'un crawler web capable d'arpenter internet à la recherche d'information ciblées en évaluant leur pertinence par rapport au domaine concerné, et de gérer le volume des données ainsi collectées.
Le troisième verrou de cette étape est l'hétérogénéité des données collectées. La thèse de David Werner a proposé de solutionner ce verrou par l'utilisation de facettes et de deux modèles : unificateur et intégrateur.

Les données collectées par la première étape sont ensuite converties en **connaissances** à l'étape suivante. Pour cela les méthodes d'Analyse Formelle de Concepts ont été mises en oeuvre car elle permettent de générer des concepts et leur hiérarchie permettant d'une part de regrouper les données ayant du sens ensemble, mais aussi de pouvoir convertir cette structure en ontologie lors de la quatrième étape du processus.
Les verrous rencontrés pour adapter la FCA aux données collectées ont été :

- la prise en considération de la multidimensionnalité de la FCA qui a été réalisée par la création de l'outil MARM avec ses preuves mathématiques. Le projet Eurostars PSDP, m'a permis le recrutement en post doctorat d'Alexandre Bazin pour que nous travaillons sur cette problématique.

- le volume qui est en cours d'achèvement par l'adaptation de Quality Cover en n-dimensions afin de limiter le nombre de concepts en sortie tout en maximisant la couverture de la connaissance conservée. Le projet FUI WineCloud m'a permis de recruter en post doctorat Amira Mouakher qui a conçu Quality Cover pendant sa thèse.

- Enfin le dernier verrou : celui du flou est un verrou qui n'a pas encore été abordé. Un projet débutant en 2020 s'intéressera à cette problématique.

La troisième étape de cette approche présente le moteur de l'**Intuition Artificielle**, basé sur un paradigme de graphes. Le premier verrou de cette approche a été de définir / créer l'Intuition Artificielle. Pour lever ce verrou, je présente une approche formelle basée sur un polynôme de matrices permettant de mettre en évidence les relations les plus fortes n'existant pas dans le graphe de connaissances et ainsi de souligner que quelque chose ne va pas. Enfin je présente des algorithmes pour implémenter cette approche. Ils sont fonctionnels et ont été tester avec succès sur DBLP et sur un jeu de données médicales OrphaData. Cependant ce verrou n'est pas marqué comme complètement levé, dans le sens où d'autres algorithmes (non encore évalués dans l'approche) peuvent être utilisés pour diminuer la complexité.

Enfin la dernière étape consiste à **comprendre et exprimer** l'Intuition Artificielle. En effet cette dernière identifie une incohérence et cherche à la transmettre à l'utilisateur. Pour cela elle doit pouvoir comprendre les relations impliquées, et ensuite elle doit les expliquer. Les ontologies sont idéales pour ce genre de traitement. Cependant, pour comprendre, la machine doit connaître les connaissances qui ont servit à l'Intuition Artificielle donc l'ontologie doit être construite à partir de ces connaissances. Un premier verrou s'impose : automatiser le processus de construction de l'ontologie à partir des connaissances. Ce verrou se décompose en trois éléments : les deux premiers sont résolus, à savoir la transposition des concepts et diagramme de Hasse en classes et leurs relations de subsomption ainsi que la conversion des règles d'association en règles d'inférence. Cependant la qualification des caractéristiques relationnelles et des restrictions (existentielle et universelle) à partir du treillis n'ont pas encore été formalisées. Cette étape est prévue dans un projet en cours de rédaction.
Une problématique de qualification des concepts a du être préalablement résolu. Un projet industriel QAPE, m'a permis le recrutement en post doctorat de Quentin Brabant, et j'ai accepté d'encadrer un stage de thèse de l'université de Sfax pour Wided Selmi pour lever ce verrou.
Le troisième verrou de cette partie, porte sur la véracité des conclusions de l'Intuition Artificielle. Pour cela nous avons montré dans la thèse de Yoan Chabot que nous pouvons appuyer les règles d'inférence de l'ontologie sur des fonctions formelles.
Enfin le verrou de l'inférence concerne la création d'un moteur d'inférence multidimensionnel flou de façon à ce que toutes les étapes du processus soient cohérentes. Malheureusement ce genre de moteur d'inférence n'existe pas, mais un projet incluant deux post doctorants va débuter en 2020 pour créer ce moteur.

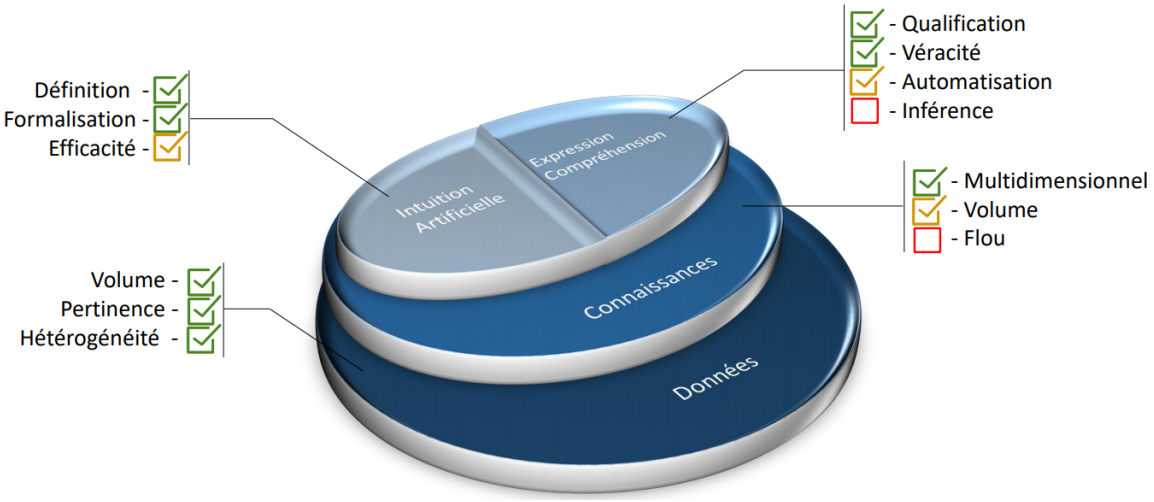

Figure 13.1 – Les étapes vers l'Intuition Artificielle et la résolution de leurs verrous

La voie de l'Intuition Artificielle est déjà bien tracée. Une partie du chemin a été parcourue et la suite est en bonne partie balisée. Progressivement ce domaine va se solidifier et la lecture de ce manuscrit pourra peut être aider ceux qui voudraient y contribuer ou confronter leurs idées à celles déjà émises.

V

ANNEXES

14

PROJETS DE RECHERCHE

L'HDR démontrant la capacité à diriger des recherches, je présente ici les projets pour lesquels j'avais (ou j'ai) la responsabilité scientifique. A ce titre je m'assure du recrutement et du suivi des post doctorants, je définis et valide les axes de recherches et assure la relecture et rédaction des articles scientifiques. D'autres collègues participent à ces projets qui sollicitent souvent plus d'un domaine de compétence. A ce titre leur expertise est sollicitée.

14.1/ PROJETS SCIENTIFIQUE AVEC FINANCEMENTS PUBLICS

14.1.1/ WINECLOUD - PROJET FUI

Titre du projet : Traçabilité et analyse Big Data sur toute la chaîne de valeur du vin

Durée du projet : 2017-2020

Consortium : Orange, Les Caves de Lugny, L'Université Bourgogne Franche-Comté et Photon Lines. Labelisé par le pôle de compétitivité Vitagora (région Bourgogne Franche Comté) et le pôle de compétitivité Images et Réseaux (région Bretagne).

Budget : 3.6 millions euros

Projet : Une des particularités de ce projet est de proposer une plateforme Big Data dédiée, confidentielle et sécurisée, ouverte à tous les acteurs de la chaîne de valeur vitivinicole, centralisant les données de traçabilité sur tout le cycle de vie du produit, favorisant les échanges entre producteurs et consommateurs mais aussi permettant d'optimiser les process de production ou de distribution. Cette plateforme doit permettre à chacun des acteurs de la chaîne de :

- gérer sa traçabilité (et celle de ses associés) de façon confidentielle et/ou partagée ;
- permettre aux professionnels de la filière d'alimenter la plateforme avec ses propres connaissances et données vitivinicoles, grâce à la boucle ouverte (agilité permettant à chacun d'adhérer à tout moment) ;

- intégrer les données liées à la vie du produit après la vente, et ce grâce à la remontée d'informations consommateurs.

Implication : Dans ce projet nous utilisons des méthodes de fouilles de données afin de comprendre et prédire les facteurs déclenchant de certains problèmes dans les vignes telles que les maladies ou les ravageurs.

Publications : [Hugol-Gential et al., 2019b,Mouakher et al., 2019,Belkaroui et al., 2019, Hugol-Gential et al., 2019a, Hugol-Gential et al., 2019c, Belkaroui et al., 2018]

14.1.2/ PSDP (PREDICTIVE SMART DATA PLATFORM) - EUROSTAR PROJECT

Titre du projet : Développement d'une plateforme prédictive de données intelligente pour une adaptabilité en temps réel

Durée du projet : 2016-2019

Consortium : Kernix Software et Erowz

Budget : 280 000 euros

Projet : Des volumes massifs de données structurées et non structurées sont créés à chaque interaction et le défi de les exploiter pour prédire des événements et des comportements constitue une opportunité majeure pour l'industrie et Internet. Pour acquérir un avantage concurrentiel et maximiser la valeur client, les entreprises doivent être en mesure d'exploiter le potentiel du Big Data et de l'utiliser de manière intelligente. Ce projet a offert une plateforme de données intelligente prédictive de nouvelle génération aux principaux responsables des systèmes d'information et des services techniques des grandes entreprises industrielles ainsi qu'aux grands propriétaires de sites Web. La plateforme rassemble de grandes quantités de données hétérogènes provenant de différents silos de données, capteurs, objets connectés, réseaux sociaux, etc. La cohérence des données est assurée au niveau du stockage, grâce à une technologie basée sur les graphes, apportant une nouvelle valeur grâce à l'analyse approfondie des relations et des corrélations. Cela permet également de débloquer de nouvelles utilisations prédictives qui sont impossibles autrement, en offrant de riches fonctionnalités pour les algorithmes prédictifs et l'apprentissage automatique.

Implication : Nous avons mis en place un système de profilage dynamique des petites annonces et des utilisateurs de telle sorte que ces profils s'influencent mutuellement.

Publications : [Bazin et al., , Bazin et al., 2018]

14.1.3/ QAPEMINING

Titre du projet : Développement d'une plateforme comparative d'assurance

Durée du projet : 2018-2020

Consortium : QAPE

Budget : 380 000 euros

Projet : QAPE veut développer son propre graphe sémantique du domaine de l'assurance. Ce graphe, initialement programmé pour représenter les conditions contenues dans les tableaux de garantie d'assurance maladie, sera ensuite étendu à d'autres domaines de l'assurance. En outre, le projet explore des sources de données ouvertes sur le Web (”opendata“) afin de créer des profils de personnes selon leur conditions d'assurance. Enfin, le projet identifie les liens de causalité entre les profils et les éléments contenus dans les tableaux de garantie afin de vérifier si la couverture de risque décrite dans son tableau de garantie est appropriée.

Implication : Ce projet nécessite beaucoup de corrélations de données pour comprendre et prédire si des types de soins peuvent en entrainer d'autres.

14.1.4/ SMARTTEAM PROJECT

Titre du projet : Développement d'un outil pour bâtir une équipe de travail

Durée du projet : 2015

Consortium : Mythmakers (Lauréat de la bourse de la French Tech)

Budget : 15 000 euros

Projet : Développer un outil de profilage sémantique du comportement des membres souhaitant participer à des projets de manière collaborative. L'objectif était de construire les groupes les plus complets en termes de récupération des compétences en associant raisonnement sémantique et métaheuristique.

Implication : Des techniques de fouille de données par recherche de corrélations ont été mis en oeuvre.

14.1.5/ WEBDRONE

Titre du projet : Identification de produits de luxe contrefaits

Durée du projet : 2014

Consortium : Webdrone, Département de recherche VISION du LE2I

Budget : 15 000 euros

Projet : Dans le cadre de la croissance et de la diffusion du Web, il y a une augmentation du nombre d'images sur Internet: la possibilité d'identifier ces images est donc un problème majeur. L'objectif était, d'une part, d'identifier des objets et, d'autre part, d'identifier leur marque. Pour cela, nous nous sommes appuyés sur des techniques d'analyse d'images. Parmi ces techniques, nous nous sommes intéressés à la représentation d'images grâce à *sacs de mots visuels*, à la méthode de la machine à vecteurs de support (SVM) et l'Analyse Formelle de Concept.

Implication : La combinaison des sacs de mots visuels et l'analyse de concepts formels est originale car elle n'avait jamais été faite auparavant et que les résultats étaient très intéressants

Publications : [Benezeth et al., 2015]

15

THÈSES ENCADRÉES

Dans tous mes encadrements j'ai notamment participé à l'écriture des articles soumis, même si je ne le note pas à chaque description ci-dessous. Chaque semaine nous avons une réunion et les étudiants produisent un rapport d'activité. Chaque développement et état de l'art donne lieu à un rapport technique.

15.1/ MOHAMMED DOUAD-TAOUFIK

SUJET : CONCEPTION AUTOMATIQUE D'ONTOLOGIE À PARTIR DE DONNÉES BINAIRES ET FLOUES DONT LA CONNAISSANCE EST EXTRAITE PAR UNE MÉTHODE D'ANALYSE FORMELLE DE CONCEPTS (AFC).

Dans le cadre de nos travaux, nous nous intéressons plus particulièrement à l'étude des logiques descriptives et de l'analyse formelle de concepts (AFC) afin de créer le lien entre ces deux représentations (proches) de données. En effet, le premier vise à représenter à l'aide de formules la connaissance d'un domaine de métier dans un univers de discours. Tandis que le second vise la découverte de connaissances à l'aide d'un outil mathématique basé sur des mathématiques ensemblistes. Les ontologies sont construites par des experts possédant ces connaissances et permettent l'inférence sur ces données pour déduire de nouvelles connaissances. L'AFC génère des concepts à partir de données et les hiérarchise mais ne permet pas d'inférence. Cette hiérarchie suit cependant un modèle de subsomption similaire à la celui des ontologies. L'intérêt de cette thèse est donc de construire une ontologie non pas à partir des connaissances d'un expert, mais à partir des connaissances issues des données grâce à l'AFC. Malheureusement l'étudiant n'a pas été honnête et n'était intéressé que pour une présence en France et n'a pas travaillé sur sa thèse. Nous avons donc du renoncer à cette thèse après de nombreuses tentatives pour le motiver et négociations via l'école doctorale.

ENCADREMENT

Co-encadrement (50%) de la thèse de Mohamed Douad-Taoufik sur une conception automatique d'ontologie à partir de données binaires et floues dont la connaissance est extraite par une méthode d'analyse formelle de concepts (AFC). Directeur : Christophe Cruz. (Commencée en 2015 abandonnée en 2016).

Financement de son employeur. Mon implication porte sur la partie AFC et le suivi hebdomadaire de recherche, ainsi que des démarches administratives.

15.2/ MARWAN BATROUNI

SUJET : ARCHITECTURE BASÉE SUR LA SÉMANTIQUE POUR UN CADRE DE CONSTRUCTION DE SCENARII.

Le modèle prédictif est typiquement l'un des défis les plus avancés du data mining. Toute inexactitude dans la construction du modèle va impacter la qualité de la prédiction. Une particularité du Big Data dans ce contexte n'est pas seulement le volume des données à traiter mais l'extrême diversité et rapidité d'évolution des sources de données qui l'alimentent. Le modèle prédictif s'appuie sur la notion de résolution de scénarii. Or dans ce domaine, la notion de scénario n'est pas formellement définie. Notre objectif est de définir formellement cette notion de façon a pouvoir proposer une plateforme de prédiction pouvant s'adapter à tout contexte.

ENCADREMENT

Co-encadrement (50%) de la thèse de Marwan Batrouni sur un modèle sémantique intelligent de simulation de scenarii dans le domaine du BigData. Directeur : Christophe Nicolle. (Soutenue en 2018).

Financement personnel. Il travaille pour une compagnie d'assurance (Vertafore) aux Etats Unis et mène sa thèse en parallèle. Mon implication porte d'une part sur le suivi administratif et sur le maintien des lignes directrices du projet de thèse car le doctorant a tendance à dévier en allant de pistes intéressantes en pistes prometteuses. Je dirige également la partie formelle du projet sur la définition de la notion de scénario et d'évolution du système par transitions d'états.

15.3/ THOMAS HASSAN

SUJET : APPROCHE BIG DATA ET WEB SÉMANTIQUE POUR LA FOUILLE ET LA CLASSIFICATION DE DONNÉES WEB AUTOMATIQUE.

Au cours de l'année 2014-2015, la société First-Eco était particulièrement intéressée quant à l'opportunité d'améliorer le système de recommandation issu de la thèse de David Werner par l'ajout d'un nouveau web service permettant de crawler le web pour la recherche et le recoupement des nouvelles économiques. De nouvelles contraintes ont été fixées pour l'élaboration de l'approche : le prototype devant fournir une solution d'analyse de données passant l'échelle pour déterminer la Valeur des données dans un contexte Big Data. La connexion entre les outils de Big Data et le Web sémantique, adapte le processus pour traiter ces questions. Pour valider l'approche, le système est évalué en fonction de deux critères : la qualité des résultats (données pertinentes extraites par le processus), et de performance. Le deuxième point est un aspect crucial et justifie la nécessité d'une architecture Big Data.

ENCADREMENT

Co-encadrement (50%) de la thèse de Thomas Hassan sur une approche Big Data et Web Sémantique pour la fouille et la classification de données web automatique. Directeur : Christophe Cruz. (Soutenue en novembre 2017).

Financement régional par une bourse JCE (Jeune Chercheur Entrepreneur). Ma contribution au-delà de l'administratif porte sur la gestion des données et les méthodes de paramétrage automatique et d'évaluation de la plateforme développée.

15.4/ YOAN CHABOT

SUJET : CONSTRUCTION, ENRICHMENT AND SEMANTIC ANALYSIS OF TIMELINES : APPLICATION TO DIGITAL FORENSICS.

Cette thèse s'intéresse à la construction et l'analyse de chronologies d'évènements pour répondre à des questions posées sur des actions passées, présentes ou futures. Elle permet de déterminer les circonstances des évènements ainsi que leurs causes et leurs effets. L'analyse d'évènements est un problème complexe trouvant des applications dans de nombreux domaines et l'étude de chronologies est un problème tout particulièrement significatif pour le domaine de la criminalistique informatique. Ce dernier a pour objectif d'assister les enquêteurs de la police judiciaire dans la résolution d'enquêtes pour lesquelles les nouvelles technologies ont été utilisées comme vecteur ou pris pour cible. Lors d'une enquête, les investigateurs tentent de reconstruire, à partir des informations laissées dans une scène de crime par les protagonistes, les évènements survenus lors d'un incident. Cette étape est primordiale pour une enquête numérique pour déterminer la nature de l'incident et ses caractéristiques et comprendre les circonstances de l'incident et la responsabilité des protagonistes à partir de la description des évènements passés.

ENCADREMENT

Co-encadrement (40%) de la thèse de Yoan Chabot en co-tutelle avec l'University College of Dublin « Construction, enrichment and semantic analysis of timelines: Application to digital forensics ». Directeurs : Christophe Nicolle (France) et Tahar Kechadi (Irlande). (Soutenue en novembre 2015).

Financement régional par moitié et par l'UCD pour l'autre moitié. Ce doctorant étant très autonome, mon encadrement a été principalement scientifique. Le système corrèle des événements par rapprochement d'objets, de sujets, de données temporelles... Mon encadrement s'est porté majoritairement sur la formalisation de ces corrélations. Par ailleurs, nous sommes souvent sollicités comme experts ou partenaires sur ces problématiques. Je m'investis donc dans ces rôles. Yoan a été embauché avant même sa soutenance chez Orange Lab.

15.5/ DAVID WERNER

SUJET : INDEXATION ET RECOMMANDATION D'INFORMATIONS : VERS UNE QUALIFICATION PRÉCISE DES ITEMS PAR UNE APPROCHE ONTOLOGIQUE, FONDÉE SUR UNE MODÉLISATION MÉTIER DU DOMAINE.APPLICATION À LA RECOMMANDATION D'ARTICLES ÉCONOMIQUES.

La gestion efficace de grandes quantités d'informations est devenue un enjeu de plus en plus important pour les systèmes d'information. De nouvelles sources d'information quotidienne apparaissent sur le web. On peut facilement trouver ce qu'on veut si on cherche un article, une vidéo ou un artiste en particulier. Cependant, il devient très difficile, voire impossible, de tout explorer pour découvrir de nouveaux contenus. Les systèmes de recommandation sont des outils logiciels qui visent à aider les humains pour faire face à la surcharge d'information. Le travail présenté dans cette thèse propose une architecture de recommandation efficace des nouvelles économiques. Notre approche ontologique repose sur un modèle pour la caractérisation précise des éléments basés sur un vocabulaire contrôlé. L'ontologie contient un vocabulaire formel modélisant une vue du domaine de connaissances. En collaboration avec la société Actualis SARL, ce travail a conduit à la commercialisation d'un nouveau produit très compétitif, First ECO Pro'fil.

ENCADREMENT

Co-encadrement (50%) de la thèse de David Werner sur l'indexation et recommandation d'informations : vers une qualification précise des items par une approche ontologique fondée sur une modélisation métier du domaine. Directeur : Christophe Cruz. (Soutenue en juin 2015).
Financement CIFRE pour la société Actualis. David était un excellent ingénieur mais la rédaction a toujours été compliquée. Mon implication a donc beaucoup été à le former à cette tâche et le motiver sur ce point. Dans cette thèse, comme les autres, je l'ai orienté sur la formalisation des méthodes, et notamment sur les notions de similarités asymétriques.

16

SUPERVISION DE THÈSES

16.1/ PATRICIA LOPEZ-CUEVA

SUJET : DEBUGGING EMBEDDED MULTIMEDIA APPLICATION EXECUTION TRACES THROUGH PERIODIC PATTERN MINING.

Cette thèse en collaboration avec la société STMicroelectronics s'est intéressée à l'identification de patterns périodiques dans des traces d'exécution de processus. Ces traces sont un outil puissant très utilisé pour corriger des systèmes embarqués. Cependant le volume de ces traces est très important rendant leur analyse très complexe. Afin de proposer une solution, nous nous sommes intéressés à la découverte de comportement périodiques des applications par des méthodes de pattern mining dans les traces d'exécution. Malheureusement les définitions existantes ne prenaient pas en compte des espaces de taille variable dans la périodicité. Dans cette thèse nous avons proposé une nouvelle définition des patterns périodiques fréquents permettant de lever ce verrou. Nous avons également proposé un opérateur de fermeture qui réduit de manière significative le nombre de patterns périodiques obtenus. Plusieurs expérimentations ont été menées sur des jeux de données synthétiques pour démontrer cette réduction de patterns et nous avons découvert des comportements anormaux par l'application de nos méthodes sur des applications embarquées.

SUPERVISION

Supervision de la thèse de Patricia Lopez-Cueva sur l'étude de patterns périodiques sur des traces d'exécution d'applications multimédia embarquées. Directeur : Jean-François Méhaut. (Soutenue en juillet 2013).
Financement CIFRE pour la société STMicroElectronics. Cet encadrement s'est attaché à la rédaction d'articles et à la définition formelle de la périodicité cyclique d'un événement quelque soit la variabilité des espaces entre deux cycles. En sortant de sa thèse Patricia a été embauchée chez Thales Alenia Space, où elle est ingénieure R&D.

16.2/ SOFIANE LAGRAA

SUJET : NEW MP-SOC PROFILING TOOLS BASED ON DATA MINING TECHNIQUES.

La miniaturisation des composants électroniques a conduit à l'introduction de systèmes électroniques complexes qui sont intégrés sur une seule puce avec multiprocesseurs, dits Multi-Processor System-on-Chip (MPSoC). La majorité des systèmes embarqués récents sont basées sur des architectures massivement parallèles MPSoC, d'où la nécessité de développer des applications parallèles embarquées. La conception et le développement d'une application parallèle embarquée devient de plus en plus difficile notamment pour les architectures multiprocesseurs hétérogènes ayant différents types de contraintes de communication et de conception tels que le coût du matériel, la puissance et la rapidité. Un défi à relever par de nombreux développeurs est le profilage des applications parallèles embarquées afin qu'ils puissent passer à l'échelle sur plusieurs cœurs possibles. Cela est particulièrement important pour les systèmes embarqués de type MPSoC, où les applications doivent fonctionner correctement sur de nombreux cœurs. En outre, la performance d'une application ne s'améliore pas forcément lorsque l'application tourne sur un nombre de cœurs encore plus grand. La performance d'une application peut être limitée en raison de multiples goulot d'étranglement notamment la contention sur des ressources partagées telles que les caches et la mémoire. Cela devient contraignant et une perte de temps pour un développeur de faire un profilage de l'application parallèle embarquée et d'identifier des goulots d'étranglement dans le code source qui diminuent la performance de l'application.

SUPERVISION

Pour surmonter ces problèmes, dans cette thèse, nous proposons trois méthodes automatiques qui détectent les instructions du code source qui ont conduit à une diminution de performance due à la contention et à l'évolutivité des processeurs sur une puce. Les méthodes sont basées sur des techniques de fouille de données exploitant des gigaoctets de traces d'exécution de bas niveau produites par les platesformes MPSoC. Sofiane a commencé sa thèse à la fin de mon contrat post doctoral. Ma contribution a portée sur certaines de ces techniques de fouille.

16.3/ BENJAMIN NEGREVERGNE

SUJET : UN ALGORITHME DE FOUILLE DE DONNÉES GÉNÉRIQUE ET PARALLÈLE POUR ARCHITECTURE MULTI-COEURS.

Dans le domaine de l'extraction de motifs, il existe un grand nombre d'algorithmes pour résoudre une large variété de sous problèmes sensiblement identiques. Cette variété d'algorithmes freine l'adoption des techniques d'extraction de motifs pour l'analyse de données. Dans cette thèse, nous proposons un formalisme qui permet de capturer une large gamme de problèmes d'extraction de motifs. Pour démontrer la généralité de ce formalisme, nous l'utilisons pour décrire trois problèmes d'extraction de motifs : le problème d'extraction d'itemsets fréquents fermés, le problème d'extraction de graphes relationnels fermés ou le problème d'extraction d'itemsets graduels fermés. Ce formal-

isme nous permet de construire ParaMiner qui est un algorithme générique et parallèle pour les problèmes d'extraction de motifs. ParaMiner est capable de résoudre tous les problèmes d'extraction de motifs qui peuvent ^tre décrit dans notre formalisme. Pour obtenir de bonne performances, nous avons généralisé plusieurs optimisations proposées par la communauté dans le cadre de problèmes spécique d'extraction de motifs. Nous avons également exploité la puissance de calcul parallèle disponible dans les archi- tectures parallèles. Nos expériences démontrent qu'en dépit de la généricité de ParaMiner ses performances sont comparables avec celles obtenues par les algorithmes les plus rapides de l'état de l'art. Ces algorithmes bénécient pourtant d'un avantage important, puisqu'ils incorporent de nombreuses optimisations spéciques au sous problème d'extraction de motifs qu'ils résolvent.

SUPERVISION

Sous la direction de Marie-Christine Rousset et de Alexandre Termier. Soutenue le 29-11-2011 à Grenoble , dans le cadre de École doctorale mathématiques, sciences et technologies de l'information, informatique (Grenoble) , en partenariat avec Laboratoire d'Informatique de Grenoble (équipe de recherche) . Je suis arrivée au laboratoire sur la fin de la thèse, et j'ai aidé Benjamin à la conception du POC et participé à la relecture de la thèse.

16.4/ HAMID MIRISAEE

SUJET : FOUILLE D'ITEMS ET D'ITEMSETS REPRÉSENTATIFS AVEC DES MÉTHODES DE DÉCOMPOSITION DE MATRICES BINAIRES ET DE SÉLECTION D'INSTANCES

Dans cette thèse, nous nous intéressons à la recherche d'"items" et d'"itemsets" d'intérêt via la décomposition de matrice binaire (Binary Matrix Factorization, BMF) et à la recherche d'objets représentatifs. Pour cela, nous étudions l'état de l'art des techniques de décomposition matricielle. Nous établissons, dans le premier Chapitre, un lien entre BMF et le problème de programmation binaire quadratique sans contraintes (Unconstrained Binary Quadratic Programming, UBQP) afin d'utiliser les algorithmes et heuristiques existant dans la littérature pour UBQP et les appliquer à BMF.Nous proposons dans le Chapitre 2 une nouvelle heuristique adaptée au calcul de BMF. Cette technique efficace optimise les solutions de BMF ligne par ligne (ou colonne par colonne) en inversant 1 bit à chaque fois. En utilisant le lien établi dans le Chapitre 2 qui nous permet d'appliquer les algorithmes et heuristiques d'UBQP à BMF, nous comparons la méthode proposée (1-opt-BMF) avec les heuristiques spécialisées pour UBQP (1-opt-UBQP) ainsi que les heuristiques classiques (1-opt-Standard). Nous montrons ensuite, en théorie et en pratique, l'efficacité de 1-opt-BMF sur une large variété de données publiques. Dans le Chapitre 3, nous nous intéressons au problème de la recherche des itemsets représentatifs en utilisant BMF et 1-opt-BMF. Pour cela, nous considérons dans un premier temps le lien entre le problème de "frequent itemset mining" et BMF, et proposons une nouvelle méthode que nous appelons "Decomposition Itemset Miner" (DIM). Une série d'expérience montre la qualité des résultats obtenus et l'efficacité de notre méthode.Enfinf, nous nous intéressons, dans le Chapitre 4, à la recherche d'objets représentatifs (qui donnent une

vue globale sur les données) dans des données de grandes dimensions. Nous examinons les méthodes disponibles dans la littérature en donnant les avantages et les inconvénients de chacune. Ensuite, nous défnissons mathématiquement le problème de sélection d'instance (Instance Selection Problem: ISP) et présentons trois variantes à ce problème ainsi que leur solutions. Dans les expériences, nous montrons que, bien qu'ISP puisse surpasser les autres méthodes dans certains cas, il vaut mieux le considérer en général comme une technique complémentaire dans le cadre de la recherche des objets représentatifs.

SUPERVISION

Sous la direction de Frédéric Pétrot et de Alexandre Termier. Soutenue le 13-06-2014 à Grenoble , dans le cadre de École doctorale mathématiques, sciences et technologies de l'information, informatique (Grenoble) , en partenariat avec Techniques de l'informatique et de la microélectronique pour l'architecture des systèmes intégrés (Grenoble) (laboratoire) et de Laboratoire d'informatique de Grenoble (laboratoire). Durant cette thèse j'ai contribué à une partie mathématique de la décomposition de matrice.

17

Encadrement de Master Recherche

En initiation à la recherche et master recherche mon encadrement s'intéresse à former les étudiants à la conception d'un état de l'art et sa rédaction qui est publiée. Le stage de M2R ayant pour objectif de se poursuivre en thèse, au moins un des verrous identifiés dans l'état de l'art doit être levé. Pour cela, comme avec les doctorants, nous avons au moins une réunion hebdomadaire et un rapport d'activité hebdomadaire. Malgré mon impossibilité à pouvoir encadrer de nouvelles thèses, je continue à fournir des sujets de stage et à les encadrer.

17.1/ Thomas Hassan

Sujet : Capture et l'analyse de flux pour les systèmes de recommandation (2014)

Ce travail poursuit la thèse de David Werner. Thomas Hassan a commencé par un état de l'art produit et publié lors d'un stage d'initiation à la recherche, puis ce stage de master recherche a introduit sa thèse. Le sujet de ce stage de M2R s'intéresse au fait que la génération massive de contenu Web a créé de nouveaux défis en matière de gestion de cette masse de données. Ces nouveaux enjeux sont les axes de Big Data : le volume croissant de données, la vélocité des modifications des données, et de leur grande variété de sources et formats. Dans ce contexte, trouver et analyser de grands volumes de données est impossible à réaliser manuellement. Nous avons décrit les besoins des systèmes de recommandation basés sur le contenu, et exposé des problèmes de passage à l'échelle des raisonneurs pour répondre à ces besoins. Nous avons proposé une approche de classification automatique des documents Web composé de cinq étapes. Les modifications apportées à l'architecture préalable ont permis de distribuer l'analyse des documents via la phase d'indexation et de distribuer la charge de calcul du raisonneur par l'utilisation de règles de classification simples. Le prototype met en œuvre les éléments architecturaux proposés, et est utilisé comme une interface entre les modules. Le prototype est fonctionnel et intègre, d'une part, le traitement avancé de la langue à partir de techniques de recherche d'information, et d'autre part, l'inclusion de règles SWRL, améliorant ainsi l'expressivité du système sans surcharger le raisonneur. L'indexation et la vectorisation permettent d'extraire les éléments pertinents, tandis que la hiérarchie peut générer des connaissances à partir des sources et les visualiser. Ce prototype servira de

base à l'élaboration d'une solution appliquée à une connaissance de l'entreprise dans un contexte Big Data.

17.2/ ALDRIC MANCEAU

SUJET : IDENTIFICATION D'IMAGES DE CONTREFAÇON DE PRODUIT DE LUXE (2014)

De par la croissance et l'étendue du web, on constate une augmentation du nombre d'images sur Internet : ainsi être en mesure d'identifier ces images est un besoin majeur. L'objectif de ce stage est, d'une part, d'identifier les objets et d'autre part, d'identifier leur marque. Pour cela, nous nous sommes appuyés sur les techniques d'analyse d'images. Parmi ces techniques, nous nous sommes intéressés à la représentation des images par des sacs de mots visuels. Les images sont alors représentées par des histogrammes représentant le nombre d'occurrences des mots visuels. En première étape, nous avons combiné la méthode de machine à support de vecteurs (SVM) combinée avec des sacs de mots visuels. SVM est une approche classique de la classification dans l'état de l'art. Il a servi de modèle de référence pour notre nouvelle approche. La deuxième étape s'est basée sur la combinaison de mots visuels sacs avec l'analyse formelle de concepts. Celle-ci est basée sur la notion de similarité d'attributs. Cette combinaison des deux méthodes n'avait jamais été faite préalablement pourtant ses résultats sont très intéressants.

17.3/ BEHROOZ OMIDVAR-TEHRANI

SUJET : EXPLORATION SÉMANTIQUE DE PATTERNS FRÉQUENTS (2012)

Les données utilisateurs telles que les traces d'usage des smartphones et le Web social sont souvent décrites par des informations socio-démographiques (ex., âge, sexe, métier, etc.) et de leurs activités (ex., donner un avis sur un restaurant, voter, critiquer un film, etc.). L'analyse des données utilisateurs intéresse beaucoup les études de la population, le marketing en-ligne, les recommandations et l'analyse des données à grande échelle. Cependant, les outils d'analyse des données utilisateurs sont encore très limités. Ces travaux analysent les données utilisateurs en formant des groupes d'utilisateurs. Cela diffère de l'analyse des utilisateurs individuels et aussi des analyses statistiques sur une population entière. Compte tenu des données utilisateurs brutes, obtenir les groupes d'utilisateurs en optimisant une ou plusieurs dimensions de qualité.

18

Encadrement de stage d'initiation à la Recherche

18.1/ Chahrazed Taouli

Analyse des Outils de data mining pour l'analyse d'images de perfusion tumorale en TEP clinique et pré-clinique (2016)

L'objet de cette étude est de constituer un état de l'art pour identifier les méthodes de data mining et machine learning connues et exploitées dans le domaine pour le traitement de telles images. L'état de l'art évalue les motivations de ces choix, leur pertinence (selon les auteurs) voire les autres outils suggérés comme pertinents pour l'analyse de ces données.

18.2/ Joe Raad

Capture et l'analyse de flux pour les systèmes de recommandation (2014)

L'objet de cette étude est de constituer un état de l'art sur les méthodes de comparaison entre les sources d'information textuelles dans le but de déterminer si elles portent sur un sujet similaire. La première catégorie de méthodes se concentre sur la recherche de nouvelles définitions de mesures sémantiques, et la seconde sur l'amélioration de l'identification de paraphrase, et enfin la dernière catégorie s'intéresse à l'amélioration des techniques d'extraction d'événements.

Les stages en M2R d'Aldric Manceau et Thomas Hassan sont des prolongements de leurs stages d'initiation à la recherche (2013)

19

SUPERVISIONS POST DOCTORALES

19.1/ QUENTIN BRABANT

Quentin est post-doctorant au sein du CIAD. Il a rejoint le laboratoire en avril 2019 dans le cadre du projet Kovers, mené en partenariat avec la start-up QAPE. Son rôle est d'appliquer des méthodes d'IA symbolique et de fouille de données au profilage de souscripteurs de mutuelles et à la prédiction de dépenses de santé futures. Ce travail s'appuie sur les données du SNIIRAM concernant les dépenses de m'assurance maladie, et notamment sur la sous-partie publiquement accessible de ces données : OpenDAMIR.

19.2/ RAMI BELKAROUI

Rami a été recruté dans le cadre du projet FUI WineCloud, en collaboration avec plusieurs partenaires (R-Tech Solutions, Orange, la Cave de Lugny, Photon Lines, etc). Ce projet propose de construire une plateforme ayant des outils d'aide à la décision viticole. Cette dernière se base sur la formalisation des savoir-faire des viticulteurs afin de construire un modèle sémantique se basant sur une ontologie métier. Rami a en charge la partie recherche impliquant les ontologies.

19.3/ AMIRA MOUAKHER

Amira a été recrutée dans le cadre du projet FUI WineCloud, mais sur la partie fouille de données. Le modèle ontologique est enrichi par les données de capteurs hétérogènes ainsi que les règles d'association (plus précisément règles de classification). Ce modèle nous permet d'avoir un système de monitoring permettant d'anticiper les événements pouvant survenir à la vigne et de lancer des alertes et des recommandations aux viticulteurs. Amira doit extraire les règles de raisonnement de l'ontologie et y intégrer les données issues de capteurs.

19.4/ NEIMEH LALEH

Neimeh a été recrutée sur le contrat Eurostars PSDP (Predictive Smart Data Platform). Elle a travaillé sur un état de l'art des mesures de similarité mais suite a des ennuis de

santé, elle nous a quitté.

19.5/ ALEXANDRE BAZIN

Alexandre a été recruté dans le cadre du projet Eurostars PSDP (Predictive Smart Data Platform) à la suite de Neimeh. L'objectif était d'établir un profilage dynamique de visiteurs d'un site de petites annonces et d'un profilage dynamique de ces annonces. L'objectif étant de proposer une un système de recommandation où les visiteurs influencent le profil des annonces et réciproquement. Pour cela nous avons construit un outil multi-dimensionnel scalable. Différents travaux se sont intéressés à cette problématique. En effet, les algorithmes déjà proposés génèrent un très grand nombre de règles, parfois inutiles et ne présentant pas de corrélations intéressantes. Pour palier à ce problème, nous avons pensé à appliquer un algorithme qui extrait des bases génériques. Cet algorithme propose la possibilité d'avoir un nombre compacte de règles d'association à partir desquelles ont peut reconstruire tout l'ensemble. A partir de ces règles, on ne gardera que les règles ayants un certain seuil de confiance ou même un seuil d'une certaine mesure de corrélation (bond, lift, any-confidence, all-confidence, etc).

20

ENCADREMENT DE STAGE DE THÈSE

20.1/ WIDED SELMI

SUJET : LABELISATION DE CONCEPTS FORMELS

Dans le cadre de l'analyse formelle de concepts, les correspondances forment des concepts qui sont des ensembles d'objets possédants les mêmes attributs et inversement. Cependant ces concepts n'ont pas de sémantique associée, i.e. pas d'étiquette permettant d'indiquer le sens du concept et ses relations aux autres concepts. Wided a pour objectif d'établir un processus de labelisation.

21

ENSEIGNEMENT

21.1/ LIEUX D'EXERCICE

J'ai débuté les enseignements en même temps que ma thèse dans diverses universités dont voici la liste :

- Depuis 2013 : IUT Métier du Multimédia et de l'Internet de l'Université de Bourgogne
- 2012-2013 : IUT Services et Réseaux de Communication de l'Université de Lorraine
- 2011-2012 : École Nationale Supérieure d'Informatique et de Mathématiques Appliquées (Ensimag) de l'Institut Polytechnique de Grenoble
- 2006-2010 : IUT informatique de l'université de Strasbourg

21.2/ MATIÈRES ENSEIGNÉES

Titulaire d'une formation en informatique, j'ai toujours donné des cours dans ce domaine. J'ai pu le diversifier avec l'introduction de l'enseignement de la fouille de données correspondant à mon coeur de recherche. Voici la liste des matières enseignées :

- Ingénierie logicielle
- Algorithme et Programmation
- Développement Web
- Intégration Web
- Javascript
- JQuery
- Jeux et intelligence artificielle
- Data Mining
- Base de données
- Analyse et conception de systèmes d'information (1 Merise)

- Compression et stockage
- Système et réseau
- Suivi de Stage
- Tutorat

21.3/ RESPONSABILITÉS

Enseigner signifiant faire partie d'une équipe d'enseignement, j'ai collaboré aux tâches qui entourent les enseignements (et leurs évaluations) à proprement parler. Voici une liste de responsabilités que j'ai pu assumer : J'ai également pris en charge certaines responsabilités :

- Création (en collaboration) de 2 modules de réseau et un module de JQuery
- Responsabilité des modules de Jquery, et de compression et stockage de la licence ATC (Activité et Technique de Communication).
- Responsabilité du module d'ouverture scientifique
- Responsabilité d'au moins 2 modules d'informatique chaque année en IUT MMI
- Animation des journées scientifiques pour la découverte de la science
- Organisation des journées portes ouvertes à l'université et participation aux congrès destinés à faire connaitre l'iut et son programme
- responsable des projets tutorés de l'iut MMI pendant 1 an
- responsable des emplois du temps de l'iut MMI depuis 2 ans
- Elue au Conseil Scientifique et de la Recherche de l'IUT Dijon-Auxerre pendant 2 ans
- Diffusion de mon domaine de recherche en master par des enseignements de data mining : Master ENSIMAG (2011) à Grenoble et Master Smart City à dijon (2019) et en IUT : responsabilité du module OS201 « Jeux et Intelligence Artificielle »
- Participations à des comités de sélection pour des postes MCF en 27ème section

BIBLIOGRAPHY

[kNN, 1951] (1951). **Discriminatory analysis, nonparametric discrimination: Consistency properties**. Randolph Field.

[Abbassi et al., 2007] Abbassi, Z., et Mirrokni, V. S. (2007). **A recommender system based on local random walks and spectral methods**. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 102–108. ACM.

[Abeer Al-Arfaj, 2015] Abeer Al-Arfaj, A. A.-S. (2015). **Ontology construction from text: Challenges and trends**. *International Journal of Artificial Intelligence and Expert Systems (IJAE), Volume (6)*.

[Aggarwal et al., 2006] Aggarwal, C. C., Han, J., Wang, J., et Yu, P. S. (2006). **A framework for on-demand classification of evolving data streams**. *IEEE Trans. on Knowl. and Data Eng.*, 18(5):577–589.

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., et Swami, A. (1993). **Mining association rules between sets of items in large databases**. In *ACM SIGMOD record*, volume 22, pages 207–216. ACM.

[Agrawal et al., 1996] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., et Verkamo, A. I. (1996). **Fast discovery of association rules**. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI/MIT Press.

[Agrawal et al., 1994] Agrawal, R., et Srikant, R. (1994). **Fast algorithms for mining association rules**. *IBM Research, Proceedings of the 20th VLDB Conference, Santiago, Chile*.

[Albert et al., 1999] Albert, R., Jeong, H., et Barabási, A. (1999). **Internet: Diameter of the world-wide web**. *Nature*, 401(6749):130–131.

[Alvarez et al., 2011a] Alvarez, M. A., Qi, X., et Yan, C. (2011a). **A shortest-path graph kernel for estimating gene product semantic similarity**. *Journal of biomedical semantics*, 2(1):3.

[Alvarez et al., 2011b] Alvarez, M. A., et Yan, C. (2011b). **A graph-based semantic similarity measure for the gene ontology**. *Journal of bioinformatics and computational biology*, 9(06):681–695.

[Anderson, 2007] Anderson, M. (2007). **Artificial intuition: A new possible path to artificial intelligence**. *Intuition and logic*. 4 cites: https://scholar.google.com/scholar?cites=13240944404456399 79&as$_sdt$ = 2005&$sciodt = 0,5$&$hl = en$.

[Aussenac-Gilles, 2005] Aussenac-Gilles, N. (2005). **Méthodes ascendantes pour l'ingénierie des connaissances**. thesis, Université Paul Sabatier - Toulouse III.

[Barbut et al., 1970] Barbut, M., et Monjardet, B. (1970). **Ordre et classification – Algèbre et combinatoire**. Hachette, Paris.

[Batet et al., 2011] Batet, M., Sánchez, D., et Valls, A. (2011). **An ontology-based measure to compute semantic similarity in biomedicine**. *Journal of biomedical informatics*, 44(1):118–125.

[Batet et al., 2010] Batet, M., Sanchez, D., Valls, A., et Gibert, K. (2010). **Exploiting taxonomical knowledge to compute semantic similarity: an evaluation in the biomedical domain**. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 274–283. Springer.

[Bazin et al., 2018] Bazin, A., et Bertaux, A. (2018). **k-Partite Graphs as Contexts**. In *The 14th International Conference on Concept Lattices and Their Applications (CLA2018)*, Olomouc, Czech Republic.

[Bazin et al.,] Bazin, A., Gros, N., Bertaux, A., et Nicolle, C. **Condensed representations of association rules in n-ary relations**. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* ***(Impact Factor 2.775- Quartile Q1)***.

[Baziz et al., 2007] Baziz, M., Boughanem, M., Pasi, G., et Prade, H. (2007). **An information retrieval driven by ontology from query to document expansion**. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, pages 301–313. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.

[Belkaroui et al., 2018] Belkaroui, R., Bertaux, A., Labbani, O., Hugol-Gential, C., et Nicolle, C. (2018). **Towards events ontology based on data sensors network for viticulture domain**. In *Proceedings of the 8th International Conference on the Internet of Things*, IOT '18, pages 44:1–44:7, New York, NY, USA. **ACM**.

[Belkaroui et al., 2019] Belkaroui, R., Mouakher, A., Bertaux, A., Labbani, O., Hugol-Gential, C., et Nicolle, C. (2019). **Winecloud: Une ontologie dévénements pour la modélisation sémantique des données de capteurs hétérogènes**. In *Revue des Nouvelles Technologies de l'Information. EGC 2019, vol. RNTI-E-35*, pages 379–380.

[Belohlavek, 2011] Belohlavek, R. (2011). **What is a fuzzy concept lattice? ii**. *Rough Sets, Fuzzy Sets, Data Mining and Granular Computing*, pages 19–26.

[Benaicha, 2017] Benaicha, M. (2017). **Identification des concepts pour la ré-ingénierie des ontologies**. Master's thesis, Université du Québec à Chicoutimi, Chicoutimi.

[Benezeth et al., 2015] Benezeth, Y., Bertaux, A., et Manceau, A. (2015). **Bag-of-word based brand recognition using Markov Clustering Algorithm for codebook generation**. In ***IEEE*** *International Conference on Image Processing (ICIP)*, Québec, France.

[Benner et al., 1987] Benner, P., et Tanner, C. (1987). **How expert nurses use intuition**. *AJN The American Journal of Nursing*, 87:23–34.

[Berka et al., 2010] Berka, P., et Rauch, J. (2010). **Machine learning and association rules**. *Czech Science Foundation*.

[Bi et al., 2011] Bi, W., et Kwok, J. T. (2011). **Multi-label classification on tree- and dag-structured hierarchies**. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 17–24, USA. Omnipress.

[Billhardt et al., 2002] Billhardt, H., Borrajo, D., et Maojo, V. (2002). **A context vector model for information retrieval**. *Journal of the American Society for Information Science and Technology*, 53(3):236–249.

[Birkhoff, 1948] Birkhoff, G. (1948). **Lattice Theory**, volume XXV of *Colloquium Publications*. American Mathematical Society, New York.

[Bobadilla et al., 2013] Bobadilla, J., Ortega, F., Hernando, A., et Gutiérrez, A. (2013). **Recommender systems survey**. *Knowledge-Based Systems*, 46:109 – 132.

[Bodenreider et al., 2005] Bodenreider, O., Aubry, M., et Burgun, A. (2005). **Non-lexical approaches to identifying associative relations in the gene ontology**. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 91. NIH Public Access.

[Brandes et al., 2003] Brandes, U., Gaertler, M., et Wagner, D. (2003). **Experiments on graph clustering algorithms.** In Battista, G. D., et Zwick, U., editors, *ESA*, volume 2832 of *Lecture Notes in Computer Science*, pages 568–579. Springer.

[Budanitsky et al., 2006] Budanitsky, A., et Hirst, G. (2006). **Evaluating wordnet-based measures of lexical semantic relatedness**. *Computational Linguistics*, 32(1):13–47.

[Bělohlavek, 1999] Bělohlavek, R. (1999). **Fuzzy galois connections**. *Math. Logic Quaterly*.

[Bělohlavek et al., 2005] Bělohlavek, R., et Vychodil, V. (2005). **What is a fuzzy concept lattice?** In *3rd Int. Conference on Concept Lattices and Their Applications (CLA 2005)*, pages 34–45.

[Bulskov et al., 2002] Bulskov, H., Knappe, R., et Andreasen, T. (2002). **On measuring similarity for conceptual querying**. In *International Conference on Flexible Query Answering Systems*, pages 100–111. Springer.

[Burusco Juandeaburre et al., 1994] Burusco Juandeaburre, A., et Fuentes-González, R. (1994). **The study of the l-fuzzy concept lattice**. *Mathware & soft computing. 1994 Vol. 1 Núm. 3 p. 209-218*.

[Carbonnel et al., 2017] Carbonnel, J., Huchard, M., Miralles, A., et Nebut, C. (2017). **Feature model composition assisted by formal concept analysis**. In *ENASE: Evaluation of Novel Approaches to Software Engineering*, pages 27–37. SciTePress.

[Cerf et al., 2008] Cerf, L., Besson, J., Robardet, C., et Boulicaut, J.-F. (2008). **Data-peeler: Constraint-based closed pattern mining in n-ary relations**. In *proceedings of the 2008 SIAM International conference on Data Mining*, pages 37–48. SIAM.

[Cerri et al., 2014] Cerri, R., Barros, R. C., et De Carvalho, A. C. P. L. F. (2014). **Hierarchical multi-label classification using local neural networks**. *J. Comput. Syst. Sci.*, 80(1):39–56.

[Chabot et al., 2014] Chabot, Y., Bertaux, A., Nicolle, C., et Kechadi, M.-T. (2014). **A complete formalized knowledge representation model for advanced digital forensics timeline analysis**. *Digit. Investig.* ***(Index ISI WoS - IF=1.648) (Quartile Q2)***, 11(S2):S95–S105.

[Chabot et al., 2015] Chabot, Y., Bertaux, A., Nicolle, C., et Kechadi, T. (2015). **An ontology-based approach for the reconstruction and analysis of digital incidents timelines**. *Digit. Investig.* ***(Index ISI WoS IF=1.648) (Quartile Q2)***, 15(C):83–100.

[Chaplot et al., 2018] Chaplot, D. S., et Salakhutdinov, R. (2018). **Knowledge-based word sense disambiguation using topic models**. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[Chebotarev et al., 2006a] Chebotarev, P., et Shamis, E. (2006a). **The matrix-forest theorem and measuring relations in small social groups**. *arXiv preprint math/0602070*.

[Chebotarev et al., 2006b] Chebotarev, P., et Shamis, E. (2006b). **On proximity measures for graph vertices**. *arXiv preprint math/0602073*.

[Chen et al., 2006] Chen, E., Chen, X., Sheu, P. C.-Y., et Qian, T. (2006). **An evolutionary computational method for n-connection subgraph discovery**. *2012 IEEE 24th International Conference on Tools with Artificial Intelligence*, 0:169–178.

[Chen et al., 2017] Chen, F., Lu, C., Wu, H., et Li, M. (2017). **A semantic similarity measure integrating multiple conceptual relationships for web service discovery**. *Expert Systems with Applications*, 67:19–31.

[Chen et al., 2012] Chen, Y.-L., et Chiu, Y.-T. (2012). **Vector space model for patent documents with hierarchical class labels**. *Journal of Information Science*, 38(3):222–233.

[Cimiano et al., 2005] Cimiano, P., et Völker, J. (2005). **Text2Onto - A Framework for Ontology Learning and Data-driven Change Discovery**. pages 227–238.

[Cimiano et al., 2006] Cimiano, P., Völker, J., et Studer, R. (2006). **Ontologies on Demand? - A Description of the State-of-the-Art, Applications, Challenges and Trends for Ontology Learning from Text**. *Information, Wissenschaft und Praxis*, 57(6-7):315–320.

[Couto et al., 2003] Couto, F. M., Silva, M. J., et Coutinho, P. M. (2003). **Implementation of a functional semantic similarity measure between gene-products**.

[Cross et al., 2011] Cross, V., et Yu, X. (2011). **Investigating ontological similarity theoretically with fuzzy set theory, information content, and tversky similarity and empirically with the gene ontology**. In *International Conference on Scalable Uncertainty Management*, pages 387–400. Springer.

[Damasio, 1994] Damasio, A. R. (1994). **Descartes' Error. Emotion, reason and the human brain.** Avon Books, New York.

[Damasio, 1999] Damasio, A. R. (1999). **The Feeling of What Happens: Body and Emotion in the Making of Consciousness**. Harcourt Brace and Co.

[Dane et al., 2007] Dane, E., et Pratt, M. (2007). **Exploring intuition and its role in managerial decision making**. *Academy of Management Review*, 32.

[d'Aquin et al., 2012] d'Aquin, M., et Noy, N. F. (2012). **Review: Where to publish and find ontologies? a survey of ontology libraries**. *Web Semant.*, 11:96–111.

[de Gemmis et al., 2015] de Gemmis, M., Lops, P., Semeraro, G., et Musto, C. (2015). **An investigation on the serendipity problem in recommender systems**. *Information Processing & Management*, 51(5):695–717.

[De Knijff et al., 2013] De Knijff, J., Frasincar, F., et Hogenboom, F. (2013). **Domain taxonomy learning from text: The subsumption method versus hierarchical clustering**. *Data Knowl. Eng.*, 83:54–69.

[Dhillon et al., 2003] Dhillon, I. S., Mallela, S., et Modha, D. S. (2003). **Information-theoretic co-clustering**. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 89–98, New York, NY, USA. ACM.

[Diligenti et al., 2000] Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. L., et Gori, M. (2000). **Focused crawling using context graphs**. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 527–534, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Dong et al., 2014] Dong, H., et Hussain, F. (2014). **Self-adaptive semantic focused crawler for mining services information discovery**. *Industrial Informatics, IEEE Transactions on*, 10:1616–1626.

[Dorogovtsev et al., 2002] Dorogovtsev, S. N., et Mendes, J. F. F. (2002). **Evolution of networks**. In *Adv. Phys*, pages 1079–1187.

[Dreyfus et al., 1989] Dreyfus, H. L., et Dreyfus, S. E. (1989). **Computers in the human context: Information technology, productivity, and people**. chapter Why Computers May Never Think Like People, pages 125–143. MIT Press, Cambridge, MA, USA.

[Dundas et al., 2011] Dundas, J., et Chik, D. (2011). **Implementing human-like intuition mechanism in artificial intelligence**. *CoRR*, abs/1106.5917.

[Duquenne et al., 1986] Duquenne, V., et Guigues, J.-L. (1986). **Famille minimale d'implications informatives résultant d'un tableau de données binaires**. *Math. et Sci. Hum.*, 24(95):5–18.

[Ehrig et al., 2004] Ehrig, M., Haase, P., Hefke, M., et Stojanovic, N. (2004). **Similarity for Ontologies - a Comprehensive Framework**. In *In Workshop Enterprise Modelling and Ontology: Ingredients for Interoperability, at PAKM 2004*.

[Epstein et al., 1996] Epstein, S., Pacini, R., Denes-Raj, V., et Heier, H. (1996). **Individual differences in intuitive–experiential and analytical–rational thinking styles**. *Journal of personality and social psychology*, 71:390–405.

[Euzenat et al., 2007] Euzenat, J., Shvaiko, P., et others (2007). **Ontology matching**, volume 18. Springer.

[Faloutsos et al., 2004] Faloutsos, C., McCurley, K. S., et Tomkins, A. (2004). **Fast discovery of connection subgraphs**. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 118–127, New York, NY, USA. ACM.

[Faloutsos et al., 1999] Faloutsos, M., Faloutsos, P., et Faloutsos, C. (1999). **On power-law relationships of the internet topology**. In *SIGCOMM '99: Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, pages 251–262, New York, NY, USA. ACM.

[Fayyad et al., 1996] Fayyad, U. M., Piatetsky-Shapiro, G., et Smyth, P. (1996). **From data mining to knowledge discovery: an overview**. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., et Uthurusamy, R., editors, *Advances in knowledge discovery and data mining*, pages 1–34. American Association for Artificial Intelligence, Menlo Park, CA, USA.

[Feldman et al., 1998] Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., et Zamir, O. (1998). **Text mining at the term level**. In Żytkow, J. M., et Quafafou, M., editors, *Principles of Data Mining and Knowledge Discovery*, pages 65–73, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Fernández et al., 2009] Fernández, M., Overbeeke, C., Sabou, M., et Motta, E. (2009). **What makes a good ontology? a case-study in fine-grained knowledge reuse**. pages 61–75.

[Fouss et al., 2007] Fouss, F., Pirotte, A., Renders, J.-M., et 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).

[Frantz, 2003] Frantz, R. (2003). **Herbert simon. artificial intelligence as a framework for understanding intuition**. *Journal of Economic Psychology*, 24:265–277.

[Gabrilovich et al., 2007] Gabrilovich, E., et Markovitch, S. (2007). **Computing semantic relatedness using wikipedia-based explicit semantic analysis.** In *IJcAI*, volume 7, pages 1606–1611.

[Galbrun et al., 2012] Galbrun, E., et Miettinen, P. (2012). **From black and white to full color: extending redescription mining outside the Boolean world**. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(4):284–303.

[Galbrun et al., 2017] Galbrun, E., et Miettinen, P. (2017). **Redescription Mining**. SpringerBriefs in Computer Science. Springer International Publishing.

[Gamallo, 2017] Gamallo, P. (2017). **Compositional semantics using feature-based models from wordnet**. *SENSE 2017*, page 1.

[Ganesan et al., 2012] Ganesan, V., Swaminathan, R., et Thenmozhi, M. (2012). **Similarity measure based on edge counting using ontology**. *International Journal of Engineering Research and Development*, 3(3):40–44.

[Ganter et al., 2001] Ganter, B., et Kuznetsov, S. (2001). **Pattern structures and their projections**. In *Proceedings of the 9th International Conference on Conceptual Structures (ICCS 2001)*, LNCS 2120, pages 129–142. Springer.

[Ganter et al., 1997] Ganter, B., et Wille, R. (1997). **Applied lattice theory: Formal concept analysis**. In *General Lattice Theory, G. Grätzer editor, Birkhäuser*. Preprints.

[Ganter et al., 2012] Ganter, B., et Wille, R. (2012). **Formal Concept Analysis: Mathematical Foundations**. Springer Science & Business Media. Google-Books-ID: hNwqBAAAQBAJ.

[Gao et al., 2015] Gao, J.-B., Zhang, B.-W., et Chen, X.-H. (2015). **A wordnet-based semantic similarity measurement combining edge-counting and information content theory**. *Engineering Applications of Artificial Intelligence*, 39:80–88.

[Garla et al., 2012] Garla, V. N., et Brandt, C. (2012). **Semantic similarity in the biomedical domain: an evaluation across knowledge sources**. *BMC bioinformatics*, 13(1):261.

[Getoor et al., 2005] Getoor, L., et Diehl, C. P. (2005). **Link mining: A survey**. *SIGKDD Explor. Newsl.*, 7(2):3–12.

[Gigerenzer, 2007] Gigerenzer, G. (2007). **Gut feelings: The intelligence of the unconscious**.

[Girotra et al., 2013] Girotra, M., Nagpal, K., Minocha, S., et Sharma, N. (2013). **Comparative survey on association rule mining algorithms**. *International Journal of Computer Applications*.

[Gobet et al., 2009] Gobet, F., et Chassy, P. (2009). **Expertise and intuition: A tale of three theories**. *Minds and Machines*, 19:151–180.

[Gori et al., 2007] Gori, M., Pucci, A., Roma, V., et Siena, I. (2007). **Itemrank: A random-walk based scoring algorithm for recommender engines.** In *IJCAI*, volume 7, pages 2766–2771.

[Guérif, 2006] Guérif, S. (2006). **Réduction de dimension en apprentissage numérique non supervisé**. In *Handbook on Ontologies*.

[Haase et al., 2005] Haase, P., et Stojanovic, L. (2005). **Consistent Evolution of OWL Ontologies**. In Gómez-Pérez, A., et Euzenat, J., editors, *The Semantic Web: Research and Applications*, number 3532 in Lecture Notes in Computer Science, pages 182–197. Springer Berlin Heidelberg.

[Haav, 2004] Haav, H.-M. (2004). **A semi-automatic method to ontology design by using fca.** 110.

[Harispe et al., 2013] Harispe, S., Ranwez, S., Janaqi, S., et Montmain, J. (2013). **Semantic measures for the comparison of units of language, concepts or instances from text and knowledge representation analysis**. *A Comprehensive Survey and a Technical Introduction to Knowledge-based Measures Using Semantic Graph Analysis, LGI2P/EMA Research Center, Parc scientifique, France*.

[Hazman et al., 2011] Hazman, M., R.ElBeltagy, S., et Rafea, A. (2011). **A survey of ontology learning approaches**. *International Journal of Computer Applications*.

[Hearst et al., 1998] Hearst, M., Dumais, S., Osman, E., Platt, J., et Scholkopf, B. (1998). **Support vector machines**. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

[Heryaningsih et al., 2018] Heryaningsih, N. Y., et Khusna, H. (2018). **Development of syntax of intuition-based learning model in solving mathematics problems**. *Journal of Physics: Conference Series*, 948:012018.

[Hirst et al., 1998] Hirst, G., St-Onge, D., et others (1998). **Lexical chains as representations of context for the detection and correction of malapropisms**. *WordNet: An electronic lexical database*, 305:305–332.

[Hliaoutakis, 2005] Hliaoutakis, A. (2005). **Semantic similarity measures in mesh ontology and their application to information retrieval on medline**. *Master's thesis*.

[Hliaoutakis et al., 2006] Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E. G., et Milios, E. (2006). **Information retrieval by semantic similarity**. *International journal on semantic Web and information systems (IJSWIS)*, 2(3):55–73.

[Hodgkinson et al., 2008] Hodgkinson, G., et Langan-Fox, J. and, S.-S. E. (2008). **Intuition: a fundamental bridging construct in the behavioural sciences**. *The British Journal of Psychology*, 99(1):1–27.

[Hughes et al., 2007] Hughes, T., et Ramage, D. (2007). **Lexical semantic relatedness with random graph walks.** In *EMNLP-CoNLL*, pages 581–589.

[Hugol-Gential et al., 2019a] Hugol-Gential, C., Simon, M., Bertaux, A., Narsis, O., Belkaroui, R., Mouakher, A., et Nicolle, C. (2019a). **Une ontologie de la culture de la vigne : des savoirs académiques aux savoirs d'expérience**. In *Recherches en communication*, pages 111–129.

[Hugol-Gential et al., 2019b] Hugol-Gential, C., Simon, M., Bertaux, A., Narsis, O., Belkaroui, R., Mouakher, A., et Nicolle, C. (2019b). **Une ontologie de la vigne au verre: la terminologie professionnelle au regard des savoirs mobilisés en viticulture**. *Recherches en Communication*, 48(48).

[Hugol-Gential et al., 2019c] Hugol-Gential, C., Simon, M., Bertaux, A., Narsis, O., Belkaroui, R., et Nicolle, C. (2019c). **Traçabilité et analyse bigdata sur la chaîne de valeur vitivinicole. de la science à l'expérience, la pluralité des savoirs en viticulture : le cas des maladies**. In *Univigne*.

[Hulten et al., 2001] Hulten, G., Spencer, L., et Domingos, P. (2001). **Mining time-changing data streams**. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 97–106, New York, NY, USA. ACM.

[Héon et al., 2013] Héon, M., et Nkambou, R. (2013). **G-OWL : Vers un langage de modélisation graphique, polymorphique et typé pour la construction d'une ontologie dans la notation OWL**.

[Jain et al., 2010] Jain, S., et Bader, G. D. (2010). **An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology**. *BMC bioinformatics*, 11(1):562.

[Jeh et al., 2002] Jeh, G., et Widom, J. (2002). **Simrank: a measure of structural-context similarity**. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM.

[Jiang et al., 2013] Jiang, C., Coenen, F., et Zito, M. (2013). **A survey of frequent subgraph mining algorithms**. *Knowledge Eng. Review*, 28(1):75–105.

[Jiang et al., 2003] Jiang, G., Ogasawara, K., Endoh, A., et Sakurai, T. (2003). **Context-based ontology building support in clinical domains using formal concept analysis**. *International Journal of Medical Informatics*, 71(1):71 – 81.

[Jiang et al., 1997a] Jiang, J. J., et Conrath, D. W. (1997a). **Semantic similarity based on corpus statistics and lexical taxonomy**. *arXiv preprint cmp-lg/9709008*.

[Jiang et al., 1997b] Jiang, J. J., et Conrath, D. W. (1997b). **Semantic similarity based on corpus statistics and lexical taxonomy**. *arXiv preprint cmp-lg/9709008*.

[Jiang et al., 2015] Jiang, Y., Zhang, X., Tang, Y., et Nie, R. (2015). **Feature-based approaches to semantic similarity assessment of concepts using wikipedia**. *Information Processing & Management*, 51(3):215–234.

[Jung, 1923] Jung, C. (1923). **Psychological types**. Harcourt, Brace, Oxford, England.

[Kahneman, 2003] Kahneman, D. (2003). **A perspective on judgment and choice: mapping bounded rationality.** *Am Psychol*, 58(9):697–720.

[Kahneman, 2011] Kahneman, D. (2011). **Thinking, fast and slow**. Farrar, Straus and Giroux, New York.

[Karypis et al., 1996] Karypis, G., et Kumar, V. (1996). **Parallel multilevel k-way partitioning scheme for irregular graphs**. In *Proceedings of the 1996 ACM/IEEE Conference on Supercomputing*, Supercomputing '96, Washington, DC, USA. IEEE Computer Society.

[Kaytoue et al., 2011] Kaytoue, M., Kuznetsov, S. O., Napoli, A., et Duplessis, S. (2011). **Mining gene expression data with pattern structures in formal concept analysis**. *Information Sciences*, 181(10):1989–2001.

[Kempe et al., 2003] Kempe, D., Kleinberg, J., et Tardos, E. (2003). **Maximizing the spread of influence through a social network**. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, pages 137–146, New York, NY, USA. ACM.

[Klein, 1998] Klein, G. (1998). **Sources of power : how people make decisions**. MIT Press, Cambridge, Mass. ; London.

[Kolata, 1982] Kolata, G. B. (1982). **How can computers get common sense?** *Science*, 217 4566:1237–8.

[Kolotygin et al., 2017] Kolotygin, A., et Akayomov, A. (2017). **Artificial intuition**. *US Patent App. 14/833,505*. Query date: 2019-09-05.

[Koren et al., 2007] Koren, Y., North, S. C., et Volinsky, C. (2007). **Measuring and extracting proximity graphs in networks**. *ACM Trans. Knowl. Discov. Data*, 1(3).

[Landauer et al., 1997] Landauer, T. K., et Dumais, S. T. (1997). **A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.** *Psychological review*, 104(2):211.

[Leacock et al., 1998a] Leacock, C., et Chodorow, M. (1998a). **Combining local context and wordnet similarity for word sense identification**. *WordNet: An electronic lexical database*, 49(2):265–283.

[Leacock et al., 1998b] Leacock, C., et Chodorow, M. (1998b). **Combining local context and WordNet similarity for word sense identification**. *WordNet: An electronic lexical database*, 49(2):265–283.

[Leeuwen et al., 2015] Leeuwen, M. v., et Galbrun, E. (2015). **Association Discovery in Two-View Data**. *IEEE Transactions on Knowledge and Data Engineering*, 27(12):3190–3202.

[Lehmann et al., 1995] Lehmann, F., et Wille, R. (1995). **A triadic approach to formal concept analysis**. *Conceptual structures: applications, implementation and theory*, pages 32–43.

[Lesk, 1986] Lesk, M. (1986). **Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone**. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. Citeseer.

[Leskovec et al., 2006] Leskovec, J., et Faloutsos, C. (2006). **Sampling from large graphs**. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 631–636, New York, NY, USA. ACM.

[Leskovec et al., 2010] Leskovec, J., Lang, K. J., et Mahoney, M. (2010). **Empirical comparison of algorithms for network community detection**. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 631–640, New York, NY, USA. ACM.

[Li et al., 2010] Li, B., Wang, J. Z., Feltus, F. A., Zhou, J., et Luo, F. (2010). **Effectively integrating information content and structural relationship to improve the go-based similarity measure between proteins**. *arXiv preprint arXiv:1001.0958*.

[Li et al., 2003] Li, Y., Bandar, Z. A., et 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 et al., 2006] Li, Y., McLean, D., Bandar, Z. A., O'shea, J. D., et Crockett, K. (2006). **Sentence similarity based on semantic nets and corpus statistics**. *IEEE transactions on knowledge and data engineering*, 18(8):1138–1150.

[Lin, 1998] Lin, D. (1998). **An information-theoretic definition of similarity.** In *Icml*, volume 98, pages 296–304. Citeseer.

[Lin et al., 1998] Lin, D., et others (1998). **An information-theoretic definition of similarity.** In *Icml*, volume 98, pages 296–304.

[Lin et al., 2010] Lin, J., et Dyer, C. (2010). **Data-Intensive Text Processing with MapReduce**. Morgan and Claypool Publishers.

[Lin, 2013] Lin, J. J. (2013). **Monoidify! monoids as a design principle for efficient mapreduce algorithms**. *CoRR*, abs/1304.7544.

[Lops et al., 2011] Lops, P., de Gemmis, M., et Semeraro, G. (2011). **Content-based recommender systems: State of the art and trends.** In Ricci, F., Rokach, L., Shapira, B., et Kantor, P. B., editors, *Recommender Systems Handbook*, pages 73–105. Springer.

[Luxenburger, 1991] Luxenburger, M. (1991). **Implications partielles dans un contexte**. *Mathématiques, informatique et sciences humaines*, 29(113):35–55.

[M et al., 2004] M, O., V, S., et J., S. (2004). **Ontology design with formal concept analysis. concept lattices and their applications**. *CLA Conference proceedings*.

[Maedche et al., 2001] Maedche, A., et Staab, S. (2001). **Comparing ontologies-similarity measures and a comparison study**. AIFB.

[Makhalova et al., 2017] Makhalova, T., et Nourine, L. (2017). **An incremental algorithm for computing n-concepts**. In *Formal Concept Analysis for Knowledge Discovery. Proceedings of International Workshop on Formal Concept Analysis for Knowledge Discovery (FCA4KD 2017), Moscow, Russia, June 1, 2017.* CEUR-WS. org.

[Mao et al., 2002] Mao, W., et Chu, W. W. (2002). **Free-text medical document retrieval via phrase-based vector space model.** In *Proceedings of the AMIA Symposium*, page 489. American Medical Informatics Association.

[Marker, 2002] Marker, D. (2002). **Model theory: an introduction**. Springer Science & Business Media.

[Mazandu et al., 2011] Mazandu, G. K., et Mulder, N. J. (2011). **It-gom: An integrative tool for ic-based go semantic similarity measures**. Technical Report, Technical report, University of Cape Town (South Africa).

[Mazandu et al., 2013] Mazandu, G. K., et Mulder, N. J. (2013). **Information content-based gene ontology semantic similarity approaches: toward a unified framework theory**. *BioMed research international*, 2013.

[McCarthy, 1989] McCarthy, J. (1989). **Mathematical logic in artificial intelligence**.

[McCormick, 2004] McCormick, R. (2004). **Issues of learning and knowledge in technology education**. *International Journal of Technology and Design Education*, 14(1):21–44.

[Meijer et al., 2014] Meijer, K., Frasincar, F., et Hogenboom, F. (2014). **A semantic approach for extracting domain taxonomies from text**. *Decision Support Systems*, 62.

[Meng et al., 2013] Meng, L., Huang, R., et Gu, J. (2013). **A review of semantic similarity measures in wordnet**. *International Journal of Hybrid Information Technology*, 6(1):1–12.

[Messai, 2009] Messai, N. (2009). **Analyse de concepts formels guidée par des connaissances de domaine : Application à la découverte de ressources génomiques sur le web**. Thèse de doctorat, Université Henri Poincaré. Nancy 1.

[Mihalcea et al., 2006] Mihalcea, R., Corley, C., Strapparava, C., et others (2006). **Corpus-based and knowledge-based measures of text semantic similarity**. In *AAAI*, volume 6, pages 775–780.

[Mikolov et al., 2013] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., et Dean, J. (2013). **Distributed representations of words and phrases and their compositionality**. In *Advances in neural information processing systems*, pages 3111–3119.

[Minsky, 2006] Minsky, M. (2006). **The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind**. Simon & Schuster, Inc., New York, NY, USA.

[Moeller et al., 2009] Moeller, R., et Haarslev, V. (2009). **Tableau-based reasoning**. In *Handbook on Ontologies*.

[Morgado et al., 2008] Morgado, L., et Gaspar, G. (2008). **Towards background emotion modeling for embodied virtual agents**. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '08,

pages 175–182, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Motik et al., 2009] Motik, B., Shearer, R., et Horrocks, I. (2009). **Hypertableau reasoning for description logics**. *J. Artif. Int. Res.*, 36(1):165–228.

[Mouakher et al., 2019] Mouakher, A., Belkaroui, R., Bertaux, A., Labbani, O., Hugol-Gential, C., et Nicolle, C. (2019). **An ontology-based monitoring system in vineyards of the burgundy region**. In *28th IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE-2019)*, page to appear. IEEE.

[Myers, 2002] Myers, D. G. (2002). **Intuition: Its powers and perils**. Yale University Press, New Haven, CT, US.

[Napoli, 2006] Napoli, A. (2006). **A smooth introduction to symbolic methods in knowledge discovery**. In *Categorization in Cognitive Science*. Elsevier.

[Navigli, 2009] Navigli, R. (2009). **Word Sense Disambiguation: A Survey**. *ACM Comput. Surv.*, 41(2):10:1–10:69.

[Newman, 2003] Newman, M. E. J. (2003). **The structure and function of complex networks**. *SIAM Review*, 45(2):167–256.

[NF et al., 2009] NF, N., NH, S., PL, W., B, D., M, D., N, G., C, J., DL, R., MA, S., CG, C., et MA, M. (2009). **Bioportal: ontologies and integrated data resources at the click of a mouse**. *PMC free article, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2732233/*.

[Ngan et al., 2006] Ngan, L. D., Hang, T. M., et Goh, A. E. S. (2006). **Semantic similarity between concepts from different owl ontologies**. In *Industrial Informatics, 2006 IEEE International Conference on*, pages 618–623. IEEE.

[Nguyen et al., 2011] Nguyen, K.-N. T., Cerf, L., Plantevit, M., et Boulicaut, J.-F. (2011). **Multidimensional association rules in boolean tensors**. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 570–581. SIAM.

[Ning L, 2010] Ning L, Guanyu L, L. S. (2010). **Using formal concept analysis for maritime ontology building. international forum on information technology and applications**. *International Forum on Information Technology and Applications*.

[Noulas et al., 2012] Noulas, A., Scellato, S., Lathia, N., et Mascolo, C. (2012). **A random walk around the city: New venue recommendation in location-based social networks**. In *Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (socialcom)*, pages 144–153. Ieee.

[Olsson et al., 2011] Olsson, C., Petrov, P., Sherman, J., et Perez-Lopez, A. (2011). **Finding and explaining similarities in linked data.** In *STIDS*, pages 52–59.

[Oosterhuis, 1990] Oosterhuis, K. (1990). **Artificial Intuition: arbeiten am Computer**. Aedes. 3 cites: https://scholar.google.com/scholar?cites=13240536698865291663&as$_sdt$ = 2005&$sciodt$ = 0, 5&hl = en.

[Othman et al., 2008] Othman, R. M., Deris, S., et 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.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., et Winograd, T. (1998). **The pagerank citation ranking: Bringing order to the web**. Technical report;, Stanford University.

[Pasquier et al., 1999a] Pasquier, N., Bastide, Y., Taouil, R., et Lakhal, L. (1999a). **Discovering frequent closed itemsets for association rules**. In *International Conference on Database Theory*, pages 398–416. Springer.

[Pasquier et al., 1999b] Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L., et others (1999b). **Efficient mining of association rules using closed itemset lattices**. *Information systems*, 24(1):25–46.

[Paul et al., 2012] Paul, R., Groza, T., Zankl, A., et Hunter, J. (2012). **Semantic similarity-driven decision support in the skeletal dysplasia domain**. In *International Semantic Web Conference*, pages 164–179. Springer.

[Peixoto et al., 2015] Peixoto, R., Cruz, C., et Silva, N. (2015). **Semantic hmc: Ontology-described hierarchy maintenance in big data context**. In Ciuciu, I., Panetto, H., Debruyne, C., Aubry, A., Bollen, P., Valencia-García, R., Mishra, A., Fensel, A., et Ferri, F., editors, *On the Move to Meaningful Internet Systems: OTM 2015 Workshops*, pages 492–501, Cham. Springer International Publishing.

[Peixoto et al., 2016] Peixoto, R., Hassan, T., Cruz, C., Bertaux, A., et Silva, N. (2016). **An unsupervised classification process for large datasets using web reasoning**. In *Proceedings of the International Workshop on Semantic Big Data*, SBD '16, pages 9:1–9:6, New York, NY, USA. **ACM**.

[Pekar et al., 2002] Pekar, V., et Staab, S. (2002). **Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision**. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[Peng X, 2007] Peng X, Z. W. (2007). **An incremental and fca-based ontology construction method for semantics-based component retrieval**. *Seventh International Conference on Quality Software*.

[Pennington et al., 2014] Pennington, J., Socher, R., et Manning, C. D. (2014). **Glove: Global vectors for word representation.** In *EMNLP*, volume 14, pages 1532–1543.

[Penrose, 1989] Penrose, R. (1989). **The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics**. Oxford University Press, Inc., New York, NY, USA.

[Pirró, 2009] Pirró, G. (2009). **A semantic similarity metric combining features and intrinsic information content**. *Data & Knowledge Engineering*, 68(11):1289–1308.

[Pirró et al., 2010] Pirró, G., et Euzenat, J. (2010). **A feature and information theoretic framework for semantic similarity and relatedness**. *The Semantic Web–ISWC 2010*, pages 615–630.

[Pirró et al., 2008] Pirró, G., et Seco, N. (2008). **Design, implementation and evaluation of a new semantic similarity metric combining features and intrinsic information content**. *On the Move to Meaningful Internet Systems: OTM 2008*, pages 1271–1288.

[Pittet, 2014] Pittet, P. (2014). **OntoVersionGraph : a change management methodology dedicated to formal ontologies and their user views in a collaborative context : application to SHOIN (D) ontologies**.

[Priya et al., 2015] Priya, M., et Kumar, C. A. (2015). **A survey of state of the art of ontology construction and merging using formal concept analysis**. *Indian Journal of Science and Technology, Vol 8(24)*.

[Prokopchuk, 2017] Prokopchuk, Y. (2017). **Algorithms of artificial intuition for implementation of strong ai**. *Construction, materials science, mechanical . . . .* Query date: 2019-09-03.

[Rada et al., 1989] Rada, R., Mili, H., Bicknell, E., et Blettner, M. (1989). **Development and application of a metric on semantic nets**. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30.

[Rais et al., 2016] Rais, M., et Lachkar, A. (2016). **Biomedical word sense disambiguation context-based: improvement of SenseRelate method**. In *2016 International Conference on Information Technology for Organizations Development (IT4OD)*, pages 1–6. IEEE.

[Ramage et al., 2009] Ramage, D., Rafferty, A. N., et Manning, C. D. (2009). **Random walks for text semantic similarity**. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, pages 23–31. Association for Computational Linguistics.

[Ranganathan, 1933] Ranganathan, S. (1933). **Colon Classification**. *Asia Madras, London*.

[Ranwez et al., 2006] Ranwez, S., Ranwez, V., Villerd, J., et Crampes, M. (2006). **Ontological distance measures for information visualisation on conceptual maps**. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 1050–1061. Springer.

[Resnik, 1995] Resnik, P. (1995). **Using information content to evaluate semantic similarity in a taxonomy**. *arXiv preprint cmp-lg/9511007*.

[Resnik et al., 1999] Resnik, P., et others (1999). **Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language**. *J. Artif. Intell. Res.(JAIR)*, 11:95–130.

[Resnik, 1993] Resnik, P. S. (1993). **Selection and information: a class-based approach to lexical relationships**. *IRCS Technical Reports Series*, page 200.

[Ricci et al., 2011] Ricci, F., Rokach, L., et Shapira, B. (2011). **Introduction to recommender systems handbook.** In Ricci, F., Rokach, L., Shapira, B., et Kantor, P. B., editors, *Recommender Systems Handbook*, pages 1–35. Springer.

[Rodríguez et al., 2003] Rodríguez, M. A., et Egenhofer, M. J. (2003). **Determining semantic similarity among entity classes from different ontologies**. *IEEE transactions on knowledge and data engineering*, 15(2):442–456.

[Rohit, 2016] Rohit, S. (2016). **Association rule mining algorithms: Survey**. *International Research Journal of Engineering and Technology*.

[Salton et al., 1988] Salton, G., et Buckley, C. (1988). **Term-weighting approaches in automatic text retrieval**. *Inf. Process. Manage.*, 24(5):513–523.

[Sánchez et al., 2012a] Sánchez, D., et Batet, M. (2012a). **A new model to compute the information content of concepts from taxonomic knowledge**. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 8(2):34–50.

[SáNchez et al., 2013] SáNchez, D., et Batet, M. (2013). **A semantic similarity method based on information content exploiting multiple ontologies**. *Expert Systems with Applications*, 40(4):1393–1399.

[Sánchez et al., 2012b] Sánchez, D., Batet, M., Isern, D., et Valls, A. (2012b). **Ontology-based semantic similarity: A new feature-based approach**. *Expert Systems with Applications*, 39(9):7718–7728.

[Sanderson et al., 1999] Sanderson, M., et Croft, B. (1999). **Deriving concept hierarchies from text**. *22nd annual international ACM SIGIR conference on Research and development in information retrieval*, page 206–213.

[Sanfeliu et al., 1983] Sanfeliu, A., et Fu, K.-S. (1983). **A distance measure between attributed relational graphs for pattern recognition**. *IEEE transactions on systems, man, and cybernetics*, (3):353–362.

[Santos et al., 2009] Santos, A. P., et Rodrigues, F. (2009). **Multi-label hierarchical text classification using the acm taxonomy**.

[Sarkar et al., 2008] Sarkar, P., Moore, A. W., et Prakash, A. (2008). **Fast incremental proximity search in large graphs**. In *Proceedings of the 25th international conference on Machine learning*, pages 896–903. ACM.

[Saruladha et al., 2010] Saruladha, K., Aghila, G., et Raj, S. (2010). **A survey of semantic similarity methods for ontology based information retrieval**. In *Machine Learning and Computing (ICMLC), 2010 Second International Conference on*, pages 297–301. IEEE.

[Satuluri et al., 2009] Satuluri, V., et Parthasarathy, S. (2009). **Scalable graph clustering using stochastic flows: Applications to community discovery**. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 737–746, New York, NY, USA. ACM.

[Schlicker et al., 2006] Schlicker, A., Domingues, F. S., Rahnenführer, J., et Lengauer, T. (2006). **A new measure for functional similarity of gene products based on gene ontology**. *BMC bioinformatics*, 7(1):302.

[Searle, 1990] Searle, J. R. (1990). **Is the Brain's Mind a Computer Program?** *Scientific American*, 262:26–31.

[Shet et al., 2012] Shet, K., Acharya, U. D., et others (2012). **A new similarity measure for taxonomy based on edge counting**. *arXiv preprint arXiv:1211.4709*.

[Silla et al., 2011] Silla, Jr., C. N., et Freitas, A. A. (2011). **A survey of hierarchical classification across different application domains**. *Data Min. Knowl. Discov.*, 22(1-2):31–72.

[Singh et al., 2013] Singh, J., Saini, M., et Siddiqi, S. (2013). **Graph based computational model for computing semantic similarity**. *Emerging Research in Computing, Information, Communication and Applications, ERCICA*, 2013:501–507.

[Sirin et al., 2007] Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., et Katz, Y. (2007). **Pellet: A practical owl-dl reasoner**. *Web Semantics*, 5(2):51–53.

[Sloman, 1971] Sloman, A. (1971). **Interactions between philosophy and artificial intelligence: the role of intuition and non-logical reasoning in intelligence**. In *IJCAI 1971*.

[Snelting, 2000] Snelting, G. (2000). **Software reengineering based on concept lattices**. In *Software Maintenance and Reengineering, 2000. Proceedings of the Fourth European*, pages 3–10. IEEE.

[Stanovich et al., 2000] Stanovich, K. E., et West, R. F. (2000). **Individual differences in reasoning: Implications for the rationality debate?** *Behavioral and Brain Sciences*, 23:645–665.

[Stojanovic et al., 2001] Stojanovic, N., Maedche, A., Staab, S., Studer, R., et Sure, Y. (2001). **Seal: a framework for developing semantic portals**. In *Proceedings of the 1st international conference on Knowledge capture*, pages 155–162. ACM.

[Studer et al., 1998] Studer, R., Benjamins, V. R., et Fensel, D. (1998). **Knowledge engineering: principles and methods. data knowl eng 25(1-2):161-197**. *Data Knowledge Engineering*, 25:161–197.

[Sun et al., 2016a] Sun, Y., Bie, R., et Zhang, J. (2016a). **Measuring semantic-based structural similarity in multi-relational networks**. *International Journal of Data Warehousing and Mining (IJDWM)*, 12(1):20–33.

[Sun et al., 2016b] Sun, Y., Lu, C., Bie, R., et Zhang, J. (2016b). **Semantic relation computing theory and its application**. *Journal of Network and Computer Applications*, 59:219–229.

[Sun et al., 2016c] Sun, Z., Han, J., et Hao, H.-W. (2016c). **Structural feature-based event clustering for short text streams**. In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3252–3257. IEEE.

[Sure et al., 2009] Sure, Y., Staab, S., et Studer, R. (2009). **Ontology engineering methodology**. *Springer-Verlag Berlin Heidelberg*.

[Sussna, 1993] Sussna, M. (1993). **Word sense disambiguation for free-text indexing using a massive semantic network**. In *Proceedings of the second international conference on Information and knowledge management*, pages 67–74. ACM.

[Sy et al., 2012] Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., et Ranwez, V. (2012). **User centered and ontology based information retrieval system for life sciences**. *BMC bioinformatics*, 13(Suppl 1):S4.

[Ta et al., 2012] Ta, C. D. C., et Tuoi, P. T. (2012). **Improving the formal concept analysis algorithm to construct domain ontology**. *2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 74–78.

[Taieb et al., 2014] Taieb, M. A. H., Aouicha, M. B., et Hamadou, A. B. (2014). **Ontology-based approach for measuring semantic similarity**. *Engineering Applications of Artificial Intelligence*, 36:238–261.

[Tang S, 2010] Tang S, C. Z. (2010). **Using the format concept analysis to construct the tourism information ontology**. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery*.

[Tao et al., 2009] Tao, W., et He, P. (2009). **Intuitive learning and artificial intuition networks**. *2009 Second International Conference on ....* 7 cites: https://scholar.google.com/scholar?cites=17782377369203984806&as$_sdt$ = 2005&$sciodt$ = 0, 5&hl = en.

[Tsarkov et al., 2006] Tsarkov, D., et Horrocks, I. (2006). **Fact++ description logic reasoner: System description**. In *Proc. of the Int. Joint Conf. on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 292–297. Springer.

[Tsoumakas et al., 2007] Tsoumakas, G., et Katakis, I. (2007). **Multi-label classification: An overview**. *Int J Data Warehousing and Mining*, 2007:1–13.

[Tsukiyama et al., 1977] Tsukiyama, S., Ide, M., Ariyoshi, H., et Shirakawa, I. (1977). **A new algorithm for generating all the maximal independent sets**. *SIAM Journal on Computing*, 6(3):505–517.

[Turney et al., 2010] Turney, P. D., et Pantel, P. (2010). **From frequency to meaning: Vector space models of semantics**. *Journal of artificial intelligence research*, 37:141–188.

[Tversky, 1977] Tversky, A. (1977). **Features of similarity.** *Psychological review*, 84(4):327.

[Urbani, 2013] Urbani, J. (2013). **Three laws learned from web-scale reasoning**. *AAAI Fall Symposium - Technical Report*, pages 76–79.

[Urbani et al., 2011] Urbani, J., van Harmelen, F., Schlobach, S., et Bal, H. (2011). **Querypie: Backward reasoning for owl horst over very large knowledge bases**. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., et Blomqvist, E., editors, *The Semantic Web – ISWC 2011*, pages 730–745, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Valtchev et al., 2004] Valtchev, P., Missaoui, R., et Godin, R. (2004). **Formal concept analysis for knowledge discovery and data mining: The new challenges**. In *ICFCA*, volume 2961, pages 352–371. Springer.

[van Dongen, 2000] van Dongen, S. (2000). **A Cluster algorithm for graphs**. PhD thesis, Centrum voor Wiskunde en Informatica.

[Vandenbussche et al., 2009] Vandenbussche, P.-Y., et Charlet, J. (2009). **Méta-modèle général de description de ressources terminologiques et ontologiques**. In *IC 2009 - 20èmes Journées Francophones d'Ingénierie des Connaissances*, volume 20, page à paraître, Hammamet, Tunisia. 12 pages.

[Varelas et al., 2005] Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E. G., et Milios, E. E. (2005). **Semantic similarity methods in wordnet and their application to information retrieval on the web**. In *Proceedings of the 7th annual ACM international workshop on Web information and data management*, pages 10–16. ACM.

[Virtanen, 2003] Virtanen, S. (2003). **Clustering the chilean web**. In *Proceedings of the First Conference on Latin American Web Congress*, LA-WEB '03, pages 229–, Washington, DC, USA. IEEE Computer Society.

[Von Luxburg et al., 2010] Von Luxburg, U., Radl, A., et Hein, M. (2010). **Hitting and commute times in large graphs are often misleading**. *arXiv preprint arXiv:1003.1266*.

[Voutsadakis, 2002] Voutsadakis, G. (2002). **Polyadic concept analysis**. *Order*, 19(3):295–304.

[Wang et al., 2012] Wang, J., Xie, D., Lin, H., Yang, Z., et Zhang, Y. (2012). **Filtering gene ontology semantic similarity for identifying protein complexes in large protein interaction networks**. *Proteome science*, 10(1):S18.

[Webber et al., 2013] Webber, J., Robinson, I., et Eifrem, E. (2013). **Graph databases**.

[Wenyin et al., 2010] Wenyin, L., Quan, X., Feng, M., et Qiu, B. (2010). **A short text modeling method combining semantic and statistical information**. *Information Sciences*, 180(20):4031–4041.

[Widmer et al., 1996] Widmer, G., et Kubat, M. (1996). **Learning in the presence of concept drift and hidden contexts**. *Mach. Learn.*, 23(1):69–101.

[Wilson et al., 2008] Wilson, T. D., et Bar-Anan, Y. (2008). **The unseen mind**. *Science*, 321(5892):1046–1047.

[Wu et al., 1994a] Wu, Z., et Palmer, M. (1994a). **Verbs semantics and lexical selection**. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

[Wu et al., 1994b] Wu, Z., et Palmer, M. (1994b). **Verbs semantics and lexical selection**. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.

[Yamaguchi, 2001] Yamaguchi, T. (2001). **Acquiring conceptual relationships from domain-specific texts**. pages 13–18.

[Yao et al., 2012] Yao, Y., Tong, H., Xu, F., et Lu, J. (2012). **Subgraph extraction for trust inference in social networks.** In *ASONAM*, pages 163–170. IEEE Computer Society.

[Yazdani et al., 2013] Yazdani, M., et Popescu-Belis, A. (2013). **Computing text semantic relatedness using the contents and links of a hypertext encyclopedia**. *Artificial Intelligence*, 194:176–202.

[Yu et al., 2014] Yu, H.-F., Jain, P., Kar, P., et Dhillon, I. S. (2014). **Large-scale multi-label learning with missing labels**. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–593–I–601. JMLR.org.

[Zadeh, 1965] Zadeh, L. A. (1965). **Fuzzy sets**. *Information Control*, 8:338–353.

[Zhang et al., 2009] Zhang, Y., Zhang, L., Nie, G., et Shi, Y. (2009). **A survey of interestingness measures for association rules**. In *International Conference on Business Intelligence and Financial Engineering, 2009. BIFE'09.*, pages 460–463. IEEE.

[Zheng et al., 2016] Zheng, W., Zou, L., Peng, W., Yan, X., Song, S., et Zhao, D. (2016). **Semantic sparql similarity search over rdf knowledge graphs**. *Proceedings of the VLDB Endowment*, 9(11):840–851.

[Zhong et al., 2002] Zhong, J., Zhu, H., Li, J., et Yu, Y. (2002). **Conceptual graph matching for semantic search**. In *International Conference on Conceptual Structures*, pages 92–106. Springer.

[Zhu et al., 2017] Zhu, G., et Iglesias, C. A. (2017). **Computing semantic similarity of concepts in knowledge graphs**. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85.

[Škopljanac Mačina et al., 2014] Škopljanac Mačina, F., et Blašković, B. (2014). **Formal concept analysis – overview and applications**. *Procedia Engineering*, 69:1258 – 1267. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

21

PUBLICATIONS

Revues

2019 Bazin, Alexandre, Nicolas Gros, Aurélie Bertaux et Christophe Nicolle. "Condensed Representations of Association Rules in n-ary Relations". *IEEE Transactions on Knowledge and Data Engineering (TKDE)* ***(Impact Factor 2.775- Quartile Q1)***, nº to appear (2019).

2019 Hugol-Gential, Clémentine, Marie Simon, Aurélie Bertaux, Ouassila Narsis, Rami Belkaroui, Amira Mouakher et Christophe Nicolle. "Une ontologie de la vigne au verre : la terminologie professionnelle au regard des savoirs mobilisés en viticulture". *Recherches en Communication* 48, nº 48 (2019).

2018 Batrouni, Marwan, Aurélie Bertaux et Christophe Nicolle. "Scenario analysis, from BigData to black swan". *Computer Science Review* - ***(Quartile Q1)*** 28 (2018) : 131–139. ISSN : 1574-0137. doi :https://doi.org/10.1016/j.cosrev.2018.02.001. http://www.sciencedirect.com/science/article/pii/S157401371730151X.

2016 Peixoto, Rafael, Thomas Hassan, Christophe Cruz, Aurélie Bertaux et Nuno Silva. "Hierarchical Multi-Label Classification Using Web Reasoning for Large Datasets". *Open Journal Of Semantic Web*, 2016. doi :10.19210/1006.3.1.1. https://hal.archives-ouvertes.fr/hal-01356375.

2015 Chabot, Yoan, Aurélie Bertaux, Christophe Nicolle et Tahar Kechadi. "An Ontology-Based Approach for the Reconstruction and Analysis of Digital Incidents Timelines". *Digital Investigation* - ***(Index ISI WoS IF=1.648) (Quartile Q2)***, Digital Investigation, Special Issue on Big Data and Intelligent Data Analysis, 2015, 18. http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=S2rbTXTFkN7h2fBpngU&page=1&doc=1.

2014 Chabot, Yoan, Aurélie Bertaux, Christophe Nicolle et Tahar Kechadi. "A Complete Formalized Knowledge Representation Model for Advanced Digital Forensics Timeline Analysis". Fourteenth Annual DFRWS Conference, Denver, USA, *Digital Investigation* - ***(Index ISI WoS - IF=1.648) (Quartile Q2)*** 11, nº 2 (2014) : S95–S105. doi :10.1016/j.diin.2014.05.009. http://www.journals.elsevier.com/digital-investigation/.

2012 Martin, Sylvain, Aurélie Bertaux, Florence Le Ber, Elodie Maillard et Gwenaël Imfeld. "Seasonal Changes of Macroinvertebrate Communities in a Stormwater Wetland Collecting Pesticide Runoff From a Vineyard Catchment (Alsace, France)". *Archives of Environmental Contamination and Toxicology*- ***(Index ISI WoS IF=2,012) (Quartile Q1)*** 62, nº 1 (2012) : 29–41. doi :10.1007/s00244-011-9687-6. https://www.researchgate.net/journal/0090-4341_Archives_of_Environmental_Contamination_and_Toxicology.

2010 Bertaux, Aurélie, Florence Le Ber, Agnès Braud et Michèle Trémolières. "Mining Complex Hydrobiological Data with Galois Lattices". *International Journal of Computing & Information Sciences* 7, nº 2 (2010) : 63–77.

Conferences

2019 Belkaroui, Rami, Amira Mouakher, Aurélie Bertaux, Ouassila Labbani, Clémentine Hugol-Gential et Christophe Nicolle. "WINECLOUD : Une ontologie dévénements pour la modélisation sémantique des données de capteurs hétérogènes". In *Revue des Nouvelles Technologies de l'Information. EGC 2019, vol. RNTI-E-35*, 379–380. 2019.

2019 Hugol-Gential, Clémentine, Marie Simon, Aurélie Bertaux, Ouassila Narsis, Rami Belkaroui, Amira Mouakher et Christophe Nicolle. "Une ontologie de la culture de la vigne : des savoirs académiques aux savoirs d'expérience". In *Recherches en communication*, 111–129. 2019.

2019 Hugol-Gential, Clémentine, Marie Simon, Aurélie Bertaux, Ouassila Narsis, Rami Belkaroui et Christophe Nicolle. "Traçabilité et analyse BigData sur la chaîne de valeur vitivinicole. De la science à l'expérience, la pluralité des savoirs en viticulture : le cas des maladies". In *Univigne*. 2019.

2019 Mouakher, Amira, Rami Belkaroui, Aurélie Bertaux, Ouassila Labbani, Clémentine Hugol-Gential et Christophe Nicolle. "An Ontology-Based Monitoring System in Vineyards of the Burgundy Region". In *28th IEEE International Conference on Enabling Technologies : Infrastructure for Collaborative Enterprises (WETICE-2019)*, to appear. IEEE, 2019.

2018 Batrouni, Marwan, Aurélie Bertaux et Christophe Nicolle. “Analyse Ontologique de scénario dans un contexte Big Data”. In *Extraction et Gestion des Connaissances, EGC 2018, Paris, France, January 23-26, 2018*, 387–388. 2018. http://editions-rnti.fr/?inprocid=1002414.

2018 Batrouni, Marwan, Steven Finch, Scott Wilson, Aurélie Bertaux et Christophe Nicolle. “Intelligent Cloud Storage Management for Layered Tiers”. In *Cooperative Design, Visualization, and Engineering - 15th International Conference, CDVE 2018, Hangzhou, China, October 21-24, 2018, Proceedings*, 33–43. 2018. doi :10.1007/978-3-030-00560-3_5. https://doi.org/10.1007/978-3-030-00560-3%5C_5.

Juin 2018 Bazin, Alexandre, et Aurélie Bertaux. “k-Partite Graphs as Contexts”. In *The 14th International Conference on Concept Lattices and Their Applications (CLA2018)*. Olomouc, Czech Republic, juin 2018. https://hal.archives-ouvertes.fr/hal-01964756.

2018 Belkaroui, Rami, Aurélie Bertaux, Ouassila Labbani, Clémentine Hugol-Gential et Christophe Nicolle. “Towards Events Ontology Based on Data Sensors Network for Viticulture Domain”. In *Proceedings of the 8th International Conference on the Internet of Things*, 44 :1–44 :7. IOT ’18. Santa Barbara, California : **ACM**, 2018. ISBN : 978-1-4503-6564-2. doi :10.1145/3277593.3277619. http://doi.acm.org/10.1145/3277593.3277619.

2017 Hassan, T., C. Cruz et A. Bertaux. “Predictive and evolutive cross-referencing for web textual sources”. In *2017 Computing Conference*, 1114–1122. 2017. doi :10.1109/SAI.2017.8252230.

2017 Hassan, Thomas, Christophe Cruz et Aurélie Bertaux. “Ontology-based Approach for Unsupervised and Adaptive Focused Crawling”. In *Proceedings of The International Workshop on Semantic Big Data*, 2 :1–2 :6. SBD ’17. Chicago, Illinois : ACM, 2017. ISBN : 978-1-4503-4987-1. doi :10.1145/3066911.3066912. http://doi.acm.org/10.1145/3066911.3066912.

2016 Peixoto, Rafael, Thomas Hassan, Christophe Cruz, Aurélie Bertaux et Nuno Silva. “An Unsupervised Classification Process for Large Datasets Using Web Reasoning”. In *Proceedings of the International Workshop on Semantic Big Data*, 9 :1–9 :6. SBD ’16. San Francisco, California : ACM, 2016. ISBN : 978-1-4503-4299-5. doi :10.1145/2928294.2928301. http://doi.acm.org/10.1145/2928294.2928301.

2015 Benezeth, Yannick, Aurélie Bertaux et Aldric Manceau. “Bag-of-word based brand recognition using Markov Clustering Algorithm for codebook generation”. In ***IEEE** International Conference on Image Processing (ICIP)*. Québec, France, 2015.

2015 Chabot, Yoan, Aurélie Bertaux, Christophe Nicolle et Tahar Kechadi. “De la scène de crime aux connaissances : représentation d’évènements et peuplement d’ontologie appliqués au domaine de la criminalistique informatique”. In *Extraction et Gestion des Connaissances 2015*, sous la direction d’Editions RNTI. Revue des Nouvelles Technologies de l’Information. Luxembourg, Luxembourg, 2015.

2015 Chabot, Yoan, Aurélie Bertaux, Christophe Nicolle et Tahar Kechadi. “Représentation sémantiquement riche d’évènements pour le domaine de la criminalistique informatique”. In *Extraction et Gestion des Connaissances*. Revue des Nouvelles Technologies de l’Information. Luxembourg, Luxembourg, 2015.

2015 Peixoto, Rafael, Thomas Hassan, Christophe Cruz, Aurélie Bertaux et Nuno Silva. “Semantic HMC : A Predictive Model Using Multi-label Classification for Big Data”. In *Trustcom/BigDataSE/ISPA, 2015 IEEE*. Helsinki, France, 2015. doi :10.1109/Trustcom.2015.578. https://hal.archives-ouvertes.fr/hal-01356367.

2015 Peixoto, Rafael, Hassan Thomas, Christophe Cruz, Aurélie Bertaux et Nuno Silva. “Semantic HMC for Business Intelligence using Cross-Referencing”. In *14th International Conference on Informatics in Economy - **(Index ISI WoS)***. Bucharest, Romania, 2015. http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=S2rbTXTFkN7h2fBpngU&excludeEventConfig=ExcludeIfFromFullRecPage&page=1&doc=3.

2015 Raad, Joe, Aurélie Bertaux et Christophe Cruz. “A Survey on how to Cross-Reference Web Information Sources”. In *Science and Information Conference*, sous la direction de **IEEE**, 609–618. Londres, United Kingdom, 2015. doi :10.1109/SAI.2015.7237206.

2015 Thomas, Hassan, Rafael Peixoto, Christophe Cruz, Aurélie Bertaux et Nuno Silva. “Analyse Sémantique du Big Data par Classification Hiérarchique Multi-Label”. In *extraction et Gestion des Connaissances*. Revue des Nouvelles Technologies de l’Information. Luxembourg, Luxembourg, 2015.

2015 Werner, David, Thomas Hassan, Aurelie Bertaux, Christophe Cruz et Nuno Silva. “Semantic-Based Recommender System with Human Feeling Relevance Measure”. In *Intelligent Systems in Science and Information 2014 : Extended and Selected Results from the Science and Information Conference 2014*, sous la direction de Kohei Arai, Supriya Kapoor et Rahul Bhatia, 177–191. Cham : Springer International Publishing, 2015. ISBN : 978-3-319-14654-6. doi :10.1007/978-3-319-14654-6_11. http://dx.doi.org/10.1007/978-3-319-14654-6_11.

2014 Chabot, Yoan, Aurélie Bertaux, Tahar Kechadi et Christophe Nicolle. “Reconstruction et analyse sémantique de chronologies cybercriminelles”. In *Extraction et Gestion des Connaissances 2014*, sous la direction d’Editions RNTI, t. RNTI-E-26, 521–524. Revue des Nouvelles Technologies de l’Information. Rennes, France, 2014.

2014 Chabot, Yoan, Aurélie Bertaux, Christophe Nicolle et Tahar Kechadi. “A Complete Formalized Knowledge Representation Model for Advanced Digital Forensics Timeline Analysis”. In *Fourteenth Annual DFRWS Conference - **(Index ISI WoS)(Quartile Q2)***, t. 11, S95–S105. Digital Investigation 2. Denver, United States : Elsevier, 2014. doi :10.1016/j.diin.2014.05.009. http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=S2rbTXTFkN7h2fBpngU&excludeEventConfig=ExcludeIfFromFullRecPage&page=1&doc=4.

2014 Chabot, Yoan, Aurélie Bertaux, Christophe Nicolle et Tahar Kechadi. “Automatic Timeline Construction and Analysis For Computer Forensics Purposes”. In ***IEEE** Joint Intelligence & Security Informatics Conference 2014 (IEEE JISIC2014) - **(Index ISI WoS)***, 4. La Haye, Netherlands, 2014.

2014 David, Werner, Hassan Thomas, Christophe Cruz, Aurélie Bertaux et Nuno Silva. “Using DL-Reasoner for Hierarchical Multilabel Classification applied to Economical e-News”. In *SAI, Science and Information Conference, 2014 - **(Index ISI WoS)***, 313–320. Science and Information Conference (SAI), 2014. london, United Kingdom : **IEEE**, 2014. doi :10.1109/SAI.2014.6918205. http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=S2rbTXTFkN7h2fBpngU&excludeEventConfig=ExcludeIfFromFullRecPage&page=1&doc=5.

2014 Thomas, Hassan, David Werner, Aurélie Bertaux et Christophe Cruz. “profile refinement in ontology-based recommander systems for economical e-news”. In *IE, The 14th International Conference on Informatics in Economy - **(Index ISI WoS)***. Bucharrest, Romania, 2014.

2014 Werner, David, Christophe Cruz et Aurélie Bertaux. “Evaluation de la pertinence dans un système de recommandation sémantique de nouvelles économiques”. In *EGC - Fouille de données complexes*. EGC - Fouille de données complexes. Rennes, France, 2014.

2014 Werner, David, Nuno Silva, Christophe Cruz et Aurélie Bertaux. “An Ontology-Based Recommender System using Hierarchical Multiclassification for Economical e-News”. In *Science and Information Conference*. Londres, United Kingdom : **IEEE/Springer**, 2014.

2013 Tehrani, Behrooz Omidvar, Sihem Amer-Yahia, Alexandre Termier, Aurélie Bertaux, Éric Gaussier et Marie-Christine Rousset. “Towards a Framework for Semantic Exploration of Frequent Patterns”. In *IMMoA 2013 - International Workshop on Information Management in Mobile Application (in conjunction with VLDB 2013)*, sous la direction de Thierry Delot, Sandra Geisler, Sergio Ilarri et Christoph Quix, 1075 : 7–14. Http ://ceur-ws.org/Vol-1075/ - ISSN : 1613-0073. Riva del Garda, Trento, Italy : CEUR-WS, 2013.

2012 López-Cueva, Patricia, Aurélie Bertaux, Alexandre Termier, Jean-François Méhaut et Miguel Santana. “Debugging Multimedia Application Traces through Periodic Pattern Mining”. In *EMSOFT 2012, part of ESWEEK - Embedded Systems Week - **Index ISI WOS***, 13–22. Session 1A : Testing and Characterization of Embedded Software. Tampere, Finland : **ACM**, 2012. doi :10.1145/2380356.2380366. http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=8&SID=P1IhTi7rQL2fRvFrK9x&page=1&doc=5.

2012 López-Cueva, Patricia, Aurélie Bertaux, Alexandre Termier, Jean-François Méhaut et Miguel Santana. “Periodic Pattern Mining of Embedded Multimedia Application Traces”. In *EMC 2012 - 7th International Conference on Embedded and Multimedia Computing- **(Index ISI WoS)***, 181 : 29–37. Lecture Notes in Electrical Engineering (LNEE) / Engineering. Gwangju, South Korea : **Springer**, 2012. doi :10.1007/978-94-007-5076-0_4. http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=5&SID=S2rbTXTFkN7h2fBpngU&excludeEventConfig=ExcludeIfFromFullRecPage&page=1&doc=6.

2009 Bertaux, Aurélie, Florence Le Ber et Agnès Braud. “Correspondances de Galois pour la manipulation de contextes flous multi-valués”. In *Extraction et Gestion de Connaissances*, 193–198. RNTI. Strasbourg, France : Cépaduès, 2009.

2009 Bertaux, Aurélie, Florence Le Ber, Agnès Braud et Michèle Trémolières. “Identifying ecological traits : a concrete FCA-based approach”. In *7th International Conference on Formal Concept Analysis - ICFCA 2009*, sous la direction de S. Rudolph S. Ferré, 5548 : 224–236. LNAI. Darmstadt, Germany : **Springer**, 2009.

2009 Bertaux, Aurélie, Florence Le Ber, Pulu Li et Michèle Trémolières. “Combiner treillis de Galois et analyse factorielle multiple pour l’analyse de traits biologiques”. In *XVIème Recontres de la Société Francophone de Classification - SFC 2009*, 117–120. Grenoble, France, 2009.

2007 Bertaux, Aurélie, Agnès Braud et Florence Le Ber. “Mining Complex Hydrobiological Data with Galois Lattices”. In *International Workshop on Advances in Conceptual Knowledge Engineering (ACKE’07)*, sous la direction d’A.M.Tjoa et R.R. Wagner, 519–523. Regensburg, Germany, 2007.

2007 Deruyver, Aline, Yann Hodé et Aurélie Bertaux. “Graph Consistency Checking for Automatic Image Segmentation Driven by Knowledge”. In *International Conference on Human Machine Interaction*, 246–251. Timimoun, Algeria, 2007.

2006 Korczak, J., et Aurélie Bertaux. “Extension de l’algorithme CURE aux fouilles de données volumineuses”. In *extraction et Gestion des Connaissances*, t. RNTI-E-6, 547–548. Revue des Nouvelles Technologies de l’Information. Villeneuve d’Ascq, France, 2006.

2006 Korczak, Jerzy, et Aurélie Bertaux. “Fouille d’images IRMf : algorithme CURE”. In *Extraction et Gestion des Connaissances*, 107–117. Revue des Nouvelles Technologies de l’Information. lille, France, 2006.

2005 Korczak, Jerzy, et Aurélie Bertaux. “Fouille d’images IRMf : algorithme CURE”. In *Conférence Extraction et Gestion des connaissances (EGC)*, 107–117. Paris, France, 2005.

Books

2014 Chabot, Yoan, Aurélie Bertaux, Tahar Kechadi et Christophe Nicolle. “Event Reconstruction : A State of the Art”. In *Handbook of Research on Digital Crime, Cyberspace Security, and Information Assurance*, 15. IGI Global, 2014.

2014 Werner, David, Christophe Cruz, Aurélie Bertaux et Nuno Silva. “An Ontology-Based Recommender System using Hierarchical Multiclassification for Economical e-News”. In *Studies in Computational Intelligence*. **Springer** book series, 2014.

International workshops and posters

2014 Hassan, Thomas, Rafael Peixoto, Christophe Cruz, Aurlie Bertaux et Nuno Silva. *Semantic HMC for Big Data Analysis*. 2014 **IEEE** International Conference on Big Data. Poster, 2014.

2014 Werner, David, Christophe Cruz et Aurélie Bertaux. *Evaluation de la pertinence dans un système de recommandation sémantique de nouvelles économiques*. Extraction et Gestion des Connaissances. Poster, 2014.

2007 Bertaux, Aurélie, Agnès Braud et Florence Le Ber. *Mining Complex Hydrobiological Data with Galois Lattices*. International Workshop on Advances in Conceptual Knowledge Engineering, 2007.

List of Figures

List of Tables

LISTE DES DÉFINITIONS

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/