Utilisation de proximités sémantiques pour améliorer la recherche et le rendu d'information

Sylvie Ranwez, Vincent Ranwez, Mohameth-François Sy, Jacky Montmain, Michel Crampes

Utilisation de proximités sémantiques pour améliorer la recherche et le rendu d'information

Sylvie Ranwez[1], Vincent Ranwez[2], Mohameth-François Sy[1],
Jacky Montmain[1] et Michel Crampes[1]

[1] LGI2P, Ecole des Mines d'Alès
Site EERIE, Parc scientifique G. Besse, F – 30 035 Nîmes, France
{sylvie.ranwez, mohameth.sy, jacky.montmain,
michel.crampes}@mines-ales.fr
[2] Institut des Sciences de l'Evolution (UMR 5554 CNRS),
Université Montpellier II, CC 064, 34 095 Montpellier Cedex 05, France
vincent.ranwez@univ-montp2.fr

Résumé : Pour exploiter efficacement des corpus documentaires toujours plus volumineux, les moteurs de recherche doivent évoluer. Leurs limites actuelles concernent principalement le fait que la mesure de la pertinence d'un document par rapport à une requête est souvent non-explicite et que l'interaction avec la liste des réponses est limitée.
Nous proposons une méthode et un environnement de requêtage basés sur les ontologies, qui utilisent des opérateurs d'agrégation pour calculer une mesure de pertinence globale, fonction de la proximité sémantique des documents du corpus avec chaque concept de la requête d'une part, et des préférences de l'utilisateur, d'autre part. Nous construisons ensuite une carte sémantique qui reflète la pertinence des documents sélectionnés et explicite leur adéquation avec la requête. Cette interface homme/machine laisse envisager un processus de requêtage itératif et interactif.

Mots-clés : Recherche d'information, proximité sémantique, interaction, cartes conceptuelles, opérateurs d'agrégation.

1 Introduction

Alors que la numérisation en masse et l'accès à un nombre toujours plus élevé de documents bouleversent nos méthodes de veille et de recherche documentaire, une évolution des outils d'indexation et de recherche d'information s'impose. Durant la dernière décennie, les ontologies ont pris une place prépondérante dans les modèles d'indexation de corpus et les techniques de requêtage. Considérant les systèmes basés sur les ontologies, deux tendances se dégagent. La première veut exploiter totalement la sémantique exprimée dans les ontologies ; il en résulte un langage de requêtage souvent abscons, et donc difficile à adopter par un utilisateur néophyte. La seconde approche, au contraire, prône un langage plus intuitif, mais réduit souvent l'ontologie à un simple vocabulaire contrôlé. Un défaut supplémentaire commun à ces deux types d'outils est le manque d'organisation et de justification des résultats. En effet, la

plupart d'entre eux se contentent de présenter à l'utilisateur un ensemble de documents, qui, dans le meilleur des cas, sont triés en fonction d'une mesure de pertinence globale non explicite. Par conséquent, l'utilisateur ne saurait tirer profit de cette mesure pour affiner sa requête dans un processus de recherche itératif.

Cet article propose une approche alternative. La méthode de requêtage présentée suppose que le système dispose d'une ontologie du domaine et d'une base de documents, chacun étant indexé par un ensemble de concepts de l'ontologie. Les termes des requêtes considérées sont donc des concepts ; le problème de la désambigüisation ne se pose donc pas. Sur la base d'une mesure de similarité entre les concepts d'une ontologie, notre système propose de voir la pertinence d'une requête comme l'agrégation des mesures de similarité entre chacun des concepts de la requête et ceux indexant le document. L'opération d'agrégation permet d'intégrer les préférences de l'utilisateur dans la définition de la pertinence de la réponse à la requête. Cette valeur de pertinence est utilisée pour sélectionner et ordonner les documents. Elle permet de favoriser les documents indexés par les concepts exacts de la requête, sans exclure les autres, dès lors qu'ils sont indexés par des concepts proches (relations d'hyperonymie, d'hyponymie). Le système explicite le choix des résultats retenus en détaillant pour chaque document, sous forme de pictogramme, les scores relatifs à chacun des concepts de la requête. Ces pictogrammes sont ensuite positionnés sur une carte sémantique en fonction des scores globaux des documents associés. Ainsi, l'utilisateur en visualisant d'un seul coup d'œil à la fois les documents pertinents et la raison de leur pertinence peut ajuster sa requête en la reformulant ou en modulant l'opérateur d'agrégation. L'apport majeur de cette approche est donc l'interactivité accrue entre l'utilisateur et le système, basée d'une part sur la capacité du système à justifier de ses propositions, d'autre part, sur une définition de la pertinence qui repose sur l'utilisation conjointe des concepts de similarité sémantique et de préférences de l'utilisateur.

Après une présentation de la problématique et de différentes approches concernant les mesures de similarité sémantique et les modèles d'agrégation en recherche d'information (section 2), cet article propose un modèle d'appariement document-requête (section 3). Une mise en œuvre de ce modèle dans un environnement de requêtage est présentée dans la section 4.

2 Problématique et état de l'art

Un système de recherche d'information (SRI) englobe le stockage, la représentation et l'accès à une collection de documents (ou corpus). Dans un contexte applicatif précis, un utilisateur exprime ses besoins sous forme de requêtes, et obtient en retour une liste de documents jugés pertinents. Il souhaite alors que le système diminue le *'silence'* (proportion des documents pertinents non retrouvés) et le *'bruit'* (proportion de documents non pertinents parmi ceux renvoyés) (Prié, 1999) – p. 37.

Des langages de requêtes complexes, SPARQL étant certainement le plus répandu *(http://www.w3.org/TR/rdf-sparql-query/)*, ont été proposés pour exploiter pleinement la plus-value apportée par l'utilisation d'ontologies. Cependant, la complexité de ces

langages nécessite une forte adaptation de l'interface si on veut que ces solutions soient directement utilisables par des utilisateurs quelconques. Par ailleurs, la tâche d'indexation de documents est plus ou moins fastidieuse suivant la richesse sémantique visée. Elle se limite donc souvent à une association, parfois pondérée, de concepts. Dans la suite, nous nous focaliserons sur les SRI permettant une recherche simple et conviviale parmi des documents indexés par un ensemble de concepts.

Les SRI mettent généralement en œuvre trois processus (Belkin *et al.*, 1992) : i) un processus d'indexation qui vise à fournir une représentation compacte et la plus expressive possible des documents et des requêtes ; ii) un processus d'appariement permettant de sélectionner des documents pertinents par rapport à une requête ; et iii) un processus de reformulation des requêtes qui est généralement mis en œuvre comme intermédiaire aux deux précédents.

Plusieurs modèles de SRI ont été développés proposant un cadre théorique aussi bien pour l'indexation que pour la mesure de la pertinence permettant de sélectionner et de classer des documents (Van Rijsbergen, 1979). Dans la plupart des cas, cette mesure consiste à déterminer un score communément appelé RSV (*Retrieval Status Value*). (Farah et Vanderpooten, 2007) considèrent que le calcul du RSV fait appel à deux niveaux d'agrégation : i) le calcul du poids des termes de la requête par rapport à un document (généralement, il s'agit d'une mesure statistique de l'importance d'un terme dans un document donné) et ii) l'utilisation d'un opérateur d'agrégation permettant de calculer un score synthétique, utilisé pour estimer la pertinence globale des documents.

2.1 Requêtage booléen et ses généralisations

Les requêtes booléennes constituent une des formes les plus simples de requête. Cependant, des études ont montré que même les opérateurs booléens simples (ET, OU, NOT) sont rarement utilisés (Jansen *et al.*, 2000), et lorsqu'ils le sont, c'est parfois à contre sens et donc contre productif (Jansen, 2000; Lucas et Topi, 2004). En fait, même si les utilisateurs ont une idée sur le fait que les termes soient nécessairement présents (requête conjonctive) ou un seul d'entre eux (requête disjonctive), ils ne le précisent généralement pas au système.

Les normes Lp de Minkowski-Hölder fournissent un cadre théorique qui permet d'exprimer cette notion à l'aide d'un seul paramètre (Salton et McGill, 1986). De plus elles sont parfaitement adaptées au cas où les termes sont pondérés. Ces poids peuvent être liés à la fréquence d'apparition des termes dans un corpus, e.g. TF-IDF (Salton et McGill, 1986), ou traduire une indexation suivant un modèle d'ensemble flou. Dans ce dernier cas, la pondération représente le degré d'appartenance du terme au document.

Les opérateurs d'agrégation vus précédemment peuvent entraîner des pertes d'informations (Schamber, 1994). En effet, il est difficile de distinguer deux documents ayant le même score de pertinence : l'utilisateur peut vouloir retrouver un document par l'effet combiné de tous les termes de sa requête et non par l'effet d'un terme fortement lié au document. De plus, ils exploitent très peu de ressources sémantiques (thésaurus, ontologie), et ne permettent donc pas de retrouver des documents indexés par des termes sémantiquement proches de ceux de la requête. En

effet, ces opérateurs agrègent uniquement les poids d'un sous-ensemble des termes indexant les documents : ceux qui apparaissent tels quels dans la requête. Cette remarque introduit une seconde problématique : la reformulation de requêtes.

2.2 Reformulation de requêtes

La reformulation d'une requête est une étape intermédiaire entre le processus d'indexation et celui d'appariement. Elle vise à compléter la requête d'un utilisateur en postulant que celui-ci n'est généralement pas capable d'exprimer clairement ses besoins. La reformulation consiste à rajouter/supprimer des termes ou à réévaluer les poids des termes présents. Dans le cas de l'ajout de termes, ces derniers peuvent provenir directement de la collection ou bien de documents pertinents sélectionnés par l'utilisateur, on parle alors de réinjection de pertinence (*relevance feedback*) (Crouch et Yang, 1992). Si le processus est automatique on parlera de pseudo réinjection de pertinence (Boughanem *et al.*, 1999). La reformulation peut aussi être basée sur le vocabulaire issu de ressources externes telles que des ontologies ou des thésaurus (Andreasen, 2003).

Une stratégie classique de reformulation consiste à injecter dans la requête les hyponymes des termes de la requête. Cette méthode permet de compléter de manière intéressante les approches booléennes vues précédemment. En effet, il devient alors possible de renvoyer des documents qui ne sont pas annotés par les termes de la requête et donc d'éviter les silences. Cette approche mixte est, par exemple, utilisée par PUBMED *(http://www.ncbi.nlm.nih.gov/pubmed)* et GOFish (Berriz *et al.*, 2003). Toutefois, aucune différence n'étant faite entre les termes ajoutés et les termes initiaux, la perte d'information pour l'utilisateur est encore accrue. N'étant pas conscient des changements survenus concernant sa requête, il lui devient difficile de savoir rapidement en quoi les documents sélectionnés correspondent à sa requête.

2.3 Mesures de similarité sémantique

L'utilisation de mesures de similarité sémantique permet d'aller encore plus loin que la simple reformulation. En effet, il devient non seulement possible de sélectionner des documents indexés par des concepts absents de la requête mais également de trier ces documents en fonction de leur adéquation à la requête. L'approche décrite dans cet article s'appuie sur ces mesures de similarités, et nous les détaillons donc davantage. Certaines de ces mesures satisfaisant les axiomes d'une distance, des concepts similaires sont également des concepts proches (au sens de cette distance). Nous utiliserons donc indifféremment les termes de proximité ou de similarité sémantique dans la suite de cet article.

Plusieurs mesures de similarité entre objets d'une ontologie ont été proposées dans des contextes différents. On peut, cependant, classer de telles mesures en deux grandes catégories : i) celles utilisant la structure de réseaux sémantiques comme espace métrique (mesures de type *intensionnel*), ii) et celles introduisant des mesures statistiques pour évaluer le contenu informationnel de concepts par le moyen d'instances de concepts ou d'occurrences de termes exprimant un concept dans un corpus (mesures de type *extensionnel*) (Resnik, 1999).

Les mesures de type *intensionnel* utilisent la hiérarchie de concepts pour évaluer la similarité entre concepts. Ces hiérarchies sont vues comme des graphes orientés dont les nœuds correspondent à des concepts et les arcs marquent des relations de subsomption (*is-a*).

Pour évaluer la similarité entre deux concepts C_1 et C_2, (Rada *et al.*, 1989) proposent de considérer $dist(C_1,C_2)$ comme étant le plus court chemin, en termes d'arcs, entre C_1 et C_2. Si tous les arcs de ce chemin vont dans le même sens, l'un des concepts subsume l'autre. À l'opposé, lorsque ce chemin comporte plusieurs changements d'orientation, la relation entre les concepts devient ténue. (Hirst et St Onge, 1998) proposent donc une généralisation de cette similarité, $\pi_{HO}(C_1, C_2)$, en considérant une mesure de la longueur d'un chemin P reliant C_1 et C_2 qui prend en compte à la fois son nombre d'arcs $lg(P)$ et de changements d'orientation $nbC(P)$:

$$\pi_{HO}(C_1, C_2) = \min_{\mathrm{P}=(\mathrm{C1}\rightarrow\mathrm{C2})} \mathrm{lg(P)} + \mathrm{K} * \mathrm{nbC(P)} \quad (1)$$

Le facteur K permet d'ajuster l'impact du nombre de changements de direction. Dans le cas où K=0, on retrouve la proximité de (Rada *et al.*, 1989). À l'opposé, un K très grand impose un nombre minimal de changements de direction et donc un chemin passant par le plus petit ancêtre commun (*least common ancestor*) de C_1 et C_2 noté $lca(C_1,C_2)$ ou par leur plus grand descendant commun. Le *lca* joue donc un rôle important dans plusieurs mesures de similarité. Dès 1994, (Wu et Palmer, 1994) avaient d'ailleurs proposé de l'utiliser dans ce cadre. Cependant en se focalisant sur le *lca*, leur mesure négligeait le fait que des concepts possèdent ou ne possèdent pas des descendants communs. (Zargayouna et Salotti, 2004) en ont proposé une variante qui intègre cette information.

Une limite importante des mesures basées sur les longueurs de chemin dans un graphe *is-a* est due au fait que ses arcs ne représentent pas tous des degrés de généralisation équivalents. La distance proposée dans (Ranwez *et al.*, 2006) prend en compte cette information en s'appuyant sur le nombre de descendants de chaque concept (*ancEx* dénote l'ensemble des ancêtres exclusifs et *desc* les descendants).

$$d_{ISA}(a,b) = |desc(ancEx(a,b)) \cup desc(a) \cup desc(b) - desc(a) \cap desc(b)| \quad (2)$$

Nous avons prouvé que cette mesure satisfait les axiomes d'une distance. Cette approche s'inscrit dans la lignée des méthodes cherchant à évaluer le contenu informationnel d'un concept de manière intensionnelle (sans corpus).

Les mesures de type extensionnel sont généralement basées sur la notion de contenu informationnel d'un concept définie par (Resnik, 1999). L'information d'un concept C_i est donnée par la probabilité $P(C_i)$ d'avoir ce concept ou ses descendants utilisés dans un corpus :

$$\mathrm{IC}(C_i) = -\log(P(C_i)) \quad (3)$$

$P(C_i)$ est le rapport entre le nombre d'instances de C_i ou de ses descendants et le nombre total d'instances dans le corpus. Selon cette approche, plus deux concepts partagent de l'information, plus ils sont similaires.

La mesure de similarité entre deux concepts C_1 et C_2 proposée par (Resnik, 1999) s'appuie sur le concept le plus informatif qui subsume C_1 et C_2 et que l'on notera $\mathrm{MICA}(C_1,C_2)$: $\pi_{Resnik} = IC(MICA(C_1, C_2))$. Notez que $\mathrm{MICA}(C_1,C_2)$ n'est pas forcément un *lca* de C_1 et C_2. (Lin, 1998) définit une variante de cette mesure :

$$\pi_{lin}(C_1, C_2) = \frac{2*IC(\,MICA(C_1,C_2)\,)}{IC(C_1)+IC(C_2)} \quad (4)$$

Cette mesure présente la similarité comme un degré probabiliste de chevauchement des concepts descendants.

3 Requêtage basé sur la proximité sémantique

Dans cette section, nous proposons la mise en œuvre des mesures de similarité sémantique entre concepts d'une ontologie de domaine, dans le contexte d'un système de requêtage. Nous nous plaçons dans le cas où documents et requêtes sont indexés par des concepts issus d'une ontologie d'un domaine et nous focalisons sur le processus d'appariement documents/requêtes en proposant un modèle de pertinence à deux niveaux. À partir d'une mesure de similarité simple et intuitive entre deux concepts d'une ontologie, nous fournissons une extension pour déterminer une mesure de proximité entre un concept et un groupe de concepts. Ensuite, nous utilisons une méthode d'agrégation floue pour combiner les mesures de proximité entre les concepts d'une requête et les documents d'une collection, afin de ranger ces derniers par ordre de pertinence, relativement aux préférences de l'utilisateur. Par souci de rapidité de traitement, y compris dans le cas d'une ontologie de grande taille, la mesure de similarité sémantique entre deux concepts mise en œuvre utilise l'indice de Jaccard entre leurs descendants. L'objectif poursuivi ici est double : i) il s'agit, en cas de silence, d'ajouter de manière automatique d'autres concepts (hyponymes, hyperonymes) proches de ceux de la requête pour améliorer le *rappel* du système ; ii) de présenter à l'utilisateur les documents jugés pertinents en justifiant leur sélection (documents obtenus par correspondance exacte, ou par ajout d'hyponymes ou d'hyperonymes).

3.1 Similarités sémantiques entre concepts et groupes de concepts

Plusieurs mesures de similarité sémantique entre concepts d'une ontologie ont été proposées (voir la section 2). Cependant le choix d'une mesure, dans notre stratégie, impacte fortement trois aspects de notre système : i) la pertinence des documents sélectionnés ; ii) le *rappel* du système ; iii) la compréhension du modèle de jugement des documents par l'utilisateur.

Nous utilisons une mesure de similarité simple en partant du fait qu'un terme de la requête ne peut être remplacé que par un de ses hyponymes ou un de ses hyperonymes. Une telle mesure nous permet d'expliciter pour chaque document retourné par le système, comment la requête a été reformulée. Formellement, notons *hypo*(*C*) l'ensemble des hyponymes d'un concept. Nous définissons la proximité sémantique entre deux concepts C_1 et C_2 comme étant nulle si aucun d'entre eux n'est hyponyme de l'autre et correspondant à l'index de Jaccard entre leurs hyponymes dans le cas contraire :

$$\pi_{JD}(C_1, C_2) = \begin{cases} \frac{hypo(C_1) \cap hypo(C_2)}{hypo(C_1) \cup hypo(C_2)} & si\ C_1 \in hypo(C_2)\ ou\ C_2 \in hypo(C_1) \\ 0 & sinon \end{cases} \quad (5)$$

Nous pouvons noter que :

- $\pi_{JD}(C_1, C_1) = 1$
- $\pi_{JD}(C_1, C_2) < 1$, pour tout C_1 différent de C_2
- $\pi_{JD}(C_1, C_2) > 0$, pour tout C_1 et C_2 ayant une relation d'hyponymes

Cette mesure peut être vue comme une variation de la similarité proposée par (Lin, 1998)(équation 3) avec une estimation du contenu informationnel d'un concept basé sur le nombre d'hyponymes comme proposé dans (Seco *et al.*, 2004).

À l'instar de mesures de similarités entre deux concepts, plusieurs mesures de similarités ont aussi été proposées dans le contexte de deux groupes de concepts. En particulier concernant la Gene Ontologie (GO) où des mesures de similarités entre documents permettent de déterminer les gènes impliqués dans un même processus biologique aidant par là à la prédiction d'interactions entre protéines (Pesquita *et al.*, 2009). Tandis que de telles mesures entre documents doivent être symétriques, celles mises en œuvre dans le processus d'appariement document/requête se doivent d'être asymétriques. En effet, s'il semble raisonnable de pénaliser un document parce qu'un concept d'une requête est absent de son index, l'ignorer parce qu'il est indexé par un concept absent de la même requête ne le serait pas. Etant donné une similarité entre deux concepts, une similarité entre un document et un concept de la requête pourrait être définie comme étant le maximum des similarités entre ce concept et ceux indexant le document. Cette stratégie nous permet d'avoir une mesure simple et intuitive de la proximité entre chaque concept d'une requête et un document.

Formellement, si π désigne une similarité entre deux concepts d'une ontologie O, Q_i (respectivement D_i) le i^{ème} concept de l'index de la requête Q (respectivement du document D) et $|Q|$ (respectivement$|D|$) la taille de l'index de la requête Q (respectivement du document D), alors la similarité entre un concept de la requête Q_t et D peut être donnée par $\pi(Q_t, D) = max_{0 \le i \le |D|}\, \pi(Q_t, D_i)$.

3.2 Mesure de proximité entre un document et une requête

Après avoir déterminé les similarités entre chaque concept d'une requête et un document, l'étape suivante consiste à les combiner en un score unique évaluant la pertinence globale du document par rapport à la requête. Il s'agit donc d'intégrer un modèle des préférences de l'utilisateur à la notion de proximité sémantique pour définir la pertinence globale d'un document vis-à-vis d'une requête.

Evaluer le score de pertinence d'un document par rapport à une requête, permet de pouvoir comparer des documents et de pouvoir justifier que l'un soit préféré plutôt qu'un autre. Il s'agit clairement d'une situation de représentation de préférences qui est une notion centrale dans le cadre de la théorie décisionnelle (Modave et Grabisch, 1998) et qui consiste à déterminer une fonction d'utilité U, de telle sorte que pour chaque alternative D, D' dans une liste $\mathcal{D}$ d'alternatives d'intérêts, $D \succeq D'$ (i.e. D est préférée à D') si $U(D) \geq U(D')$. Quand la liste des alternatives est de dimension

n, i.e $D = \prod_{i=1}^{n} Q_i$ alors un modèle répandu consiste à utiliser le modèle décomposable proposé par (Krantz *et al.*, 1971) : $U(q_1,..,q_n) = h(u_1(q_1),..,u_n(q_n))$

Avec $u_t(.)$ une fonction retournant un réel et h un opérateur d'agrégation satisfaisant les conditions suivantes :

- h est continu ;
- h(0, 0,…, 0) = 0 et h(1, 1,…, 1) = 1 ;
- $\forall(a_i, b_i) \in [0,1]^2$, si $a_i \geq b_i$ alors $h(a_1, \ldots, a_n) \geq h(b_1, \ldots, b_n)$.

Dans notre contexte, le produit cartésien *D* correspond aux concepts d'une requête tandis que $u_t(.)$ correspond à $\pi(Q_t, D)$, la proximité entre un concept d'une requête et un document définie plus haut. Nous obtenons ainsi un modèle d'agrégation permettant de combiner les scores entre concepts d'une requête et d'un document pour déterminer le score de pertinence de ce dernier selon l'expression des préférences de l'utilisateur.

Les opérateurs d'agrégation peuvent traiter de termes pondérés, mais dans notre cas la pondération des termes d'une requête peut prêter à confusion. En effet, demander à l'utilisateur d'affecter, *a priori*, à chaque terme de sa requête un poids rend la formulation de celle-ci plus difficile tandis que le faire automatiquement peut compliquer la traçabilité de la prise de décision dans le système. Dans cette étude, nous considérons tous les termes de la requête comme étant de même importance. Il existe 3 types d'agrégation :

- conjonctions (de type ET), $h(u_1(q_1),..,u_n(q_n)) \leq \min(u_1(q_1),..,u_n(q_n))$;
- disjonctions (de type OU), $h(u_1(q_1),..,u_n(q_n)) \geq \max(u_1(q_1),..,u_n(q_n))$;
- compromis $\min(u_1(q_1),..,u_n(q_n)) \leq h(u_1(q_1),..,u_n(q_n)) \leq \max(u_1(q_1),..,u_n(q_n))$.

Notre objectif est de fournir à l'utilisateur un moyen intuitif de choisir l'opérateur utilisé qui modélise l'expression de ses préférences dans l'évaluation d'un score global de pertinence entre une requête et un document. Nous nous focalisons uniquement sur les opérateurs de compromis parce qu'ils correspondent au comportement décisionnel commun qui consiste à affecter un score global, compris entre les valeurs min et max des scores élémentaires (convexité). Pour cela nous nous appuyons sur les familles d'opérateurs de Yager (Yager, 1979) qui fournissent une fonction paramétrée pour représenter les opérateurs de compromis :

$$Y_m(\pi(Q_1,D),..,\pi(Q_{|Q|},D)) = \left(\left(\sum_{t=1}^{|Q|} \pi(Q_t,D)^q\right) / |Q|\right)^{1/q}, \; q \in \mathbb{R} \qquad (6)$$

Afin d'avoir une idée du type d'agrégation qu'elle permet, on peut considérer quelques une de ces valeurs remarquables :

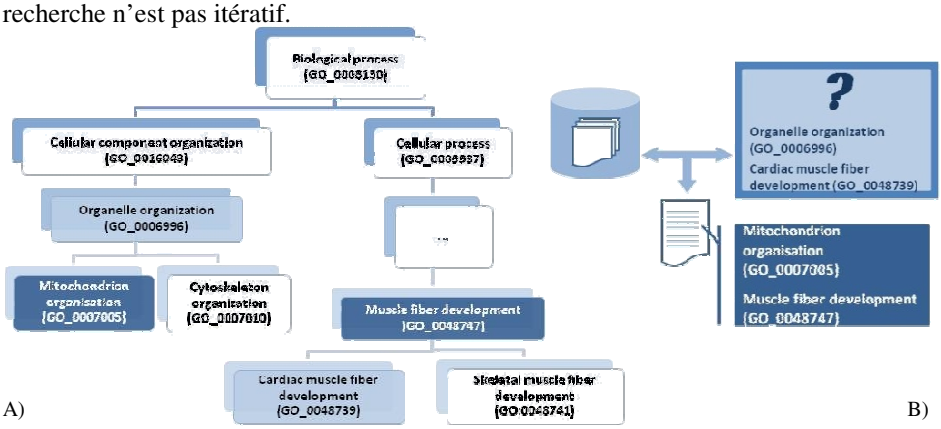
- $Y_{m_1}(.)$ moyenne arithmétique,
- $Y_{m_{-1}}(.)$ moyenne harmonique,
- $q \to 0$, moyenne géométrique,
- $q \to +\infty$, max (généralisation du OU)
- $q \to -\infty$, min (généralisation du ET)

Le recours à une famille paramétrique d'opérateurs de compromis permet de ramener le choix de l'opérateur au seul choix du paramètre *q*. Dans notre interface, l'utilisateur n'a alors qu'à déplacer un curseur pour faire tendre sa requête vers un "ET" ou un "OU" généralisé ou bien tolérer des effets compensatoires. Cet opérateur d'agrégation

couplé à l'utilisation de proximité sémantique permet de proposer un outil de requêtage efficace et interactif, qui est détaillé dans la section suivante.

4 Résultats

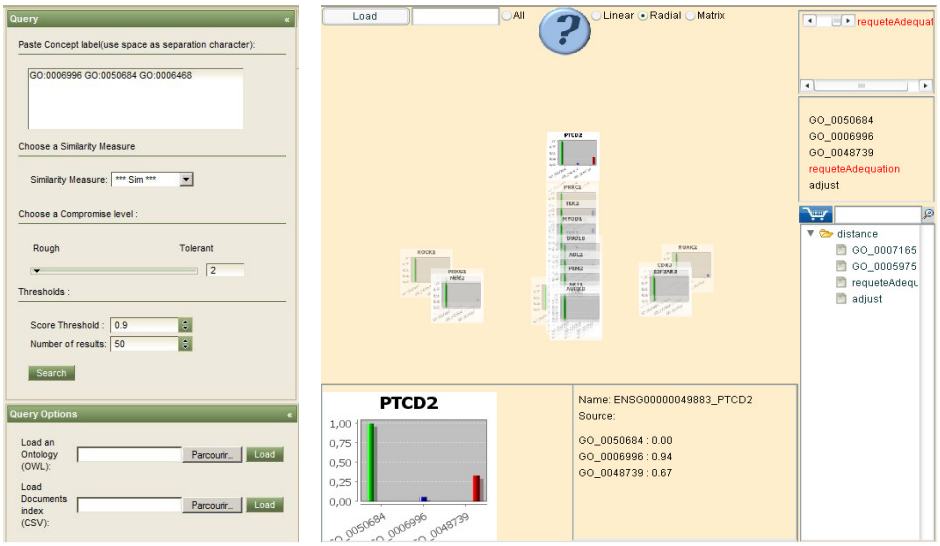
La notion de similarité sémantique est souvent utilisée pour la recherche de documents proches (Pesquita *et al.*, 2009). Nous proposons dans cet article une manière originale de l'utiliser pour le requêtage. Malgré les points communs de ces deux tâches, il existe au moins deux différences essentielles qui font que l'on ne peut pas simplement rechercher les documents similaires à ceux d'un document fictif que l'on indexerait par les termes de la requête. Dans le cas du requêtage, la proximité sémantique doit être d'une part asymétrique et d'autre part compréhensible par l'utilisateur pour qu'il puisse interagir efficacement avec le SRI. Dans le cas de recherche de documents similaires, la proximité sémantique doit être symétrique et l'explication concernant les documents renvoyés n'est pas cruciale car le processus de recherche n'est pas itératif.



A) B)

Fig. 1 – Utiliser une ontologie pour éviter les *silences* dans un système basé sur une ontologie de domaine A) et un corpus indexé, interrogé par un ensemble de concepts B).

Les systèmes de reformulation incluant l'ensemble des hyponymes des termes de la requête peuvent être vus comme une première étape vers un outil de requêtage sémantique. En effet, cette approche s'apparente à l'utilisation d'une similarité sémantique élémentaire où la similarité entre un concept de la requête et un concept indexant un document est 1 si le second est un hyponyme du premier et 0 sinon. Nous proposons une approche plus fine qui permet d'éviter les silences en identifiant des documents annotés par des hyponymes ou des hyperonymes proches des concepts présents dans la requête, comme illustré sur l'exemple de Fig 1. Lors d'une requête contenant les concepts "*Organelle organization (GO_0006996)*" et "*Cardiac muscle fiber development (GO_0048739)*" notre système sélectionne un document annoté par les concepts "*Mitochondrion organisation*" et "*Muscle fiber development*" voire (avec un score plus faible) un document annoté par "*Cellular component organization*" si cela est nécessaire.



A) B)

Fig. 2 – Interface d'interrogation A) et rendu visuel sur les cartes sémantiques B)

Fig. 2 –montre le formulaire de requêtage proposé à l'utilisateur. Il lui suffit de saisir ses concepts d'intérêts (assisté par une complétion utilisant l'ontologie) et de positionner le curseur de manière à indiquer si sa requête est plutôt de type conjonctif ou disjonctif. Le système identifie ensuite les documents pertinents et génère pour chacun d'entre eux un pictogramme explicitant les liens entre son indexation et chaque concept de la requête. Cette information est représentée sous la forme d'un histogramme, dont chaque barre représente un concept de la requête, sa hauteur correspondant à son score de pertinence et sa couleur variant, suivant que cette valeur est obtenue par un concept hyponyme, hyperonyme ou identique à celui de la requête. Ces pictogrammes sont ensuite positionnés sur une carte sémantique de sorte que leur distance physique au pictogramme symbolisant la requête (point d'interrogation) soit proportionnelle à leur degré de pertinence. L'utilisateur peut ainsi identifier visuellement les documents les plus pertinents et la raison de leur adéquation à sa requête. Pour valider notre approche, nous collaborons actuellement avec des biologistes moléculaires sur un corpus correspondant à l'ensemble des gènes humains indexés par des concepts de la Gene Ontology (qui contient environ 30 000 concepts). Fig. 2 –B) est un exemple du type de résultats que nous proposons. La visualisation utilise ici le principe de sonde sémantique qui fait l'objet d'une autre communication.

5 Conclusion et perspectives

L'approche décrite dans cet article est un pas important vers un outil de requêtage capable d'exploiter la richesse sémantique des ontologies tout en conservant une simplicité d'utilisation. C'est également un des premiers systèmes qui permet d'informer l'utilisateur des raisons pour lesquelles les documents ont été retenus en s'appuyant sur des pictogrammes intuitifs. Le fait de positionner l'ensemble de ces

pictogrammes sur une carte sémantique ouvre des possibilités d'interaction avec l'utilisateur notamment pour l'aider à identifier et sélectionner les documents qu'il préfère et à reformuler et préciser sa requête.

Il est clair que sur cet aspect, nous sous-exploitons encore les possibilités offertes par cette carte sémantique. Nous prévoyons de coupler de manière plus forte l'outil de requêtage et celui de visualisation. Une piste possible est l'utilisation de pondération sur les termes de la requête. Les opérateurs d'agrégations se généralisent assez simplement au cas pondérés. Par contre, il est difficile de demander, *a priori*, à l'utilisateur de fournir de telles valeurs. Nous pensons donc procéder en deux étapes d'abord faire une requête classique pour identifier un premier ensemble de documents et les positionner sur la carte sémantique. Puis permettre à l'utilisateur de changer les poids de chaque terme de la requête (via des curseurs). Ces changements de pondérations étant immédiatement répercutés sur le positionnement des documents, grâce à l'utilisation de *sondes sémantiques* (Crampes et DeOliveira-Kumar, 2010). Une fois que l'utilisateur a trouvé des pondérations lui permettant de faire émerger les documents qu'il préfère, il peut alors relancer (d'un simple clic) sa requête avec ces nouvelles pondérations.

Références

ANDREASEN, T. (2003). An approach to knowledge-based query evaluation. *Fuzzy Sets and Systems, 140*(1), 75-91.

BELKIN, N., INGWERSEN, P. & PEJTERSEN, A. M. (1992, June 21-24). *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, Copenhagen, Denmark.

BERRIZ, G. F., WHITE, J. V., KING, O. D. & ROTH, F. P. (2003). GoFish finds genes with combinations of Gene Ontology attributes. *Bioinformatics, 19*(6), 788-789.

BOUGHANEM, M., CHRISMENT, C. et SOULE-DUPUY, C. (1999). Query modification based on relevance back-propagation in an ad hoc environment. *Information Processing & Management, 35*(2), 121-139.

CRAMPES, M. & DEOLIVEIRA-KUMAR, J. (2010). Semantic Search on Heterogeneous Database Sources Using Visual Probes. *Special issue of Journal of Web Semantics, (to appear).*

CROUCH, C., J. & YANG, B. (1992). *Experiments in automatic statistical thesaurus construction.* Paper presented at the 15th annual international ACM SIGIR conference on Research and development in information retrieval, Copenhagen, Denmark.

FARAH, M. & VANDERPOOTEN, D. (2007, 28–30 mars). *L'Agrégation en Recherche d'Information. Une revue critique des principaux modèles théoriques de Recherche d'Information.* Paper presented at the CORIA 2007, Saint Etienne, France.

HIRST, G. & ST ONGE, D. (1998). Lexical Chains as representation of context for the detection and correction malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database and some of its applications (Language, Speech, and Communication).* Cambrige, MA, USA: The MIT Press.

JANSEN, B. J. (2000). The effect of query complexity on Web searching results. *Inf. Res., 6*(1).

JANSEN, B. J., SPINK, A. & SARACEVIC, T. (2000). Real life, real users, and real needs: a study and analysis of user queries on the web. *Information Processing & Management, 36*(2), 207-227.

KRANTZ, D. H., LUCE, R. D., SUPPES, P. & TVERSKY, A. (1971). *Foundations of measurement* (Vol. 1: Additive and polynomial representations): Academic Press, New York.

LIN, D. (1998). *An Information-Theoretic Definition of Similarity.* Paper presented at the Fifteenth International Conference on Machine Learning.

LUCAS, W. & TOPI, H. (2004). Training for Web search: Will it get you in shape? *Journal of the American Society for Information Science and Technology, 55*(13), 1183-1198.

MODAVE, F. & GRABISCH, M. (1998, July). *Preference representation by a Choquet integral: Commensurability hypothesis.* Paper presented at the 7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'98), Paris, France.

PESQUITA, C., FARIA, D., FALCAO, A. O., LORD, P. & COUTO, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Comput Biol, 5*(7), e1000443.

PRIÉ, Y. (1999). *Modélisation de documents audiovisuels en Strates Interconnectées par les Annotations pour l'exploitation contextuelle.* Ph.D., Institut National des Sciences Appliquées de Lyon, Lyon.

RADA, R., MILI, H., BICKNELL, E. & BLETTNER, M. (1989). *Development and application of a metric on semantic nets* (Vol. 19). New York, NY, ETATS-UNIS: Institute of Electrical and Electronics Engineers.

RANWEZ, S., RANWEZ, V., VILLERD, J. & CRAMPES, M. (2006). Ontological distance measures for information visualisation on conceptual maps. *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops, Pt 2, Proceedings, 4278*, 1050-1061.

RESNIK, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research, 11*, 95-130.

SALTON, G. & MCGILL, M. J. (1986). *Introduction to Modern Information Retrieval*: McGraw-Hill, Inc.

SCHAMBER, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology, 29*, 3-48.

SECO, N., VEALE, T. & HAYES, J. (2004). An intrinsic information content metric for semantic similarity in WordNet. *Ecai 2004: 16Th European Conference on Artificial Intelligence, Proceedings, 110*, 1089-1090.

VAN RIJSBERGEN, C. J. (1979). *Information Retrieval*: Butterworth-Heinemann.

WU, Z. & PALMER, M. (1994). *Verbs semantics and lexical selection.* Paper presented at the 32nd meeting on Association for Computational Linguistics, Las Cruces, New Mexico.

YAGER, R. R. (1979). Possibilistic decision making. *IEEE Trans. on Systems, Man and Cybernetics*(9), 388-392.

ZARGAYOUNA, H. & SALOTTI, S. (2004, 5-7 mai). *Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML.* Paper presented at the 15èmes journées francophones d'ingénierie des connaissances IC2004, Lyon, France.