Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents

Mohameth-François Sy

▶ **To cite this version:**

Mohameth-François Sy. Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents. Recherche d'information [cs.IR]. Université Montpellier II - Sciences et Techniques du Languedoc, 2012. Français. NNT : . tel-00822516

HAL Id: tel-00822516

https://theses.hal.science/tel-00822516

Submitted on 14 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NUMERO D'IDENTIFICATION

ACADEMIE DE MONTPELLIER

UNIVERSITE DE MONTPELLIER II

SCIENCES ET TECHNIQUES DU LANGUEDOC

Utilisation d'ontologies comme support à la recherche et à la navigation dans une collection de documents

THESE

Présentée et soutenue le 11/12/2012/ pour l'obtention du grade de

Docteur de l'Université Montpellier II

Discipline : Informatique

Formation Doctorale : Informatique

Ecole Doctorale : Information, Structures, Systèmes

Par

Mohameth François Sy

Composition du jury :

Nathalie Aussenac-Gilles	Directeur de Recherche	CNRS, IRIT	Rapporteur
Moussa LO	Maitre de Conférence, HDR	Université Gaston Berger	Rapporteur
Michel Beigbeder	Maitre-assistant	ENS des Mines de Saint-Etienne	Examinateur
Patrice Bellot	Professeur	LSIS, Marseille	Examinateur
Michel Crampes	HDR	ENS des Mines d'Alès	Directeur de thèse
Sylvie Ranwez	Maitre-assistant	ENS des Mines d'Alès	Examinateur
Vincent Ranwez	Professeur	Montpellier SupAgro	CoDirecteur de thèse
Jacky Montmain	Professeur	ENS des Mines d'Alès	Examinateur

A Mody Sy

Remerciements

Je tiens avant tout à exprimer ma profonde gratitude à mes encadrants et tout particulièrement à Sylvie Ranwez et à Vincent Ranwez. Durant ces trois années de thèse, ils m'ont accompagné par leur conseil ainsi que par leur soutien constant. Je les remercie aussi pour la confiance qu'ils m'ont témoignée tout au long de cette thèse. Ce fut réellement un grand plaisir de travailler avec eux. Je remercie aussi Michel Crampes, mon directeur de thèse, pour ses conseils avisés et pour son soutien lors des moments clés.

Je remercie Jacky Montmain pour sa disponibilité, pour son approche pédagogique et pour son apport scientifique à cette thèse. J'ai particulièrement apprécié de travailler avec lui.

Je remercie Armelle Regnault et Patrick Augereau. Les travaux que nous avons menés ensemble ont permis d'enrichir ma thèse.

Je remercie Nathalie Aussenac-Gilles ainsi que Moussa Lo pour avoir accepté d'être rapporteur de ma thèse. Les jugements et les remarques qu'ils ont fournis à l'égard de mon travail ont été pertinents et très enrichissants.

Je remercie Patrice Bellot pour avoir examiné mes travaux et présidé le jury de ma thèse. Je joins à ces remerciements Michel Beigbeder.

Un grand merci à François Trousset, toujours prompt à mener une discussion scientifique enrichissante.

Je remercie tous les doctorants du *LGI2P* pour avoir partagé avec moi d'excellents moments durant ces trois années. Mes remerciements vont tout particulièrement à Benjamin Duthil et à Mambaye Lo mes deux collègues de bureau toujours enclin à rendre service. Merci également à Abdelhak Imoussaten avec qui la bonne humeur et le rire sont toujours de rigueur. Je n'oublie pas Sébastien Harispe avec qui j'ai eu des discussions mutuellement enrichissantes.

Un grand merci à Valérie Roman, à Claude Badiou et à Françoise Andre dont l'aide fut précieuse concernant toutes mes démarches administratives.

Merci à Françoise Armand pour son aide précieuse dans ma recherche bibliographique et lors des tâches d'impression du manuscrit de ma thèse.

Mention spéciale à Ansata Dada Baldé, mon épouse, pour son soutien de tous les instants.

Résumé

Les ontologies offrent une modélisation des connaissances d'un domaine basée sur une hiérarchie des concepts clefs de ce domaine. Leur utilisation dans le cadre des Systèmes de Recherche d'Information (SRI), tant pour indexer les documents que pour exprimer une requête, permet notamment d'éviter les ambiguïtés du langage naturel qui pénalisent les SRI classiques.

Les travaux de cette thèse portent essentiellement sur l'**utilisation d'ontologies lors du processus d'appariement durant lequel les SRI ordonnent les documents** d'une collection en fonction de leur pertinence par rapport à une requête utilisateur. Nous proposons de calculer cette pertinence à l'aide d'une stratégie d'agrégation de scores élémentaires entre chaque document et chaque concept de la requête. Cette agrégation, simple et intuitive, intègre un modèle de préférences dépendant de l'utilisateur et une mesure de similarité sémantique associée à l'ontologie. L'intérêt majeur de cette approche est qu'elle permet d'expliquer à l'utilisateur pourquoi notre SRI, OBIRS, estime que les documents qu'il a sélectionnés sont pertinents. Nous proposons de renforcer cette justification grâce à une visualisation originale où les résultats sont représentés par des pictogrammes, résumant leurs pertinences élémentaires, puis disposés sur une carte sémantique en fonction de leur pertinence globale.

La Recherche d'Information étant un processus itératif, il est nécessaire de permettre à l'utilisateur d'interagir avec le SRI, de comprendre et d'évaluer les résultats et de le guider dans sa reformulation de requête. Nous proposons une **stratégie de reformulation de requêtes conceptuelles** basée sur la transposition d'une méthode éprouvée dans le cadre de SRI vectoriels. La reformulation devient alors un problème d'optimisation utilisant les retours faits par l'utilisateur sur les premiers résultats proposés comme base d'apprentissage. Nous avons développé une heuristique permettant de s'approcher d'une requête optimale en ne testant qu'un sous-espace des requêtes conceptuelles possibles. Nous montrons que l'identification efficace des concepts de ce sous-espace découle de deux propriétés qu'une grande partie des mesures de similarité sémantique vérifient, et qui suffisent à garantir la connexité du voisinage sémantique d'un concept.

Les modèles que nous proposons sont validés tant sur la base de performances obtenues sur des jeux de tests standards, que sur la base de cas d'études impliquant des experts biologistes.

Mots-clés : Ontologies, Recherche d'Information, Reformulation, Carte sémantique, Visualisation

Thèse effectuée au Centre de Recherche LGI2P de l'Ecole des Mines d'Alès, Site de Nîmes, Parc Scientifique Georges Besse, 30 035 Nîmes cedex 1

Abstract

Ontology based information retrieval

Domain ontologies provide a knowledge model where the main concepts of a domain are organized through hierarchical relationships. In *conceptual* Information Retrieval Systems (IRS), where they are used to index documents as well as to formulate a query, their use allows to overcome some ambiguities of classical IRSs based on natural language processes.

One of the contributions of this study consists in the use of ontologies within IRSs, in particular to **assess the relevance of documents with respect to a given query**. For this matching process, a simple and intuitive aggregation approach is proposed, that incorporates user dependent preferences model on one hand, and semantic similarity measures attached to a domain ontology on the other hand. This matching strategy allows justifying the relevance of the results to the user. To complete this explanation, semantic maps are built, to help the user to grasp the results at a glance. Documents are displayed as icons that detail their elementary scores. They are organized so that their graphical distance on the map reflects their relevance to a query represented as a probe.

As Information Retrieval is an iterative process, it is necessary to involve the users in the control loop of the results relevancy in order to better specify their information needs. Inspired by experienced strategies in vector models, we propose, in the context of conceptual IRS, to **formalize ontology based relevance feedback**. This strategy consists in searching a conceptual query that optimizes a tradeoff between relevant documents closeness and irrelevant documents remoteness, modeled through an objective function. From a set of concepts of interest, a heuristic is proposed that efficiently builds a near optimal query. This heuristic relies on two simple properties of semantic similarities that are proved to ensure semantic neighborhood connectivity. Hence, only an excerpt of the ontology *dag* structure is explored during query reformulation.

These approaches have been implemented in OBIRS, our ontological based IRS and validated in two ways: automatic assessment based on standard collections of tests, and case studies involving experts from biomedical domain.

Keywords: Ontologies, Information Retrieval, Relevance feedback, Semantic maps, Visualization

Table des matières

Table des figures

Liste des tableaux

Chapitre 1 : Introduction générale

1.1. Contexte de la thèse

La croissance exponentielle des données électroniques disponibles les rend quasiment inutilisables sans outils efficaces pour trouver la « bonne » information, i.e. l'information pertinente et exploitable dans un contexte donné. Ainsi, à l'heure actuelle, ce n'est souvent plus le manque d'information qui pose problème mais, la difficulté à retrouver, organiser et visualiser l'information pertinente dans un contexte spécifique. Or, surmonter ces difficultés est crucial dans les processus de prise de décision (gouvernance d'instituts par exemple), ou dans le domaine de l'intelligence économique (système de veille).

Depuis plusieurs années, l'équipe KID (Knowledge and Image Analysis for Decision) du centre de recherche LGI2P (Laboratoire de Génie Informatique et d'ingénierie de Production) de l'Ecole Nationale Supérieure des Mines d'Alès (équipe à laquelle j'ai été rattaché pendant mes travaux de thèse) accompagne des collectifs de chercheurs dans l'animation de leurs communautés, en particulier autour de la toxicologie nucléaire (projet ToxNuc[1], avec comme partenaire principal le CEA, Commissariat à l'Energie Atomique), du soutien à l'innovation industrielle (dans le cadre du Carnot M.I.N.E.S.[2]) ou encore de la recherche d'informations biomédicales (plateformes AvieSan[3] des différents instituts thématiques multi-organismes de l'Inserm – ITMO). Une plateforme collaborative dédiée et intuitive a été proposée aux partenaires de chacun de ces projets. Ces plateformes ont pour but la gestion de collectifs de scientifiques, ainsi que la gestion des ressources (publications scientifiques, rapports de recherche, documents administratifs…) qu'ils souhaitent partager. Du fait du grand nombre de ces ressources, des problèmes liés à leur gestion et à l'optimisation de leur exploitation se sont posés, notamment autour de quatre tâches principales : la navigation, l'indexation, la recherche et la visualisation d'information. On observe également une forte demande d'outils performants en recherche d'information dans le cadre de nos collaborations avec l'ISEM

[1] www.toxcea.fr
[2] Méthodes INnovantes pour l'Entreprise et la Société : http ://www.communaute.carnot-mines.eu/
[3] E.g. pour l'ITMO Immunologie, Hématologie et Pneumologie (IHP) : https ://ihp.aviesan.fr

(Institut des Sciences de l'Evolution de Montpellier) et l'IRCM (Institut de Recherche en Cancérologie de Montpellier).

Nous proposons, dans cette thèse et en réponse aux problématiques de nos partenaires, un modèle de **recherche d'information utilisant des ontologies de domaine** aussi bien dans les fonctions de **recherche** de documents que de **visualisation** de ceux-ci.

Il s'agit, non seulement, de restituer des résultats pertinents en réponse à des requêtes mais aussi de pouvoir expliciter la pertinence de ces résultats afin de guider l'utilisateur lors de futures recherches d'information. Du fait que plusieurs de nos partenaires sont focalisés sur le domaine biomédical, nous nous sommes naturellement tournés vers l'usage d'ontologies et de thésaurus de ce même domaine (*MeSH*[4], *Gene Ontologie*[5]…) pour valider nos travaux. Ce choix a également été conforté par le fait que les ontologies citées sont relativement complètes et, étant utilisées et mises à jour par une large communauté scientifique, elles offrent une bonne couverture de la connaissance du domaine. Cependant, notre approche est générique et reste valable dès lors qu'une ontologie de domaine est disponible. Nos propositions ont été validées et implémentées au sein de l'application Web *OBIRS*[6] (*Ontology based Information Retrieval System*) et de ses déclinaisons (*OBIRS-feedback* et *CoLexIR*).

Dans la section suivante, nous allons préciser le positionnement de nos contributions au regard de la Recherche d'Information dont les processus de base sont définis.

1.2. L'utilisateur au cœur du scénario de la Recherche d'Information : la boucle de pertinence

La **Recherche d'Information** (RI) a fait l'objet de nombreuses définitions tout au long de son histoire. Considérons deux définitions : une parmi les plus anciennes et une autre parmi les plus récentes. Gerard Salton a défini le système SMART de la manière suivante (Salton 1968) :

"The SMART retrieval system takes both documents and search requests in unrestricted English, performs a complete content analysis automatically, and retrieves those documents which most nearly match the given request."

Pour les auteurs de (Manning et al. 2008), la tâche principale d'un Système de Recherche d'Information (SRI) consiste à sélectionner, à partir de larges collections d'objets (généralement des documents textuels sous forme électronique), ceux qui sont susceptibles de répondre aux besoins en information d'un utilisateur :

"Information Retrieval is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)."

[4] MeSH (Medical Subject Headings) : http ://www.ncbi.nlm.nih.gov/mesh/
[5] Gene Ontology : http ://www.geneontology.org/
[6] OBIRS : http ://www.ontotoolkit.mines-ales.fr/ObirsClient/

Comme nous pouvons le constater, ces définitions, quoiqu'informelles et séparées de près d'un demi-siècle, sont quasiment identiques. Au regard de celles-ci, la Recherche d'Information traite, donc, de l'organisation, du stockage (généralement dans un ordinateur sous forme électronique), de la recherche et de la sélection d'informations pertinentes par rapport à un besoin en information d'un utilisateur. Au cours d'une session de recherche, correspondant à l'ensemble des interactions entre un SRI et un utilisateur concourant à la satisfaction d'un besoin en information de celui-ci, le scénario de base suivant, adapté de (Dominich 2008) est à l'œuvre (c.f. Figure 1.1) :

- dans un premier temps, un utilisateur Ut (chercheur, touriste, internaute, etc.) a un besoin en information (par exemple, articles scientifiques traitant d'un certain sujet, une page Web fournissant des informations sur une ville,…). Le besoin en information est abstrait et donc subsiste en l'état de pensée et il est important d'en distinguer deux aspects. En effet, le premier concerne l'*objet* du besoin, i.e. un article dans un journal scientifique pour un chercheur par exemple. Le second aspect concerne les préférences souvent implicites de l'utilisateur sur l'objet de son besoin en information. Ce cas survient, par exemple, en considérant que deux chercheurs peuvent rechercher des journaux scientifiques mais préférer des domaines différents du fait de leurs spécialisations distinctes, i.e. un mathématicien et un informaticien par exemple. Aussi, un même journal peut avoir une importance relative qui diffère selon le domaine considéré. Ces exemples simples montrent que deux composantes sont à l'œuvre lorsqu'un utilisateur a un besoin en information : *l'objet* (pouvant être communs à plusieurs utilisateurs) et des *préférences* sur celui-ci, spécifiques à chaque utilisateur ;
- le besoin en information est pris en compte dans un SRI et est traduit sous les formes exploitables (en utilisant un langage adéquat) d'une requête concernant sa composante *objet* d'une part, et d'un modèle utilisateur pour sa composante *préférences* d'autre part. Dans la suite et par souci de mise en œuvre, nous considérons qu'une requête doit regrouper à la fois l'objet de la recherche ainsi que le modèle de préférences de l'utilisateur. Il existe donc une différence à faire entre le besoin en information d'un utilisateur, informel et intangible (sous forme de pensée) et sa prise en compte formelle par un SRI sous la forme d'une requête qui peut être traitée par un ordinateur. Il est très difficile d'établir une telle correspondance du fait du double enjeu, pour l'utilisateur, de formaliser son besoin en information et pour le SRI de le "comprendre" ;
- le SRI sélectionne un ensemble d'éléments (résultats) dans une collection d'objets (articles de journaux, pages Web,…) en réponse à la requête soumise et suivant un modèle de pertinence défini ;
- si l'utilisateur Ut juge satisfaisant les résultats qui lui sont retournés, la session de recherche s'arrête. Dans le cas contraire, il peut chercher à reformuler itérativement la requête correspondant à son besoin en information au travers de jugements sur les résultats (c.f. Figure 1.1) ou de changement de paramètres de la requête, participant ainsi à la construction de la réponse pertinente du SRI. Il s'agit d'une étape d'évaluation mettant en jeu l'utilisateur (son modèle de préférence), donc un facteur

externe au SRI. Dans un tel cas, il apparaît que l'interaction entre l'utilisateur et le SRI joue un rôle central dans la capacité des SRI à fournir des résultats pertinents vu que l'utilisateur est au final le seul juge de cette pertinence. Même s'il est très difficile d'appréhender et de contrôler les facteurs liés à l'utilisateur, il est important de pouvoir les gérer et les intégrer dans les différents processus d'un SRI.

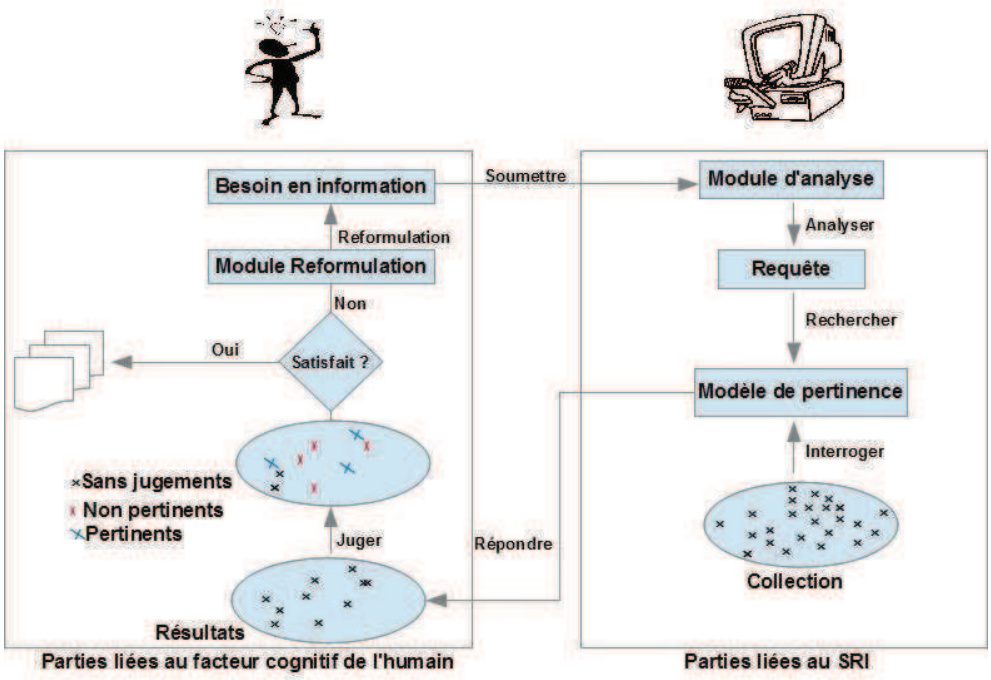


Figure 1.1 : Scénario de base d'une session de recherche d'information faisant intervenir des facteurs cognitifs humains et une partie logicielle correspondant au SRI. Parmi les résultats retournés par le SRI, l'utilisateur sélectionne ceux qui sont pertinents (croix bleues) et ceux qui ne le sont pas (croix rouges).

Le scénario illustré dans la Figure 1.1 montre qu'il y a une interaction nécessaire entre l'utilisateur et son activité cognitive, d'une part, et le système logiciel correspondant au SRI d'autre part pour aboutir à la satisfaction d'un besoin en information. Cette nécessité se traduit, dans un SRI, par une interface intelligente pour la présentation de l'information à l'utilisateur et pour la compréhension de celle-ci. Il s'agit clairement, pour les interfaces des SRI, de jouer un rôle de médiateur. Tout part de l'utilisateur pour revenir à lui puisqu'il est le juge en dernier ressort décidant de la fin d'une session de recherche une fois son besoin en information satisfait. Nous appelons *boucle de pertinence* chaque itération dans une session de recherche. Ce scénario souligne aussi l'importance de considérer les deux composantes (objets de la recherche et préférences sur ceux-ci) d'un besoin en information d'un utilisateur comme constituant de sa requête. La mise en œuvre de ces deux aspects conduit donc à une démarche d'automatisation cognitive dont l'objectif est la recherche et le développement d'outils qui aident l'opérateur humain à rapidement percevoir et comprendre une situation, l'analyser et décider des actions à mener. Les modules et outils que nous proposons doivent faciliter l'interaction homme-machine dans le scénario précédent afin d'optimiser la pertinence et la fiabilité des résultats restitués par le SRI, en tirant profit des capacités cognitives de l'utilisateur et calculatoires de la machine pour une recherche d'information efficiente et

contrôlée. Nous allons introduire dans le paragraphe suivant la traduction, dans les différentes tâches d'un SRI, du scénario de base illustré dans la Figure 1.1.
Nous pouvons dire qu'un SRI vise à diminuer le *'silence'* (absence de documents pertinents dans les résultats du SRI) et le *'bruit'* (proportion de documents non pertinents parmi ceux fournis) et cela relativement au besoin en information d'un utilisateur tel que défini dans le scénario précédent. Pour arriver à un tel résultat, trois processus sont généralement mis en œuvre (Belkin et al. 1992; Manning et al. 2008) :

i) **un processus d'indexation** qui vise à fournir un modèle de représentation et de description, le plus compact et expressif possible, d'une granule d'information souvent désigné par *document*, terme générique pouvant désigner tout type de documents ainsi que leurs sous parties (passages dans un texte par exemple) ;

ii) **un processus d'appariement** (modèle de pertinence) permettant de sélectionner des documents pertinents par rapport à une requête. Une telle pertinence est évaluée sur la base d'une fonction d'appariement appelé *RSV* (*Retrieval Status Value*) induisant un ordre (éventuellement partiel) de pertinence dans l'espace des documents relativement à une requête. A chaque document de la collection est associé un score généralement normalisé et compris entre 0 et 1 ;

iii) **un processus de reformulation des requêtes** permettant de prendre en compte des retours sur les résultats fournis ou des précisions concernant les paramètres d'une requête pour améliorer celle-ci. La Figure 1.2 montre l'architecture globale d'un SRI intégrant les 3 processus précédents.

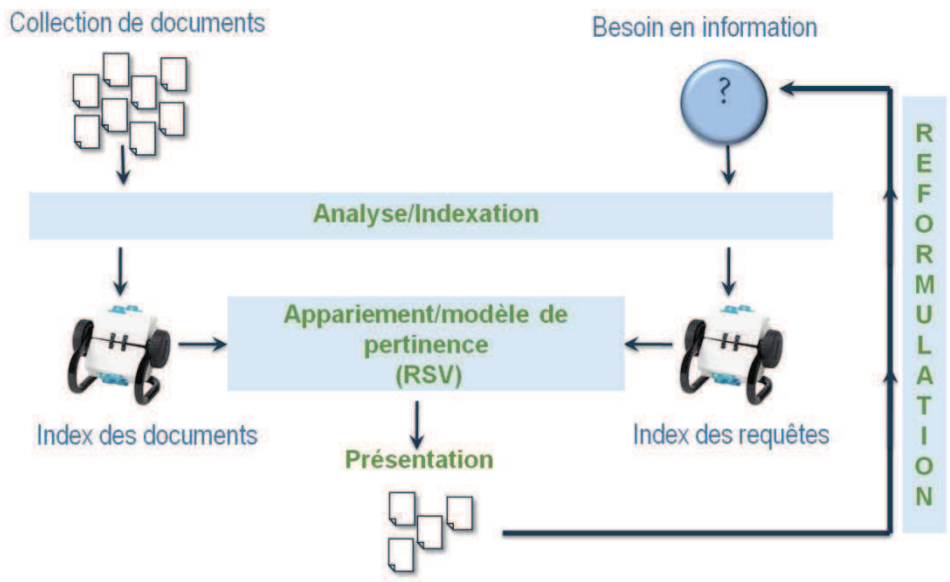


Figure 1.2 : Architecture d'un système de Recherche d'Information

Pour mettre en œuvre ces trois processus et suivant le cadre théorique utilisé pour l'indexation et l'appariement, plusieurs modèles pour les SRI sont à distinguer (Rijsbergen 1979). Il s'agit principalement :

- des modèles basés sur la **théorie des ensembles** dont le représentant le plus connu est le modèle booléen basique ou étendu (Salton et al. 1983). Dans sa version basique, il s'agit d'un des premiers modèles de RI où les unités d'indexation sont vues comme des

prédicats de la logique du premier ordre et où les documents et les requêtes sont représentés par une combinaison logique des unités d'indexation ;

- des **modèles algébriques** dont le premier fut le **modèle vectoriel** (Salton et al. 1975; Salton & Buckley 1988). Dans ces modèles, les documents et requêtes sont représentés par des vecteurs de poids exprimés dans la base canonique de l'espace des unités d'indexation. Le poids d'un terme quantifie son pouvoir discriminant et est lié à la fréquence du terme. La pertinence d'un document relativement à une requête est généralement déterminée par leur similarité évaluée en fonction de l'angle que forment leurs deux vecteurs dans l'espace vectoriel d'indexation ;
- des **modèles probabilistes** pour lesquels la pertinence d'un document vis-à-vis d'une requête est vue comme une probabilité de pertinence document-requête. Il s'agit des modèles ayant suscité le plus de travaux de recherche ces dix dernières années. Ils se fondent sur l'hypothèse que les unités d'indexation et leurs fréquences dans un document ou dans une collection de documents peuvent être représentées par des distributions de probabilités (Ponte & Croft 1998; Amati & Rijsbergen 2002; Clinchant & Gaussier 2009).

Ces trois modèles constituent des cadres généraux dans lesquels plusieurs déclinaisons sont possibles. De manière implicite, les modèles de RI tentent de modéliser une certaine conception de la pertinence d'un document pour un utilisateur ayant un besoin d'information. Il s'agit, pour la RI, d'essayer de capter, et éventuellement d'automatiser, le processus de décision à l'œuvre lorsqu'un utilisateur est confronté à un besoin en information et se voit offrir plusieurs choix. De nombreux travaux ont souligné la nature multidimensionnelle de la pertinence (Mizzaro 1997, 1998; Borlund 2003; Pereira et al. 2012) appelant à prendre en compte plusieurs critères dans l'évaluation de la pertinence d'un document. Dès lors, l'objectif principal de la recherche d'information peut être reformulé comme suit :

La Recherche d'Information se propose de choisir à partir d'un ensemble de documents ceux répondant au mieux aux critères d'un utilisateur supposés pris en compte dans une requête représentant son besoin en information dans ses deux composantes : objets de la recherche et modèle de préférences.

En considérant les *documents* comme des *alternatives*, le parallèle avec la théorie de l'utilité multi attribut (*MAUT : Multi Attribute Utility Theory*) devient possible (Wong et al. 1991; Farah & Vanderpooten 2006) comme en attestent les modèles d'agrégation (Farah et al. 2006) en RI où *MAUT* est envisagée pour modéliser les préférences d'utilisateurs sous une forme analytique.

Outre cette catégorisation fondée sur le modèle de pertinence, les SRI peuvent aussi être regroupés en deux groupes suivant le mode d'indexation (Haav & Lubi 2001). Il s'agit des SRI dits *classiques* qui représentent documents et requêtes par des *termes* et des SRI, dits *conceptuels*, qui se rapportent, pour la construction de l'indexation d'un document, aux *sens* des *termes* ainsi qu'aux relations qui les lient. L'émergence de cette dernière catégorie, favorisée par le développement du Web sémantique, permet de lever certaines limites de l'indexation par mots. Il s'agit, dans un premier temps, de limitations liées à la polysémie (un mot revêt plusieurs sens) et à la synonymie (deux mots se rapportant à un même sens). Dans

un deuxième temps, l'exploitation de mesures de similarité ou de proximité sémantiques, basée sur des relations entre concepts (synonymie, spécialisation, généralisation,...), permet de diminuer le silence des SRI.
Le Web sémantique est un projet de recherche élaboré il y a une dizaine d'années visant à permettre l'exploitation et le partage des ressources du web par les applications informatiques et les opérateurs humains (Berners-Lee et al. 2001). Le but du Web sémantique n'est pas de se substituer au Web classique mais de l'enrichir d'une couche sémantique. Pour arriver à un tel résultat, des propositions de technologies, de formats et de structures de données (notamment les ontologies) ont été faites. Une ontologie peut être vue comme une conceptualisation formelle (revêtant un caractère conventionnel) d'une connaissance d'un domaine.

Les travaux de cette thèse se focalisent sur la **recherche d'information conceptuelle** concernant l'unité d'indexation, sur les **modèles d'agrégation** concernant le modèle de pertinence et sur la justification de la pertinence d'un document au travers d'une **visualisation originale**. Notre contribution se situe dans les processus d'appariement et de reformulation d'un SRI de même qu'au niveau de la présentation des résultats. Elle s'inscrit donc dans le cadre du scénario de base de la recherche d'information (c.f. Figure 1.1) en prenant en compte un modèle conceptuel de connaissances (modèle de l'objet) et un modèle de préférences dans l'évaluation de la pertinence tout en fournissant une visualisation des résultats d'une requête axée sur une meilleure interaction de l'utilisateur avec le SRI au travers de fonctionnalités de justification et d'explication des résultats à plusieurs niveaux.

Une fois fixé le cadre de nos contributions, il convient d'éliciter dans le paragraphe suivant les hypothèses fortes qui les sous-tendent.

Dans un premier temps, notre modèle de pertinence suppose l'existence d'une indexation conceptuelle de chaque document de la collection dans laquelle la recherche s'effectue. Il est clair que l'efficacité de notre système, et de nombre de SRI conceptuels, dépend de la qualité de l'étape d'indexation conceptuelle. Cependant, la tâche d'indexation, incluant l'extraction de concepts à partir des documents et la désambigüisation éventuelle de ceux-ci, n'est pas traitée dans cette thèse. Il existe, notamment dans le domaine biomédical, de nombreuses collections de documents indexés par des concepts sur la base d'ontologies ou de thésaurus tels que la *Gene Ontology* et le *MeSH*. Parmi celles-ci, nous citerons la collection *MuCHMORE*[7] et les annotations de gènes par la *Gene Ontology*[8]. De telles collections nous ont non seulement permis de tester nos approches avec des indexations conceptuelles mais aussi de déployer nos outils chez nos partenaires intervenant dans le domaine biomédical. Dans un deuxième temps, il est difficile, pour un utilisateur, de formuler directement une requête conceptuelle. Cela est dû, essentiellement, à son manque de connaissances des éléments présents dans l'ontologie. Dès lors, il devient nécessaire de l'assister dans cette tâche en lui permettant de saisir sa requête au moyen de termes que le SRI va relier (de manière automatique ou semi automatique) à des concepts de l'ontologie en procédant en un alignement entre termes saisis

[7] Multilingual Concept Hierarchies for Medical Information Organization and Retrieval : http ://muchmore.dfki.de/

[8]Par exemple la base de données ENSEMBL : http ://www.ensembl.org/index.html

et concepts d'une ontologie. Pour cela, plusieurs techniques existent allant d'une simple exploitation des labels des concepts, jusqu'à l'exploitation de leurs composantes lexicales si celles-ci sont disponibles.

Dans cette partie, nous avons situé nos contributions dans le cadre du scénario de base de la recherche d'information et introduit l'implémentation d'un tel scénario que nous proposons. Dans la section suivante, nous allons décliner les sous objectifs que nous nous sommes fixés dans nos contributions.

1.3. Objectifs de la thèse

La réflexion menée pendant ces trois années de thèse a permis de soulever de nombreuses questions. A la lumière du cadre de nos contributions (voir l'encadré dans la section 1.2), les principales questions qui nous ont guidés dans notre progression sont présentées ici.

Une des premières questions auxquelles on est confronté, lorsque l'on se place dans le cadre du scénario de base de la recherche d'information que nous avons mis en exergue précédemment, est celle du choix du modèle de description pour décrire au mieux les objets sur lesquels portent les requêtes des utilisateurs. Un tel modèle repose essentiellement sur une unité d'indexation que l'on veut la plus expressive et concise possible ce qui revient à choisir la bonne granularité pour la représenter. Etant donné que les ontologies se proposent d'organiser les connaissances d'une communauté dans un domaine autour de sens réputés sans ambiguïtés et d'expliciter les relations entre sens, nous avons entrepris de répondre à la question suivante :

Les ontologies de domaine conduisent-elles à une amélioration de la pertinence des SRI lorsque leurs éléments de base, i.e. les concepts (sens), sont choisis comme unité d'indexation ? Parmi la multitude d'informations contenues dans les ontologies de domaines, lesquelles sont pertinentes pour la Recherche d'Information ?

Il s'agit clairement de montrer, à travers ces deux questions, qu'un modèle de pertinence basée sur une indexation conceptuelle (laquelle est faite de concepts comme unité d'indexation) permet de meilleures performances en termes de pertinence relativement à une indexation par simple termes. Nous reviendrons dans le chapitre suivant sur la différence entre termes et concepts.

Etant donnés les critères d'un utilisateur exprimés dans sa requête et un document d'une collection, un modèle d'agrégation combine plusieurs degrés de pertinence élémentaires (un pour chaque critère relativement au document) en vue d'obtenir un score mesurant la pertinence globale du document. Dans un tel cas, nous avons soulevé la question de l'évaluation des scores élémentaires à l'aide de fonctions d'utilité dès lors que les requêtes et les documents sont indexés par des concepts issus d'une ontologie de domaine :

Comment les ontologies, en tant qu'espaces conceptuels dans lesquels des métriques peuvent être déployées, peuvent-elles aider à mieux évaluer les degrés de pertinence élémentaires (ou utilités élémentaires) d'un document relativement à la requête d'un utilisateur ?

Une fois les fonctions d'utilité définies, il faut les combiner en essayant de capter, dans l'opérateur d'agrégation retenu, les préférences de l'utilisateur ainsi que sa stratégie de décision. Au-delà de la nature mathématique des opérateurs d'agrégation (conjonction, disjonction ou de compromis), nous nous intéressons, dans cette thèse, à leur pouvoir explicatif pour permettre à l'utilisateur de comprendre pourquoi des documents lui sont proposés par le SRI. Nous essayons donc de répondre aux questions suivantes :

Quel opérateur d'agrégation pour combiner les critères de pertinence choisis et permettre d'éviter un effet boîte noire du SRI ?

Comment procéder à une justification des résultats obtenus pour mettre en lumière les éléments ayant conduit à de bonnes performances ou au contraire à de mauvaises ?

Les réponses aux questions précédentes constituent un premier pas vers une présentation des résultats d'une recherche par l'utilisateur centrée sur leur compréhension. Nous proposons d'aller plus loin en considérant que la manière de présenter les résultats influe fortement sur la perception qu'aura l'utilisateur de la pertinence relative des documents qui lui sont fournis. Il est donc important de prendre en considération les capacités de traitement des utilisateurs dans les stratégies de présentation des résultats (interface homme/machine). Ce constat nous a conduits à poser la question suivante :

Quelles stratégies de visualisation des résultats de RI permettent d'expliciter au mieux, à l'utilisateur, l'adéquation des documents retournés ?

Comment les ontologies peuvent elles aider dans la mise en œuvre de telles techniques ?

Une meilleure interface de visualisation permet une meilleure perception et compréhension des résultats d'une recherche par un utilisateur. Cela lui permet, dans le cadre d'une session de recherche, d'injecter des informations supplémentaires dans le SRI en précisant son besoin en information. Cette interaction est souvent permise dans les SRI à travers une tâche de reformulation de requêtes. Il s'agit de l'une des applications majeures des ontologies dans le cadre des SRI conceptuels. Du fait de la taille, toujours plus grande, des ontologies, il est important de proposer des solutions algorithmiques efficaces assurant la réactivité du SRI lors de la tâche de reformulation.

Quelles solutions algorithmiques efficaces pour une stratégie de reformulation de requête conceptuelle mettant en œuvre des jugements utilisateurs et une ontologie ?

Les modèles et techniques présentés dans ce manuscrit viennent apporter des éléments de réponse à ces questions et ouvrir de nouvelles perspectives de recherche.

Au regard des questions que nous venons de soulever, nos objectifs sont les suivants :

- mise en place d'un modèle de pertinence mettant en œuvre une ontologie de domaine modélisant la collection de documents cible et un modèle de préférences simple représentant l'utilisateur ;
- conception d'un modèle de justification et de diagnostic des résultats d'une requête afin d'expliquer à l'utilisateur l'adéquation des documents qui lui ont été restitués avec sa requête. Ce modèle repose sur deux aspects. Premièrement, il s'agit d'un modèle de diagnostic de la contribution de chaque élément d'une requête au score d'un document. Deuxièmement, nous proposons la mise en œuvre d'une stratégie de visualisation originale des résultats d'une requête ;
- dans le cadre d'un SRI conceptuel, nous proposons une stratégie rapide et efficiente de reformulation de requêtes exploitant des jugements des utilisateurs ainsi qu'une ontologie de domaine.

1.4. Organisation du manuscrit

Le présent manuscrit est organisé de la manière suivante.

Le Chapitre 2 présente un état de l'art concernant la Recherche d'Information dont l'un des objectifs principaux est de justifier la nécessité de disposer d'un modèle de connaissance pour une description précise et non ambigüe des documents et des requêtes utilisateurs, le but étant l'amélioration de la pertinence des résultats. Le deuxième objectif est de montrer, qu'étant donné que, dans la boucle de pertinence, l'utilisateur est seul juge des résultats, il est nécessaire de le prendre en compte dans le modèle de pertinence à travers un modèle de préférence, aussi simple soit-il, à intégrer dans sa requête. Par ailleurs, nous montrons qu'il est nécessaire de le considérer à nouveau lorsqu'il s'agit de présenter les résultats de sa recherche au travers d'une visualisation favorisant une meilleure interaction. Ces deux objectifs s'appuient d'abord sur une critique des approches classiques de RI (c.f. 2.2) reposant sur une indexation basée sur des termes. Ce diagnostic justifie le fondement des approches conceptuelles de RI reposant sur la notion d'ontologie que nous introduisons dans la section 2.3. Un état de l'art des approches de RI conceptuelle est fourni (c.f. 2.4). Dans la section 2.5, nous étudions la prise en compte de l'utilisateur à différents niveaux dans le SRI. Le premier niveau (c.f. 2.5.1) concerne le modèle de pertinence en RI pour lequel nous introduisons un modèle d'agrégation qui synthétise les préférences de l'utilisateur sous une forme analytique facilement exploitable. Le deuxième niveau concerne la visualisation (c.f. 2.5.2) des résultats d'une recherche, notamment des fonctionnalités de justification et de diagnostic en accord avec les modes cognitifs de l'utilisateur. Enfin, le troisième niveau concerne les stratégies de reformulation de requêtes permettant à un utilisateur de préciser son besoin en information (c.f. 2.5.3). Pour finir, le Chapitre 2 traite des différentes approches existantes pour l'évaluation des SRI et de leurs principales métriques (c.f. 2.6).

Le Chapitre 3 montre l'utilisation que nous avons faite d'une ontologie de domaine dans le cadre d'un SRI. Il s'agit de l'approche *OBIRS* consistant en un modèle d'agrégation à trois niveaux (c.f. 3.3) allant de l'expression d'une requête conceptuelle utilisateur (incluant un modèle de préférence simple) au calcul du score de pertinence d'un document selon un opérateur d'agrégation en passant par l'utilisation de mesures de proximités sémantiques comme fonctions d'utilité. Le modèle de diagnostic et de justification que nous proposons est

ensuite détaillé (c.f. 3.4). Dans la section 3.5, nous présentons le prototype d'*OBIRS* ainsi que deux validations de notre approche par des experts. La première concerne la recherche de gènes tandis que la seconde concerne la recherche de documents scientifiques traitant du Cancer. Pour chacune de ces applications, une étude de cas est fournie.

Le Chapitre 4 présente, pour finir, notre stratégie de reformulation de requête conceptuelle, OBIRS-*feedback*. Il s'agit d'une combinaison d'une méthode de réinjection de pertinence explicite (relevance feedback) et d'une méthode d'expansion conceptuelle de requêtes mettant en œuvre une ontologie de domaine. L'approche repose sur des propriétés simples mais suffisantes de nombreuses familles de mesures de similarité sémantiques pour proposer des heuristiques rapides pour la reformulation de requêtes conceptuelles. Dans la section 4.2 l'approche OBIRS-*feedback* est détaillée. Pour finir, nous menons une validation d'OBIRS-*feedback* en utilisant le corpus *MuCHMORE* et suivant le protocole TREC.

Enfin, nos contributions sont discutées dans le chapitre de conclusion (c.f. Chapitre 5) au regard des perspectives qu'elles ouvrent et des valorisations dont elles font l'objet auprès de nos partenaires.

Chapitre 2 : Modèles de connaissances et modèles de préférences pour une Recherche d'Information conceptuelle et interactive

2.1. Introduction

L'objectif premier d'un Système de Recherche d'Information (SRI) est de sélectionner, à partir d'une source d'informations, celles pertinentes relativement à un besoin en information exprimé par un utilisateur. Pour cela, plusieurs fonctionnalités sont nécessaires pour permettre le stockage, l'organisation, la recherche et la visualisation de ces informations afin qu'elles soient exploitables par un (groupe d') utilisateur(s). Afin de mettre en œuvre de telles fonctionnalités, et par là rendre opérationnel le scénario de base de la RI tel que présenté dans la section 1.2, plusieurs modèles ont été proposés. Il s'agit principalement : i) de stratégies d'indexation (pour estimer les unités d'indexation pertinentes pour la description d'un document) et de pondération (pour estimer l'importance des unités d'indexation dans un document) ; ii) de modèles de pertinence pour évaluer la pertinence des documents relativement à une requête d'un utilisateur ; iii) et d'approches pour capter et exploiter les préférences d'un utilisateur à différents niveaux dont la formulation des besoins (durant le processus d'appariement), la visualisation des résultats et la reformulation.

Depuis une dizaine d'années, l'essor des ontologies comme modèle de représentation des connaissances d'un domaine, lié à l'émergence du Web sémantique, a suscité un fort développement des approches conceptuelles en Recherche d'Information. Dans celles-ci, les unités d'indexation se rapportent à des sens (concepts) exprimés dans les documents et les requêtes et non plus à des mots contenus dans ceux-ci à l'instar des approches classiques en RI. Ce renouveau permet de dépasser certaines limites des modèles d'indexation par mots clés dues entre autres à l'ambiguïté du langage naturel. Aujourd'hui, de nombreux formalismes d'ontologies sont disponibles et leur degré de maturité permet de les exploiter dans des applications réelles. Cette thèse s'inscrit dans la lignée des approches conceptuelles pour la recherche d'information et l'état de l'art proposé dans ce chapitre détaille les principaux modèles qui peuvent être utilisés dans les différentes phases du scénario de base de la RI. Nous y décrivons l'apport de l'usage des ontologies.

La première section (c.f. 2.2) rappelle les principes de base de la recherche d'information ainsi que les modèles classiques qui la fondent. Les principales limites de ces modèles sont ensuite énoncées justifiant ainsi la nécessité d'explorer les approches conceptuelles de RI. La section 2.3 introduit la notion d'ontologie avec notamment les différents types d'ontologies existantes et leurs différents degrés de formalisation. Nous détaillons ensuite les avantages qu'offre l'introduction des ontologies dans les différents processus de la RI.

Plusieurs approches de RI conceptuelles ont été mises en œuvres dans l'état de l'art. Les principales sont introduites dans la section 2.4 où nous présentons notamment : i) l'utilisation des ontologies comme espace d'indexation des documents d'une collection et des stratégies de pondération de concepts existantes, de même que ii) les principales mesures de similarité et de proximité sémantique qui sont à la base des modèles d'appariement entre documents et requête.

Le scénario de base de la RI ne s'arrête pas aux unités d'indexation et aux modèles de pertinence. L'utilisateur y joue un rôle central. La section 2.5 étudie la manière dont

l'utilisateur est inclus dans la boucle de pertinence : i) en montrant comment sa stratégie de décision peut être prise en compte à travers des modèles d'agrégation et de préférences, ii) en étudiant les stratégies de visualisation dans les SRI conceptuels lui permettant de mieux interagir avec le système et de comprendre les résultats qui lui sont présentés, iii) et en intégrant, à travers différentes stratégies de reformulation, les précisions qu'il fournit concernant son besoin en information découlant de cette compréhension.

Enfin, il est important de pouvoir évaluer un SRI en le comparant d'une part à des systèmes existants afin d'apprécier au mieux ses apports, et en évaluant son utilisation (facilité d'appropriation, ergonomie, usages…) d'autre part. La section 2.6 présente, sans être exhaustive, les principales métriques d'évaluation des SRI généralement mises en œuvre.

2.2. Approches classiques de Recherche d'Information et leurs limites

A la lumière du scénario de base de la RI (c.f. Figure 1.1, page 4) et de sa déclinaison en terme de différents processus dans l'architecture d'un SRI (c.f. Figure 1.2 à la page 5), nous pouvons formellement, et avec (Dominich 2008), définir un SRI comme étant une fonction qui associe à chaque requête utilisateur, Q_{Ut} appartenant à l'ensemble des requêtes possibles H_Q, un ensemble D_{res}^{Q} de documents au moins partiellement ordonnés et inclus dans une collection D de documents :

$$\begin{aligned} SRI : H_Q &\rightarrow D \\ Q_{Ut} &\mapsto D_{res}^{Q} \end{aligned} \qquad \textbf{(2.1)}$$

Ainsi, on définit la fonction d'appariement *RSV* (*Retrieval Status Value*) comme suit :

$$RSV : H_Q \times D \rightarrow [0,1] \qquad \textbf{(2.2)}$$

A chaque document d_j est associé un score $RSV(Q_{Ut}, d_j)$ représentant sa pertinence relativement à une requête Q_{Ut}. Aussi, l'inégalité $RSV(Q_{Ut}, d_j) \geq RSV(Q_{Ut}, d_i)$ implique que le document d_j est au moins aussi pertinent que le document d_i au regard de la requête Q_{Ut}. Dans la suite et en l'absence d'ambiguïté, le terme *document* pourra désigner, à la fois, une granule d'information et sa représentation formelle, appelée *index du document*, qu'un SRI peut manipuler. De même, le terme *requête* pourra désigner un besoin en information ou sa représentation formelle constituée à la fois de l'objet du besoin et des préférences d'un utilisateur qui lui sont associées.

Les documents exploités, i.e. analysés puis représentés par un SRI, peuvent être de différentes natures : textes, données multimédia, gènes, etc. Historiquement, les premiers types de documents à avoir été numérisés étaient de nature textuelle, expliquant ainsi la prépondérance qui leur est donnée dans les différents modèles de pertinence. La Figure 2.1 (A) montre quelques exemples de types de documents (gènes, vidéos, etc.). Au-delà de la nature, un SRI peut considérer plusieurs vues sur des documents correspondant à différents niveaux de description. Dans le cas d'un document textuel (c.f. Figure 2.1 B), il peut s'agir, par exemple,

d'une vue par le contenu, d'une vue logique (structure XML pour un fichier HTML par exemple) ou d'une vue par des attributs (auteur, date,…).

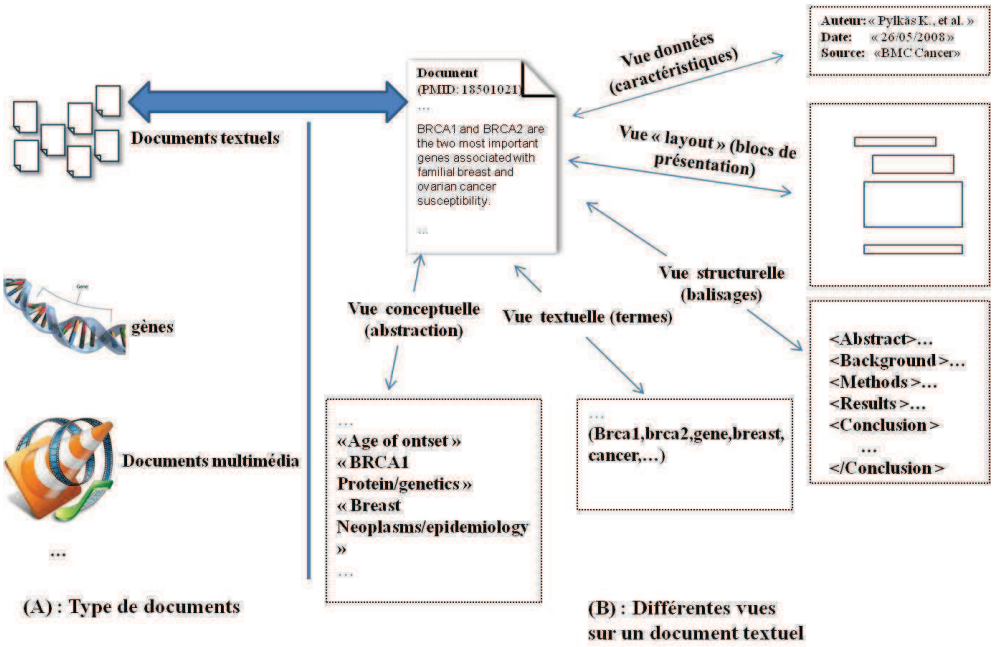


Figure 2.1 : (A) Exemples de types de documents pouvant être pris en compte dans un SRI et (B) de différentes vues possibles d'un document textuel

La définition de la RI donnée par l'équation (**2.1**) masque certaines spécificités de la recherche d'information et lui confère un large rayon d'action qu'il convient de mieux délimiter. En effet, la RI a pour but d'informer l'utilisateur sur l'existence et l'emplacement de documents en rapport avec son besoin en information. Elle ne vise pas à indiquer la manière dont son besoin en information est couvert dans les documents. La RI se différencie donc des systèmes de recherche de données (recherche SQL dans une base de données relationnelle par exemple), ainsi que des systèmes d'extraction d'information (IE). Le Tableau 2.1 synthétise les principales différences entre les systèmes de RI et ceux de recherche de données bien que leur frontière ne soit pas entièrement étanche, notamment dans le cas du modèle booléen que nous introduisons dans la section 2.2.1.1.

	RI	Recherche de données
Appariement	Approché	Exact
Spécification d'une requête	Incomplète	Complète
Eléments recherchés	Eléments pertinents	Recherche exhaustive
Sensibilité aux erreurs de syntaxe	Insensible	Sensible

Tableau 2.1 : Comparaison entre un système de Recherche d'Information et un système de recherche de données (Rijsbergen 1979)

Une fois le champ d'action de la RI défini, nous allons introduire les principales familles de modèles de Recherche d'Information existantes dans la section suivante. Nous utilisons, pour cela, les notations du Tableau 2.2 par souci d'uniformisation de la présentation tout en sachant qu'il ne s'agit nullement de notations standards acceptées par tous (il n'en existe pas d'ailleurs à notre connaissance). Pour chaque approche de RI que nous présentons, nous allons indiquer

le formalisme qu'elle met en œuvre pour représenter documents et requêtes (son modèle d'indexation) ainsi que celui qui est utilisé par la fonction d'appariement *RSV*.

Notation	Description
$d_j = \{t_r, r = 1..dl_j\}$	Un document d_j indexé par un nombre dl_j de termes t_r
$D = \{d_j, j = 1..\|D\|\}$	Collection de documents d_j
$Q = \{t_r, r = 1..\|Q\|\}$	Une requête Q
$tf_{t_r}^{Q}$	Fréquence du terme t_r dans la requête Q
$tf_{t_r}^{d_j}$	Fréquence du terme t_r dans le document d_j
$Ntf_{t_r}^{d_j}$	Version normalisée dans [0,1] de $tf_{t_r}^{d_j}$
$df_{t_r} = \sum_{d_j} tf_{t_r}^{d_j}$	Somme des fréquences de t_r dans la collection D
idf_{t_r}	Fréquence inversée du terme t_r dans la collection, mesurant son pouvoir discriminant. Elle dépend de df_{t_r}. Par exemple : $idf_{t_r} = log\frac{\|D\|}{df_{t_r}} + 1$
$FD_{t_r} = \sum_{d_j} Id\left(tf_{t_r}^{d_j} > 0\right)$, avec $Id\left(tf_{t_r}^{d_j} > 0\right) = 1$ si $tf_{t_r}^{d_j} > 0$ et 0 sinon.	Fréquence documentaire de t_r : le nombre de documents indexés par t_r
$dlAv$	Longueur moyenne des documents

Tableau 2.2 : Principales notations et leur description en Recherche d'Information classique

2.2.1. Principaux modèles de pertinence en Recherche d'Information

2.2.1.1. Modèles booléens

Le modèle booléen est l'un des premiers et des plus simples modèles de RI. Il est basé sur la théorie des ensembles. Dans ce modèle, une requête Q est une expression booléenne reliant des termes par les trois connecteurs logiques *ET* ($\wedge$), *OU* ($\vee$) et *NON* ($\neg$). En pratique, on pré-calcule pour chaque terme t_r l'ensemble D_{t_r} des documents le contenant (*inverted file*). L'évaluation d'une requête s'effectue, ensuite, par application d'opérations sur ces ensembles. Par exemple, considérons les termes suivants issus du thésaurus *MeSH*, ainsi que les documents qu'ils indexent :

$D_{tumor} = \{d_1, d_2, d_3, d_4\}$
$D_{brca1} = \{d_1, d_2, d_3\}$
$D_{leukemia} = \{d_1\}$

La requête $Q = tumor \wedge (brca1 \vee \neg leukemia)$ se traduit de manière ensembliste en $D_{res}^{Q} = D_{tumor} \cap \left(D_{brca1} \cup (D \backslash D_{leukemia})\right)$.

La notion de pertinence est modélisée comme une propriété binaire des documents (le *RSV* d'un document est soit 0 soit 1) :

$$RSV(Q, d_j) = \begin{cases} 1 \text{ si } d_j \in D_{res}^{Q} \\ 0 \; sinon \end{cases} \quad \textbf{(2.3)}$$

Dans notre exemple, le résultat de la requête Q ainsi formée correspond à $\{d_1, d_2, d_3, d_4\}$.

Il apparait que la pertinence au sens de ce modèle est rigide. Il n'existe pas de notion de réponse partielle. Son avantage principal est sa simplicité ainsi que sa transparence. En effet, l'utilisateur se voit retourner en réponse à sa requête des documents qui correspondent exactement à la requête formulée. Cependant, ce modèle présente plusieurs limites. En effet, il est basé sur un critère de décision binaire (adéquation exacte) ayant pour conséquence le fait que les résultats ne peuvent pas être ordonnés. Le nombre de résultats retournés est souvent très grand ou très petit, d'où des difficultés pour les analyser. Par ailleurs, il a été montré que les opérateurs booléens sont très rarement (Jansen 2000) ou très mal (Lucas & Topi 2004) utilisés par les utilisateurs pour exprimer leurs requêtes dans le cadre des moteurs de recherche sur le Web.

2.2.1.2. Modèle vectoriel

Dans le modèle vectoriel, documents et requêtes sont tous deux représentés par un vecteur de poids exprimé dans l'espace euclidien engendré par l'ensemble des termes t_r d'indexation (Salton et al. 1975, 1988). La pondération de chaque terme quantifie son importance dans un document et dans une requête. Dans ce cadre, un document d_j (respectivement une requête$Q$) est formellement représenté par un vecteur $\overrightarrow{d_j} = \left(w_{1j}, w_{2j}, \dots, w_{df_{t_r}j}\right)^T$ (respectivement par $\vec{Q} = \left(w_{1Q}, w_{2Q}, \dots, w_{df_{t_r}Q}\right)^T$). La pertinence d'un document et d'une requête se réduit alors au calcul de la similarité entre deux vecteurs $\overrightarrow{d_j}$ et $\vec{Q}$ dans un espace euclidien : il s'agit de trouver les documents dont le vecteur s'approche le plus du vecteur requête. Différentes mesures de similarité entre vecteurs ont été exploitées dans le cadre de ce modèle dont :

- le cosinus :

$$RSV\left(\vec{Q}, \overrightarrow{d_j}\right) = cos(\alpha) = \frac{\sum_{i=1}^{df_{t_r}} w_{iQ} * w_{ij}}{\left(\sum_{i=1}^{df_{t_r}} w_{iQ}\right)^{1/2} * \left(\sum_{i=1}^{df_{t_r}} w_{ij}\right)^{1/2}} \quad \textbf{(2.4)}$$

- le coefficient de Jaccard :

$$RSV\left(\vec{Q}, \overrightarrow{d_j}\right) = \frac{\sum_{i=1}^{df_{t_r}} w_{iQ} * w_{ij}}{\sum_{i=1}^{df_{t_r}} {w_{iQ}}^2 + \sum_{i=1}^{df_{t_r}} {w_{ij}}^2 - \sum_{i=1}^{df_{t_r}} w_{iQ} * w_{ij}} \quad \textbf{(2.5)}$$

- le coefficient de Dice :

$$RSV\left(\vec{Q}, \overrightarrow{d_j}\right) = \frac{2 * \sum_{i=1}^{df_{t_r}} w_{iQ} * w_{ij}}{\sum_{i=1}^{df_{t_r}} {w_{iQ}}^2 * \sum_{i=1}^{df_{t_r}} {w_{ij}}^2} \quad \textbf{(2.6)}$$

Le cosinus est la mesure de similarité entre vecteurs de documents et de requêtes la plus utilisée. La Figure 2.2 montre un exemple de représentation d'un document $\overrightarrow{d_j}$ et d'une requête $\vec{Q}$ dans un espace formé de deux termes $\overrightarrow{t_1}$ et $\overrightarrow{t_2}$. Dans ce cas, la pertinence est représentée par $cos(\alpha)$.

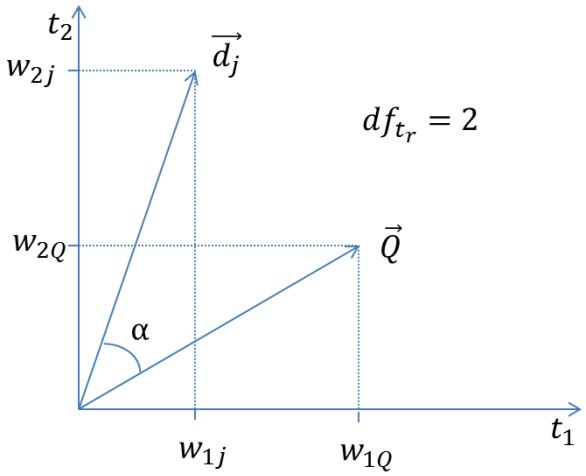


Figure 2.2 : Représentation d'un document et d'une requête dans un espace d'indexation à deux dimensions

Dans le cadre vectoriel, un document est retrouvé même s'il ne satisfait pas complètement la requête et un terme de poids nul, dans le document, joue toujours un rôle en ce sens qu'il diminue le *RSV*. Souvent les vecteurs sont normalisés, c'est-à-dire que chacune de leurs composantes est divisée par leur norme pour obtenir un vecteur de normes unitaires, dans le but de ne pas favoriser les documents longs.

Il est important de souligner les hypothèses sur lesquelles se fonde le modèle vectoriel et qui constituent en même temps ses principaux désavantages. En effet, dans ce modèle les termes indexant les documents sont considérés comme indépendants, c'est-à-dire que la présence simultanée de deux ou plusieurs termes dans un document n'implique pas forcément leur corrélation. Cette hypothèse a pour conséquence le fait que les termes d'indexation constituent une base dans l'espace des termes d'indexation. La considération de la dépendance des termes est susceptible de faire baisser les performances du SRI du fait que celle-ci est généralement locale (dans un document par exemple) et non globale à la collection. Malgré ces réserves et sa relative simplicité, le modèle vectoriel reste le modèle le plus utilisé, notamment pour la recherche de documents textuels.

Le succès du modèle vectoriel dépend plus de la stratégie de pondération des termes mise en œuvre que de la structure de l'espace vectoriel adopté. Très tôt, la pondération des termes est exposée comme étant un problème non trivial (Salton 1969). Ces poids peuvent être binaires (0 ou 1), ou prendre des valeurs réelles, généralement dans l'intervalle [0,1], dans le cas d'une indexation *floue*. Le rôle joué par une pondération non booléenne est multiple en permettant de renvoyer des documents répondant approximativement aux requêtes et d'ordonner les résultats. Dès lors, le modèle de pondération devient un composant essentiel d'un SRI. La pondération d'un terme d'une requête peut provenir de l'indexation des documents (lorsqu'il y

a une occurrence de ce terme dans le document) ou être précisée par l'utilisateur lors de la formulation de sa requête. Mais, la plupart des approches de Recherche d'Information se basent sur la statistique d'occurrence des termes d'indexation dans une collection de documents pour évaluer leur importance dans les documents ou dans les requêtes. Ces méthodes tirent leur origine de *la loi de Zipf* (Zipf 1949) et de *la conjecture de Luhn* (Luhn 1957). La loi de Zipf stipule que si nous classons, par ordre de fréquences décroissantes, les termes qui apparaissent dans un texte, alors il apparait que la fréquence d'un mot est inversement proportionnelle à son rang de classement dans la liste :

$$rang * frequence \approx constante \quad \textbf{(2.7)}$$

Utilisant la loi de Zipf, la conjecture de Luhn émet une hypothèse sur l'informativité des termes dans un document. La relation entre la fréquence et le rang des termes permet de sélectionner les termes représentatifs d'un document en éliminant les plus fréquents (supposés non représentatifs pour la description du document) et les très rares.

Sur la base de ces deux considérations, plusieurs techniques de pondération ont vu le jour. La plus simple et la plus utilisée est celle basée sur le facteur $tf_{t_r}^{d_j}$ (pondération locale utilisant la fréquence du terme t_r dans le document considéré) et le facteur idf_{t_r} (pondération globale relative à la fréquence du terme t_r dans la collection). Le facteur $tf_{t_r}^{d_j}$ peut être utilisé tel quel ou sa forme normalisée $Ntf_{t_r}^{d_j}$.

Le facteur idf_{t_r} prend en compte le fait qu'un terme qui apparaît souvent dans la collection ne doit pas avoir le même impact qu'un terme moins fréquent. La formule $tf_{t_r}^{d_j} * idf_{t_r}$ constitue une approximation de l'importance d'un terme dans un document. Cependant, elle ne prend pas en compte la longueur des documents ce qui peut être problématique.

2.2.1.3. Modèles booléens étendus

Ce modèle fut développé suite au constat que le modèle vectoriel ne permettait pas de faire des requêtes booléennes (Salton et al. 1983). L'objectif est de permettre l'ordonnancement des résultats dans le modèle booléen. Pour ce faire, il faut assouplir la condition d'appartenance d'un terme à un ensemble (qu'elle ne soit plus binaire). Le concept d'appartenance partielle à un ensemble est introduit ainsi que le concept de pondération des termes d'indexation. Une telle extension combine des caractéristiques des modèles vectoriel et booléen. En effet, elle permet d'exprimer une requête dont les termes sont séparés par des « *ET* » ou par des « *OU* » tout en utilisant un espace vectoriel pour prendre en compte la pondération de ces termes dans l'indexation des documents considérés et permet donc de classer ces documents par pertinence.

Considérons une requête simple constituée de deux termes t_1 et t_2. Dans ce cas, les vecteurs de poids décrivant l'indexation des documents sont réduits aux deux coordonnées correspondant à ces termes. On réduit ainsi la dimension de l'espace vectoriel utilisé en se focalisant sur ces dimensions pertinentes pour la requête. L'objectif est ensuite d'identifier les

documents dont les coordonnées sur ces deux axes sont aussi grandes que possible. On peut envisager cela sous deux angles différents : soit comme une recherche des documents dont l'indexation (w_1, w_2) est aussi proche que possible du vecteur (1,1) de cet espace ; soit comme une recherche des documents dont l'indexation est aussi éloignée que possible du vecteur (0,0) de cet espace. Ces deux manières d'envisager le problème correspondent à ce que (Salton et al. 1983) appelle le « *ET* étendu » et le « *OU* étendu » et les *RSV* correspondants se calculent (sur notre exemple) de la manière suivante :

$$RSV\left(Q_{ET}, d_1\right) = 1 - \sqrt{\frac{(1-w_1)^2 + (1-w_2)^2}{2}} \qquad \textbf{(2.8)}$$

$$RSV\left(Q_{OU}, d_1\right) = \sqrt{\frac{(w_1)^2 + (w_2)^2}{2}} \qquad \textbf{(2.9)}$$

2.2.1.4. Modèles probabilistes

Ces modèles abordent le problème de la RI dans un cadre probabiliste. Le premier modèle de RI entrant dans ce cadre fut proposé dans (Maron & Kuhns 1960). L'intuition derrière ce modèle réside dans le constat que la RI est un processus incertain et imprécis. L'imprécision réside dans l'expression des besoins et l'incertitude dans la représentation des informations, d'où l'usage des probabilités pour tenter de mesurer cette incertitude et cette imprécision.

Le modèle probabiliste se propose de mesurer l'adéquation d'un document vis-à-vis d'une requête par la probabilité de pertinence de ce document vis-à-vis de cette requête. Un critère d'ordre sur l'ensemble des documents est ainsi proposé, il est désigné par "*probability ranking principle"* (*PRP*) (Robertson 1977).

Dans cette vision probabiliste, il n'y a pas "des documents plus ou moins pertinents" mais "des documents dont on est plus ou moins sûr qu'ils sont pertinents". Dans le cadre du *PRP*, pour un document d_j, et une requête Q il n'y a que deux possibilités qui se présentent :

- R : d_j est pertinent pour Q
- $\overline{R}$: d_j est non pertinent pour Q

On est dans le cas de deux classes considérées comme des variables. Il s'agit, dans la fonction d'évaluation, de choisir les documents dont la probabilité d'appartenance à la classe des documents pertinents, $P(R|d_j)$, est supérieure à leur probabilité d'appartenir à la classe des documents non pertinents, $\mathrm{P}(\overline{R}|d_j)$. Dans un tel contexte, le *RSV* est donné par la formule suivante :

$$RSV(Q, d_j) = \frac{\mathrm{P}\left(\mathrm{R}|d_j\right)}{\mathrm{P}\left(\overline{R}|d_j\right)} \qquad \textbf{(2.10)}$$

L'application de la règle de Bayes donne :

$$\mathrm{P}\left(\mathrm{R}|d_j\right)=\frac{\mathrm{P}\left(d_j \mid R\right)\mathrm{P}\left(R\right)}{\mathrm{P}\left(d_j\right)} \qquad \mathrm{P}\left(\overline{R} \mid d_j\right)=\frac{\mathrm{P}\left(d_j \mid \overline{R}\right)\mathrm{P}\left(\overline{R}\right)}{\mathrm{P}\left(d_j\right)} \qquad \textbf{(2.11)}$$

Avec $\mathrm{P}(d_j \mid R)$ (respectivement $\mathrm{P}(d_j \mid \overline{R})$) la probabilité d'observer le document d_j sachant qu'il est pertinent (respectivement sachant qu'il n'est pas pertinent). $\mathrm{P}(\mathrm{R}|d_j)$ (respectivement $\mathrm{P}(\overline{R} \mid d_j)$) est la probabilité que le document d_j soit pertinent (respectivement non pertinent) sachant sa description. Si l'on réécrit l'équation (2.10) en substituant les termes par leur formulation dans l'équation (2.11), on obtient :

$$RSV\left(Q,d_j\right)=\frac{P(d_j \mid R)P(R)}{P(d_j)}*\frac{P(d_j)}{P(d_j \mid \overline{R})P(\overline{R})}=\frac{P(d_j \mid R)}{P(d_j \mid \overline{R})}*\frac{P(R)}{P(\overline{R})} \qquad \textbf{(2.12)}$$

Tous les documents ont a priori la même probabilité $P\left(R\right)$ d'être pertinents. $\frac{P\left(R\right)}{P\left(\overline{R}\right)}$ est donc le même pour tous les documents du corpus, l'ordonnancement des documents peut donc s'effectuer avec :

$$RSV\left(Q,d_j\right)=\frac{P(d_j \mid R)}{P(d_j \mid \overline{R})} \qquad \textbf{(2.13)}$$

On montre alors que :

$$RSV\left(Q,d_j\right)=\sum_{t_r \in Q \cap d_j}\left(\log\left(\frac{p_r(1-q_r)}{q_r(1-p_r)}\right)\right) \qquad \textbf{(2.14)}$$

où p_r (respectivement q_r) est la probabilité qu'un terme t_r de la requête apparaisse dans un document d_j sachant que celui-ci est pertinent (respectivement non pertinent).

Le problème se ramène dès lors à l'estimation de p_r et de q_r. En pratique, ces probabilités sont estimées à partir de bases d'apprentissage où les documents pertinents sont connus. On peut alors estimer la valeur de p_r par la proportion de documents pertinents (dans la base d'apprentissage) contenant le terme t_r. On estime q_r de manière analogue en évaluant la proportion de documents non pertinents contenant t_r. Pour plus de détails sur le *PRP*, le lecteur pourra consulter (Boughanem 2008).

Plusieurs approches probabilistes de RI existent et il est impossible d'en faire une représentation exhaustive. Nous allons en détailler quelques unes parmi les plus représentatives dans l'état de l'art. Nous ne traiterons pas par exemple des modèles de langues (Ponte et al. 1998) ni de la famille des modèles *DFR* (*Divergence From Randomness*) (Amati et al. 2002). Nous présentons principalement le modèle *BM25* (Robertson et al. 1996). Notons que notre objectif n'est pas de proposer une nouvelle stratégie de pondération mais nous voulons mettre en exergue les composantes principales des modèles de pondération probabilistes.

Le modèle BM25

Dans le cadre du *PRP*, si des collections d'apprentissage de documents pertinents/non pertinents ne sont pas disponibles, il est difficile d'estimer les paramètres des modèles. (Robertson & Walker 1994) introduit la notion de termes élites pour pallier cet inconvénient. De leurs travaux découle la formule *BM25* (2.15 **)** constituant l'un des modèles de RI les plus utilisés et dont la fonction *RSV* s'appuie notamment sur les fréquences d'apparition des termes de la requête dans le document considéré $\left(tf_{t_r}^{d_j}\right)$ et dans la collection $\left(FD_{t_r}\right)$:

$$RSV(Q,D) = \sum_{t_r \in Q \cap D} \left(\frac{tf_{t_r}^{d_j}}{tf_{t_r}^{d_j} + k\left((1-b) + b * \frac{dl_j}{dlAv}\right)} * \log\left(\frac{|D| - FD_{t_r} + 0.5}{FD_{t_r} + 0.5}\right) * tf_{t_r}^{Q} \right) \qquad \textbf{(2.15)}$$

k est un paramètre d'influence de la fréquence $tf_{t_r}^{d_j}$ et b est un paramètre d'influence de la normalisation $\left(\frac{dl_j}{dlAv}\right)$ de la longueur du document d_j. Dans cette équation (2.15), nous pouvons exhiber, pour le terme t_r, la composante de pondération locale $\left(Ntf_{t_r}^{d_j} = \frac{tf_{t_r}^{d_j}}{tf_{t_r}^{d_j} + k\left((1-b) + b * \frac{dl_j}{dlAv}\right)} \right)$ relative à sa fréquence $tf_{t_r}^{d_j}$ dans d_j et celle de pondération globale $\left(\log\left(\frac{|D| - FD_{t_r} + 0.5}{FD_{t_r} + 0.5}\right)\right)$ correspondant au facteur idf_{t_r}(c.f. section 2.2.1.2) relative à la fréquence de t_r dans la collection D de documents. Il s'agit donc de stratégies de pondération de termes qu'il est possible d'utiliser dans les modèles vectoriels par exemple. Il faut noter que la pondération locale n'est là que pour procéder à une normalisation des longueurs des documents.

Au-delà du modèle de pertinence mis en œuvre, les modèles de Recherche d'Information peuvent être catégorisés suivant l'unité d'indexation sur laquelle ils se fondent. Une telle considération conduit à deux types de modèles de RI dont un, dit *classique*, qui utilise des termes comme unité d'indexation et l'autre, dit *conceptuel*, qui utilise des concepts issus de ressources sémantiques (Haav et al. 2001). Le travail présenté dans ce mémoire s'inscrit dans la continuité des SRI conceptuels. La section suivante détaille les limites des approches classiques de RI ayant conduits à l'émergence des approches conceptuelles.

2.2.2. Limites des modèles "*sac de termes*" : de la nécessité d'un modèle de connaissances en Recherche d'Information

Dans les approches dites "classiques" de RI, l'unité d'indexation est constituée par des termes (éventuellement composés). La majeure partie des SRI utilisent historiquement de telles unités d'indexation. C'est le cas notamment du système SMART (Salton 1969). Ces indexations sont obtenues au moyen d'une analyse lexicale permettant d'extraire de chaque document les termes que l'on estime les plus discriminants pour décrire son contenu. Les degrés d'importance de tels termes sont captés par un système de pondération généralement

basé sur la statistique d'usage de ceux-ci dans la collection (e.g. estimé par le coefficient *TF/IDF* (Salton et al. 1975)).

Le modèle de représentation (indexation) de documents le plus répandu dans les modèles de RI classiques est celui du *sac de termes* (Figure 2.3). Dans cette représentation, un document est vu comme un ensemble de jetons pondérés représentant chacun un terme extrait du document. L'hypothèse fondamentale de cette approche est celle de l'indépendance des termes d'indexation.

Figure 2.3 : Le modèle classique de "*sac de termes*" utilisé pour indexer les documents dans un SRI

Evaluer la pertinence d'un document par rapport à une requête nécessite de déterminer les termes de la requête présents dans l'index du document. Dans les SRI classiques, les modèles de pertinence peuvent reposer sur un appariement exact ou sur une similarité entre chaîne de caractères représentant les termes d'indexation. La distance d'édition, par exemple, peut être utilisée pour mesurer une telle similarité comme étant une fonction du nombre minimal de caractères à supprimer, insérer ou remplacer pour passer d'un terme à un autre. Ainsi lorsque l'on soumet une requête, les SRI classiques retrouvent les documents indexés par les termes de la requête ou par certaines de leurs variations lexicales (*tumerous* au lieu de *tumor* par exemple). Cependant, ils souffrent d'un certain nombre de défauts affectant leurs performances. En effet, les SRI classiques ne retrouvent pas des documents ayant, dans leur index, des synonymes des termes de la requête (*carcinoma* au lieu de *tumor* dans la Figure 2.4 par exemple). Cette limitation est la plus commune selon (Haav et al. 2001; Bhagdev et al. 2008; Giunchiglia et al. 2009). Au-delà de la synonymie, plusieurs autres relations entre termes peuvent exister. La méronymie, traduisant une relation d'une partie à un tout (*pied* est un *méronyme* de *corps*), en est un exemple de même que la relation de spécialisation (une *voiture est un véhicule*). Dans les approches classiques, aucune distance entre chaîne de caractères ne permet de trouver une similarité entre *voiture* et *véhicule* bien que ces deux termes soient liés : un document traitant de *voiture* pouvant être pertinent, jusqu'à une certaine mesure, pour une recherche portant sur *véhicule*.

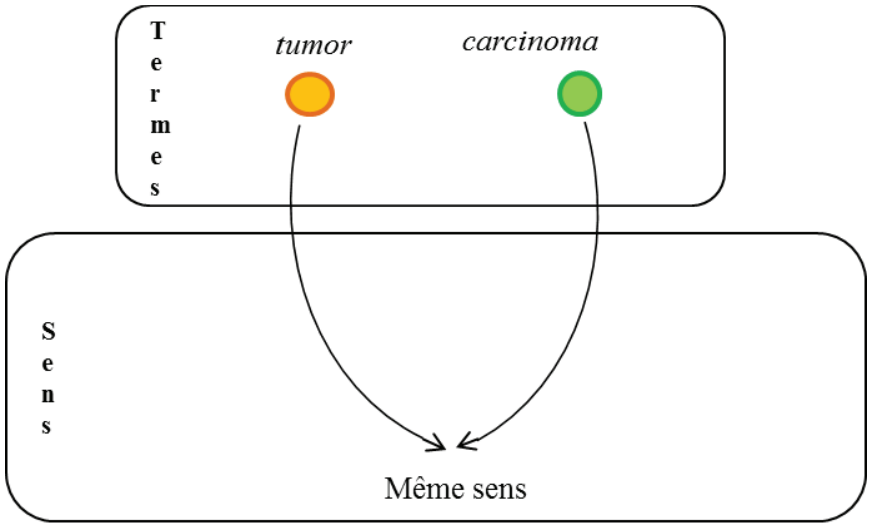


Figure 2.4 : Une illustration du problème de synonymie des approches « sacs de termes » en RI. Les termes *tumor* et *carcinoma* sont synonymes.

Par ailleurs, du fait de l'ambigüité intrinsèque au langage naturel, les SRI classiques souffrent d'un problème de polysémie. En effet, un terme peut revêtir plusieurs sens en fonction de son contexte d'utilisation. Dans la Figure 2.5, le terme *java* peut renvoyer au *café*, à un *language de programmation* ou à *l'île* indonésienne. Un SRI classique sélectionnera donc un document dès lors que son index contient un terme de la requête, même si le sens de celui-ci dans le contexte du document est différent du sens que l'utilisateur avait à l'esprit en formulant sa requête. Tous ces problèmes conduisent à un manque de précision des SRI classiques (Stokoe et al. 2003).

Pour lever les limites relatives à la non prise en compte de relations entre termes et à l'ambigüité due à la polysémie, plusieurs solutions endogènes aux approches classiques de RI ont été envisagées. Elles concernent principalement l'exploitation des statistiques de cooccurrences des termes d'indexation pour tenter d'en préciser le sens (désambigüiser) en exploitant celui (supposé non ambigu) d'autres termes co-occurrents. La relation de cooccurrence est déterminée en utilisant un contexte qui est soit une partie du document (un paragraphe par exemple), soit le document en entier, voire l'ensemble de la collection. Les différents travaux dans ce sens, dont ceux de (Peat & Willett 1991), montrent que de telles solutions ne permettent pas de résoudre efficacement les problèmes soulevés du fait qu'une cooccurrence de termes suggère comme statistiquement probable un lien entre eux sans en préciser la nature. En plus, l'existence de relations entre des termes ne dépend pas de leur présence conjointe dans un corpus mais elle préexiste (dans un domaine ou dans une langue) à la constitution de celui-ci. Il n'est pas raisonnable, donc, de faire et défaire des relations entre deux mêmes termes au gré de corpus ou de modèles statistiques utilisés pour étudier leur cooccurrence.

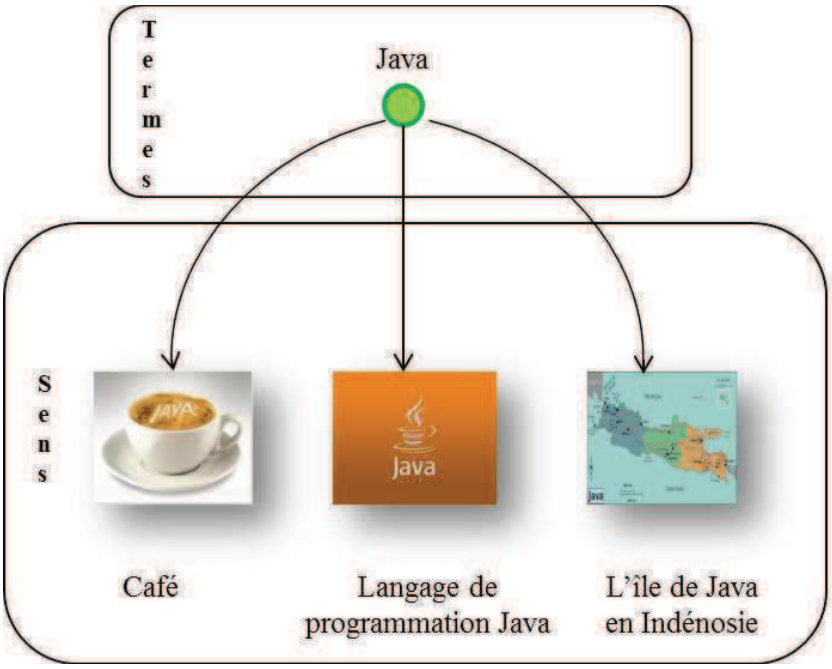


Figure 2.5 : Trois sens associés au terme "java"

Plus généralement, le manque d'efficacité en termes de précision et de rappel des approches classiques émane de la non-prise en compte de l'aptitude d'un humain à manipuler des idées ou des objets ainsi que leurs relations explicites ou implicites. Aptitude qui est innée pour interpréter les documents dont il est l'auteur. Considérons le dialogue suivant entre un professeur et un élève afin d'illustrer la stratégie mise en œuvre par un humain pour répondre à une question :

- Le professeur : « Quel est le dernier livre que tu as lu ?»
- L'élève : « J'ai lu le dernier essai de Léopold Sedar Senghor »

La réponse de l'élève ne repose pas uniquement sur les mots de la question du professeur. L'élève a mobilisé ses connaissances en littérature d'une part et la catégorisation (non nécessairement exhaustive) des choses et idées dans ce domaine d'autre part. Les deux processus suivants sont à l'œuvre :

- Capacité d'identification : aptitude à identifier l'objet "livre" et l'action de "lire".
- Capacité de catégorisation, de spécialisation et de généralisation : pouvoir de déduction et d'inférence permettant de savoir et de retenir qu'un "essai" (ou un "roman", etc.) est un livre.

Pour l'humain, la capacité d'identifier les choses et de les catégoriser émane d'un apprentissage continu et d'une mémorisation grâce à des interactions avec son environnement (sa famille, l'école, etc.). D'un autre côté, la communication entre le professeur et l'élève n'a été possible que parce qu'ils partagent la connaissance relative à l'identification et la catégorisation des types de livres (essais, romans,…) et de l'action de lire. Si tel n'avait pas été le cas, i.e. l'hypothétique ignorance de l'élève sur les différents types de livres, le dialogue suivant aurait pu avoir lieu :

- Le professeur : "Quel est le dernier livre que tu as lu ?"
- Elève : Peut-on considérer un essai comme étant un livre ?
- Professeur : Oui
- L'élève : "J'ai lu le dernier essai de Léopold Sedar Senghor"

Il s'agit d'un rééquilibrage (alignement) qui tend à mettre au même niveau de langage et de connaissance le professeur et l'élève.

Ce dialogue simple nous montre que l'humain a besoin de connaissances partagées (référentiel de connaissances) et suscitant un consensus avec son environnement pour communiquer et se faire comprendre efficacement. Les systèmes de recherche d'information classiques, en se limitant aux termes et à leurs statistiques d'usages dans les collections de documents, passent à côté des idées et des connaissances explicites ou implicites que les auteurs ont voulu transmettre dans ces documents. Pour surmonter ces limites, une solution est donc de se rapporter aux sens associés aux termes comme unité d'indexation et de les rendre explicites et exploitables. L'hypothèse fondamentale d'une telle approche est de considérer que le contenu des documents ainsi que les besoins en information des utilisateurs sont mieux décrits par les abstractions conceptuelles des entités réelles que par les mots et les termes qu'ils renferment (Baziz et al. 2007; Dragoni et al. 2012). Un point de vue cognitif du monde est donc privilégié. Les SRI qui reposent sur un tel point de vue sont caractérisés de conceptuels.

Lorsque la collection dans laquelle s'effectue la recherche est relative à un domaine donné, les SRI conceptuels nécessitent donc la disponibilité d'un inventaire, plus ou moins exhaustif et exploitable, des concepts (sens) de ce domaine ainsi que de leurs relations. Les ontologies, dont la popularisation a été accélérée par l'avènement, à l'orée des années 2000, du Web sémantique (Berners-Lee et al. 2001), offrent une telle structuration des connaissances. Dans la section suivante, nous allons les définir et étudier leur mise en œuvre dans les SRI.

2.3. Les ontologies comme modèle de connaissance : vers la RI conceptuelle

Issue de l'évolution de plusieurs modèles de connaissances plus ou moins formalisés, la notion d'ontologie en informatique a connu un succès certain dans de nombreuses applications, au-delà même des communautés historiques qui l'ont vu naître : IC (Ingénierie des Connaissances) et IA (Intelligence Artificielle). En témoigne son utilisation intensive dans les sciences de la vie (Oliver et al. 2009) et en bio-informatique notamment (Pesquita et al. 2009). La définition des ontologies, leur formalisation et leur utilisation ont rapidement évolué, en grande partie poussées par le développement du Web sémantique et des applications dédiées. Leur pouvoir de description a apporté un renouveau en allant au-delà de l'utilisation de simples termes pour l'indexation des documents et la formulation de requêtes. Elles peuvent intervenir à différentes phases du processus de RI : lors de l'indexation, lors du calcul de pertinence, mais également lors de la restitution (visuelle) des résultats.

Cette section commence par définir la notion d'ontologie à travers ses héritages multiples et présente le *MeSH* et la *Gene Ontology* (*GO*), deux ontologies sur lesquelles se sont basés nos travaux (c.f. section 2.3.1). Ensuite, nous détaillons les avantages liés à l'utilisation des

ontologies dans les différentes tâches qui constituent la RI (c.f. section 2.3.2). Nous tirons de cette étude, la structure d'ontologie que nous adoptons dans le cadre de cette thèse.

2.3.1. Définition de la notion d'ontologie de par ses héritages multiples

Pour comprendre ce qu'est une ontologie, et par là justifier la multitude de définitions existantes, il est important de considérer la diversité de ses héritages. Le premier d'entre eux est philosophique. En effet, selon (Guarino & Welty 2000), la notion d'**O**ntologie est une branche de la philosophie qui concerne l'étude de *l'être* en tant qu'être et de ses propriétés. Le deuxième est l'héritage logique (Sowa 2000) où l'**o**ntologie est vue comme un ensemble de concepts qui sont reliés par des relations de plusieurs types (et notamment par la relation de spécialisation "*is-a*" induisant une taxonomie qui organise les connaissances). Ces concepts sont considérés comme de futurs prédicats logiques permettant de raisonner par classification. Le troisième est l'héritage terminologique qui a longtemps étudié les différences entre termes et concepts et l'articulation entre le langage et la connaissance (Aussenac-Gilles 2008). Le quatrième, enfin, concerne le domaine de la gestion documentaire dans le sillage des langages et thésaurus documentaires fondés sur une structuration hiérarchisée.

Plusieurs définitions des ontologies ont été proposées mais nous pouvons rappeler ici la définition de Studer (Studer et al. 1998) qui combine celle de Gruber (Gruber 1993) et celle de (Borst 1997) :

"*An ontology is a formal, explicit specification of a shared conceptualization."*

Nous traduisons cette phrase de la manière suivante :

"Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée."

Une telle définition met en exergue le fait qu'une ontologie doit être formelle (tout en ne précisant pas le degré de formalisation) et refléter une conceptualisation consensuelle au sein d'une communauté. Il s'agit donc d'une définition formelle, normalisée, partagée et non nécessairement exhaustive des connaissances d'un domaine ainsi que de leurs abstractions. Selon (Guarino et al. 2009), une conceptualisation peut être définie comme une structuration formelle d'une réalité (des objets par exemple) ou d'une partie de celle-ci, telle qu'elle est perçue et organisée par un agent indépendamment :

- du vocabulaire utilisé pour exprimer cette réalité ;
- de l'occurrence réelle d'une situation spécifique à cette réalité.

Par exemple, deux synonymes partagent la même conceptualisation (c.f. Figure 2.4).

De la notion de conceptualisation, découle la différence entre une ontologie et une classification. En effet, une classification ne s'occupe que de l'accès à des objets, organisés certes, mais représentés de manière syntaxique sans précision sur leur nature et celle des liens qui les unissent. La Figure 2.6, adaptée de (Sowa 2000), montre les processus d'abstraction, de désignation et de représentation d'un objet (livre de Léopold Sédar Senghor nommé *Liberté1* dans cet exemple) en œuvre chez l'humain. Il s'agit d'une composition de trois

triangles de sens (*meaning triangles*). Un triangle de sens permet de voir le rapport entre un objet (un livre dans cet exemple), le concept pour le *penser* et le terme (*Liberté1*) utilisé pour le désigner. Le terme *Liberté1*, qui désigne l'objet livre, est conceptualisé comme étant un mot que l'on code par le symbole "*Liberté1*". Le processus de représentation du symbole "*Liberté1*" du nom de l'objet livre, suit le même schéma de conceptualisation pour aboutir à un niveau physique (les bits codant une chaîne de caractères par exemple).

Comme nous venons de le voir, la notion de concept est au centre des ontologies. Il convient, dès lors, de bien la définir. Selon (Gandon 2002), un concept est une notion généralement exprimée par un signe et qui représente un groupe d'objets ou d'êtres partageant des caractéristiques et des propriétés permettant de les identifier comme appartenant à un même ensemble. Il s'agit donc d'une définition en intension. Par exemple, l'intension du concept "*Livre*" inclut le fait que c'est un "*document écrit formant une unité et conçu comme tel* (selon wikipedia)". L'extension d'un concept concerne l'ensemble des objets ou *êtres* dont il représente une abstraction. Si nous reprenons l'exemple du concept "*Livre*", une de ses extensions pourrait être le livre de "*Liberté1"* représenté dans la Figure 2.6. Les relations (représentant une notion d'association) entre les concepts d'une ontologie peuvent être définies suivant le même schéma, i.e. en intension (ensemble d'attributs caractérisant l'ensemble des manifestations possibles de la relation) et en extension (l'ensemble des occurrences de la relation). Un terme peut correspondre à un mot ou à un groupe de mots "non vides", donc utile, (différent des articles par exemple), pouvant être de différents types grammaticaux. Un groupe de mots peut aussi former une unité sémantique. Dès lors, un terme ne s'oppose pas toujours à un concept. Cependant, nous considérons dans ce manuscrit une séparation nette entre termes (issus des documents) et concepts (modélisés dans une ontologie).

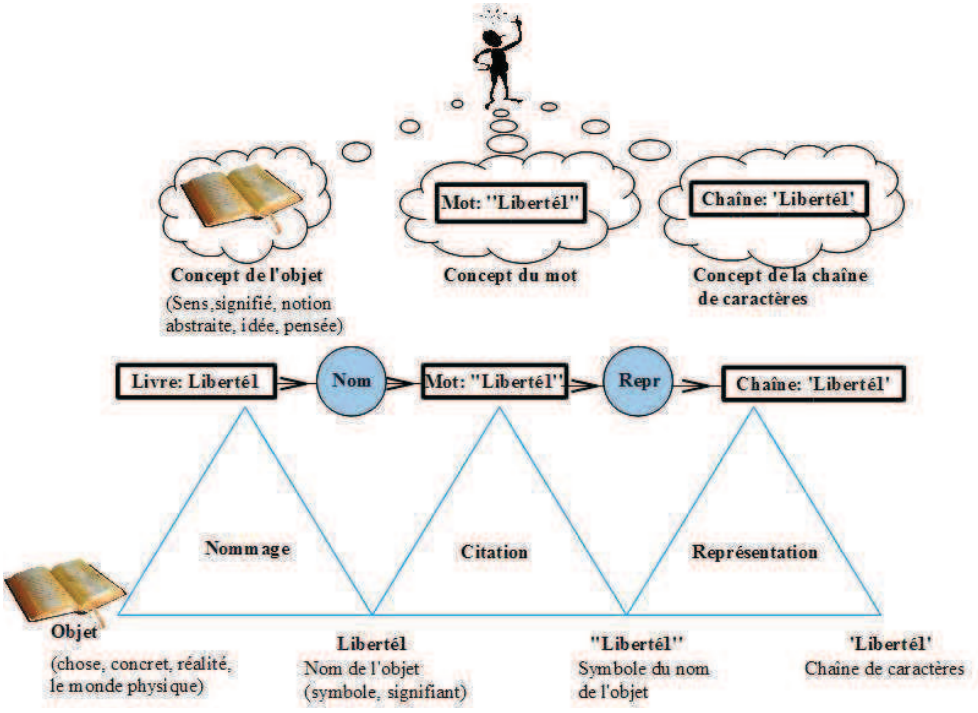


Figure 2.6 : Illustration de l'utilisation des signes pour nommer des objets relativement à une conceptualisation. Cette illustration utilise une composition de trois triangles de sens (*meaning triangles*) de Ogden and Richards adaptée de (Sowa 2000) .

Historiquement, plusieurs modèles plus ou moins précis et formels ont été proposés dans l'optique de représenter les connaissances d'un domaine. Dans (Lassila & McGuinness 2001; Guarino et al. 2009), les auteurs fournissent un continuum de tels modèles suivant leur degré de précision :

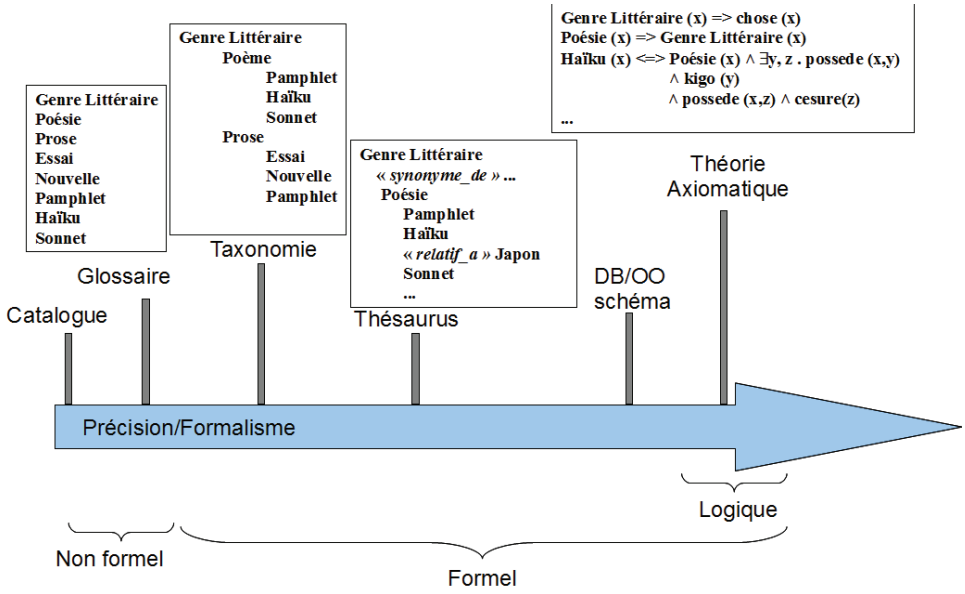


Figure 2.7 : Continuum des structures de connaissances suivant leur degré de formalisation (O. Lassila et al. 2001; N. Guarino et al. 2009).

La Figure 2.7 montre un exemple de formalisation de sous-domaines dans les genres littéraires. Le degré de précision évolue d'un vocabulaire contrôlé (sans organisation et de manière non formelle) vers un formalisme logique (OWL[9] par exemple) permettant des inférences. Le glossaire d'un livre peut être considéré comme un vocabulaire contrôlé. Pour une taxonomie, une hiérarchie simple (spécialisation) est introduite par rapport au vocabulaire contrôlé. Dans notre exemple, la *Poésie* est un *Genre littéraire* de même qu'un *Sonnet* est un poème. Un thésaurus va au-delà d'une taxonomie en introduisant des sujets connexes (relation "*relatif_a*" dans l'exemple). Une ontologie va encore plus loin en considérant les objets d'un domaine pour en extraire les unités de sens (concepts) et en explicitant leurs relations ainsi que leurs contraintes.

Dans ce manuscrit, nous adoptons la définition suivante d'une ontologie telle qu'elle est conçue dans la communauté IC (Ingénierie des Connaissances) :

Définition 2.1 : (Maedche & S. Staab 2001; Aussenac-Gilles 2008) (*définition de la structure d'une ontologie*) Une ontologie Θ peut être formellement définie par $\Theta := \{C, R, H_C, Rel, Ax\}$, dans lequel :

- C est un ensemble de concepts organisés sous la forme d'une taxonomie (hiérarchie) H_C à travers des relations taxonomiques orientées (*is-a* par exemple) permettant des héritages multiples. La relation *is-a* (*est-un*) est souvent désignée par *hyponymie* ou *hyperonymie* suivant que l'on considère le concept qui est spécialisé ou celui qui est généralisé.

[9] OWL : Web Ontology Language (http ://www.w3.org/TR/owl-features/)

- R est un ensemble de relations non taxonomiques entre concepts définies par leur domaine et co-domaine.
- $Rel : R \times C \times C \rightarrow \{0,1\}$ associe à chaque relation non taxonomique dans R , l'ensemble des couples de concepts satisfaisant cette relation. Si $r \in R$ est une relation non taxonomique et $(c_x, c_y) \in C^2$ deux concepts, alors $Rel(r, c_x, c_y) = 1$ s'il existe dans Θ la relation r entre c_x et c_y et $Rel(r, c_x, c_y) = 0$ sinon.
- Ax est un ensemble d'axiomes logiques permettant d'inférer des faits implicites et exprimés dans un langage logique adapté telle que la logique de description.

Une ontologie n'est donc pas seulement un thésaurus bien que dans la littérature certains les confondent parfois. Un thésaurus peut être vu comme un cas particulier d'ontologie. Une ontologie peut aussi être définie comme étant de domaine lorsqu'elle modélise des connaissances spécifiques à une communauté. C'est le cas notamment du domaine biomédical (thésaurus *MESH*, *Gene Ontology*), ou agricole avec *AGROVOC*[10]. Une ontologie peut aussi être *générale* lorsqu'elle traite de notions "*universelles*" comme *DBpedia* (Mendes et al. 2012).

Le thésaurus *MeSH* et la *Gene Ontology* nous ont particulièrement intéressés dans le cadre de nos applications. Cela se justifie d'une part par la grande quantité de données biomédicales disponibles et décrites à l'aide de ces deux ressources conceptuelles et par nos collaborations avec des biologistes utilisant ces données d'autre part. Nous avons donc utilisé ces deux ontologies pour évaluer nos contributions tant par des biologistes que de manière systématique à l'aide de collections de documents biomédicaux.

Le *MeSH* (*Medical Subject Headings*) est un thésaurus (vocabulaire contrôlé et structuré) de la *NLM* (*National Library of Medicine*[11]) utilisé dans l'indexation des articles publiés dans la base de documents scientifiques *PubMed*[12]. Dans sa version 2010, le *MeSH* contient 25 603 concepts organisés hiérarchiquement dans une structure de *dag* (graphe acyclique direct) avec héritages multiples. Ces concepts (descripteurs) sont regroupés en 16 catégories regroupant différents aspects du domaine biomédical (anatomie, maladies, médicaments, personnel médical, etc.). Des relations comme la synonymie, la subsomption et celle de *part-of* (partie de) sont disponibles. La Figure 2.8 montre un extrait de la réduction du MeSH aux relations *is-a*. Nous avons utilisé le *MeSH* lors des évaluations de notre environnement *OBIRS* (c.f. section 3.3) et de sa déclinaison OBIRS-*feedback*. En effet, le *MeSH* est l'ontologie de support de l'indexation du corpus de tests *MuchMORE* (c.f. section 4.3).

La Gene Ontologie (GO) a été conçue initialement, pour unifier le vocabulaire d'un domaine auquel différents experts contribuent : biologistes, médecins, physiciens, chimistes, etc. A l'origine, cette ontologie n'en était pas formellement une, même si elle en avait le nom. Elle servait de pivot entre des vocabulaires, souvent différents, utilisés par les différentes communautés pour désigner les mêmes choses. A partir du moment où elle a servi de support pour des traitements automatisés, il a fallu désambiguïser, organiser, compléter, amender ce

[10] http ://aims.fao.org/standards/agrovoc/about
[11] http ://www.nlm.nih.gov/
[12] http ://www.ncbi.nlm.nih.gov/pubmed

vocabulaire pour atteindre un certain degré de formalisation. Aujourd'hui, c'est une des ontologies les plus utilisée au plan international, pour l'annotation de gènes (http://www.geneontology.org/) et elle compte plus de 35 000 concepts. On y retrouve tous les concepts utiles pour indexer des gènes, organisés en trois parties : les *processus biologiques* dans lesquels ces gènes sont impliqués (*biological process*), les *fonctions moléculaires* dans lesquelles ils interviennent (*molecular functions*) et les *composants cellulaires* où ils apparaissent (*cellular component*). Du fait de son utilisation massive, cette ontologie est devenue un standard utilisé pour l'annotation par la plupart des bases de données de gènes, en particulier, la base de données Ensembl, que nous utilisons (c.f. section 3.5). La Figure 2.9 montre un extrait de la Gene Ontology.

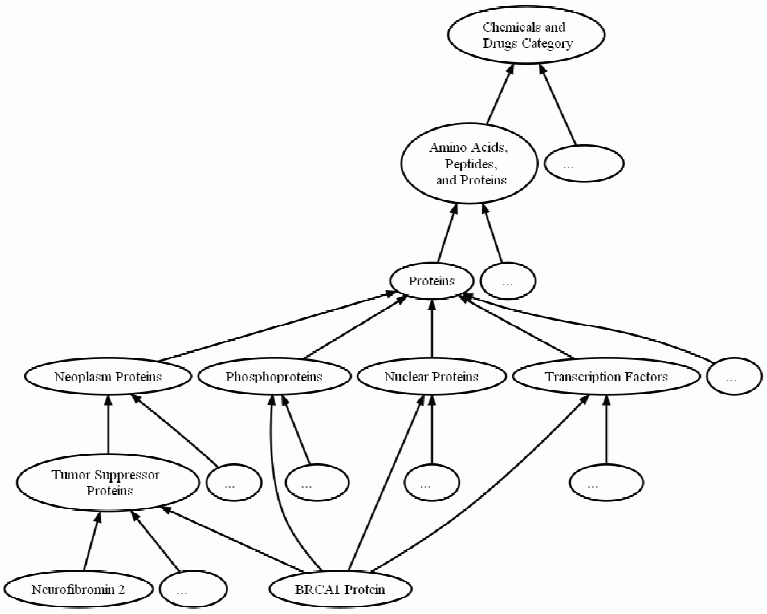


Figure 2.8 : Extrait de la catégorie "*Chemicals and Drugs Category*" du thésaurus *MeSH*

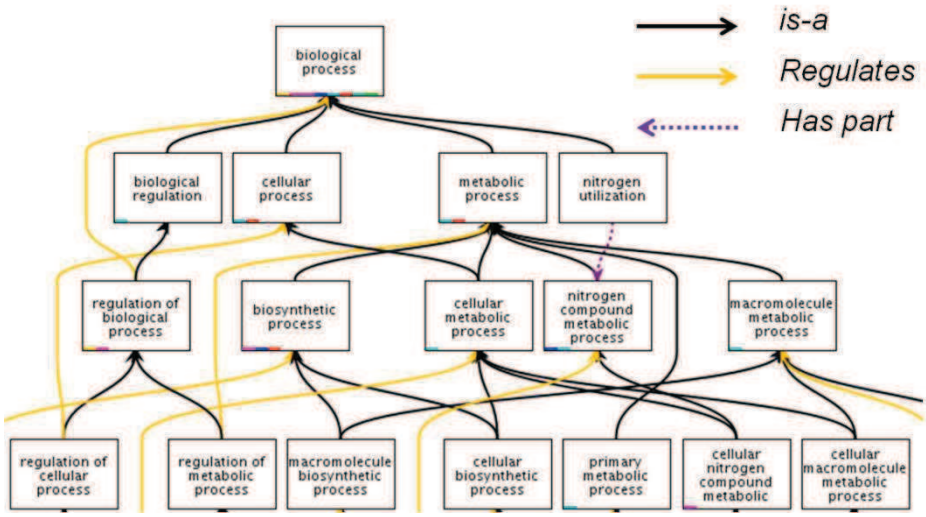


Figure 2.9 : Extrait de la branche *biological process* de la Gene Ontology

2.3.2. Quels apports des ontologies pour la Recherche d'Information ?

Avec une ontologie comme modèle de connaissance formel, nous disposons d'un moyen, potentiellement puissant, de représentation des connaissances d'un domaine et pouvant intervenir dans plusieurs processus d'un SRI. Dans la suite, nous présentons les principaux apports des ontologies à ces différents processus :

- **Amélioration de l'indexation.** Les concepts de l'ontologie peuvent se substituer aux termes dans les indexations des documents et des requêtes. Ainsi ces représentations se rapportent aux sens et permettent de distinguer les occurrences des termes se rapportant à des sens différents (*polysémie*) mais aussi de regrouper les termes *synonymes* ou les différentes formulations d'une même idée. Dans ce contexte, les ontologies sont considérées comme une version élaborée et formelle des thésaurus et langages documentaires ; leur formalisation permettant d'élargir les possibilités de caractérisation des documents (en terme de contenu informationnel) et des besoins en information des utilisateurs. Une ontologie est vue comme un langage pivot entre l'utilisateur qui a un besoin en information et le SRI. Il est important de distinguer une indexation conceptuelle d'une annotation sémantique. En effet, les deux notions sont souvent confondues alors qu'il subsiste une différence essentielle entre elles. L'indexation conceptuelle découle d'une tradition documentaire et vise donc la classification des objets tandis que l'annotation sémantique aboutit à la production de faits (généralement sous forme de triplets RDF[13]) destinés à enrichir une base de connaissances et à permettre un raisonnement. Les types de relations sémantiques considérés sont dès lors différents et représentent les relations taxonomiques pour l'indexation conceptuelle ou toute autre relation pour l'annotation sémantique. Dans cette thèse, nous nous intéressons aux indexations conceptuelles des documents et des requêtes.
- **Amélioration des fonctions d'appariement.** L'espace conceptuel ainsi fourni, permet de mettre en œuvre des mesures de similarité sémantique exploitant, notamment, l'organisation taxonomique des concepts. Ces mesures permettent d'indiquer dans quelle mesure deux concepts sont proches et par extension dans quelle mesure une requête et un document indexés par des concepts le sont. Les modèles de pertinence basés sur une ontologie (Baziz et al. 2007; Dragoni et al. 2012) disposent d'une sémantique explicite entre les unités d'indexation contrairement aux approches classiques de RI où l'on tente de la déterminer à travers des modèles statistiques et probabilistes.
- **Pivot pour l'intégration de données.** Il s'agit de l'intégration au niveau conceptuel et sémantique de sources de données hétérogènes de par leur format de représentations par exemple. Servant de langage pivot, les ontologies permettent d'unifier la représentation et l'interrogation de telles ressources moyennant un coût d'indexation. Cette intégration peut être centralisée reposant ainsi sur une seule ontologie (pouvant éventuellement résulter de la fusion d'autres ontologies) comme dans (Jalabert 2007).

[13] Ressource Description Framework : http ://www.w3.org/RDF/

Elle peut aussi être décentralisée suivant différents entrepôts disposant chacun d'une ontologie (Gandon et al. 2008).

- **Support pour l'expansion de requêtes.** Il s'agit de la possibilité d'obtenir des concepts sémantiquement proches de ceux de la requête (par synonymie ou hyponymie par exemple). L'ontologie est alors utilisée comme une source d'évidences externes et globales dans le cadre de la tâche de reformulation de requête. Ce point est l'une des contributions de ce manuscrit et fera l'objet d'une longue discussion dans le Chapitre 4 et d'un état de l'art.
- **Amélioration de la visualisation.** Les ontologies peuvent également être utilisées comme guide sémantique lors de la visualisation des résultats d'un SRI. Nous verrons, par exemple qu'il est possible d'utiliser, pour cette visualisation, des cartes sémantiques[14]. Pour faciliter la compréhension de l'utilisateur, ces cartes doivent être aussi explicites que possible. Par exemple, il est possible de tirer parti de l'espace conceptuel fourni par les ontologies, pour proposer une disposition des résultats telle que la distance physique entre eux soit proportionnelle à leur distance sémantique dans l'espace conceptuel. Ces cartes peuvent être interactives, afin de permettre à l'utilisateur de sélectionner certains éléments, de fournir un retour (*feedback*) au système, de modifier certains paramètres (les poids des concepts de la requête, par exemple) et d'en observer dynamiquement les effets ou encore pour changer de vue sur les éléments du corpus qui lui sont présentés. Nous exploitons cette possibilité dans les solutions logicielles proposées dans le Chapitre 3 et le Chapitre 4.
- **Assistance pour la navigation dans une collection de documents.** Du fait qu'elle hiérarchise les concepts d'un domaine, une ontologie peut permettre une navigation plus intelligente et intuitive au sein d'une collection de documents indexés par ces concepts. Cette exploration peut s'effectuer à l'aide de mesures de similarité entre documents (exploitant la hiérarchie) ou en choisissant de privilégier certains types de relations entre ceux-ci (Villerd 2008).

En reprenant l'architecture d'un SRI comme elle a été décrite dans la Figure 1.2 à la page 5, l'usage des ontologies dans le cadre de la RI peut être représenté comme dans la Figure 2.10.

D'après la Définition 2.1, une ontologie renferme potentiellement de nombreuses relations avec des sémantiques différentes. Cet aspect soulève la question de l'adéquation des ressources sémantiques utilisées dans le cadre des différentes tâches de la RI. Les relations taxonomiques, formant la hiérarchie H_C sont privilégiées pour la fonction d'appariement d'un SRI (Vallet et al. 2005; Dinh & Tamine 2011; Dragoni et al. 2012). La raison découle de la nature "*sûre*" (au sens de maîtrisée) des relations taxonomiques. Dans la tâche d'expansion de requête, (Rada et al. 1991; Baziz et al. 2003; Khoo & Na 2007) ont mené des études sur l'impact de la prise en compte de relations sémantiques de différentes natures (*hyponymie, hyperonymie, méronymie, holonymie, synonymie, etc.*). Leurs conclusions, obtenues par rapport à des collections de tests dont CLEF2001 (Cross Language Evaluation Forum) et

[14] Nous appelons carte conceptuelle ou sémantique la représentation à l'écran d'un ensemble de données organisées en fonction de différents critères sémantiques

MEDLINE indexée par le thesaurus *MESH*, montrent que la prise en compte de relations autres que taxonomiques introduit du bruit dans les résultats des SRI conceptuels. Dans le

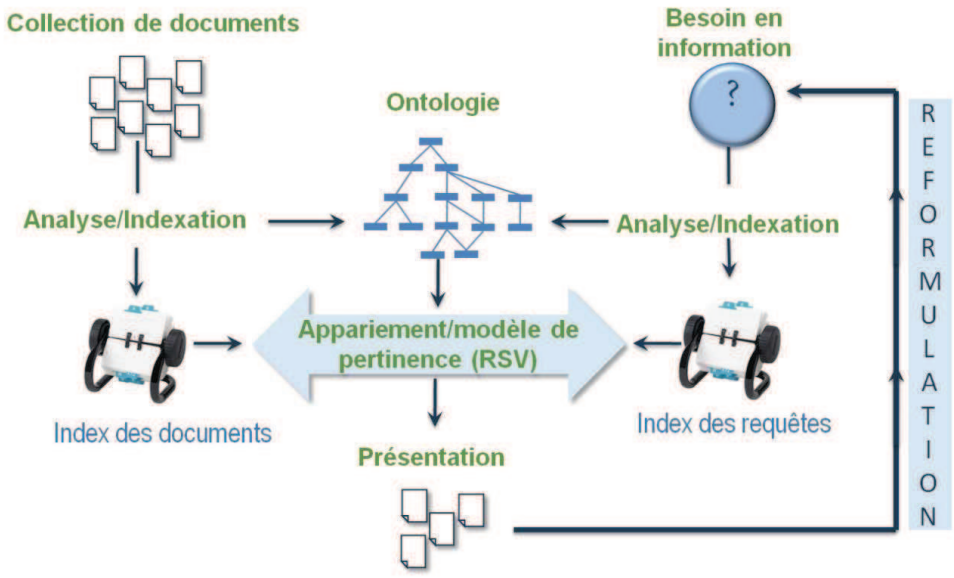


Figure 2.10 : Architecture d'un SRI utilisant une ontologie à travers ses différentes composantes

sillage de ces études, nous adoptons une démarche "*prudente*" (Aussenac-Gilles 2008) concernant la structure d'ontologie à considérer. Nous considérons donc la définition suivante d'une restriction d'une ontologie de domaine aux seules relations de subsomption *is-a*.

Définition 2.2 : (*restriction aux relations is-a d'une ontologie* Θ) La restriction Θ_{DAG} aux seules relations *is-a* d'une ontologie de domaine Θ est définie par $\Theta_{DAG} = (C, H_{isa})$ avec H_{isa} l'ensemble des couples de concepts reliés par une relation de subsomption.

Notons que la Définition 2.2 suppose que tous les concepts de C soient présents dans la hiérarchie H_C. Cela assure que tous soient également présents dans Θ_{DAG}.

La restriction aux relations de subsomption d'une ontologie peut être représentée par un graphe acyclique direct (*dag* pour *directed acyclic graph*) dont les nœuds sont les concepts de C et les arcs sont les liens *is-a*. Dans la littérature relative aux graphes, les arcs sont orientés des nœuds parents aux nœuds enfants tandis que dans la littérature relative aux ontologies c'est le contraire qui est la règle pour respecter la sémantique du lien *is-a*. Dans ce manuscrit, nous nous conformons à l'orientation des arcs utilisée dans le domaine des ontologies, i.e. des *enfants* vers les *parents*. La Figure 2.8 représente une vue d'une partie de la restriction aux relations de subsomption du thésaurus *MeSH*.

Il ressort de cette partie que les ontologies peuvent constituer des modèles de connaissances valables pour les approches de RI conceptuelle bien que leur utilisation doive se faire de manière "prudente" dépendant des tâches de RI visées. Le succès de cette utilisation n'est donc pas immédiat et dépend de plusieurs facteurs que nous nous proposons d'explorer, dans la section suivante, à travers un état de l'art sur les stratégies d'indexation conceptuelle et leur exploitation dans les modèles de pertinence des SRI conceptuelle.

2.4. Recherche d'Information conceptuelle

Mettre en œuvre une ontologie comme modèle de connaissances dans le cadre de la RI nécessite d'abord d'adopter une représentation conceptuelle des documents de la collection. Une évolution du processus d'indexation des SRI est donc nécessaire pour prendre en compte les éléments d'une ontologie comme indiqué dans la Figure 2.10. Cette évolution concerne le type d'unités d'indexation, leur structuration ainsi que leur pondération. Cette section propose un tour d'horizon de ces problématiques.

2.4.1. Indexation conceptuelle : structure d'indexation et pondération

Déterminer les concepts d'une ontologie à même de décrire de manière précise le contenu d'un document dans son ensemble n'est pas une tâche triviale. Cette indexation conceptuelle peut se faire de manière manuelle, par des experts notamment, ou de manière automatique grâce à des outils dits d'extraction de concepts à partir de documents.

Le type de documents à indexer influe sur la nature des outils à considérer. Les gènes, dans les *annotations GO*[15] par exemple, sont généralement indexés par des concepts de la *Gene Ontology* de manière manuelle par des biologistes sur la base de leur expertise. Lorsque les documents considérés sont de type textuel, comme c'est le cas historiquement dans les SRI, de nombreux outils automatique ou semi-automatique existent pour l'extraction de concepts à partir de textes. Dans le domaine biomédical, nous pouvons citer l'outil *MetaMap*[16] qui exploite le méta thésaurus *UMLS* comme source de données pour identifier des concepts dans un article scientifique. Dans cette partie, nous allons nous limiter aux documents de type textuel. Notre contribution ne se situe pas dans la tâche d'extraction de concepts à partir de documents. Nous allons présenter de manière succincte les notions indispensables à la compréhension d'un modèle d'indexation.

Selon (Dinh & Tamine 2012) et (Bannour & Zargayouna 2012), le schéma classique d'une tâche d'indexation d'un document textuel est constitué : i) d'une étape de lemmatisation (filtrage des mots ou groupes de mots non significatifs dont les articles) à l'aide de l'outil *Tree Tagger* (Schmid 1994) par exemple, ii) d'une étape d'ancrage des termes, obtenus après lemmatisation, dans les concepts et relations de l'ontologie en résolvant les problèmes d'ambigüité qui sont récurrents comme le souligne (Alexopoulou et al. 2009), iii) et d'une étape de pondération et de structuration des concepts extraits. Analysons cette dernière étape dans un premier temps.

Différentes structurations d'index ont été proposées. L'une des plus simples est le "*sac de concepts*" (Dinh et al. 2012) où chaque concept est pris indépendamment des autres. Une extension du modèle "*sac de concepts*" est le modèle vectoriel conceptuel où l'espace des concepts issus d'une ontologie est considéré pour représenter aussi bien les documents que les requêtes. Ce modèle correspond simplement au modèle vectoriel classique avec une différence dans les stratégies de pondération permettant de calculer les vecteurs poids des

[15] http://www.geneontology.org/GO.annotation.shtml
[16] http://metamap.nlm.nih.gov/

documents et des requêtes. Un tel modèle est proposé dans (Hliaoutakis et al. 2006) et dans (Ventresque et al. 2008). Pour relâcher l'hypothèse d'indépendance des dimensions d'un vecteur, intrinsèque aux modèles vectoriel et "*sacs de concepts*", un facteur de renforcement entre concepts proches peut être mis en œuvre (Hliaoutakis et al. 2006). Pour deux concepts proches au sens de l'ontologie (par exemple le concept a et son hyponyme d dans la Figure 2.11), leur importance dans un vecteur document ou requête se renforce alors mutuellement. Nous verrons dans la suite comment de telles importances sont évaluées.

Contrairement au modèle vectoriel conceptuel, le modèle "*sacs de concepts"* ne prend en compte que les concepts effectivement extraits des documents et des requêtes. En effet, dans le cadre d'une ontologie de grande taille telle que la *Gene Ontology*, prendre chaque concept comme une dimension conduit à des vecteurs de documents et de requêtes très creuses et de très grandes tailles pouvant poser des problèmes combinatoires. Pour diminuer le nombre de dimensions des vecteurs de concepts représentant les documents et les requêtes, (Dragoni et al. 2012) pose la question du choix d'une base de représentation adéquate en terme de taille dans l'espace des concepts de l'ontologie. Il propose d'adapter les travaux de (Pereira & Tettamanzi 2006) en considérant l'ensemble des concepts n'ayant pas d'hyponymes, i.e. les feuilles, comme une base dans laquelle les documents et les requêtes sont exprimés. En considérant le *dag* représenté dans la Figure 2.11, leur stratégie consiste à effectuer une coupe sur les feuilles. Il s'agit donc d'une opération de projection qui induit forcément une perte d'information et qui peut être coûteuse lorsque l'ontologie contient beaucoup de concepts feuilles.

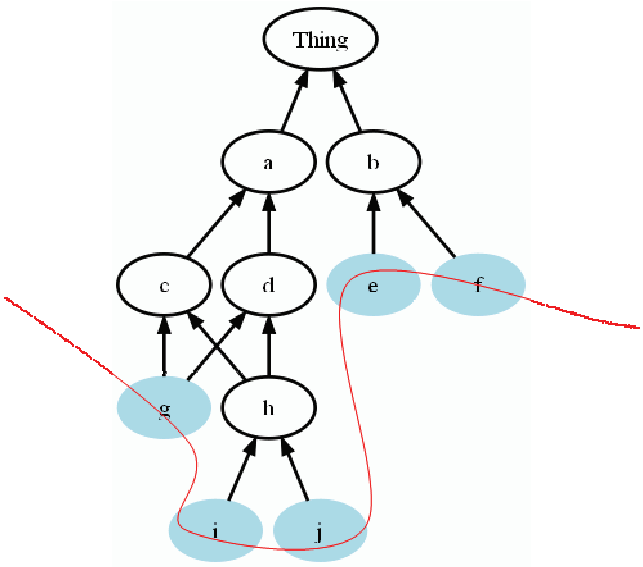


Figure 2.11 : Coupe sur les feuilles d'une structure de *dag* (graphe acyclique direct) représentant la restriction aux relations *is-a* d'une ontologie de domaine. Le fil traverse les feuilles constituant les concepts formant la base de l'espace vectoriel conceptuel.

Dans (Baziz et al. 2005), une structure de noyau sémantique, qui est en fait un réseau sémantique, est utilisée pour représenter un document et une requête. Leur appariement est effectué sur la base d'un degré d'implication entre deux noyaux sémantiques. Cette structuration d'indexation relâche l'hypothèse d'indépendance sur laquelle se fondent les approches vectorielles conceptuelles en considérant les relations qui lient les concepts.

L'avantage d'une telle représentation est d'être plus ou moins exhaustive dans la description d'un document en considérant aussi bien des concepts que les relations qui les lient. Cependant, (Dragoni et al. 2012) souligne que ce type de structure d'indexation est coûteuse en terme de temps de calcul lors de l'évaluation de la pertinence entre un document et une requête.

Nous mettons en œuvre, dans nos contributions, le modèle "*sac de concepts*". Ce modèle est simple et nous permet de mettre l'accent sur les seuls concepts d'intérêt de l'utilisateur dans sa requête et sur les concepts extraits des documents. Toujours est-il que l'estimation de la représentativité d'un concept est l'élément le plus déterminant. C'est l'objet des modèles de pondération.

Une fois un ensemble de concepts extraits d'un document et une structure d'indexation choisie, il est important de considérer les importances relatives des concepts extraits. Différentes stratégies de pondération de concepts ont donc été mises en œuvre. Elles exploitent principalement deux types d'information : i) la statistique d'occurrence des concepts dans la collection ; ii) et l'ontologie d'où sont issus les concepts.

Les stratégies de pondération exploitant la statistique d'occurrence d'un concept dans une collection constituent des adaptations des méthodes de pondération des modèles classiques telles que $tf_{t_r}^{d_j} * idf_{t_r}$ (Baziz et al. 2005; Dragoni et al. 2012) ou du modèle *BM25* (Dinh et al. 2012) déjà présentés dans la section 2.2.1. L'adaptation dont il s'agit ici repose, principalement, sur le fait que la fréquence d'occurrence dans un document d'un concept c_1 d'une ontologie Θ_{DAG} dépend de la présence de ses instances (présence lexicale) et de celles de ses hyponymes $c_2 \in Desc(c_1)$. Nous n'allons pas présenter les équations relatives à ces adaptations vues qu'elles ne sont pas très différentes des méthodes de pondération des approches classiques de RI.

Les approches statistiques de pondération telles que présentées dans la section 2.2.1 ainsi que leur adaptation dans les approches conceptuelles ont, toutes, une composante locale, relative au document, et une autre globale relative à l'ensemble de la collection. C'est le schéma $tf_{t_r}^{d_j} * idf_{t_r}$ classique. Lorsque les documents considérés par un SRI ont des tailles relativement petites et homogènes, comme c'est le cas avec des résumés d'articles scientifiques, la composante locale $\left(tf_{t_r}^{d_j}\right.$ ou sa version normalisée $\left.Ntf_{t_r}^{d_j}\right)$ devient moins déterminante. En effet, elle correspond le plus souvent à un facteur de normalisation des longueurs pour ne pas pénaliser les documents courts par rapport aux documents longs. C'est le cas notamment de *BM25* où, pour rappel, la composante de pondération locale est donnée par $Ntf_{t_r}^{d_j} = \frac{tf_{t_r}^{d_j}}{tf_{t_r}^{d_j} + k\left((1-b) + b * \frac{dl_j}{dlAv}\right)}$. Nous pouvons dire aussi la même chose lorsqu'il s'agit de documents de type gènes où, clairement, il est impossible de déterminer une pondération locale. Nos applications concernant principalement des gènes, ou des documents relativement courts et homogènes en taille (résumés d'articles scientifiques par exemple), nous ne mettons pas en œuvre de stratégies de

pondération locale. Cependant, il est important de disposer d'une stratégie de pondération globale afin de déterminer les concepts les plus informatifs dans une collection.

Concernant l'exploitation de l'ontologie dans une stratégie de pondération de concepts, (Bannour et al. 2012) considère que l'exploitation des liens sémantiques qu'elle renferme est nécessaire. Il considère, à l'instar de (Zargayouna & Salotti 2004), que les informations fournies par ces liens sont captées par des mesures de similarités ou de proximités sémantiques. Dans la section suivante, nous présentons les principales métriques mises en œuvre dans l'état de l'art pour calculer des similarités ou des proximités entre concepts. Nous montrons comment de telles similarités ou proximités peuvent permettre d'évaluer l'informativité d'un concept dans une collection en introduisant la notion de contenu informationnel.

2.4.2. Mesures de similarité sémantique pour l'évaluation du contenu informationnel de concepts

Un des apports majeurs des ontologies en Recherche d'Information, au-delà de la réduction de l'ambigüité des indexations, réside dans la possibilité d'exploiter l'espace conceptuel fourni afin de mesurer la proximité ou la similarité entre deux concepts et par extension entre un document et une requête. En ce sens, les mesures de similarité ou de proximité sémantiques constituent donc le point central des modèles de pertinences dans les SRI conceptuels.
Déterminer la similarité sémantique entre deux termes, concepts ou entités procède d'une longue tradition en psychologie et dans les sciences cognitives, domaines dans lesquels différents modèles ont été proposés (Pirro & Euzenat 2010). Par intuition, un humain trouverait que le concept « *berline* » est plus proche de « *monospace* » que du concept « *avion* » mais que « *berline* » est plus proche de « *avion* » que de « *livre* » (Gandon 2008). Les mesures de similarité et de proximité sémantiques se proposent d'exploiter l'espace conceptuel des ontologies, notamment la structure de graphe ou de réseau sémantique qu'ils induisent, pour simuler une telle intuition. Les premiers travaux concernant l'appariement conceptuel sont issus des travaux de (Quillian 1968; Collins & Loftus 1975) concernant la simulation de la définition et de la remémoration d'un concept chez un humain. Dans leurs travaux, ces deux facultés humaines sont vues comme une activation propagée à travers un réseau de liens étiquetés reliant des concepts. Parmi les types de liens qu'ils utilisent, le lien de subsomption (*is-a*) joue déjà un rôle prépondérant. Les notions de similarité et de proximité sémantiques sont souvent confondues à travers la littérature alors qu'elles sont différentes. En effet, une similarité sémantique ne considère que la relation de subsomption (*is-a*) tandis que la proximité conceptuelle prend en compte une panoplie de relations et peut revêtir plusieurs formes dont, par exemple, la complémentarité fonctionnelle (entre *marteau* et *clou*) ou la similarité fonctionnelle (entre *marteau* et *tournevis*) (Pedersen et al. 2007; Pirro et al. 2010). Les mesures de similarité sémantique sont donc des cas particuliers des mesures de proximité conceptuelle.
Les exemples précédents montrent que l'appariement entre deux concepts est généralement basé sur les caractéristiques qu'ils partagent. (Tversky 1977) propose un cadre, utilisant les notions de propriétés communes et distinctes entre deux concepts, pour évaluer la similarité sémantique entre deux concepts :

$$\pi_{Tversky}(c_1, c_2) = \alpha * comm(c_1, c_2) - \beta * diff(c_1, c_2) - \gamma * diff(c_2, c_1) \quad \textbf{(2.16)}$$

α, β et γ sont des paramètres permettant de mettre l'accent sur une des composantes de la mesure de similarité. La taxonomie d'une ontologie de domaine constitue une structure naturelle pour un tel raisonnement du fait que les concepts sont regroupés par des liens *is-a* qui sont systématiquement présents et dont le rôle est clairement explicite dans les définitions formelles des ontologies.

Plusieurs mesures de similarité entre concepts d'une ontologie ont été proposées. Elles peuvent être classées en deux grandes catégories : i) celles utilisant la structure de taxonomie d'une ontologie comme espace métrique (mesures de type *intensionnel*) et ii) 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*).

Les mesures de type *intensionnel* utilisent la hiérarchie de concepts pour évaluer la similarité entre concepts. Cette hiérarchie est vue comme un graphe acyclique direct (*directed acyclic graph* ou *dag*) (c.f. Définition 2.2 à la page 35) dont les nœuds correspondent à des concepts et les arcs à la relation de subsomption (*is-a*).

Pour évaluer la similarité entre deux concepts c_1 et c_2 de Θ_{DAG}, (Rada et al. 1989) proposent de considérer le plus court chemin (en termes de nombres d'arcs) reliant c_1 et c_2 dans Θ_{DAG}. Si tous les arcs de ce chemin vont dans le même sens, l'un des concepts subsume l'autre. A l'opposé, lorsque ce chemin comporte plusieurs changements d'orientation la relation entre les concepts devient ténue. (Hirst & St Onge 1998) propose donc une généralisation, $\pi_{HO}(c_1, c_2)$, de cette similarité 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 $nbC(P)$:

$$\pi_{HO}(c_1, c_2) = \min_{P=(c_1 \to c_2)} (lg(P) + K * nbC(P)) \quad \text{(2.17)}$$

Le facteur K permet d'ajuster l'impact du nombre de changements de direction. Dans le cas où $K = 0$, on retrouve la similarité de (Rada et al. 1989). A 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 $gcd(c_1, c_2)$ (*great common descendant*). Le *lca* joue donc un rôle important dans plusieurs mesures de similarité. Dès 1994, (Wu & Palmer 1994) avaient d'ailleurs proposé de l'utiliser dans ce cadre. Cependant en se focalisant sur le lca, leur mesure négligeait la notion symétrique de gcd et le fait que des concepts possèdent ou ne possèdent pas des descendants communs. Une limite importante des mesures basées sur les longueurs de chemin dans le graphe *is-a* est due au fait que ses arcs ne représentent pas tous des degrés de généralisation et de spécialisation équivalents. La distance sémantique π_{RW} proposée dans (Ranwez et al. 2006) prend en compte cette information en s'appuyant sur le nombre de descendants de chaque concept. Formellement, soient $S_C \subseteq C$ un ensemble de concepts de Θ_{DAG}, $Desc(S_C)$ l'ensemble des concepts c_r descendants d'au moins un concept $c_1 \in S_C$ (i.e. $(c_r, c_1) \in H_{isa}$) et $ancEx(c_1, c_2)$ l'ensemble des ancêtres exclusifs de c_1 et c_2.

$$\pi_{RW}(c_1, c_2) = \frac{}{|Desc(ancEx(c_1, c_2)) \cup Desc(\{c_1\}) \cup Desc(\{c_2\}) - Desc(\{c_1\}) \cap Desc(\{c_2\})|} \quad \textbf{(2.18)}$$

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) et adaptée de la théorie de l'information de Shannon. Le contenu informationnel d'un concept peut être interprété comme la quantité d'information qu'il exprime. Dans la plupart des cas, le contenu informationnel d'un concept c_1 est donné par la probabilité $p(c_1)$ d'avoir une occurrence de ce concept (ou de ses descendants) utilisée dans une collection. Du moment que l'occurrence d'un concept c_1 est comptée comme une occurrence de ses parents, une telle probabilité p est cumulative, croissante et monotone lorsque que l'on se déplace des feuilles vers les racines de Θ_{DAG}. Donc si $c_2 \in Desc(c_1)$ et si $p(c_i)$ est la probabilité d'occurrence du concept c_i dans la collection D, alors et $p(c_2) \leq p(c_1)$.

Définition 2.3 : (*contenu informationnel d'un concept*) Le contenu informationnel IC d'un concept c_1 de Θ_{DAG} ayant $p(c_1)$ pour probabilité d'occurrence dans une collection de documents D, est défini par :

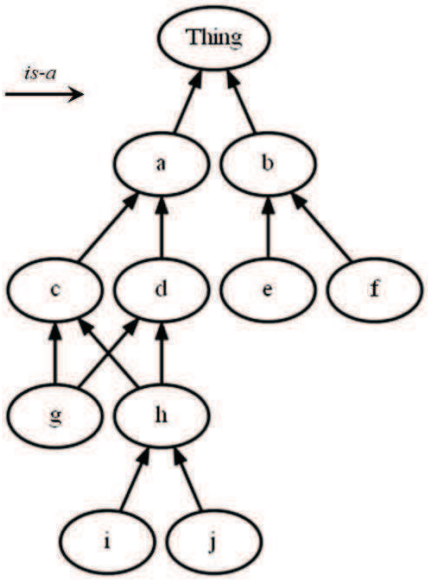
$$IC : C \to \mathbb{R}^+ \qquad c_x \mapsto -\log(p(c_x)) \quad , 0 < p(c_x) \leq 1$$

Cette définition implique que plus la probabilité $p(c_x)$ est grande, plus c_x est fréquent dans la collection et moins il est informatif. Le contenu informationnel croît donc en sens inverse de la probabilité d'occurrence. Si Θ_{DAG} a une unique racine c_{root}, alors $IC(c_{root}) = 0$.
Estimer le contenu informationnel en se basant sur la probabilité d'occurrence d'un concept nécessite de traiter une collection de documents de grande taille ce qui est gourmand en temps de calcul (Pirró 2009). D'un autre côté, une estimation correcte du contenu informationnel exige que la collection de documents considérée se rapporte au même domaine que l'ontologie. Ces problèmes ont conduit à considérer des formulations alternatives du contenu informationnel d'un concept, basées sur une ontologie. (Seco et al. 2004) proposent une mesure normalisée du contenu informationnel d'un concept c_x, qualifiée d'intrinsèque et basée sur Θ_{DAG} :

$$IC(c_x) = \frac{log\left(\frac{|Desc(c_x)|}{|C|}\right)}{log\left(\frac{1}{|C|}\right)} = 1 - \frac{log(|Desc(c_x)|)}{log(|C|)} \quad \textbf{(2.19)}$$

La quantité $|C|$ est supposée ne jamais être nulle, une ontologie étant supposée renfermer au moins un concept. La quantité $log\left(\frac{1}{|C|}\right)$ permet de normaliser la fonction IC de sorte que ses valeurs soient comprises dans $[0,1]$.

L'hypothèse de base de la formulation du contenu informationnel d'un concept, donnée par l'équation (2.19), réside dans le fait que plus un concept a d'hyponymes (de descendants) plus sa probabilité d'occurrence dans une collection est élevée et son informativité est grande (conformément à la Définition 2.3). Cette dernière est donc inversement proportionnelle au nombre d'hyponymes d'un concept. L'ontologie est considérée, suivant cette formulation, comme étant organisée de telle sorte qu'un concept n'est créé que lorsqu'il y a un besoin de le différencier d'avec l'existant (*cognitive saliency*) (Pirró 2009). Notons que si un concept c_x est une feuille dans Θ_{DAG} (i.e. $|Desc(c_x)| = \{c_x\}$), alors il a une informativité maximale ($IC(c_x) = 1$).



Concepts	Valeurs d'IC (Seco et al., 2004)
Thing	0
a	0,188
b	0,541
c	0,328
d	0,328
e	1
f	1
g	1
h	0,541
i	1
j	1

Figure 2.12 : Les valeurs de contenu informationnel de concepts d'une hypothétique taxonomie.

Combinant la notion de plus petit ancêtre commun (lca) à celle de contenu informationnel d'un concept, (Resnik 1999) introduit la notion d'ancêtre commun le plus informatif ($MICA(c_1, c_2)$) entre deux concepts c_1 et c_2.

Définition 2.4 : (*ancêtres communs les plus informatifs*) Etant donnés deux concepts $(c_1, c_2) \in C^2$, leurs ancêtres communs les plus informatifs dans Θ_{DAG} sont tels que :

$$MICA(c_1, c_2) = \left\{ \underset{c_j \in Anc(c_1) \cap Anc(c_2)}{\text{argmax}} \left(IC(c_j) \right) \right\}$$

Avec $Anc(c_x)$ les ancêtres du concept c_x. La Figure 2.13 montre un exemple de détermination du $MICA$ de deux concepts sur une structure de *dag* (graphe orienté acyclique) d'une ontologie.

Utilisant le $MICA$ de deux concepts, (Resnik 1999) propose alors la mesure de similarité sémantique suivante :

$$\pi_{resnik}(c_1, c_2) = IC\big(MICA(c_1, c_2)\big)$$

La mesure de Resnik ne satisfait pas certaines propriétés pourtant intuitives dont celle de coïncidence qui stipule que la similarité d'un concept avec lui-même est maximale. ($\pi_{resnik}(c_x, c_x) = IC(c_x)$ qui n'est pas forcement maximal).

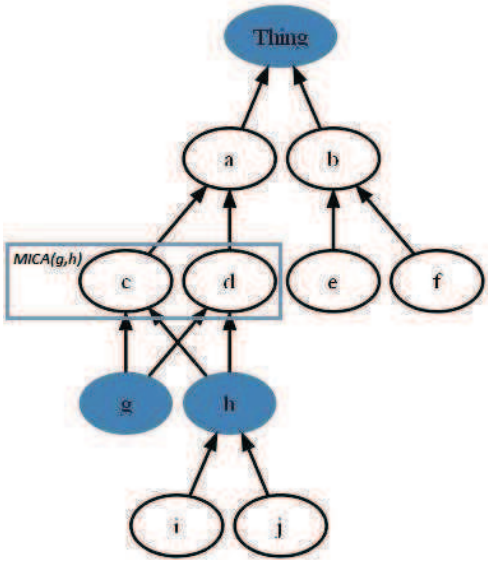


Figure 2.13 : Les ancêtres communs (c, d) les plus informatifs de deux concepts (g, h) dans un exemple hypothétique d'une structure de dag d'une ontologie de domaine

Pour contourner ces limitations, d'autres mesures de similarité sémantique basées sur le contenu informationnel et satisfaisant aux propriétés suivantes ont été proposées :

Définition 2.5 : (*mesures de similarité sémantique*) Une mesure de similarité sémantique $\pi : C \times C \rightarrow [0,1]$ est une fonction qui associe, à chaque couple de concepts $(c_x, c_y) \in C^2$ de Θ_{DAG}, une valeur réelle évaluant leur degré d'appariement et qui a les propriétés minimales suivantes :

- $\pi(c_x, c_y) \geq 0$ (positivité)
- $\pi(c_x, c_y) = \pi(c_x, c_y)$ (symétrie)
- $\pi(c_x, c_x) = 1$ (coincidence)

Cette définition appelle à quelques précisions. En effet, elle ne couvre pas toutes les mesures de similarité sémantique existantes (la propriété de symétrie n'est pas toujours respectée par exemple), mais elle met en exergue la famille de mesures qui nous intéresse dans cette thèse. Cet intérêt est justifié par les propriétés de connexité des mesures de similarités sémantiques qui découlent de la Définition 2.5 et qui sont utiles à la mise en œuvre d'une stratégie de reformulation de requêtes conceptuelles (c.f. Chapitre 4). Cette définition peut aussi bien correspondre aux mesures basées sur le contenu informationnel qu'aux mesures intensionnelles utilisant le graphe des liens de subsomption de l'ontologie.
Introduisons deux mesures de similarité sémantique largement utilisées, basées sur le contenu informationnel d'un concept et sur lesquelles se fondent en partie nos évaluations (c.f. section 4.3). Premièrement, il s'agit de la mesure de Lin (Lin 1998) définie de la manière suivante :

$$\pi_{Lin}(c_x, c_y) = \frac{2 * IC\left(MICA(c_x, c_y)\right)}{IC(c_x) + IC(c_y)} \quad \textbf{(2.20)}$$

Notons que cette mesure respecte la Définition 2.5 de même que la mesure de Jiang et al. (Jiang & Conrath 1997) qui est définie comme suit :

$$\pi_{j\&c}(c_x, c_y) = 1 - \frac{IC_x + IC_y - 2IC\left(MICA(c_x, c_y)\right)}{2} \quad \textbf{(2.21)}$$

Comme nous venons de le voir, plusieurs mesures de similarité sémantique entre concepts d'une ontologie ont été proposées dans la littérature. 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, iv) et la réactivité du système. Dans la section 3.3, nous présentons plus en détail les mesures de similarité sémantique mises en œuvre dans *OBIRS*.

2.5. Inclure l'utilisateur dans la boucle de pertinence

En tant que principal juge quant à la pertinence des documents par rapport à son besoin en information, l'utilisateur doit être intégré dans la boucle de pertinence. Il est au départ du scénario de base de la RI, lorsqu'il exprime son besoin en information, au cours du processus de recherche à travers ses préférences et à la fin de celui-ci, lorsqu'il juge les résultats qui lui sont fournis. La suite de cette section explore l'intégration qui est faite de l'utilisateur au cours du processus de RI.

2.5.1. Vers une prise en compte des préférences utilisateur : les modèles d'agrégation

Dans cette section, nous introduisons la notion d'opérateur d'agrégation ainsi que les modèles d'agrégation à l'œuvre dans les principaux modèles de recherche d'information. Une discussion est menée sur la manière dont les différentes familles d'opérateurs d'agrégation prennent en charge différentes stratégies de décision d'un utilisateur. Dans le cadre de la RI conceptuelle, nous menons une discussion sur les critères pertinents à prendre en compte dans un processus d'agrégation.

2.5.1.1. Stratégies d'agrégation à travers les modèles de Recherche d'Information

La recherche d'information peut être considérée comme un problème de décision multicritère dès lors que l'on considère la nature multidimensionnelle de la pertinence (Mizzaro 1997, 1998; Borlund 2003). Deux étapes se présentent généralement pour évaluer la pertinence d'un document par rapport à une requête : i) la définition des sources d'évidences à partir

desquelles la pertinence peut être évaluée, ii) la mise en œuvre de stratégies pour combiner de telles sources d'évidences.

Suivant cette vision de la pertinence, un modèle de recherche d'information peut être formellement défini comme un quintuplé comme indiqué dans (Farah et al. 2006) :

- $D = \{d_j,\ j = 1..|D|\}$: un ensemble de documents d_j (alternatives) à évaluer et à ordonner par rapport à une requête Q ;

- $F = \{x_i, i = 1..k\}$: un ensemble d'attributs ou de sources d'évidences que nous appellerons critères et au regard desquels la pertinence d'un document est évaluée. De tels critères doivent être ceux qui caractérisent la pertinence d'un document et doivent être choisis avec soin ;

- X_i : une fonction évaluant la performance $X_i(d_j, t_r)$ de chaque document d_j de D par rapport à un critère x_i de F et relativement à un terme t_r de la requête. Il s'agit généralement de fonctions de pondération qui mesurent le pouvoir discriminant d'un terme. Les scores $X_i(d_j, t_r)$ obtenus sont considérés comme des degrés de pertinence élémentaires ;

- $\mathcal{E}$: un ensemble de vecteurs de performances (ou profils) des différents documents de D. Le vecteur colonne (c.f. Tableau 2.3) de performances d'un document d_j pour un terme t_r d'une requête Q a pour composantes les $X_i(d_j, t_r)$, $i = 1..k$. Le vecteur ligne (c.f. Tableau 2.3) de composantes $X_i(d_j, t_r)$, $r = 1..n$ rassemble les performances élémentaires de d_j pour le critère i vis-à-vis de chacun des termes de la requête. La valeur agrégée $w_{j,t_r} = agreg_{i=1..k}\, X_i(d_j, t_r)$ mesure la performance de d_j pour le terme t_r sur l'ensemble des critères $i = 1..k$. La valeur agrégée $X_i(d_j, Q) = agreg'_{r=1..n}\, X_i(d_j, t_r)$ mesure la pertinence de d_j pour le critère x_i sur l'ensemble des termes de la requête Q. Il y a donc en jeu deux agrégations possibles : l'une selon les différents critères de pertinence retenus et l'autre selon les termes de la requête.

- A : est une fonction de rangement permettant de capter la relation d'ordre induite par les degrés de pertinence élémentaires. Cette fonction de rangement se base généralement sur le calcul du *RSV* à travers une agrégation des valeurs w_{j,t_r} ou $X_i(d_j, Q)$.

Concernant les opérateurs d'agrégation que nous présentons ici, la différenciation entre termes et concepts n'entre pas en considération. Le cadre d'agrégation présenté reste valable suivant que l'on se rapporte aux termes ou aux concepts. Nous désignons donc par termes les éléments d'une requête et nous préciserons à chaque fois qu'il est nécessaire de se situer exclusivement dans un cadre conceptuel.

Document d_j	Requête Q				
		t_1	….	t_n	Agrégation ligne sur les termes de la requête
Critères	Critère 1 : x_1	$X_1(d_j, t_1)$	….	$X_1(d_j, t_n)$	w_{j,x_1}
	. . .	. . .	. . .	. . .	. . .
	Critère k : x_k	$X_k(d_j, t_1)$	….	$X_k(d_j, t_n)$	w_{j,x_k}
Agrégation colonne (sur les critères de pertinence)		w_{j,t_1}	….	w_{j,t_n}	$RSV(Q, d_j)$

Tableau 2.3 : Stratégies d'agrégation en Recherche d'Information adapté de (Farah & Vanderpooten 2007)

Deux facteurs sont connus pour affecter considérablement la pertinence d'un document par rapport à une requête et donc l'efficacité des SRI. D'après (Farah et al. 2006; Pereira et al. 2009; Pereira et al. 2012), le premier facteur concerne l'identification de sources d'évidences pouvant contribuer à évaluer la pertinence d'un document et recouvrant éventuellement différents aspects d'une telle pertinence. Ce premier facteur aboutit à la définition de critères qui permettent, chacun, d'obtenir une relation de préférence partielle sur les documents comparés. Des priorités peuvent être introduites entre les critères (Pereira et al. 2012), de même qu'ils peuvent être regroupés en catégories (Pereira et al. 2009). Globalement, deux types de critères sont considérés en RI : i) les critères dépendant d'une requête Q de l'utilisateur, établis par exemple à partir d'analyses statistiques (fréquences des termes de la requête, etc.) sur le contenu des documents (modèles de description textuelle) et de l'analyse de la structure d'indexation du corpus (les modèles de description de structure basés sur les liens hypertexte entre documents html par exemple) ; ii) les critères indépendants d'une requête Q de l'utilisateur : il s'agit de critères relatifs aux caractéristiques autres que structurelles des documents qui sont utilisées pour mieux discriminer ceux-ci (ex : origine du document, auteur, etc.). Les auteurs de (Pereira et al. 2009) considèrent, par exemple, un critère de confiance qu'ils appellent "*reliability*".

Le second facteur, affectant l'évaluation de la pertinence d'un document, est la combinaison de ses degrés de pertinence élémentaires obtenus en l'évaluant par rapport aux critères définis (dans F). Il s'agit donc de définir les fonctions X_i et A. La majeure partie des approches de RI mettent en œuvre des opérateurs d'agrégation analytiques tels que la somme pondérée, pour effectuer une telle combinaison aboutissant au *RSV*. (Clinchant & Gaussier 2010) note que les modèles *DFR* (Amati et al. 2002), *BM25* (S. Robertson et al. 1994), ainsi que vectoriel (Salton et al. 1975) sont de la forme suivante :

$$RSV(Q, d_j) = \sum_{t_r \in Q \cap d_j} \left(a\left(tf_{t_r}^{Q}\right) * h\left(tf_{t_r}^{d_j}, dl_j, stat_{t_r}, \theta\right) \right) \qquad \textbf{(2.22)}$$

Avec θ étant un ensemble de paramètres additionnels dépendant du modèle considéré. Pour le modèle *BM25* par exemple, $\theta = \{dlAv, b, k, |D|\}$. Le paramètre $stat_{t_r}$ peut correspondre à la fréquence du terme t_r dans toute la collection (df_{t_r}) ou dans les documents où il apparait (FD_{t_r}). La fonction a est généralement la fonction identité, i.e. $a(tf_{t_r}^Q) = tf_{t_r}^Q$. La fonction h est une fonction de pondération qui est considérée de classe C^2 (deux fois dérivables) et sa signature est de la forme :

$$h : \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times D_\theta \to \mathbb{R}$$

Avec D_θ le domaine des paramètres dans θ.

Les paramètres de la fonction h correspondent d'une certaine manière à des critères sur lesquels se fonde l'évaluation de la pertinence. Notons que la contrainte $(t_r \in Q \cap d_j)$ de présence effective du terme t_r dans le document d_j peut être lissée lorsque t_r est un concept. Dans ce cas, la présence du concept t_r peut être obtenue soit à travers ses instances (présences lexicales) soit à travers celles des concepts avec lesquels il est lié par une relation de subsomption (c.f. 2.4.1).

Dans (Farah et al. 2006), un problème majeur concernant la procédure d'agrégation en RI est soulevé. En effet, l'agrégation nécessite la commensurabilité des degrés de pertinence élémentaires qui est généralement supposée sans précaution supplémentaire. Dans le modèle basique de l'équation (2.22) ci-dessus, l'agrégation ne pose pas de problème particulier parce que les degrés de pertinence élémentaires associés à chaque terme de la requête ont la même sémantique et sont nécessairement commensurables. Le problème est plus épineux lorsque, par exemple, les degrés de pertinence considérés concernent des fréquences de termes, des nombres de citations de documents ou des critères d'adéquation (*Appropriateness*) et de couverture (*Coverage*) tels que utilisés dans (Pereira et al. 2009). Une solution possible est alors d'inverser l'ordre d'agrégation : c'est-à-dire commencer à agréger les scores élémentaires de tous les termes t_r d'une requête relativement à un critère x_i fixé (w_{j,x_i}). Les données agrégées sont donc homogènes. Puis nous pouvons raisonner sur les scores w_{j,x_i} de tous les critères avec des techniques d'analyse multicritères moins contraignantes que la théorie de l'utilité multi-attributs *MAUT*, e.g. ELECTRE (Roy 1991). Le Tableau 2.3 donne un aperçu des deux stratégies d'agrégation.

Dans le cadre de la RI conceptuelle, il est possible de considérer un critère de pertinence x_1 lié à une proximité ou à une similarité sémantique entre concepts issus d'une ontologie de domaine. Dans un tel cas, on cherche à évaluer le vecteur ligne de composantes $X_1(d_j, t_r)$, $r = 1..n$ ($k = 1$ puisqu'il n'y a qu'un seul critère considéré) qui rassemble les performances élémentaires de d_j pour le critère x_1 vis-à-vis de chacun des termes t_r d'une requête. Ensuite, on a seulement besoin d'un modèle d'agrégation pour combiner les scores de pertinence élémentaires $X_1(d_j, t_r)$ sur $r = 1..n$. La question de la commensurabilité n'affecte pas cette approche du fait que les degrés de pertinences élémentaires sont calculés sur la base d'une proximité ou d'une similarité sémantique unique définie sur une ontologie unique. Dans la

suite, nous allons définir formellement les opérateurs d'agrégation et en présenter les principaux sans être exhaustifs. Nous allons cependant nous restreindre au cadre de la RI conceptuelle en considérant un seul critère comme indiqué précédemment. Donc, par souci de cohérence, nous noterons le score élémentaire $X_1(d_j, t_r)$ par $X_r(d_j, t_r)$ pour montrer que l'agrégation s'effectue sur les termes t_r de la requête.

2.5.1.2. Propriétés élémentaires des opérateurs d'agrégation

Un opérateur d'agrégation combine plusieurs valeurs de manière à en obtenir une seule qui représente au mieux l'ensemble des valeurs d'entrées en tenant compte de chacune d'elles. Cette présentation générale permet de percevoir la possibilité d'utilisation d'un tel opérateur dans de nombreux domaines. On trouve, dans la littérature, beaucoup de travaux portant sur l'agrégation de différents types de données allant d'un nombre infini ou fini de valeurs réelles (Grabisch et al. 2000), de valeurs ordinales [17] (Grabisch et al. 2000; Yager 2007), de distributions de probabilité (Nelsen 1998), d'ensembles flous (Dubois & H. Prade 1985), de sources incertaines (Dubois & Prade 2004). Les définitions qui suivent sont adaptées, dans le contexte de la RI, des travaux de (Grabisch et al. 2000) et de (Le Capitaine 2009).

Soit $D = \{d_1, d_2, \dots, d_{|D|}\}$ un ensemble fini d'alternatives parmi lesquelles un choix doit être effectué par un décideur (utilisateur) relativement à un besoin pouvant être exprimé à travers une requête. Un opérateur d'agrégation A est défini comme une fonction telle que :

$$\begin{array}{rl} A : [0,1]^n & \rightarrow [0,1] \\ \left(X_1(d_j, t_1), X_2(d_j, t_2), \dots, X_n(d_j, t_n)\right) & \mapsto \ w_{j,x_1} \end{array} \quad \textbf{(2.23)}$$

Avec $\left(X_1(d_j, t_1), X_2(d_j, t_2), \dots, X_n(d_j, t_n)\right)$ le vecteur des degrés de pertinence élémentaire de l'alternative d_j et w_{j,x_1} la pertinence globale de d_j (c.f. Tableau 2.3). On parle d'approche par critères de synthèse unique. L'opérateur d'agrégation le plus simple est la moyenne arithmétique :

$$A\left(X_1(d_j, t_1), X_2(d_j, t_2), \dots, X_n(d_j, t_n)\right) = \frac{1}{n}\sum_{r=1}^{n} X_i(d_j, t_r) \quad \textbf{(2.24)}$$

Une fois le score global de chaque alternative calculé, l'alternative optimale $d_j{}^*$ peut, par exemple, être sélectionnée comme suit :

$$d_j{}^* = \underset{d_j \in D}{\text{Argmax}}\left(A\left(X_1(d_j, t_1), X_2(d_j, t_2), \dots, X_n(d_j, t_n)\right)\right) \quad \textbf{(2.25)}$$

Etudions maintenant quelques propriétés qu'un opérateur d'agrégation doit satisfaire. Dans un premier temps, il s'agit des conditions aux bornes et de la monotonie. En effet, l'agrégation de

[17] Où seule la relation d'ordre induite par les quantités n'a de sens contrairement aux différences et distances qui ne peuvent être interprétées entre ces quantités

valeurs minimales (0) doit donner une valeur agrégée minimale de même que l'agrégation de valeurs maximales (1) doit aboutir à une valeur maximale :

$$A(0,\dots,0) = 0 \qquad \textbf{(2.26)}$$

$$A(1,\dots,1) = 1 \qquad \textbf{(2.27)}$$

L'équation (2.26 **)** revient à dire que si aucun des critères n'est satisfait alors la sortie globale ne le sera pas tandis que l'équation (2.27) signifie que si les critères sont tous satisfaits alors la sortie le sera aussi. En l'occurrence, le score d'un document d_j doit être nul (respectivement valoir 1) si aucun des termes de son index $C(d_j)$ ne correspond à un terme d'une requête (respectivement si tous les termes de la requête sont présents dans son index).

Définition 2.6 : (*Idempotence*) Un opérateur d'agrégation possède un élément idempotent $a \in [0,1]$ si :

$$\mathrm{A}(a,\dots,a) = \mathrm{a}, \mathrm{a} \in [0,1] \qquad \textbf{(2.28)}$$

L'opérateur devient idempotent si l'équation (2.28 **)** est satisfaite pour tout élément de [0,1]. Au vu des propriétés induites par les équations (2.26) et (2.27), 0 et 1 sont des éléments idempotents.

Définition 2.7 : (*monotonie*) Un opérateur d'agrégation est dit monotone si :

$$\forall r \in 1..n, \qquad X_r(d_j,t_r) \geq X_r(d_{j'},t_r) \Longrightarrow$$
$$A\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) \geq A\left(X_1(d_{j'},t_1),\dots,X_n(d_{j'},t_n)\right) \qquad \textbf{(2.29)}$$

Si une entrée augmente alors la valeur de sortie doit aussi augmenter. Signalons que désirer l'idempotence et la monotonie revient à vouloir un comportement de compensation (Grabisch et al. 1998) (voir Définition 2.12).

Définition 2.8 : (*continuité*) Un opérateur d'agrégation est continu si la fonction d'agrégation est continu au sens usuel du terme.

La continuité contraint l'opérateur à ne pas se comporter de manière chaotique.

Définition 2.9 : (*symétrie, neutralité ou commutativité*) Un opérateur d'agrégation est symétrique si :

$$A\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) = A\left(X_{\sigma(1)}(d_j,t_1),\dots,X_{\sigma(\mathrm{n})}(d_j,t_n)\right) \qquad \textbf{(2.30)}$$

σ étant une permutation dans l'ensemble $\{1,..,n\}$. Les opérateurs maximum et minimum de même que les moyennes sont symétriques contrairement aux moyennes pondérées. La symétrie est généralement requise lorsque l'on combine des critères d'importance égale.

Au-delà des propriétés mathématiques décrites ci-dessus, analysons quelques propriétés moins formelles qu'un opérateur peut prendre en considération pour modéliser la stratégie de prise de décision d'un utilisateur. Par exemple, dans le cas d'une requête multi termes $Q = \{t_1, \dots, t_n\}$, l'utilisateur peut vouloir accorder des importances différentes aux critères à travers des poids. Par ailleurs, la possibilité de pouvoir exprimer plusieurs stratégies de décision est aussi nécessaire ; face à plusieurs critères, l'utilisateur peut avoir une attitude conjonctive, disjonctive, ou de compromis. Il peut également considérer que certains critères sont absolument nécessaires pour sa satisfaction (critères de veto) ou que d'autres sont suffisants pour qu'il soit satisfait (critères d'acceptation). De telles considérations permettent, par exemple, à deux utilisateurs considérant les mêmes critères avec les mêmes importances relatives et le même vecteur de performances $\left(X_1(d_j, t_1), \dots, X_n(d_j, t_n)\right)$ par rapport à un document d_j, d'obtenir un score global différent suivant la stratégie de décision que chacun d'eux aura mis en œuvre. Il est aussi important pour un opérateur d'agrégation de permettre un lien entre les vecteurs de performances $\left(X_1(d_j, t_1), \dots, X_n(d_j, t_n)\right)$ des alternatives d_j et l'interprétation sémantique qu'impliquent les scores globaux obtenus. Par exemple, il peut être intéressant de connaître la contribution de chaque critère x_i au score global. Une telle propriété permet d'éviter que l'opérateur ne soit une "*boite noire*" ne permettant pas d'expliquer la sortie par les entrées.

Les opérateurs d'agrégation peuvent être catégorisés en plusieurs classes suivant le comportement qu'ils autorisent ou pas (Grabisch et al. 1998). Il s'agit principalement des opérateurs de conjonction, de disjonction et de compromis. Nous introduisons, dans la suite, une liste non exhaustive des différents opérateurs existants.

Définition 2.10 : (*opérateur d'agrégation conjonctif*) Un opérateur d'agrégation *A* est dit conjonctif si :

$$A\left(X_1(d_j, t_1), \dots, X_n(d_j, t_n)\right) \leq \min\left(X_1(d_j, t_1), \dots, X_n(d_j, t_n)\right) \quad \textbf{(2.31)}$$

Dans le cas d'une agrégation conjonctive, les arguments (degrés de pertinence partiels) sont combinés suivant une opération correspondant au "*et*" logique. Le score global est élevé si et seulement si tous les degrés de performances sont élevés. Un opérateur conjonctif peut être utilisé en recherche d'information lorsque l'utilisateur souhaite la satisfaction de tous les critères par chaque document. Ce qui signifie que le score global d'un document d_j ne peut excéder le degré de performance élémentaire $X_r(d_j, t_r)$ le plus faible.

Définition 2.11 : (*opérateur d'agrégation disjonctif*) Un opérateur d'agrégation *A* est dit disjonctif si :

$$A\left(X_1(d_j, t_1), \dots, X_n(d_j, t_n)\right) \geq \max\left(X_1(d_j, t_1), \dots, X_n(d_j, t_n)\right) \quad \textbf{(2.32)}$$

Il s'agit de combiner les arguments de l'opérateur en utilisant une opération de "*ou* " logique. Le score global sera faible si et seulement si tous les degrés de pertinence partiels ont des

valeurs faibles. Le degré de pertinence élémentaire le plus élevé limite le score global. Les opérateurs disjonctifs sont donc duaux des opérateurs conjonctifs.

Définition 2.12 : (*opérateur de compromis*) Un opérateur d'agrégation A est dit de compromis si :

$$\min\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) \le A\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) \le \max\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) \quad \textbf{(2.33)}$$

pour tout vecteur de performances $\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right)$ associé à une alternative d_j. Notons que désirer un opérateur d'agrégation monotone (2.29) et idempotent (2.28) revient à vouloir un opérateur de compromis. Ces derniers aboutissent à un score global compris entre le minimum des arguments et leur maximum. Ce ne sont ni des opérateurs conjonctifs ni disjonctifs. La moyenne arithmétique (2.24) est l'exemple le plus connu et utilisé des opérateurs de type compromis. Un degré de performance $X_r(d_j,t_r)$ faible (respectivement élevé) d'un critère peut être compensé par un autre score élevé (respectivement faible). Ces opérateurs sont parfois désignés comme étant des opérateurs de compensation. Parmi de tels opérateurs, nous distinguons aussi les opérateurs de moyennes ordonnées *OWA* (Yager 1988) et les moyennes quasi-arithmétiques dont nous donnons les détails dans la suite.

Considérons E un intervalle réel fini ou infini et A un opérateur d'agrégation symétrique, continu, strictement croissant et idempotent. A est une moyenne quasi-arithmétique si et seulement si il existe une fonction $f : E \to \mathbb{R}$ continue et strictement monotone telle que :

$$A\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) = f^{-1}\left(\frac{1}{n}\sum_{r=1}^{n} f\left(X_r(d_j,t_r)\right)\right) \quad \textbf{(2.34)}$$

Suivant la fonction f choisie, il est possible de couvrir un large spectre de moyennes dont le Tableau 2.4, adapté de (Le Capitaine 2009), donne une synthèse. La majeure partie des moyennes peuvent être obtenues en considérant des valeurs particulières de q dans la moyenne de Hölder comme indiqué dans le Tableau 2.4. Il faut aussi noter que lorsque q tend vers $+\infty$ alors la puissance de Hölder correspond à l'opérateur maximum tandis que lorsqu'il tend vers $-\infty$ la même puissance devient l'opérateur minimum. Quand les critères ne jouent pas un rôle symétrique, leurs importances relatives peuvent être introduites dans la famille des opérateurs d'agrégation de type moyenne. Les importances relatives entre critères sont représentées par des poids $p_r \ge 0$ et respectant la contrainte $\sum_{r=1}^{n} p_r = 1$. L'introduction des poids fait perdre aux opérateurs de type moyenne leur propriété de symétrie (2.30). En tenant compte de la pondération, nous pouvons introduire la moyenne quasi-arithmétique pondérée :

$$A\left(X_1(d_j,t_1),\dots,X_n(d_j,t_n)\right) = f^{-1}\left(\sum_{r=1}^{n} p_r f\left(X_r(d_j,t_r)\right)\right), \sum_{r=1}^{n} p_r = 1 \quad \textbf{(2.35)}$$

Fonction f	Opérateur d'agrégation	Nom
x (fonctionidentité)	$\frac{1}{n}\sum_{r=1}^{n}\left(X_r(d_j,t_r)\right)$	Moyenne arithmétique : ($q=1$ dans l'opérateur de Hölder).
$\log x$	$\sqrt[n]{\prod_{r=1}^{n} X_r(d_j,t_r)}$	Moyenne géométrique : ($q=0$ dans l'opérateur de Hölder).
x^{-1}	$\frac{1}{\frac{1}{n}\sum_{r=1}^{n}\left(\frac{1}{X_r(d_j,t_r)}\right)}$	Moyenne harmonique : ($q=-1$ dans l'opérateur de Hölder).
x^2	$\sqrt{\frac{1}{n}\sum_{r=1}^{n}\left(X_r(d_j,t_r)\right)^2}$	Moyenne quadratique : ($q=2$ dans l'opérateur de Hölder).
x^q	$\left(\frac{1}{n}\sum_{r=1}^{n}\left(X_r(d_j,t_r)\right)^q\right)^{\frac{1}{q}}, q\in\mathbb{R}^*$	Puissance (Hölder) Famille d'opérateurs de Yager
e^{qx}	$\frac{1}{q}\log\left(\frac{1}{n}\sum_{r=1}^{n} e^{qX_r(d_j,t_r)}\right)$	Exponentielle

Tableau 2.4 : Moyennes quasi-arithmétiques usuelles adaptées de (Le Capitaine 2009)

Les opérateurs non compensatoires ne considèrent que la meilleure ou la mauvaise performance dans l'évaluation d'une alternative d_j conduisant à ne considérer qu'un seul terme de la requête. Nous pouvons citer les opérateurs minimum et maximum comme opérateurs non compensatoires généralement mis en œuvre dans le cadre de la recherche d'information.

Les deux types d'agrégation (compensatoires et non compensatoires) présentent tout de même quelques limites. Considérons un exemple basique, comme indiqué dans (Farah et al. 2007), pour illustrer le comportement des opérateurs compensatoires et non compensatoires. Soient deux documents d_1, d_2 dont les vecteurs de pertinence élémentaires sont renseignés dans le Tableau 2.5 et une requête $Q=\{t_1,t_2,t_3,t_4\}$.

	t_1	t_2	t_3	t_4
d_1	0.5	0.5	0.5	0.5
d_2	0.49	0.9	0.9	0.71

Tableau 2.5 : Exemple de vecteurs de performances

En utilisant une moyenne arithmétique ($q=1$ dans la puissance de Hölder), alors $RSV(Q,d_1)=RSV(Q,d_2)=0.5$. Il apparaît que des documents ayant des vecteurs de performances très différents peuvent avoir la même évaluation globale. De même, une petite variation d'un score élémentaire peut faire changer le classement. En effet si $X_1(d_1,t_1)$ passe de 0.5 à 0.49 alors d_2 est préféré à d_1 tandis que si $X_r(d_1,t_1)=0.51$ alors d_1 devient préféré à d_2. Si nous considérons l'opérateur min (non compensatoire), alors d_1 est jugé meilleur que d_2 bien que l'on puisse concevoir de préférer d_2 du fait de la faible différence des deux valeurs minimales 0.5 et 0.49. Le fait de ne considérer qu'un seul degré de pertinence élémentaire conduit, donc, à une perte d'information non négligeable.

Nous choisissons de mettre en œuvre dans nos contributions la famille d'opérateurs de Yager qui sont des opérateurs de compromis. En effet, ils ne sont pas contraints par le maximum ni par le minimum et correspondent, suivant différentes valeurs du paramètre q, à une grande variété d'opérateurs d'agrégation.

Nous avons montré comment l'utilisateur et ses préférences peuvent être pris en compte dans le calcul de pertinence à travers un modèle d'agrégation. Voyons maintenant son implication dans la phase de reformulation.

2.5.2. Médiation entre le système de RI et l'utilisateur : nécessité d'une composante lexicale

Intégrer l'utilisateur dans la boucle de pertinence à travers ses critères et la prise en compte de sa stratégie de décision à travers un opérateur d'agrégation, n'est pas suffisant. En effet, pour l'utilisateur, pouvoir effectuer une requête est une chose, mais être capable de comprendre les résultats qui lui correspondent et en quoi ils sont pertinents n'est pas une chose aisée sans aide. Aussi, la Recherche d'Information étant un processus itératif d'après son scénario de base (c.f. section 1.2), plusieurs affinements sont généralement nécessaires pour que l'utilisateur puisse trouver les documents pertinents relativement à son besoin en information. Mais la réussite de cette démarche itérative nécessite que l'utilisateur ait une compréhension précise des résultats qui lui sont présentés afin de pouvoir identifier rapidement les documents pertinents ou indiquer de manière plus précise ses préférences au SRI. Les techniques de visualisation peuvent donc être considérées comme des éléments clés de ce processus puisque les interfaces de visualisation jouent un rôle de médiateur entre le SRI d'une part et l'utilisateur d'autre part. C'est ainsi que plusieurs spécifications caractérisant de telles interfaces de visualisation en RI ont été proposées. Parmi celles-ci, citons les deux suivantes telles que relatées dans (Aussenac-Gilles 2008) :

- spécifications cognitives : étant données les limites cognitives des utilisateurs relatives à la prise en compte d'une grande quantité d'information, il est important d'attirer rapidement leur attention sur les documents que le SRI juge pertinents, de leur permettre aussi de se concentrer sur une section spécifique de l'interface et de comprendre simplement l'incidence sur les résultats qu'ils observent de toute action qu'ils entreprennent (Wiss & Carr 1998) ;
- spécifications relatives aux couleurs : l'hypothèse qui est faite ici est de considérer que les couleurs attirent plus l'attention que du texte et qu'il est donc plus facile, pour un utilisateur, de les repérer (Cockburn & McKenzie 2001). Les couleurs peuvent mettre en évidence une sémantique traduisant, par exemple, les importances relatives des éléments affichés : une couleur verte apparaissant dans l'icône d'un document peut signifier la présence d'un terme de la requête dans son index.

La dimensionnalité (2D ou 3D par exemple) est également une caractéristique importante des interfaces de visualisation. Cependant, la plupart des SRI affiche leurs résultats dans un espace en deux dimensions. Nous nous restreignons à cette dimensionnalité et indiquons, dans la suite, comment les SRI conceptuels présentent leurs résultats avec un intérêt particulier

concernant l'usage qu'ils font des ontologies dans cette tâche. Puisqu'il s'agit d'interfaces de visualisation, il faut avoir à disposition le SRI ou, en tout cas, pouvoir y accéder à travers le Web afin d'étudier son interface. Cela justifie le fait que les SRI détaillés ici soient presque tous accessibles en ligne. Par ailleurs, nous nous concentrons, dans la présentation qui suit, aux SRI conceptuels adressant des corpus textuels tout en étant conscients que d'autres corpus (base de gènes par exemple) peuvent être indexés de manière conceptuelle. Ce choix se justifie par un souci d'uniformisation de la présentation des approches par le fait la manière de visualiser des données dépend fortement de leur nature (on ne peut mettre en évidence du texte issu d'un gène par exemple).

La manière la plus commune de visualiser des résultats d'une requête est faite à travers une liste. Dans celle-ci, chaque résultat est représenté par un bloc contenant quelques méta données dont le titre, l'origine ou les auteurs, un texte donnant un aperçu de son contenu et dans laquelle les labels associés aux concepts de la requête sont mis en exergue. C'est le cas notamment du SRI *AlvisIR*[18] (Bossy et al. 2008) du projet *Quaro* [19] permettant de rechercher des articles scientifiques indexés par une ontologie de domaine et du système GoPubMed (Doms & Schroeder 2005) permettant de rechercher des articles de *PubMed* avec un vocabulaire issu de thésaurus tel que *MeSH* ou d'ontologies comme la *Gene Ontology* pour ne citer que ceux là. Une catégorisation des résultats est fournie suivant les concepts qui y sont référencés afin de pouvoir ne visualiser qu'un sous-ensemble d'éléments ne traitant que d'un ou plusieurs concepts donnés. La Figure 2.14 montre un extrait représentatif de la visualisation classique par liste des résultats d'une recherche dans de nombreux SRI notamment dans le domaine biomédical. Une telle approche classique ne remplit pas les spécifications cognitives et celles relatives aux couleurs mis en avant tantôt. En effet, l'unique lien mis en exergue entre la requête soumise et un document affiché réside dans la mise couleur des termes apparaissant dans le document correspondant (de manière exacte ou approchée mais toujours lexicale) aux labels des concepts de la requête. Bien que la recherche soit effectuée en ayant comme unité d'indexation des concepts, l'usage des seuls labels est limité quand il s'agit de mettre en exergue la manière dont les concepts sont relatés dans un document. Il s'agit d'une tentative très limitée d'expliquer, à un utilisateur, en quoi les documents qui lui sont fournis correspondent à sa requête. Un concept peut, en effet, se manifester dans un texte à travers différents termes selon le contexte. Une autre limite concerne le fait qu'il soit difficile d'avoir une vue globale des résultats avec une visualisation sous forme de liste.

Pour aller au-delà d'une simple identification de labels de concepts lorsqu'il s'agit de justifier l'adéquation entre une requête et un document, de nombreux travaux suggèrent d'exploiter une description plus fine des documents. Il s'agit concrètement d'associer une description conceptuelle à des parties de documents appelées passages. Le système *Ontopassage* (Lin et al. 2012) propose une telle solution. Dans ce système, les documents sont fragmentés en passages qui sont relatifs à des sujets donnés auxquelles on associe des concepts issus d'une ontologie. Dans *Ontopassage*, il est possible, pour l'utilisateur, de passer d'une visualisation

[18] http://bibliome.jouy.inra.fr/alvisir/gisdemo/Index
[19] http://www.quaero.org/catalog/#/6/

classique par liste de documents à une visualisation des passages plus fine et plus précise. Il apparaît que l'élément décisif dans la stratégie de visualisation d'un document traitant de plusieurs sujets est la capacité de segmenter celui-ci afin de mettre en évidence et d'indexer, par des concepts, ses passages.

Figure 2.14 : Extrait de la liste des résultats de la requête "*Protein*" dans le SRI *Alvis* (http://bibliome.jouy.inra.fr/alvisir/gisdemo/Index). Cet extrait est représentatif des approches communes de visualisation des résultats d'une requête. Le terme "*Protein"* est signalé (couleur rouge) dans les résumés qui accompagnent chaque document.

En considérant ces deux situations, il devient clair qu'une articulation entre concepts et entités lexicales permettant de les repérer dans un texte doit être prise en compte. Une ontologie de domaine dont chaque concept et chaque relation est associé à un lexique d'étiquettes est appelé ontologie à composante lexicale (Maedche et al. 2001). Disposer d'une composante lexicale n'est pas toujours facile. En effet, la majeure partie des méthodes de construction d'ontologies de domaine ne gardent pas les informations lexicales à partir desquelles sont formés les concepts. Cette limite est connue comme étant un problème de chaînon manquant (*missing link*) entre les niveaux ontologique et lexical (Badra et al. 2011). En effet, les formalismes, tel que OWL [20], utilisés pour représenter une ontologie de domaine, se concentrent principalement sur la description intrinsèque des concepts, des relations et des contraintes logiques qui portent sur eux. Quelques initiatives[21] sont menées dans le but d'aller au-delà des systèmes de labels implémentés dans les formalismes existants afin de doter les

[20] OWL : Web Ontology Language (http ://www.w3.org/TR/owl-features/)
[21] Ontology-Lexica Community Group : http ://www.w3.org/community/ontolex/

ontologies de domaine de composantes lexicales. Ces initiatives consistent principalement en des techniques de construction (Supekar et al. 2005) et des modèles de représentation (Reymonet et al. 2007; Badra et al. 2011; Buitelaar et al. 2011; Cimiano et al. 2011) de composantes lexicales d'une ontologie de domaine. (Maedche et al. 2001) défini la composante lexicale d'une ontologie de domaine comme suit :

Définition 2.13 : (*composante lexicale d'une ontologie*) La composante lexicale L d'une ontologie de domaine $\Theta := \{C, R, H_C, H_R, Rel, Ax\}$ est définie comme étant un quadruplet $L = \{L_C, L_R, F, G\}$ où L_C et L_R sont des listes disjointes d'entrées lexicales (termes dans notre cas) respectivement associées aux concepts dans C et aux relations dans R. F (respectivement G) fournit une relation entre chaque concept (respectivement chaque relation) et sa composante lexicale L_C (respectivement L_R).

L'articulation entre concepts et entrées lexicales est intéressante dans le cadre de la RI conceptuelle car elle permet aux utilisateurs d'avoir des indications concernant la manière dont leurs requêtes conceptuelles sont exprimées dans les documents qui leur sont retournés. Nous exploitons ce lien à travers notre approche de justification des résultats de notre modèle de pertinence (c.f. section 3.4.3).

L'effort de justification à travers une visualisation des résultats plus fine constitue un pas important vers une meilleure intégration de l'utilisateur dans la boucle de pertinence. En effet, une meilleure compréhension de la pertinence des résultats qui lui sont présentés laisse présager que l'utilisateur soit plus à même d'indiquer au SRI des informations supplémentaires concernant son besoin en information, informé qu'il est sur les raisons qui ont conduit à l'éventuel échec de sa requête. La sous-section suivante propose de faire l'état de l'art des approches, appelées stratégies de reformulation, qui exploitent des retours utilisateurs pour les intégrer dans le modèle de pertinence.

2.5.3. Stratégies de reformulation de requêtes : l'utilisateur comme seul juge

La reformulation d'une requête, d'un point de vue pratique, consiste à la modifier en y rajoutant (ou supprimant) des termes et/ou en modifiant la pondération associée à chacun de ces termes. D'après (Bhogal et al. 2007) et (Carpineto & Romano 2012), le choix des termes à ajouter ou à supprimer de même que l'estimation de leurs pondérations, est effectué sur la base de sources d'évidences qui peuvent être indépendantes de l'utilisateur dans le sens où elles ne concernent que la requête initiale (approches globales) ou, au contraire, entièrement dépendantes de son appréciation des résultats qui lui sont fournis (approches locales). Les dénominations des différentes stratégies visant à modifier ou réajuster d'une manière ou d'une autre une requête ne sont pas toujours claires et se contredisent même au gré des publications. Par exemple, dans (Bhogal et al. 2007) et plus récemment dans (Carpineto et al. 2012), ces stratégies sont désignées par "méthodes d'expansion de requêtes" alors que dans (Manning et al. 2008) l'expansion de requêtes ne désigne qu'un cas particulier de modification de requête parmi d'autres. Dans la suite de cette section, nous désignons par "méthode de reformulation de requêtes" toute stratégie de modification ou de réajustement de requêtes.

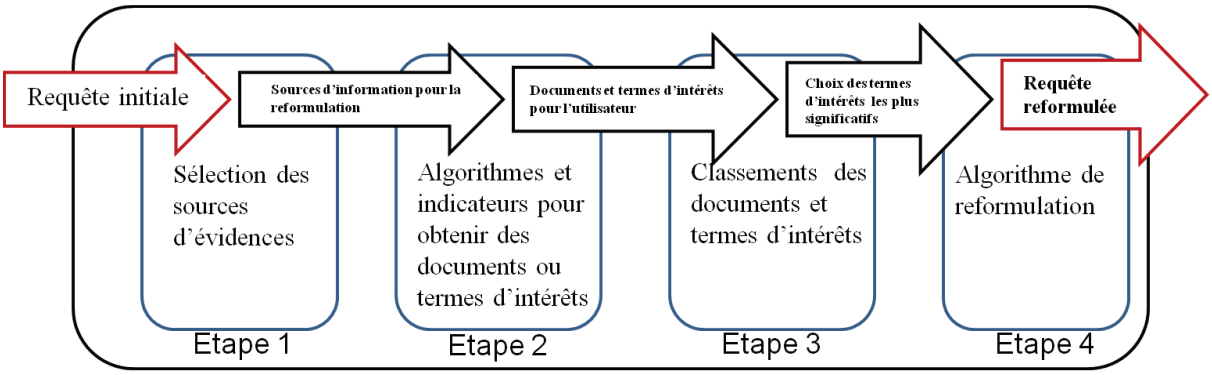


Figure 2.15 : Les étapes principales des stratégies de reformulation de requêtes adaptées de (Carpineto et al. 2012)

La reformulation de requêtes dans le cadre d'un système de recherche d'information peut être vue comme un processus formé des quatre étapes principales suivantes (c.f. Figure 2.15) telles que indiquées dans (Carpineto et al. 2012).

- **Première étape : sélection des sources d'évidences**

Il s'agit de l'étape durant laquelle sont choisies les sources d'informations à utiliser pour apporter un éclairage nouveau sur le besoin en information de l'utilisateur ayant soumis une requête initiale. Plusieurs sources sont mises en œuvre dans la littérature aboutissant à un classement des approches de reformulation en deux catégories détaillées dans la section 2.5.3.1 pour les approches globales et la section 2.5.3.2 pour les approches locales. La Figure 2.16 montre une catégorisation des approches de reformulation suivant les sources d'évidences mises en œuvre.

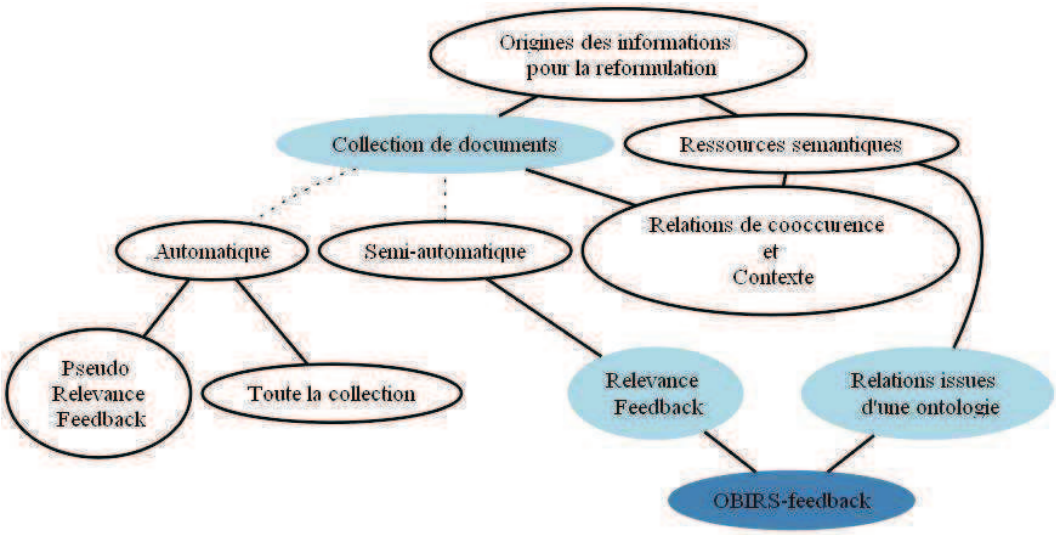


Figure 2.16 : Classification des stratégies de reformulation de requêtes suivant les sources d'informations qu'elles exploitent. Les deux liens en pointillé constituent des relations de classification, i.e. les sources de données dans une collection peuvent être obtenues de manière automatique ou semi-automatique. OBIRS-*feedback* est notre approche et elle combine une réinjection de pertinence explicite (Relevance Feedback) et une ontologie de domaine.

- **Deuxième étape : algorithmes et indicateurs pour obtenir des documents ou termes d'intérêts**

Cette étape traite des stratégies à mettre en œuvre, à partir des sources d'évidences identifiées, pour construire un ensemble de documents, ou de termes, susceptible de contribuer à une bonne reformulation. Il peut s'agir de l'exploitation de liens sémantiques modélisés à travers une ontologie de domaine (c.f. section 2.5.3.1 ci-dessous) ou d'indicateurs permettant de connaitre les documents ou termes intéressant l'utilisateur par rapport à son besoin en information courant (indicateurs inférés à partir des clics de l'utilisateur dans le cadre des méthodes de réinjection de pertinence implicite, par exemple).

- **Troisième étape : classements des documents et termes d'intérêts**

Une fois les sources d'évidences identifiées et la constitution d'un ensemble de documents ou de termes d'intérêts effectuée, il s'agit, à travers cette étape, de définir formellement une stratégie permettant de décider quand un terme est préféré à un autre. Notons que l'ensemble des documents d'intérêts peut se ramener à l'ensemble des termes les indexant. Cette étape dépend du modèle de recherche d'information sous tendant la méthode de reformulation et donc de la manière dont est modélisée la pertinence.

- **Quatrième étape : algorithme de reformulation**

Il s'agit ici de la stratégie de constitution de la requête reformulée qu'il faudra soumettre au système de recherche d'information.

Analysons maintenant les deux grandes familles de stratégies de reformulation selon les sources d'évidences.

2.5.3.1. Méthodes globales

Il s'agit des approches reformulant une requête initiale indépendamment de ses résultats. Elles opèrent donc en amont du processus de pertinence d'un système de recherche d'information. Les sources d'évidences mises en œuvre dans le choix des termes pour l'expansion peuvent être construites à partir de ressources externes. Celles-ci peuvent permettre d'identifier de nouveaux termes liés à ceux de la requête considérée (à travers une relation de synonymie, de cooccurrence par exemple mais aussi de similarité sémantique lorsqu'il s'agit de concepts). Ces approches sont généralement transparentes pour l'utilisateur. Plusieurs méthodes globales de reformulation de requête peuvent être distinguées suivant le type de ressources qu'elles utilisent.

Les méthodes utilisant des ressources sémantiques externes

Il s'agit des méthodes utilisant des ressources sémantiques externes telles des bases de données lexicales comme WordNet[22] à l'instar de (Voorhees 1994; Guarino et al. 1999; Baziz et al. 2003), des méta thésaurus biomédicaux comme *UMLS*[23] (utilisé dans *Medline* par

[22] http ://wordnet.princeton.edu/
[23] http ://www.nlm.nih.gov/research/umls/

exemple) ou des ontologies de domaine telle que la *Gene Ontologie*. La Figure 2.17 montre un exemple de reformulation où la requête "*cardiac*" est automatiquement étendue, dans *PubMed*, en rajoutant le terme "*heart*". Dans la suite, et dans un souci d'uniformisation, nous désignerons par concept les synsets issus de Wordnet ainsi que les termes issus de thésaurus biomédicaux comme *UMLS* ou le *MeSH*[24] (Medical Subject Headings).

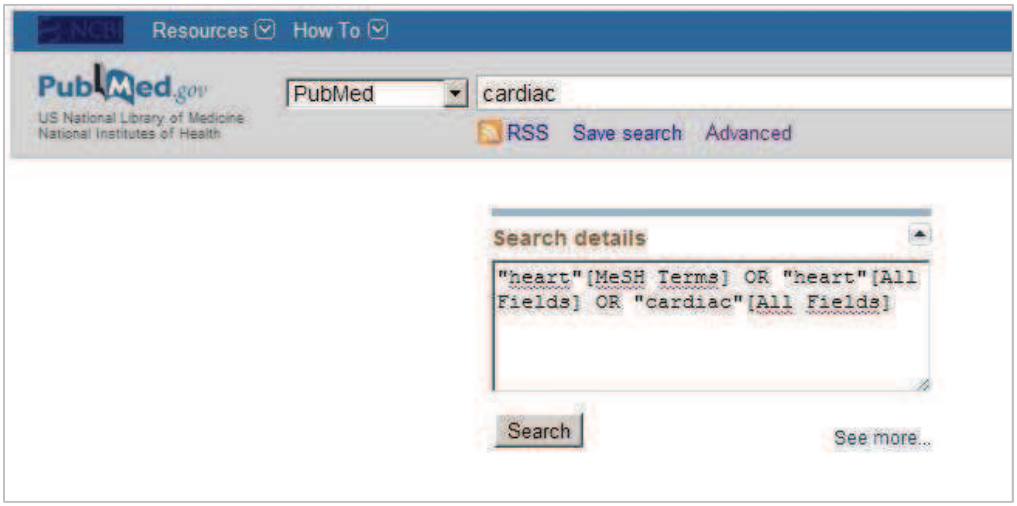


Figure 2.17 : Exemple de reformulation de requête dans *PubMed* utilisant *UMLS*

Il peut être paradoxal d'étudier la reformulation de requêtes conceptuelles, donc d'en préciser le sens, étant donné le caractère non ambigu des concepts. Mais selon (Nenad 2005), une requête conceptuelle trop générale, i.e. composée de concepts de très haut niveau dans la hiérarchie taxonomique d'une ontologie, peut conduire à des résultats bruités. Le problème pointé, dans un tel cas, est clairement lié à l'impact négatif des concepts de faible informativité (dont le contenu informationnel est faible). L'étude de la reformulation de requêtes conceptuelles est donc nécessaire.

L'usage, le plus répandu, de ressources sémantiques dans la tâche de reformulation de requêtes consiste en une stratégie d'exploration des différentes hiérarchies de concepts qu'elles renferment à travers différents types de relations sémantiques (taxonomiques ou non). Cette exploration des hiérarchies de concepts peut être vue comme une propagation des besoins en information de l'utilisateur (représentés, généralement, par une requête conceptuelle) vers d'autres concepts susceptibles d'en préciser le sens. Elle s'inspire, donc, des méthodes de *spreading activation* (Collins et al. 1975) où l'activation s'effectue le long de chemins du graphe des différentes hiérarchies de l'ontologie. Une telle propagation est en œuvre dans la stratégie d'expansion de requêtes conceptuelles de (Hliaoutakis et al. 2006) et dans les stratégies de construction d'un ensemble de concepts d'intérêts pour un utilisateur dans les travaux de (Daoud et al. 2009) et dans ceux de (Cena et al. 2011). Durant cette propagation, le rayon à explorer (la distance par rapport aux concepts de départ) est crucial et la majeure partie des travaux existants préconise un rayon limitée et donc une *expansion prudente* (Baziz

[24] http ://www.nlm.nih.gov/mesh/meshhome.html

et al. 2003). La Figure 2.18 montre l'exploration d'une ontologie hypothétique et différents rayons de propagation représentés par des ellipses en pointillés.

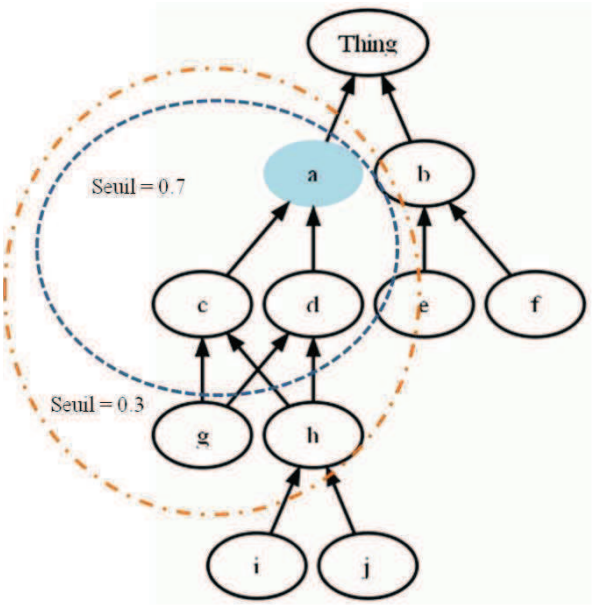


Figure 2.18 : Différents rayons dans une stratégie de propagation au travers d'une ontologie. Les rayons représentés découlent de la mesure de similarité de Lin définie dans l'équation (2.20) avec l'estimation du contenu informationnel d'un concept de Seco donnée par l'équation (2.9)

La stratégie de propagation de proche en proche, dans l'ontologie, à partir d'un concept peut être reformulée comme étant un problème d'exploration du voisinage du concept suivant un rayon donné obtenu avec une mesure de similarité sémantique. La mesure en question peut être une simple longueur d'arcs ou plus complexe comme dans la Figure 2.18 où la mesure de similarité de Lin est utilisée avec une estimation du contenu informationnel basée sur (Seco et al. 2004). Dans tous les cas, il est nécessaire d'obtenir, à l'issu de l'exploration, tous les concepts dont la similarité est supérieure à un seuil. Or rien ne garantit, dans les travaux existant, que les stratégies de propagation mises en œuvre respectent cet impératif. Dans le Chapitre 4, nous mettons en exergue les propriétés nécessaires des mesures de similarité sémantique pour atteindre un tel objectif.

Au-delà du rayon de propagation, la nature des liens à explorer est aussi importante. Dans (Rada et al. 1991), les auteurs ont mené une étude utilisant la base de données Excerpta Medica et le thésaurus *EMTREE* et indiquent la possibilité d'explorer avec succès différentes relations pour peu que celles-ci soient mentionnées dans la requête utilisateur. Dans (Baziz et al. 2003), les auteurs considèrent plusieurs relations sémantiques dans *WordNet* telles que : la synonymie, l'hypéronymie, l'hyponymie, la méronymie et son contraire l'holonymie. D'après les expériences qu'ils ont menées sur la collection CLEF2001 (Cross Language Evaluation Forum), ils sont arrivés à la conclusion que la relation *is-a* est la plus importante à prendre en compte (l'hyponymie est moins performante que l'hypéronymie) et que considérer la synonymie en plus de la relation *is-a* améliore la précision pour les premiers documents. Dès lors, considérer d'autres relations peut détériorer les performations du SRI après reformulation d'où la préconisation d'une "*expansion prudente*". Ce résultat nous conforte dans notre choix de ne considérer que la restriction aux relations *is-a* d'une ontologie. Il faut cependant nuancer

une telle conclusion. En effet, (Khoo et al. 2007) indique qu'il est possible que la prise en compte d'autres relations dans le processus d'expansion puisse améliorer les performances dès lors que celles-ci sont présentes initialement dans la requête de l'utilisateur.

Les méthodes utilisant des thésaurus générés automatiquement

Il s'agit des premières approches de reformulation de requêtes. Elles reposent essentiellement sur une analyse statistique des collections de documents pour identifier des liens entre les termes présents dans ces documents et ceux de la requête en exploitant des informations telles que leurs fréquences de cooccurrence. Des termes sont co-occurrents lorsqu'ils sont utilisés conjointement au niveau document, paragraphe ou phrase. Plusieurs travaux proposent des solutions pour utiliser ces informations de cooccurrence afin de construire des thesaurus rendant compte de la collection (Chu et al. 2002; Joho et al. 2004). Une requête peut, dès lors, être reformulée en intégrant des termes co-occurrents souvent avec les termes qu'elle contient. (Salton 1986; Peat et al. 1991) montrent que cette stratégie donne de mauvais résultats du fait de la difficulté à trouver le bon contexte pour le choix de termes co-occurrents (document, paragraphe ou phrase) et de la restriction du type de relation pris en compte : la cooccurrence ne couvre pas l'ensemble des relations possibles entre termes. Deux termes peuvent être similaires sans être co-occurrents. D'autres travaux ont essayé de corriger ces limites (Qiu & Frei 1993) en mettant en œuvre une matrice de similarités des termes et non de cooccurrence (Park & Ramamohanarao 2007).

Les stratégies fondées sur la construction automatique de thésaurus à partir de la collection de documents sont robustes mais nécessitent un corpus assez large et sont donc gourmandes en temps de calcul. De plus, le thésaurus construit doit être régénéré à chaque évolution majeure de la collection. Analysons maintenant les méthodes locales de reformulation.

2.5.3.2. Méthodes locales

Les méthodes locales d'expansion de requêtes consistent en l'exploitation d'informations dépendantes de l'utilisateur et générées à travers son activité de recherche. Elles sont, dans ce sens, dynamiques. Les sources d'évidences mises en œuvre ici concernent l'intérêt de l'utilisateur pour les résultats associés à la requête à reformuler. Il convient de bien évaluer cet intérêt en mettant en place des indicateurs appropriés. Suivant les différents types d'indicateurs qu'elles mettent en œuvre pour le choix des documents ou termes à prendre en compte, les méthodes d'expansion de requêtes locales peuvent être catégorisées en trois classes :

- Les approches de réinjection de pertinence explicite (*relevance feedback*) : il s'agit de demander à l'utilisateur de fournir les documents, ou les termes les indexant, qui sont d'intérêts pour lui afin de les exploiter pour mener la reformulation. L'utilisateur joue donc le rôle de déclencheur du processus.
- Les approches de réinjection de pertinence indirecte (*indirect relevance feedback*) : les actions de l'utilisateur à travers l'interface de recherche sont analysées pour déduire les documents ou termes auxquels il s'intéresse.

- Les approches de réinjection de pertinence implicite ou aveugle (*pseudo relevance feedback*) : dans ce cadre, les k premiers documents au sens du modèle de pertinence sont automatiquement choisis comme étant d'intérêt pour l'utilisateur et pris en compte dans le processus de reformulation.

Dans la suite, nous passons en revue quelques paramètres importants concernant la mise en œuvre des méthodes locales de reformulation.

Combien de termes ou de documents prendre en compte ?

Le processus d'expansion peut générer un grand nombre de nouveaux termes à prendre en compte. Dans (Buckley, 2004), les auteurs indiquent que le choix des termes à ajouter n'est pas crucial. Ils ajoutent les 300 premiers termes ainsi que 50 phrases issus des documents pertinents dans le cadre du modèle de Rocchio (Rocchio 1971). (Buckley, 2005) propose d'ajouter massivement des termes issus des documents que l'on sait, ou que l'on suppose être pertinents. (Voorhees 1994) a étudié l'effet du nombre de termes d'expansion dans la performance des SRI en concluant que, seul, le nombre ne suffit pas.

Réinjection de pertinence explicite : schéma général et paramètres

La réinjection de pertinence explicite ou *relevance feedback*, est un processus itératif et interactif ayant pour but d'améliorer les performances des SRI en incluant l'utilisateur dans le processus de recherche. Historiquement, la procédure de base de cette technique est la suivante :

a. l'utilisateur soumet une requête initiale simple,
b. le système de recherche d'information retourne une liste de documents ordonnés selon l'ordre de pertinence par rapport à la requête initiale,
c. l'utilisateur marque certains documents comme étant pertinents et d'autres non,
d. la requête initiale est reformulée pour déboucher sur une nouvelle requête qui va être soumise au système de recherche d'information,
e. le SRI retourne une nouvelle liste de documents. Les étapes c et d pouvant être répétées.

La Figure 2.19 montre le schéma général d'une telle stratégie.

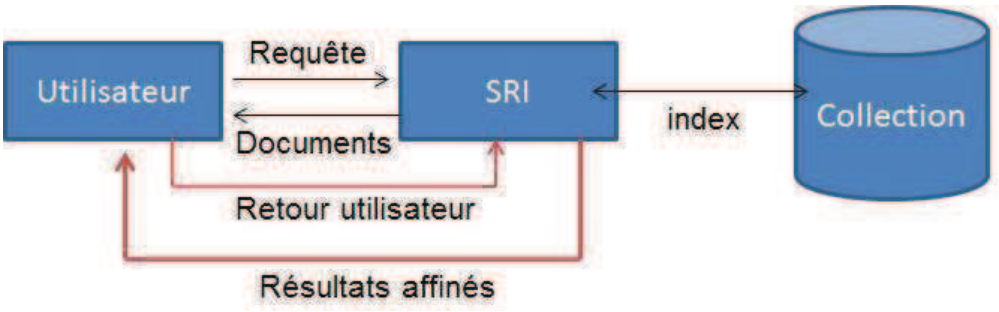


Figure 2.19 : Scénario classique de réinjection de pertinence explicite (*relevance feedback*)

Le choix des documents et des termes à considérer ainsi que des stratégies concernant la manière de les utiliser (étape e) pour reformuler une requête dépendent fortement du modèle de recherche d'information sous-tendant l'approche de réinjection de pertinence.

Dans la suite, nous introduisons l'approche de Rocchio (Rocchio 1971), l'une des stratégies de réinjection de pertinence les plus utilisées et mise en œuvre dans le cadre du modèle vectoriel, ainsi que certaines de ses variantes.

Réinjection de pertinence explicite dans le modèle vectoriel

L'algorithme standard de réinjection de pertinence dans les modèles vectoriels de recherche d'information est celui de Rocchio (Rocchio 1971). Rappelons que dans le modèle vectoriel (voir chapitre 2), les documents ainsi que les requêtes utilisateurs sont des vecteurs exprimés suivant la base canonique de l'espace des termes d'indexation. A la suite d'une requête vectorielle $\overrightarrow{Q_{init}}$, la méthode de Rocchio identifie la réinjection de pertinence comme étant la recherche d'un vecteur requête optimal, que nous noterons $\overrightarrow{Q_{opt}}$, maximisant la différence entre tous les documents pertinents DP^Q relativement à $\overrightarrow{Q_{init}}$ et tous les documents non pertinents DNP^Q relativement à $\overrightarrow{Q_{init}}$. Un tel vecteur optimal maximise de fait la similarité avec les documents pertinents, et minimise celle d'avec les documents non pertinents. Si on définit $\vec{Q}$ comme étant un vecteur appartenant à l'ensemble des vecteurs possible, alors la situation décrite précédemment peut s'écrire formellement à travers l'équation **(2.36)** en considérant le cosinus comme mesure de similarité entre deux vecteurs documents.

$$\overrightarrow{Q_{opt}} = \frac{1}{|DP^Q|}\sum_{\overrightarrow{d_i}\in DP^Q}\overrightarrow{d_i} - \frac{1}{|DNP^Q|}\sum_{\overrightarrow{d_j}\in DNP^Q}\overrightarrow{d_j} \qquad \textbf{(2.36)}$$

L'équation (2.36) traduit une configuration idéale, car l'ensemble des documents pertinents DP^Q et celui des documents non pertinents DNP^Q par rapport à une requête $\overrightarrow{Q_{init}}$ ne sont pas connus dans un contexte réel. En pratique, seule une partie des documents pertinents et non pertinents, fournie par l'utilisateur en l'occurrence, est connue. A partir d'un ensemble de documents D_{res} obtenus après la soumission d'une requête $\overrightarrow{Q_{init}}$ et dans laquelle l'ensemble D_p des documents pertinents est connu de même que l'ensemble D_{np} des documents non pertinents, Rocchio propose de rechercher une requête modifiée $\overrightarrow{Q_m}$ telle que :

$$\overrightarrow{Q_m} = \alpha\overrightarrow{Q_{init}} + \beta\frac{1}{|D_p|}\sum_{\vec{d_i}\in D_u}\vec{d_i} - \gamma\frac{1}{|D_{np}|}\sum_{\vec{d_j}\in D_{np}}\vec{d_j} \qquad \textbf{(2.37)}$$

Avec $(\alpha,\beta,\gamma) \in [0,1]^3$ des réels permettant de paramétrer la contribution de chaque terme suivant que l'on veuille que les termes du vecteur initial $\overrightarrow{Q_{init}}$ soient plus importants que ceux rajoutés.

En pratique, la nouvelle requête consiste en la requête initiale ($\alpha = 1$) en plus des termes qui différencient le plus les documents jugés pertinents et ceux jugés non pertinents. En plus, si on dispose d'assez de documents positifs et négatifs (c'est-à-dire que $|D_p \cup D_{np}|$ est assez

grand), alors les paramètres β et γ sont choisis de telle sorte que le rapport $\frac{\beta}{\gamma}$ soit élevé ($\beta \gg \gamma$). En effet, l'information que l'on obtient des documents positifs est considérée comme plus importante que celle issue des documents négatifs. Il peut arriver que des termes se retrouvent avec une pondération négative dans le vecteur final $\overrightarrow{Q_m}$. Ceux-ci sont alors ignorés, ce qui revient à les pondérer par 0.

La Figure 2.20 montre une configuration hypothétique dans laquelle l'utilisateur a fourni les documents pertinents (petit cercle) et non pertinents (croix) par rapport à une requête initiale (triangle creux).

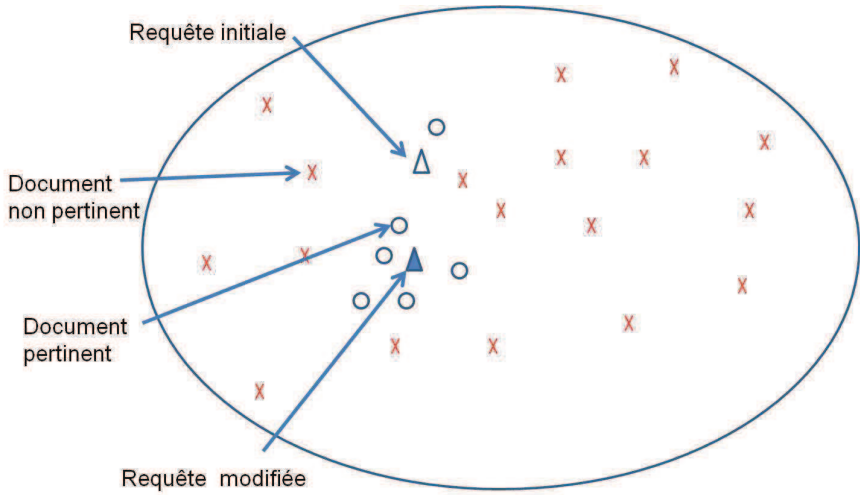


Figure 2.20 : Application de la stratégie de Rocchio sur une configuration de recherche hypothétique. La requête initiale est déplacée vers les documents marqués comme pertinents par l'utilisateur.

Cette approche de reformulation locale de Rocchio est largement utilisée. Cependant, bien qu'il soit aisé de l'implémenter dans le modèle vectoriel classique, son adaptation dans les approches conceptuelles de RI n'est pas triviale. En effet, le choix d'une requête conceptuelle (un ensemble de concepts pondérés) optimale (correspondant ici à $\overrightarrow{Q_{opt}}$) peut soulever des problèmes combinatoires dès lors que l'ontologie considérée est de grande taille. Dans une approche "sac de concepts", cela revient à chercher un sous-ensemble de concepts parmi tous les concepts de l'ontologie optimisant un indicateur inspiré de l'approche Rocchio.

Comme nous l'avons vu, deux sources d'évidences peuvent intervenir dans une stratégie de reformulation. La première concerne l'utilisateur à travers ses jugements sur les résultats qui lui sont présentés et la deuxième concerne le modèle de connaissances, indépendant de l'utilisateur, qu'il faut explorer. Nous proposons, dans le chapitre Chapitre 4, une stratégie de reformulation de requêtes combinant ces deux sources.

2.6. Evaluation en Recherche d'Information

Deux principaux critères peuvent être considérés lorsque l'on aborde l'évaluation des SRI selon (Boughanem 2008). Le premier concerne *l'efficience* du système : est-il capable de

répondre à une requête de l'utilisateur dans un temps court, en minimisant l'espace (disque) requis pour cela. Le deuxième concerne *l'efficacité* du système qui évalue sa facilité d'utilisation, la facilité de compréhension de ses résultats au travers de sa présentation mais surtout sa capacité à sélectionner des documents pertinents pour l'utilisateur. La majorité des modèles d'évaluation existants se focalisent sur la mesure de cette dernière capacité. Les raisons sont nombreuses et concernent la nature même de la notion de pertinence. La pertinence dans un SRI est généralement vue comme une correspondance entre un document et une requête ou comme une mesure de l'adéquation d'un document par rapport à une requête mais elle ne peut être réduite à cela. En effet, dans la plupart des cas, elle fait intervenir un contexte de jugement de nature subjective. Selon (Farah et al. 2006), plusieurs types de pertinence sont à distinguer. (Mizzaro 1997) met en évidence la pertinence-utilisateur liée à la perception qu'a l'utilisateur des résultats retournés par le SRI. Une telle pertinence est *subjective* car deux utilisateurs sont susceptibles d'avoir deux appréciations différentes d'un même document renvoyé en réponse à une même requête. L'appréciation d'un résultat peut aussi changer dans le temps, la connaissance de l'utilisateur concernant le domaine de recherche ayant évoluée.

La pertinence système, dite *objective*, est celle qui est la plus considérée. Elle est souvent traduite par un score synthétisant l'adéquation des documents relativement à une requête. Nous allons présenter les mesures d'évaluation entrant dans ce cadre.

L'objectif principal d'un SRI est de trouver l'ensemble des documents pertinents et d'ignorer l'ensemble des documents non pertinents relativement à une requête d'un utilisateur. Rappelons que D_{res}^{Q} est l'ensemble des documents renvoyés en réponses à une requête Q et que $D_{pert}^{Q} \subseteq D$ est l'ensemble des documents de la collection D pertinents pour Q (c.f. Figure 2.21). Notons que $D_{pert}^{Q} \cap D_{res}^{Q}$ constitue l'ensemble des documents pertinents renvoyés par le SRI.

La *précision P* mesure la proportion de documents pertinents parmi ceux renvoyés par le SRI :

$$P = \frac{\left|D_{pert}^{Q} \cap D_{res}^{Q}\right|}{\left|D_{res}^{Q}\right|} \quad \textbf{(2.38)}$$

Le *rappel* Rap mesure la proportion de documents pertinents retournés par le système relativement à l'ensemble des documents pertinents connus :

$$Rap = \frac{\left|D_{pert}^{Q} \cap D_{res}^{Q}\right|}{\left|D_{pert}^{Q}\right|} \quad \textbf{(2.39)}$$

La figure Figure 2.21 montre l'intersection possible entre tous les documents de la requête et ceux de la réponse.

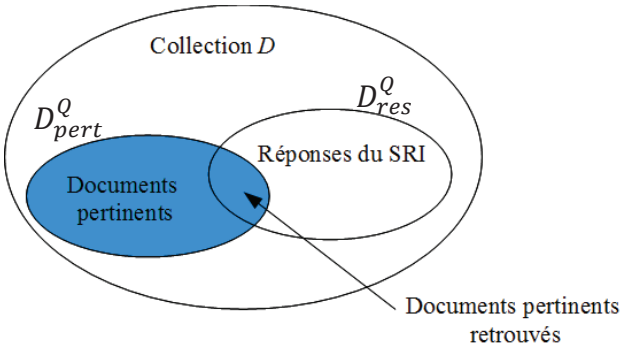


Figure 2.21 : Illustration de la précision et du rappel

D'autres mesures sont aussi utilisées, notamment, dans le cadre des campagnes TREC – *Text REtrieval Conference* (Voorhees & Harman 1997). Il s'agit des mesures de précision à k documents ainsi que de la précision moyenne. La précision à k documents, généralement notée par $P@k$, est liée à la mesure de la *précision exacte* qui est celle obtenue à la position où elle vaut le rappel. Cette position correspond à $\left|D^Q_{pert}\right|$ qui est le nombre total de documents pertinents de la requête Q. Si $D^{Q,k}_{pert} \subseteq D^Q_{pert}$ est l'ensemble des documents pertinents relativement à Q et à la position k, alors $P@k$ est définie par :

$$P@k = \frac{\left|D^{Q,k}_{pert}\right|}{k}, k \leq \left|D^Q_{pert}\right| \quad \textbf{(2.40)}$$

La précision exacte ou *R-precision* correspond à celle obtenue à la position $\left|D^Q_{pert}\right|$. Elle est nulle lorsqu'aucun document pertinent n'est trouvé et égale à 1 quand tous les documents pertinents sont retrouvés et classés en tête de la liste des résultats.

En considérant que la fonction $rank\ (d_j, Q)$ renvoie, pour chaque document d_j, sa position dans la liste des résultats de la requête Q, la précision moyenne de Q, notée *MAP* (*Mean Average Precision*), est donnée par :

$$MAP(Q) = \left(\frac{\sum_{d_j \in D^Q_{pert} \cap D^Q_{res}} \left(P@rank\ (d_j, Q)\right)}{\left|D^Q_{pert}\right|} \right) \quad \textbf{(2.41)}$$

Cette précision peut être généralisée sur un ensemble H_Q de requêtes Q en considérant sa moyenne sur cet ensemble. La précision moyenne est nulle lorsqu'aucun document pertinent n'est retrouvé pour aucune requête. Elle prend la valeur de 1 pour un SRI idéal qui renverrait tous les documents de D^Q_{pert} et les classerait en tête des résultats.

2.7. Conclusion

Dans ce chapitre, nous avons présenté comment, à travers l'état de l'art, le scénario de base de la Recherche d'Information (RI) est opérationnalisé. Pour chaque processus de ce scénario, nous avons donné un état de l'art. C'est notamment le cas concernant le processus d'indexation pour lequel les limites des approches classiques de RI ont été soulignées conduisant à l'émergence des approches conceptuelles de RI. Nous avons montré que la

construction d'une bonne réponse d'un SRI fait intervenir l'utilisateur à travers l'expression de ses préférences, son jugement sur les résultats qui lui sont présentés. Une meilleure interactivité entre l'utilisateur et le SRI doit permettre, dès lors, de trouver de manière plus rapide la bonne information. Cet état de l'art nous permet de mettre en évidence la nécessité de considérer deux composantes dans le cadre de la RI conceptuelle. La première composante comporte un modèle statistique de la collection et un modèle de connaissance décrivant de manière non ambiguë les sujets traités dans les documents. Nous considérons dans nos contributions une ontologie de domaine comme un modèle de connaissance et les similarités basées sur le contenu informationnel comme un modèle statistique des concepts dans une collection. La seconde composante concerne un modèle utilisateur modélisant ses préférences. Avec des préférences différentes, deux utilisateurs doivent avoir des résultats différents. Nous considérons dans nos contributions qu'un modèle d'agrégation de type compromis constitue un modèle de préférence. Nous étendons ce modèle en permettant une reformulation de requêtes où les jugements de l'utilisateur sont intégrés dans notre modèle de pertinence.

Le Chapitre 3 propose l'exploitation d'un modèle de connaissance (ontologie de domaine) et d'un modèle de préférences (opérateur d'agrégation) dans l'appariement d'un document par rapport à une requête et dans la visualisation des résultats d'une recherche. Le Chapitre 4 propose un modèle de reformulation de requêtes conceptuelles.

Chapitre 3 : *OBIRS*, un modèle d'agrégation en Recherche d'Information utilisant des similarités sémantiques

Publications représentatives de ce travail

Springer book chapter NTROLR 2012 : Sylvie Ranwez, Benjamin Duthil, Mohameth François Sy, Jacky Montmain, Patrick Augereau, et Vincent Ranwez. 2012. "How Ontology Based Information Retrieval Systems may Benefit from Lexical Text Analysis". In New Trends of Research in Ontologies and Lexical Resources, éd par. Alessandro Oltramari, P. Vossen, L. Qin, et Eduard Hovy. Theory and Applications of Natural Language Processing. Springer Verlag. (manuscript accepté pour publication)

BMC Bioinformatics 13 (Suppl 1) : Mohameth-François, Sy, Sylvie Ranwez, Jacky Montmain, Armelle Regnault, Michel Crampes, and Vincent Ranwez. 2011. "User Centered and Ontology Based Information Retrieval System for Life Sciences." *BMC Bioinformatics* 13 (Suppl 1) : S4.

SWAT4LS 2010 : Sylvie Ranwez, Vincent Ranwez, Mohameth-François Sy, Jacky Montmain, and Michel Crampes. 2010. "User Centered and Ontology Based Information Retrieval System for Life Sciences". In *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences,*

December 10, ed. Albert Burger, M. Scott Marshall, Paolo Romano, Adrian Paschke, and Andrea Splendiani. Vol. 698. CEUR Workshop Proceedings. Berlin, Germany : CEUR-WS.org.

IC 2010 : Sylvie Ranwez, Vincent Ranwez, Mohameth-François Sy, Jacky Montmain et Michel Crampes. 2010. "Utilisation de proximités sémantiques pour améliorer la recherche et le rendu d'information". *IC 2010, 21e Journées Francophones d'Ingénierie des Connaissances*, Nîmes, France, 9-11 juin 2010, pp.247-258.

3.1. Introduction

Dans ce chapitre, nous décrivons une approche originale de Recherche d'Information conceptuelle dénommée *OBIRS* (*Ontology Based Information Retrieval System*). Notre approche se propose d'utiliser les concepts d'une ontologie de domaine (c.f. Définition 2.2) comme langage pivot pour l'expression de requêtes conceptuelles permettant de rechercher des documents au sein d'une collection d'objets indexés conceptuellement par la même ontologie. Les tâches d'extraction de concepts (à partir de documents textuels par exemple) et de désambigüisation de ceux-ci, ne sont pas traitées dans ce mémoire. Nous considérons que ces deux tâches sont déjà effectuées. Une fois l'espace d'indexation disponible, *OBIRS* met en œuvre une stratégie d'agrégation à trois niveaux pour évaluer la pertinence, exprimée sous la forme d'un score d'un document par rapport à une requête, communément appelé *RSV* (*Retrieval Status Value*). Une stratégie de justification des résultats obtenus est mise en œuvre au travers d'une stratégie de visualisation où les résultats ne sont plus présentés dans une simple liste mais disposés sur une carte sémantique, relativement à une sonde représentant la requête utilisateur.

La première phase de notre méthode considère les concepts d'une requête utilisateur comme autant de critères que doivent satisfaire les différents documents de la collection, vus ici comme autant d'alternatives. Pour chaque critère ainsi identifié, un degré de pertinence élémentaire de chaque document est évalué en utilisant des mesures de similarité sémantique basées sur une ontologie de domaine. La dernière étape consiste à utiliser une famille d'opérateurs d'agrégation de compromis permettant d'attribuer une pertinence globale (*RSV*) aux documents en combinant leurs pertinences élémentaires tout en autorisant un mécanisme de justification pour la mise en lumière des critères les plus contributifs à leur *RSV*. Ainsi, dans notre approche, l'ontologie correspond au modèle de connaissance et les opérateurs de compromis correspondent à un modèle simple de préférences.

Une telle justification des degrés de pertinence couplée à une visualisation originale par carte sémantique des objets retrouvés participe grandement à la compréhension des résultats par l'utilisateur et permet une plus grande interaction pouvant favoriser, par exemple, les tâches de reformulation de requêtes (c.f. Chapitre 4). Dans le cas où les objets de la collection sont de types textuels, chaque document peut être segmenté relativement aux concepts de la requête pour mettre en exergue la *présence lexicale*[25] de ceux-ci dans les passages identifiés. L'objectif est de permettre une justification plus fine des résultats et un accès plus rapide aux

[25] Cette *présence lexicale* ne signifie pas que les concepts (par leurs labels) apparaissent directement dans le document, mais qu'un champ lexical qui leur est associé a été identifié dans certains passages.

informations utiles. Cette tâche de segmentation établit un lien entre les concepts de l'ontologie de domaine et les objets de la collection. Il s'agit d'une tâche d'interfaçage qui nécessite que les concepts de l'ontologie de domaine disposent d'une composante lexicale ce qui n'est pas toujours le cas. Nous montrons l'utilisation d'une telle composante lexicale au travers d'un outil que nous proposons, dénommé *CoLexIR*. Favoriser une meilleure interaction de l'utilisateur avec le système de recherche et justifier les résultats de recherche constituent l'apport principal de l'approche *OBIRS*. Ces deux principes guident le choix des différentes composantes de l'approche présentée dans ce chapitre.

Le reste du chapitre est organisé comme suit. Dans la section 0, nous donnons une vue globale de l'approche *OBIRS* en introduisant les éléments de celle-ci ainsi que les principales définitions. La section 3.3 présente, en détail, le modèle de pertinence à trois niveaux mis en œuvre dans *OBIRS*. Dans la section 3.4, un modèle de diagnostic et de justification des résultats d'une requête est détaillé. La justification des résultats d'une requête est de deux ordres : i) justification graphique en proposant une visualisation sur une carte sémantique interactive et un accès très détaillé aux passages des documents relatifs aux concepts de la requête, ii) et justification au travers du calcul de la contribution de chaque concept de la requête au score global permettant de savoir, par exemple, quel concept a le plus contribué au classement d'un document. Les applications et les études de cas réalisées, de même que les prototypes implémentés pour évaluer notre approche, sont explicités dans la section 3.5.

3.2. Présentation globale de notre approche

Le modèle de pertinence que nous présentons dans ce chapitre suppose l'existence d'une collection D de documents indexés par des concepts issus de la restriction Θ_{DAG} d'une ontologie de domaine aux seules relations de subsomption telle que définie dans la Définition 2.2. Il peut s'agir de la *Gene Ontology* (GO) ou du thésaurus *MeSH*. *OBIRS* repose donc sur un modèle de recherche d'information conceptuel, au sens de l'unité d'indexation, et adopte une représentation des documents et des requêtes utilisateurs sous forme de « sacs de concepts » éventuellement pondérés. L'existence de nombreuses collections de ressources indexées avec des concepts (*MuCHMORE*, Annotations de gènes, etc.), notamment dans le domaine biomédical, nous permet de pouvoir tester et déployer notre approche (c.f. section 3.5). Nous rappelons ici les définitions utiles pour ce chapitre.

Définition 3.1 : (*indexation conceptuelle d'un document*) Soient $\Theta_{DAG} = (C, H_{isa})$ la restriction aux relations *is-a* d'une ontologie de domaine, $\mathcal{P}(C)$ l'ensemble des parties de C et I_{c_s} l'ensemble des instances (extension) du concepts c_s dans D. L'indexation conceptuelle d'un document $d_j \in D$ par l'ontologie Θ_{DAG} est un ensemble de couples $C(d_j) = \{(c_s, w_s), s = 1..dl_j, w_s \in \mathbb{R}, c_s \in C\}$ avec c_s un concept tel qu'il existe une instance dans I_{c_s} apparaissant dans d_j et w_s un réel, mesurant l'importance du concept c_s dans d_j.

Définition 3.2 : (*requête conceptuelle*) Une requête conceptuelle Q est un sous-ensemble de l'ensemble des concepts C d'une ontologie Θ_{DAG} $\left(Q \in \mathcal{P}(C)\right)$. L'ordre des concepts n'a pas d'importance dans notre cadre.

Par extension, nous pouvons définir, une requête conceptuelle pondérée Q comme suit : $Q = \{(c_r, w_r), r = 1..n, w_r \in \mathbb{R}, c_r \in C\}$. Remarquons que, dans notre cas, les poids w_r associés aux concepts d'une requête sont fournis manuellement par un utilisateur. La sémantique de ce poids est relative à la préférence qu'un utilisateur peut avoir pour un concept plutôt que pour un autre.

Les deux définitions précédentes montrent que nous avons adopté un modèle « sacs de concepts » pour la représentation des documents et des requêtes que nous adressons. Cette représentation est souple et nous permet d'éviter d'avoir à manipuler des vecteurs de concepts ayant plusieurs dimensions creuses. Par ailleurs, les types de documents dont disposent nos partenaires et que nous adressons principalement sont des résumés d'articles scientifiques (de petites tailles donc et homogènes) et des gènes. Dès lors, comme expliqué dans la section 2.4.1, nous ne mettons pas en place un système de pondération locale des concepts. Celle-ci est réellement nécessaire lorsque les documents adressés sont de taille sensiblement différente. Concernant les gènes, il est évident qu'il n'est pas possible d'en extraire des concepts et de les pondérer localement. Dans la Définition 3.1, nous gardons cependant la pondération w_s pour conserver une approche générique.

Nous considérons également le contenu informationnel (c.f. Définition 2.3) des concepts comme une pondération globale à la collection.

La Figure 3.1 montre une vue synthétique de l'approche OBIRS.

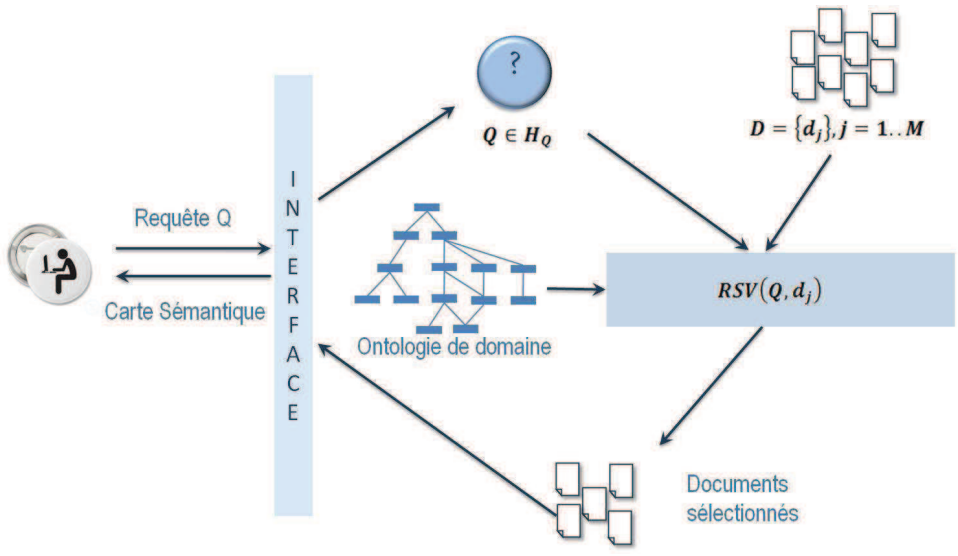


Figure 3.1 : Schéma synoptique de l'approche OBIRS

Nous proposons de décomposer le calcul de la pertinence d'un document d_j relativement à une requête Q en un processus d'agrégation en trois étapes. Tout d'abord, nous commençons par calculer la similarité entre un concept de la requête et un concept indexant d_j en utilisant une mesure de similarité sémantique π simple et intuitive (étape 1). Puis une proximité sémantique, basée sur π, est calculée entre chaque concept c_r de la requête et l'ensemble des concepts pondérés $C(d_j)$ indexant le document d_j (étape 2). Enfin, ces mesures sont

combinées dans le score global du document à travers un opérateur d'agrégation *A* (étape 3). Concernant la dernière étape, l'évaluation de la pertinence globale d'un document par rapport à une requête fournit un moyen naturel de dire pourquoi un document est préféré à un autre sur la base d'une analyse de sensibilité du *RSV*. Lorsque la requête est constituée de plusieurs concepts, il devient nécessaire de fournir une représentation de la préférence de l'utilisateur à l'un ou l'autre concept. La représentation de préférences est un sujet central dans la théorie de la décision (Modave & Grabisch 1998) et consiste généralement à trouver une fonction d'utilité U, à valeurs réelles, de telle sorte que pour toute paire d'alternatives d_l et d_j de D, $d_l \succsim d_j$ (d_l est préféré à d_j) si et seulement si $U(d_l) \geq U(d_j)$. En d'autres termes, la théorie de l'utilité multi attributs (*MAUT : Multi Attribute Utility Theory*) est un modèle de préférence. Dans notre approche, cela signifie que les modèles d'agrégation que nous utilisons sont censés capturer les préférences des utilisateurs. Suivant le modèle décomposable de préférence de (Krantz et al. 1971), la fonction d'utilité U peut être définie comme suit :

$$U(c_1, \dots, c_n) = A(u_1(c_1)), \dots, u_r(c_r)), \dots, u_n(c_n))) \quad \textbf{(3.1)}$$

$u_r : C \to [0,1], r = 1..n$ étant une fonction d'utilité pouvant être interprétée, dans le cadre de nos travaux, comme la pertinence élémentaire du document d_j relativement au concept c_r et $A : [0,1]^n \to [0,1]$ un opérateur d'agrégation qu'il convient de choisir judicieusement. Dans notre cas, la dimension n correspond au nombre de concepts des requêtes conceptuelles. La fonction d'utilité u_r joue le rôle de fonction X_r (c.f. section 2.5.1.1 à la page 44) d'évaluation de la pertinence donnant les degrés de pertinence élémentaire d'un document d_j.

Nous proposons de définir les fonctions d'utilités u_r comme des mesures de proximité sémantique entre un concept c_r et un document d_j :

$$X_r(d_j, c_r) = u_r(c_r) = \underset{(c_s, w_s) \in C(d_j)}{agreg} \left(\pi(c_r, c_s)\right) \quad \textbf{(3.2)}$$

Avec $agreg$ un opérateur d'agrégation. Plusieurs facteurs peuvent justifier le choix d'un opérateur d'agrégation *A* particulier dans le domaine de la RI. Dans l'approche *OBIRS*, nous voulons permettre à l'utilisateur d'avoir une forte interactivité avec le système et, par conséquent, éviter un effet "*boîte noire*" qui ne permet pas d'expliciter et de justifier les résultats. Le choix d'agréger des données commensurables est aussi important.

La Figure 3.2 donne un aperçu des trois étapes (du bas vers le haut) de la stratégie d'agrégation mise en œuvre dans notre approche.

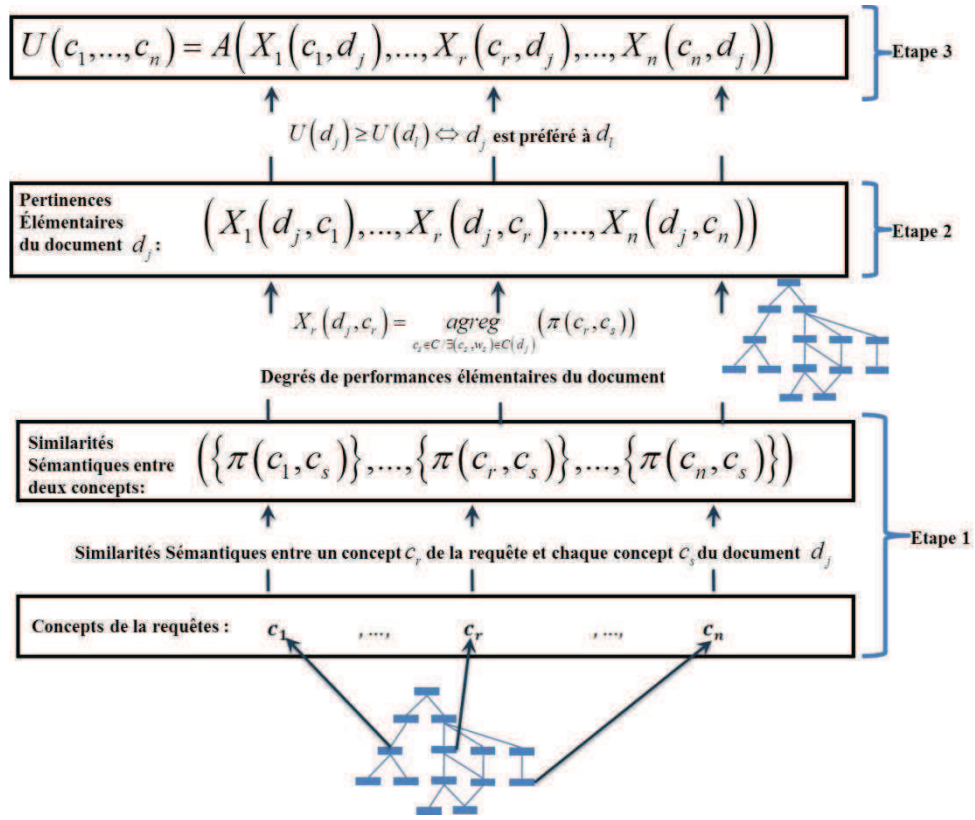


Figure 3.2 : Stratégies d'agrégation dans l'approche OBIRS en trois étapes

3.3. OBIRS : un modèle d'agrégation à trois niveaux

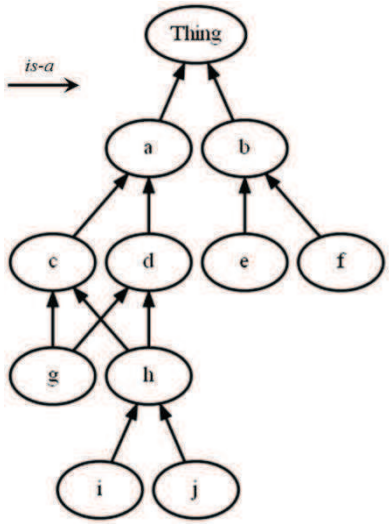
Détaillons maintenant l'approche *OBIRS* en explicitant la mise en œuvre de chacune de ses trois étapes (c.f. Figure 3.2) permettant d'évaluer la pertinence d'un document d_j, dont l'indexation conceptuelle est donnée par $C(d_j) = \{(c_s, w_s), s = 1..dl_j, w_s \in \mathbb{R}, c_s \in C\}$, relativement à une requête conceptuelle pondérée Q dont la représentation formelle est donnée par $\{(c_r, p_r), r = 1..n, p_r \in \mathbb{R}, c_r \in C\}$.

3.3.1. Similarité sémantique entre concepts d'une ontologie

La première étape de notre approche consiste à évaluer la similarité sémantique $\pi(c_r, c_s)$ entre un concept c_r de Q et un concept c_s indexant d_j. Le choix de la mesure de similarité sémantique à mettre en œuvre, durant cette étape, a un grand impact sur : i) la pertinence des documents sélectionnés, ii) le rappel du système et iii) la compréhension, par les utilisateurs, de la stratégie de sélection des documents. En outre, et afin de favoriser l'interaction de l'utilisateur avec le système, les mesures de similarités sémantiques doivent être intuitives, d'une part, de telle sorte que l'utilisateur puisse facilement les interpréter et rapides à évaluer d'autre part, de telle sorte que le SRI soit réactif même dans le cas d'ontologies de grande taille. Etant donné ces considérations, nous proposons une variante de la mesure de similarité sémantique de Lin (2.20) avec une évaluation du contenu informationnel d'un concept basée sur le nombre de ses hyponymes. Cette variante estime la similarité sémantique entre deux concepts, c_r et c_s, en considérant l'indice de Jaccard entre leurs ensembles d'hyponymes :

$$\pi_{JD}(c_r,c_s)=\begin{cases}\dfrac{|Desc(c_r)\cap Desc(c_s)|}{|Desc(c_r)\cup Desc(c_s)|} & \text{si } c_r\in Desc(c_s) ou\ c_s\in Desc(c_r)\\ 0 & \text{sinon}\end{cases} \quad \textbf{(3.3)}$$

avec $Desc(c_r)$; l'ensemble des descendants du concept c_r Notons que cette approche de la similarité sémantique entre deux concepts respecte la Définition 2.5 et donc que les valeurs $\pi_{JD}(c_r,c_s)$ sont comprises entre 0 et 1. $\pi_{JD}(c_r,c_s)=0$ si et seulement si c_r et c_s n'ont aucun hyponymes en commun tandis $\pi_{JD}(c_r,c_s)=1$ si et seulement si c_r et c_s sont identiques. La Figure 3.3 présente quelques valeurs de similarité suivant π_{JD}.



c_1	c_2	$\pi_{JD}(c_1,c_2)$
c	d	0
c	g	0,2
h	j	0,333
a	c	0,714

Figure 3.3 : Quelques valeurs de similarités sémantiques entre concepts de la restriction d'une ontologie aux relations de subsomption.

L'objectif poursuivi ici est double. Premièrement, en cas de silence, il s'agit de considérer, de manière automatique, d'autres concepts (hyponymes, hyperonymes) proches de ceux de la requête pour améliorer le rappel du système. Concrètement, un silence survient lorsqu'un concept de la requête n'est pas exactement présent dans un document d_j. Il s'agit d'un premier pas vers une stratégie d'expansion "*prudente*" d'une requête conceptuelle. En effet, la mesure de similarité π_{JD} permet de contrôler cette expansion et de la limiter aux concepts les plus proches de ceux de la requête. Deuxièmement, il est possible 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) (voir section 3.4.2). Il s'agit donc d'une première étape d'expansion globale de requêtes conceptuelles (c.f. section 2.5.3.1). Dans la Figure 3.4, nous illustrons l'expansion, induite par la mesure de similarité π_{JD} proposée, des concepts du *MeSH*, d'une requête conceptuelle (*Phosphoproteins, Neurofibromin 2*) non pondérée lorsque l'on évalue un document indexé, entre autres, par les concepts *BRCA1 Protein* (en rouge) et par *Tumor Suppressor Proteins* (en bleu).

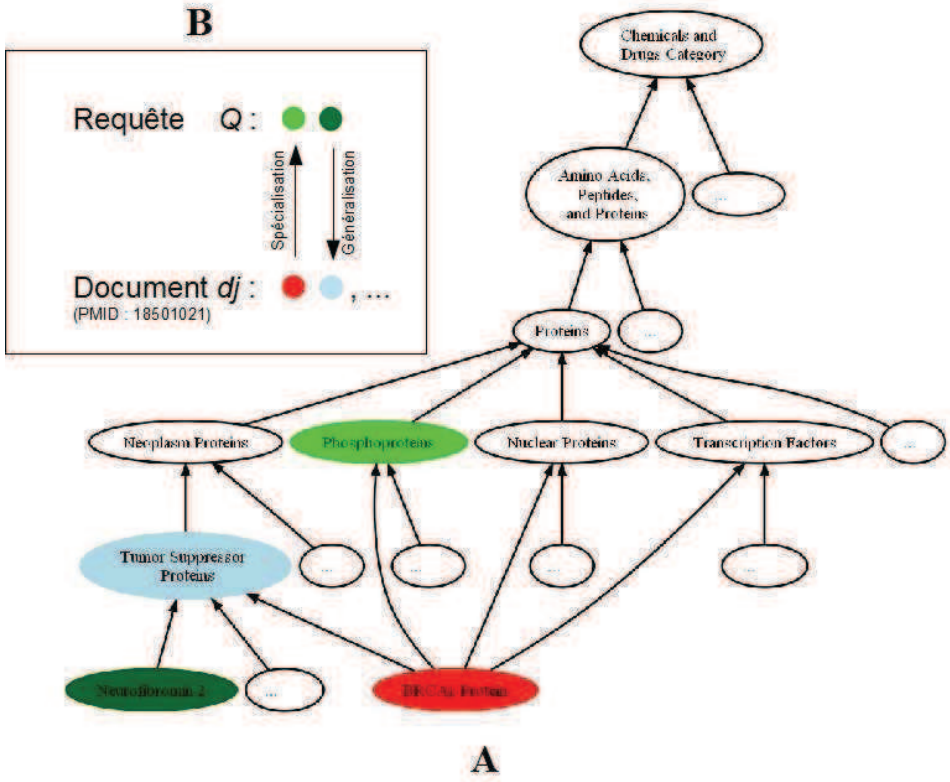


Figure 3.4 : (A) Illustration de l'utilisation d'une ontologie de domaine (ici un extrait de la catégorie "*Chimicals and* Drug *Category du MESH*") pour éviter le silence d'un SRI, concernant un document d_j indexé par les concepts en bleu et rouge, lorsqu'une requête conceptuelle Q (concepts verts) est soumise; (B) le silence est évité grâce à une expansion de Q vers son voisinage qui consiste ici au concept rouge pour l'hyponymie et au concept bleu pour l'hyperonymie.

Bien que cette évaluation de la similarité entre deux concepts soit très naturelle et ses résultats facilement compréhensibles par un utilisateur car ne faisant intervenir qu'une relation de subsomption, elle n'en demeure pas moins sévère dans le sens où deux concepts de même père (hyperonymes) ont une similarité nulle. Cependant, notre objectif est avant tout de favoriser les aspects de justification de la pertinence d'un document. D'autres mesures de similarités moins contraignantes et toujours basées sur la notion de contenu informationnel sont considérées dans notre stratégie de reformulation de requêtes conceptuelles (c.f. Chapitre 4).

Après l'évaluation de la similarité sémantique entre deux concepts, passons à l'évaluation de la proximité sémantique entre un concept d'une requête et un document d_j.

3.3.2. Pertinences élémentaires d'un document

La deuxième étape de notre approche consiste à évaluer, à partir de la similarité entre deux concepts, les pertinences élémentaires $X_r(d_j, c_r)$ d'un document d_j relativement aux concepts c_r de la requête Q (c.f. étape 2 dans la Figure 3.2). L'objectif est donc de calculer le vecteur de performances $\left(X_1(d_j, c_1), \ldots, X_r(d_j, c_r)\right)$ de chaque document d_j.

Nous proposons de considérer la pertinence élémentaire $X_r(d_j, c_r)$ comme étant la proximité sémantique entre le concept c_r de la requête Q et l'indexation conceptuelle $C(d_j)$ du document d_j. Le problème se ramène au calcul d'une proximité sémantique entre un concept et un groupe de concepts. Formellement, si $\pi(c_r, c_s)$ constitue la similarité entre c_r et c_s un concept indexant un document d_j, alors la proximité sémantique entre c_r et l'indexation conceptuelle $C(d_j)$ du document d_j est donnée par :

$$X_r(d_j, c_r) = \underset{(c_s, w_s) \in C(d_j)}{agreg} (\pi(c_r, c_s)) \quad \textbf{(3.4)}$$

Avec $agreg$ un opérateur d'agrégation. Plusieurs stratégies sont possibles concernant le choix de l'opérateur d'agrégation $agreg$. En effet, s'il est normal de pénaliser un document parce que l'un des concepts c_r (ou un autre qui lui est suffisamment proche) de la requête n'apparaît pas dans son indexation, le pénaliser parce qu'il est indexé par des concepts autres que ceux de la requête ne constitue pas une bonne stratégie. Typiquement, il s'agit de ne pas pénaliser les documents traitant de sujets multiples.

Cette considération conduit à définir la proximité sémantique entre le concept c_r et un document d_j comme étant le maximum des similarités sémantiques calculées entre c_r et chaque concept indexant d_j :

$$X_r(d_j, c_r) = \underset{(c_s, w_s) \in C(d_j)}{max} (\pi(c_r, c_s)) \quad \textbf{(3.5)}$$

Alternativement, un opérateur minimum (*min*) peut être préféré pour favoriser les documents qui traitent exclusivement du concept c_r de la requête et dont chaque concept de l'index est proche de c_r.

Considérer uniquement la sémantique, fournie par l'ontologie, dans l'évaluation de la proximité sémantique entre un concept de la requête et un document n'est pas suffisant. Le choix de l'opérateur *min* ou *max*, dans cette première étape d'agrégation, correspond à une intention de l'utilisateur (en tout cas de l'application) et ne dépend pas du modèle de connaissance utilisé.

3.3.3. Pertinence globale d'un document

Après avoir déterminé les proximités sémantiques, $X_r(d_j, c_r)$, entre chaque concept c_r de la requête Q et l'index conceptuel $C(d_j)$ d'un document d_j (c'est-à-dire l'ensemble des concepts indexant d_j), l'étape suivante (c.f. étape 3 dans la Figure 3.2) consiste à combiner les $X_r(d_j, c_r)$ en un score unique correspondant au *RSV* et qui reflète la pertinence globale de d_j par rapport à tous les concepts de Q. Il est important qu'une telle stratégie de combinaison puisse refléter la stratégie de décision de l'utilisateur en s'appuyant sur ses préférences. Ces dernières sont de deux ordres dans notre contexte. Premièrement, elles sont relatives aux critères correspondants aux concepts de la requête considérée. Il s'agit principalement de prendre en compte par l'intermédiaire d'une pondération, le fait qu'un utilisateur préfère les documents traitant d'un concept particulier de la requête plutôt qu'un autre. Deuxièmement, les préférences dont il est

question peuvent concerner la manière dont les scores élémentaires $X_r(d_j, c_r)$ sont combinés traduisant ainsi la stratégie de décision de l'utilisateur. Intégrer ces deux composantes du système de préférences de l'utilisateur permet au SRI de fournir des résultats différents à deux utilisateurs ayant des stratégies de décision différentes même si les concepts de leurs requêtes sont les mêmes. Pour une même structure de connaissance et une même mesure de similarité, la pertinence d'un résultat du SRI dépend encore de l'utilisateur et de ses besoins. Nous avons donc mis l'accent sur les opérateurs d'agrégation de type compromis. L'usage des préférences utilisateurs est donc un pré-requis de notre modèle de pertinence et nous visons l'amélioration de l'interaction entre notre système et l'utilisateur en lui fournissant un moyen simple et intuitif de représentation de telles préférences. Plus particulièrement, nous mettons en œuvre les moyennes quasi-arithmétiques pondérées définies dans la section (2.35) qui modélisent une stratégie de décision commune consistant à contraindre le score global (*RSV*) à être borné par le minimum des scores élémentaires et leur maximum (2.33). La famille d'opérateurs de Yager (c.f. Tableau 2.4) permet, en plus, de modéliser différentes stratégies de décision suivant différentes valeurs d'un seul paramètre (q) laissé au choix de l'utilisateur. Le score d'un document d_j relativement à une requête conceptuelle Q est donc donné par :

$$RSV(Q, d_j) = \left(\frac{1}{|Q|} \sum_{r=1}^{|Q|} \left(X_r(d_j, c_r) \right)^q \right)^{\frac{1}{q}}, q \in \mathbb{R}^* \qquad \textbf{(3.6)}$$

Lorsqu'une requête conceptuelle pondérée est considérée et donc que les concepts ne jouent plus un rôle symétrique (c.f. Définition 2.9), le *RSV* d'un document d_j devient :

$$RSV(Q, d_j) = \left(\sum_{r=1}^{|Q|} p_r * \left(X_r(d_j, c_r) \right)^q \right)^{\frac{1}{q}}, q \in \mathbb{R}^*, \sum_{r=1}^{|Q|} p_r = 1 \qquad \textbf{(3.7)}$$

Le Tableau 2.4 présente différents opérateurs obtenus à partir de la famille d'opérateurs de Yager suivant différentes valeurs du paramètre q. Nous les rappelons ici :

- $q = 1$, moyenne arithmétique
- $q = -1$: moyenne harmonique
- $q \to 0$: moyenne géométrique
- $q \to +\infty$: opérateur maximum (généralisation du "OU")
- $q \to +\infty$: opérateur minimum (généralisation du "ET")

En utilisant les deux stratégies correspondant aux équations (3.6) et (3.7), le choix d'un opérateur de compromis est simplement réduit au choix d'une valeur au paramètre q. L'utilisateur doit donc fournir à la fois une valeur pour le paramètre q et les poids relatifs des concepts constituant sa requête. Cette facilité nous permet de satisfaire notre objectif qui est de proposer un système simple et intuitif dans le sens où un simple curseur présenté à l'utilisateur lui permet de contrôler la valeur du paramètre q et d'indiquer s'il souhaite que la

stratégie d'agrégation tende vers un "*OU*" généralisé, vers un "*ET*" généralisé ou qu'elle tolère plus ou moins de compensation. Les poids p_r des concepts de la requête doivent être fournis en positionnant simplement un curseur pour chacun d'entre eux. La section 3.4.2 présente plus en détail la visualisation des résultats mise en œuvre.

Par souci de clarté et de simplicité, nous dénommerons par $RSV_q(Q, d_j)$ le score global d'un document d_j par rapport à une requête Q avec la préférence q de l'utilisateur.

3.4. Diagnostic et justification des résultats du modèle d'agrégation

L'un des objectifs majeurs de notre approche est d'éviter un effet boîte noire. Celui-ci survient lorsqu'aucune explication n'est fournie à l'utilisateur concernant l'évaluation, par le SRI, de la pertinence des résultats qui sont restitués. Nous proposons de mettre en œuvre deux solutions pour lever une telle limite. La première consiste en un modèle de diagnostic des résultats, facilité par le modèle analytique des préférences de l'utilisateur que constitue l'opérateur d'agrégation mis en œuvre avec l'équation (3.7). Il s'agit de mettre en exergue les concepts de la requête de l'utilisateur ayant le plus ou le moins contribués à la pertinence d'un document (section 3.4.1). La deuxième solution est d'ordre visuel et consiste à aller au-delà d'une simple présentation par liste des résultats (section 3.4.2).

3.4.1. Calcul de la contribution des concepts d'une requête

Lorsque l'on considère le modèle d'agrégation pondéré de l'équation (3.7) que nous avons mis en œuvre dans *OBIRS*, il est possible de mettre en place un modèle de diagnostic permettant de déterminer les concepts ayant contribué le plus au score élevé ou faible d'un document d_j. Dans un premier temps, considérons $\left(X_1(d_j, c_1), \dots, X_r(d_j, c_r), \dots, X_n(d_j, c_n)\right)$ un vecteur non nul et positif de degrés de pertinence élémentaires d'un document d_j relativement à une requête conceptuelle pondérée $Q = \{(c_r, p_r), r = 1..n, p_r \in \mathbb{R}, c_r \in C\}$. Rappelons que p_r correspond à la préférence de l'utilisateur au concept c_r. Dans un premier temps, nous allons estimer la contribution approximative d'un concept c_r d'une requête dans la faible valeur de pertinence d'un document d_j. Nous pouvons déterminer la dérivée partielle de la fonction $RSV_q(Q, d_j)$ par rapport au degré de pertinence élémentaire x_r du concept c_r :

$$\frac{\partial RSV_q(x_1, \dots, x_r, \dots, x_{|Q|})}{\partial x_r} = \frac{p_r * (x_r)^{q-1}}{\sum_{r=1}^{|Q|}(p_r * (x_r)^q)} * RSV_q(x_1, \dots, x_r, \dots, x_{|Q|}) \qquad \textbf{(3.8)}$$

Utilisons cette dérivée partielle pour donner le développement d'ordre 1 de la fonction à $|Q|$ variables $RSV_q(x_1, \dots, x_r, \dots, x_{|Q|})$:

$$RSV_q(x_1, \dots, x_r, \dots, x_{|Q|}) \approx RSV_q(1, \dots, 1) + \sum_{r=1}^{|Q|} \left(\frac{\partial RSV_q}{\partial x_r}(1, \dots, 1).(x_r - 1) \right)$$

$$RSV_q(1, \dots, 1) - RSV_q(x_1, \dots, x_r, \dots, x_{|Q|}) \approx \sum_{r=1}^{|Q|} C_r^{nonpertinence}, \quad \textbf{(3.9)}$$

$$C_r^{nonpertinence} = \frac{\partial RSV_q}{\partial x_r}(1, \dots, 1).(1 - x_r) = p_r * (1 - x_r)$$

Cette formule (c.f. Equation (3.9) permet d'exprimer approximativement le manque de pertinence d'un document d_j comme une somme de termes $C_r^{nonpertinence}$ tels que chacun d'eux ne dépend que d'un seul concept c_r. Notons que $\frac{\partial RSV_q}{\partial x_r}(1, \dots, 1)$ est une constante et que seul $p_r * (1 - x_r)$ dépend de c_r. La contribution de c_r, au manque de pertinence d'un document d_j, est d'autant plus grande que $C_r^{nonpertinence}$ est élevée. De plus cette contribution n'est fonction que du degré de pertinence élémentaire $x_r = X_r(d_j, c_r)$ du document d_j et de la préférence p_r de l'utilisateur pour le concept c_r. De la même manière, il est possible de déterminer les concepts c_r ayant contribué, le plus significativement, au score de pertinence d'un document d_j. Utilisons, pour cela, la formule suivante :

$$RSV_q(1, \dots, 1) - RSV_q(0, \dots, 0) + RSV_q(0, \dots, 0) - RSV_q(x_1, \dots, x_r, \dots, x_{|Q|}) =$$

$$1 + RSV_q(0, \dots, 0) - RSV_q(x_1, \dots, x_r, \dots, x_{|Q|}) \approx \sum_{r=1}^{|Q|} (p_r * (1 - x_r)),$$

$$RSV_q(x_1, \dots, x_r, \dots, x_{|Q|}) - RSV_q(0, \dots, 0) = 1 - \sum_{r=1}^{|Q|} (p_r * (1 - x_r)) = \quad \textbf{(3.10)}$$

$$\sum_{r=1}^{|Q|} (p_r * x_r) = \sum_{r=1}^{|Q|} C_r^{pertinence},$$

$$C_r^{pertinence} = p_r * x_r$$

Dans ce cas, plus $C_r^{pertinence}$ est grande, plus la contribution positive de c_r au score de pertinence de d_j est élevé.

Plus généralement, nous pouvons définir le niveau de contribution d'un concept c_r comme un pourcentage $\beta\%$ de la contribution globale et déterminer la plus petite valeur $r_0 > 0$ telle que :

$$Expl\left(RSV_q(1,\dots,1) - RSV_q(x_1,\dots,x_r,\dots,x_{|Q|}),\beta\right) \triangleq \sum_{r=1}^{r_0 \leq |Q|} \left(p_{(r)} * \left(1 - x_{(r)}\right)\right) \geq \beta\% * \sum_{r=1}^{|Q|} \left(p_r * (1 - x_r)\right) \quad \textbf{(3.11)}$$

Avec $(.)$ une permutation telle que $\left(p_{(r)} * \left(1 - x_{(r)}\right)\right) \geq \left(p_{(r+1)} * \left(1 - x_{(r+1)}\right)\right), r < |Q|$. Nous pouvons dès lors dire que les r_0 expliquent la non pertinence du document $\left(RSV_q(1,\dots,1) - RSV_q(x_1,\dots,x_r,\dots,x_{|Q|})\right)$ à hauteur de $\beta\%$. L'utilisateur peut jouer sur la précision de l'explication selon ses besoins. De la même manière, nous pouvons définir :

$$Expl\left(RSV_q(x_1,\dots,x_r,\dots,x_{|Q|}) - RSV_q(0,\dots,0),\beta\right) \triangleq \sum_{r=1}^{r_0 \leq |Q|} \left(p_{(r)} * x_{(r)}\right) \geq \beta\% * \sum_{r=1}^{|Q|} \left(p_{(r)} * x_{(r)}\right) \quad \textbf{(3.12)}$$

Les r_0 concepts ainsi sélectionnés expliquent la pertinence du document à hauteur de $\beta\%$.

L'outil de diagnostic que nous venons de présenter est un premier pas vers un système sans effet « boîte noire » dès lors que nous sommes capables d'identifier les concepts les plus contributifs positivement ou négativement au score de pertinence d'un document. L'analyse de sensibilité proposée est un moyen simple d'expliquer le résultat du calcul du *RSV* à l'utilisateur.

Pouvoir expliquer un résultat de manière calculatoire est une chose, pouvoir le faire de manière visuelle, pour l'utilisateur donc, en est une autre. La section suivante propose un modèle original de visualisation des résultats de notre modèle de pertinence.

3.4.2. Visualisation des résultats utilisant une sonde

La Recherche d'Information est souvent un processus interactif où l'utilisateur affine sa requête, sélectionne les documents qui l'intéressent et donne des indications au SRI concernant ses préférences. Pour être efficace, ce processus nécessite que l'utilisateur ait une compréhension précise concernant la pertinence des résultats qui lui sont présentés par le SRI.

Il est clair qu'une simple liste ne répond pas à nos objectifs en termes de justification et d'interactivité. Pour y répondre, notre approche sélectionne les documents pertinents, représente graphiquement leurs adéquations élémentaires aux concepts de la requête et les dispose sur une carte sémantique en fonction de leurs *RSV*.

Nous proposons de représenter chaque document d_j par un pictogramme dans lequel chaque concept de la requête est représenté par une barre verticale (voir Figure 3.5). Cet histogramme explicite, de manière synthétique, l'adéquation du document à la requête et fournit deux

informations de diagnostic intéressantes pour l'utilisateur. La première concerne le degrés de pertinence élémentaire, $X_r(d_j, c_r)$, du document d_j relativement à un concept c_r de la requête qui se traduit dans la hauteur de sa barre. Plus la barre correspondant à un concept est haute, plus son degré de pertinence élémentaire est important. La deuxième information, représentée par la couleur des barres, concerne la mise en exergue des concepts présents dans l'indexation conceptuelle d'un document ayant permis de le sélectionner. Chaque barre d'un concept c_r prend une couleur différente suivant que le concept de $C(d_j)$, le plus proche au sens de la proximité sémantique choisie (c.f. Equation (3.5), est exactement c_r (verte), un de ses hyponymes (rouge) ou un de ses hypéronymes (bleu). La barre est de couleur mauve dans les autres cas. La Figure 3.5 montre deux exemples de pictogrammes représentant deux documents tels qu'ils peuvent être visualisés dans la carte sémantique.

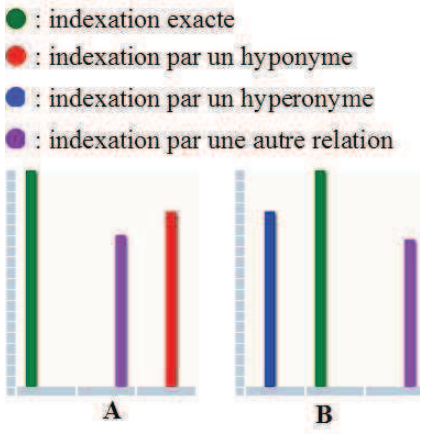


Figure 3.5 : Sémantique des couleurs utilisées pour expliquer l'appariement des concepts de la requête et ceux des documents

Notre approche sélectionne les documents pertinents et les dispose sur une carte sémantique où la requête est représentée par une « sonde sémantique ». Dans (Crampes & De Oliveira-Kumar 2010), une ***sonde sémantique*** est définie comme étant un objet particulier pour lequel un profil sémantique, pouvant être un ensemble de tags, est donné. Disposée sur une carte sémantique, une telle sonde génère un « champ sémantique » qui attire les documents avec une force d'autant plus importante qu'ils sont sémantiquement proches du profil sémantique de la sonde. Dans notre cas, les profils sémantiques des résultats et de la sonde, sont déduits de leurs indexations conceptuelles. Aussi le "champ sémantique" généré est simulé par les différentes valeurs de pertinences des documents. Il s'agit donc d'un champ scalaire où chaque point représente une valeur possible de pertinence. Les documents sont donc disposés de telle sorte que leurs distances géométriques relativement à la sonde sémantique (requête conceptuelle) soient proportionnelles à leur pertinence (*RSV*). Deux types de distances euclidiennes sont proposés à l'utilisateur pour l'affichage des résultats. La première utilise une distance euclidienne sur l'axe vertical (Figure 3.6). Les documents sur la même ligne ont une même valeur de pertinence.

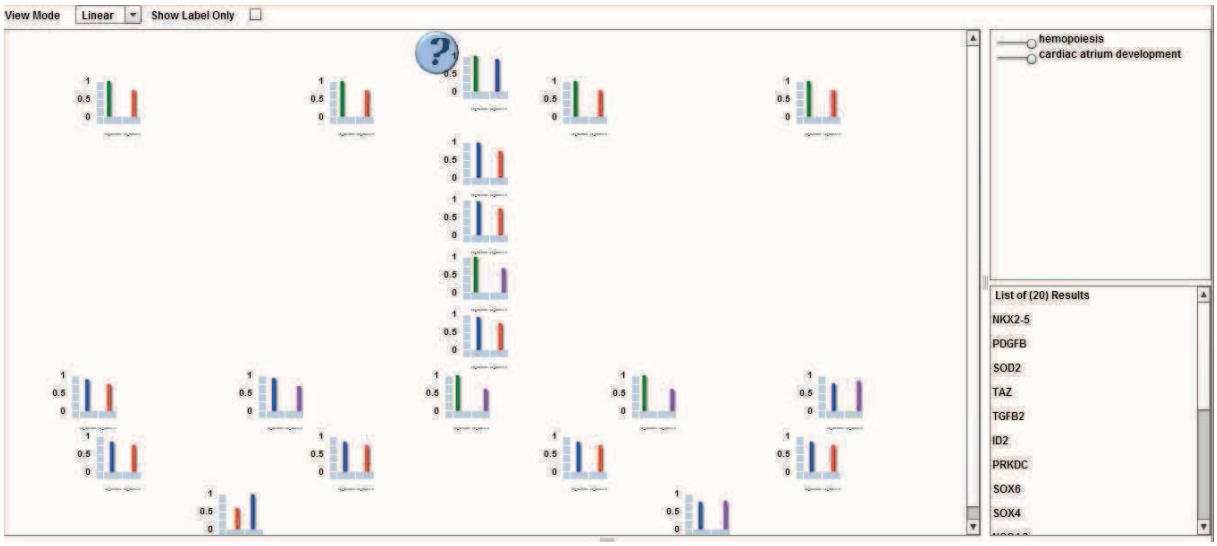


Figure 3.6 : Interface de visualisation des résultats d'une requête utilisant une sonde (point d'interrogation en haut au centre) avec un gradient linéaire des *RSV* selon l'axe vertical.

Le deuxième type de représentation utilise une distance euclidienne classique, portant sur les deux axes, ce qui conduit à un gradient radial des *RSV* (Figure 3.7). Les documents sur un même cercle (ayant la requête pour centre) ont une même valeur de pertinence.

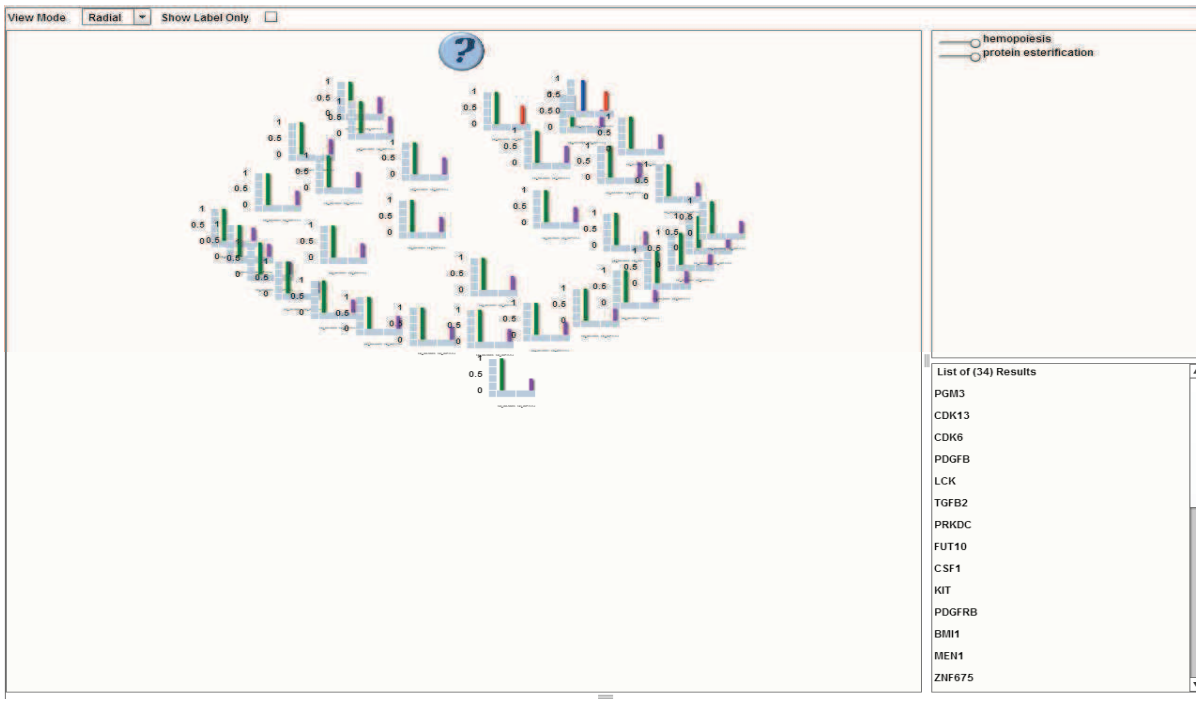


Figure 3.7 : Interface de visualisation des résultats d'une requête utilisant une sonde (point d'interrogation en haut au centre) avec un gradient radial des *RSV*.

3.4.3. Segmentation de textes comme justification fine des résultats dans le cas d'un corpus textuel

Les stratégies de justification mises en œuvre précédemment (sections 3.4.1 et 3.4.2) se focalisent sur les index conceptuels des documents et leurs similarités sémantiques relativement aux concepts et préférences utilisateurs exprimés sous forme de requêtes

conceptuelles éventuellement pondérées. Ce focus se traduit, à la fois, dans la représentation des résultats et dans la disposition graphique de ceux-ci. Cependant, aucune indication n'est fournie à l'utilisateur concernant la manière dont *sont traités les concepts* qui l'intéressent dans les documents qui lui sont présentés. Par « traiter d'un concept », nous faisons référence à la manifestation du concept au niveau lexical lorsqu'il s'agit de documents textuels (comment concrètement est abordé le concept dans le texte). Si on se réfère aux triangles de sens présenté dans la Figure 2.6, les concepts se manifestent dans un document textuel au travers de leurs signifiants ou symboles. Quand les concepts dont parlent les différents passages d'un document textuel sont connus, il est possible de mettre en exergue les passages qui traitent de concepts d'intérêts pour l'utilisateur. Prenons un exemple très général : « *Lorsque que Iseult rejoint Tristan dans la mort, un pied de vigne et un pied de lierre poussent enlacés sur les lieux du drame* » doit être un passage de *Tristan et Iseult* indiqué comme se référant au concept de l'« amour ». L'utilisateur comprendra dans l'exemple précédent que le lierre symbolise l'étreinte amoureuse. Il est souhaitable, lorsque c'est possible, que le système suggère lui-même des indications à l'utilisateur, lui indiquant les passages en lien avec sa requête. Une telle approche est particulièrement utile dans le domaine biomédical selon (Hersh 2005) où la veille scientifique constitue une part importante du métier de chercheur, et accéder rapidement à l'information constitue un enjeu crucial.

Nous proposons d'utiliser la mise en exergue de la présence lexicale des concepts d'une requête conceptuelle dans un document comme une stratégie de justification à un niveau plus fin (lexical) de l'adéquation de celui-ci avec la requête. Pour ce faire, il est nécessaire d'adjoindre à notre approche, un outil de segmentation de documents textuels guidée par une ontologie de domaine. L'analyse de l'état de l'art (c.f. section 2.5.2) montre que l'adjonction d'une composante lexicale est nécessaire pour une tâche de segmentation conceptuelle de texte (« lierre » est un élément du champ lexical rattaché à l'amour : sans la connaissance de ce lien concept-lexique, le passage de l'exemple précédent n'aurait pas été identifié comme relevant du concept « amour »). Notre objectif, dans cette partie n'est pas d'effectuer de la recherche de passages de documents comme dans le système *Ontopassage* (Lin et al. 2012) où les segments de texte sont indexés et classés relativement à la requête d'un utilisateur. Nous classons les documents uniquement en utilisant notre approche conceptuelle de RI, la segmentation associée n'a qu'une fin d'explication/justification qui vise à améliorer l'interaction homme/machine.

Nous considérons qu'une composante lexicale d'une ontologie de domaine, telle que définie dans Définition 2.13, permet d'établir un lien entre une ontologie d'où sont issus les concepts et le texte d'un document d'où sont issus les passages. Dans notre approche, nous nous limitons à construire et à attacher un "*lexique*",L_C, à chaque concept d'une ontologie de domaine Θ_{DAG} et à déterminer, dans les documents renvoyés par le SRI, les passages qui traitent de chaque concept de la requête, *i.e.* les passages dans lesquels les termes du lexique sont significativement présents. Il s'agit donc, d'une part, d'une méthode d'enrichissement d'une ontologie ne changeant pas sa structure telle que définie dans (Huang et al. 2010) et d'une méthode de segmentation de textes. Nous proposons d'adapter, pour réaliser les deux tâches précédentes (calcul des lexiques associés aux concepts et segmentation à proprement

parler), l'approche de segmentation de texte *SYNOPSIS* (Duthil et al. 2011). Le SRI conceptuel *OBIRS* couplé à l'approche de segmentation de textes *SYNOPSIS* est dénommé *CoLexIR* (*Conceptual and Lexical Information Retrieval*) (S. Ranwez et al. 2012). La Figure 3.8 schématise l'architecture de ce couplage.

Dans *CoLexIR*, la stratégie de construction d'une composante lexicale d'une ontologie de domaine et celle de segmentation de textes s'effectuent hors ligne (*offline*) en une seule fois. En effet, comme souligné dans *Ontopassage* (Lin et al. 2012), segmenter un texte relativement à une requête d'un utilisateur prend beaucoup de temps et ralentit donc l'activité de recherche. En plus, il n'est pas raisonnable de répéter, pour un concept donné, les opérations de construction de son lexique et de segmentation de textes propre au concept. En pratique, les passages et les concepts auxquels ils se rapportent sont donc stockés dans une SGBD relationnel autorisant un accès par de simples requêtes SQL. La RI conceptuelle, quant à elle, s'effectue en ligne (*inline*) car dépendant des besoins et des préférences de l'utilisateur qui varient au cours du temps. Détaillons maintenant l'adaptation que nous avons effectuée de l'approche *SYNOPSIS* que nous présentons par la même occasion.

Etant donné un concept c_r d'une requête Q, l'adaptation de l'approche *SYNOPSIS* que nous proposons tient en deux phases :

- construction automatique d'un lexique L_{c_r} associé à un concept c_r : ce lexique contient un ensemble de termes qui caractérisent c_r (sa classe, $classe_r$) et un ensemble d'autres termes qui ne le caractérisent pas (son anti-classe, $anticlasse_r$). Ces termes sont appris à partir d'un jeu de documents Web d'apprentissage obtenus grâce à un moteur de recherche comme Google ,
- extraction des passages d'un document relatifs à c_r.

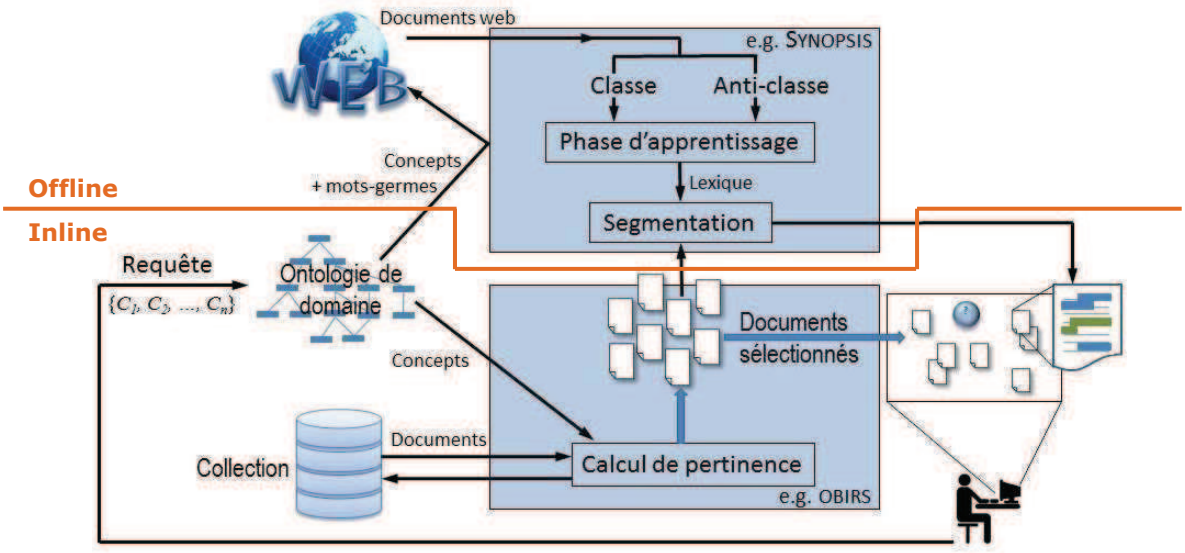


Figure 3.8 : Architecture globale du prototype *CoLexIR* combinant le prototype *OBIRS* comme SRI conceptuel et l'approche *SYNOPSIS* comme outil de segmentation de textes.

3.4.3.1. Construction d'un lexique associé à un concept d'une ontologie de domaine

Deux étapes sont nécessaires pour la construction d'un lexique associé à un concept c_r : i) l'acquisition d'un corpus d'apprentissage pertinent associé à c_r ; ii) et la détermination des termes informatifs et significatifs à partir du corpus constitué et le calcul de leur représentativité.

Dans cette problématique de segmentation contrôlée par les concepts d'une ontologie, notre contribution se situe dans l'acquisition d'un corpus d'apprentissage pertinent pour chacun des concepts. Nous avons mis en place une stratégie pour construire l'ensemble des mots germes caractérisant un concept et à partir duquel le corpus d'apprentissage est créé. Toutes les autres étapes du processus de segmentation dépendent exclusivement de l'approche *SYNOPSIS*. Nous ne faisons que les exposer pour une meilleure compréhension de *CoLexIR*.

Acquisition d'un corpus d'apprentissage relatif à un concept

L'acquisition d'un corpus d'apprentissage d'un lexique d'un concept c_r s'effectue sur la base d'un domaine dom spécifique, définissant le champ de recherche, et de mots germes caractérisant le concept d'intérêt. Le domaine en question est celui dont la connaissance est modélisée dans l'ontologie. Dans l'exemple illustré dans la Figure 3.9, l'ontologie est le *MeSH* et le domaine considéré est "*cancer*". Une fois le domaine précisé, nous déterminons un ensemble de mots caractérisant le concept c_r. Pour ce faire, nous considérons qu'il est représenté par l'ensemble de ses hyponymes traduisant, chacun, une spécialisation donnée. Suivant une telle hypothèse, l'ensemble, $germes(c_r)$, des mots germes associés à c_r est donné par :

$$germes(c_r) = \begin{cases} Desc(c_r) & si\ |Desc(c_r)| > 1 \\ Desc(c_r) \cup Anc(c_r) & sinon \end{cases}$$

Il peut arriver qu'un concept soit une feuille (i.e. $Desc(c_r) = c_r$). Dans ce cas, nous considérons que le concept est très informatif en termes de contenu informationnel (i.e $IC(c_r) = 1$ au sens de Seco (Seco et al. 2004) **(** 2.19)) et nous sélectionnons les ancêtres les moins génériques (i.e. dont l' *IC* est maximal). La figure Figure 3.9 montre un exemple de mots germes associés au concept "*dna*" issu du *MeSH*. Pour ce concept, nous avons $germes(c_r) = \{"dna", "dna, z-form", "dna, plant", \dots\}$.

Pour chaque mot germe $g_i \in germes(c_r)$, le système *SYNOPSIS* recherche sur le Web un nombre n de documents contenant à la fois g_i et le domaine ("*cancer*" par exemple). Cet ensemble de $|Desc(c_r)| * n$ documents constitue le corpus, $corpus_r$, associé à la classe du concept c_r. Dans notre adaptation, nous avons considéré $n = 300$.

De manière similaire pour construire le corpus $anticorpus_r$ associé à l'anti-classe de c_r, le système *SYNOPSIS* constitue à partir du Web un ensemble de n documents du domaine choisi ("*cancer*" par exemple) et ne contenant aucun des mots germes g_i de c_r. Considérer l'anti-classe conduit à mieux caractériser le concept. En effet, lorsqu'un terme appartient à la fois à

la classe et à l'anti-classe d'un concept alors nous pouvons raisonnablement considérer qu'il n'est pas discriminant pour celui-ci.

Une fois les corpus d'apprentissage constitués, il s'agit maintenant d'en extraire les mots significatifs par rapport au concept c_r, i.e. à chacun de ses mots germes.

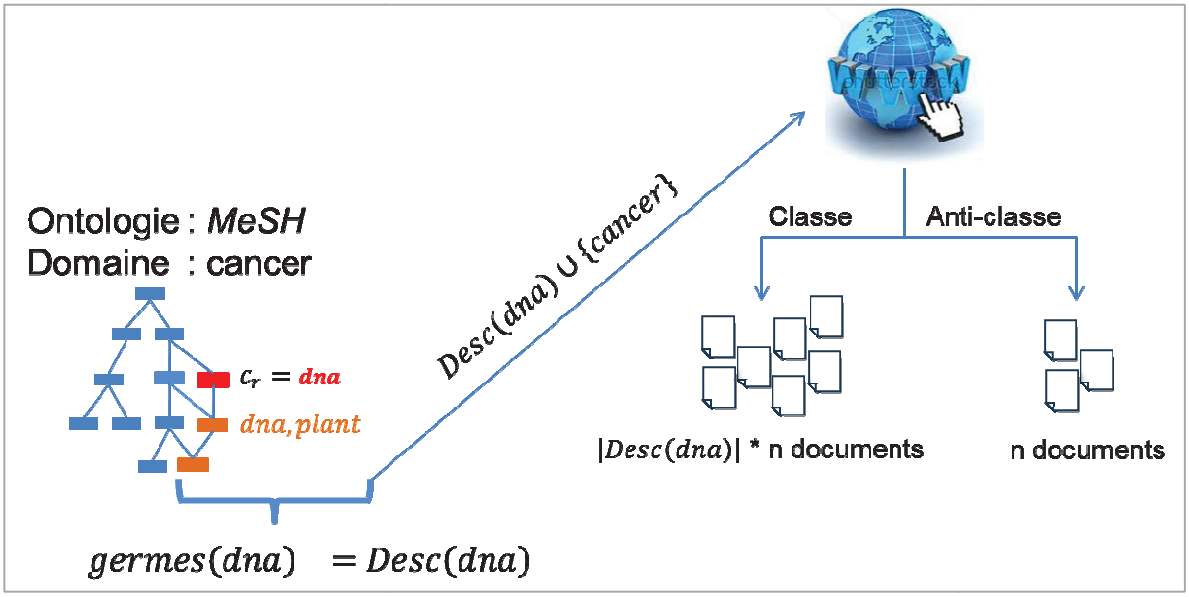


Figure 3.9 : Stratégie d'acquisition d'un corpus d'apprentissage d'un lexique d'un concept relativement à un domaine (cancer). L'union des labels du concept c_r et ceux de ses hyponymes est considéré comme l'ensemble de mots germes sur la base duquel la classe (ensemble de documents contenant un mot germe et le domaine) et l'anti-classe (ensemble de documents ne contenant aucun mot germe du concept c_r mais contenant le domaine) sont constituées.

Détermination des termes significatifs et calcul de leur représentativité

Cette étape relève de l'approche *SYNOPSIS* (Duthil et al. 2011). Etant donné un document html appartenant à l'ensemble $corpus_r$, les balises html, les données de publicité et d'autres bruits sont enlevés. Ensuite le document est transformé suivant un étiquetage morpho-syntaxique et une lemmatisation réalisés avec l'outil Tree Tagger (Schmid 1994). A ce stade, chaque document du corpus $corpus_r$ est nettoyé et ramené à une liste de termes lemmatisés, ce qui exclut les articles par exemple.

L'hypothèse fondamentale de *SYNOPSIS* est que la probabilité qu'un terme $term_j \in doc_{web}$ caractérise un concept c_r est proportionnelle à sa fréquence d'occurrence dans le voisinage immédiat d'un mot germe $g_i \in germes(c_r)$. Cette fréquence d'occurrence est évaluée dans tout le corpus d'apprentissage $corpus_r$. Afin d'évaluer cette fréquence, nous définissons la distance $dist_{nom}(mot_i, mot_j)$ qui correspond au nombre de termes de type (grammatical) « nom » (considérés ici comme les termes les plus significatifs à l'instar de (Kleiber 1996)) séparant les deux mots. Pour chaque document doc_{web} du corpus on peut alors déterminer, pour une proximité donnée sz, $occ(term_j, g_i, sz, doc_{web})$ le nombre d'occurrences d'une instance du terme $term_j$ à proximité d'une instance du mot germe g_i (i.e. le nombre d'instances de $term_j$ telles que $dist_{noun}(instance(g_i), instance(term_j)) \leq sz$).

Pour une taille de voisinage sz donnée, on peut ainsi définir un score $X(term_j, g_i, corpus)$ mesurant le lien observé entre $term_j$ et g_i dans un corpus :

$$X(term_j, g_i, corpus_j) = \sum_{doc_{web} \in corpus} occ(term_j, g_i, sz, doc_{web})$$

Nous pouvons maintenant déduire un score de représentativité de chaque terme $term_j$ par rapport à un concept c_r. Ce score, dénommé $Xc_r(term_j, sz)$, est la somme des scores du terme $term_j$ suivant les mots germes g_i au sein du corpus $corpus_r$ collecté pour la classe de ce concept :

$$Xc_r(term_j, sz) = \sum_{g_i \in germes(c_r)} X(term_j, g_i, sz, corpus_r)$$

Similairement, le score de représentativité $\bar{X}c_r(term_j)$ d'un terme $term_j$ dans l'anti-classe c_r est calculé relativement au domaine dom par :

$$\overline{Xc_r}(term_j, sz) = \sum_{g_i \in germes(c_r)} X(term_j, g_i, sz, anti\_corpus_r)$$

A partir des scores de représentativité de $term_j$ dans la classe et dans l'anti-classe du concept c_r, un score globale $Sc_r(term_j, sz)$ de représentativité est établi pour ce terme en utilisant la fonction de discrimination suivante :

$$Sc_r(term_j, sz) = \frac{\left(Xc_r(term_j, sz) - \overline{Xc_r}(term_j, sz)\right)^3}{\left(Xc_r(term_j, sz) + \overline{Xc_r}(term_j, sz)\right)^2}$$

La fonction cubique du numérateur permet une discrimination signée dans le sens où les termes $term_j$ du domaine dom qui ne sont pas représentatifs du concept c_r obtiennent un score négatif tandis que les termes représentatifs obtiennent un score positif. Cette fonction score est homogène à une fréquence. Grâce à la fonction score Sc_r, un lexique spécifique à un concept est construit, il contient tous les termes avec leur score (positif ou négatif) présents dans sa classe ou son anti-classe. C'est donc la polarité de $Sc_r(term_j, sz)$ qui déterminent l'appartenance à la classe ou l'anti classe. Nous allons étudier maintenant comment ce lexique peut être utilisé pour segmenter un document et y détecter les passages traitant d'un concept.

3.4.3.2. Extraction des passages d'un document relatives à concept

Considérons un document $d \in D$ à segmenter relativement à un concept c_r. et appartenant à la collection de documents renvoyée par notre SRI conceptuel *OBIRS*. Soit $\mathcal{F}'(noun, sz, d)$ une fenêtre glissante successivement centrée sur chaque « nom » $noun$ présent dans le texte d. A partir du lexique du concept c_r, un score est attribué à chaque fenêtre $\mathcal{F}'(noun, sz, d)$ de la manière suivante :

$$Score(\mathcal{F}'(noun, sz, d), c_r) = \sum_{term_j \in \mathcal{F}'} Sc_r(term_j, sz)$$

La fenêtre glissante $\mathcal{F}'(noun, sz, d)$ est considérée comme traitant du concept c_r si $Score(\mathcal{F}'(noun, sz, d), c_r)$ est supérieur à un seuil prédéfini.

Plus ce seuil est élevé, plus il est certain que les fenêtres retenues traitent du concept c_r. En revanche, les segments de texte retournés à l'utilisateur peuvent être peu nombreux (les segments pour lesquels la pertinence peut être garantie à un seuil donné se raréfient).

3.5. Applications et validation

Pour mettre en œuvre l'approche *OBIRS*, nous avons implémenté deux prototypes respectivement dénommé *OBIRS* et *CoLexIR*. Le premier, comme son nom l'indique, consiste en une implémentation du modèle de pertinence à trois niveaux présenté dans ce chapitre. Le second prototype adjoint à ce modèle une composante lexicale basée sur l'approche *SYNOPSIS* dans le but de l'utiliser pour la segmentation de textes telle que détaillée dans la section précédente (c.f. section 3.4.3).

Nous allons dans un premier temps présenter le prototype d'*OBIRS* que nous avons implémenté pour valider notre modèle de pertinence. L'objectif de cette présentation est d'exhiber les différents modules mis en œuvre depuis le chargement des données jusqu'à la réponse à une requête d'un utilisateur. Dans un deuxième temps, le prototype de *CoLexIR* est présenté. Pour finir, nous menons deux évaluations de nos prototypes sous forme de cas d'études concernant, d'une part, un scénario de recherche de gènes et un scénario de recherche de résumés d'articles traitant du cancer d'autre part. Notre objectif dans cette évaluation est de tester nos contributions en termes de visualisation, de justification et de diagnostic des résultats.

3.5.1. Prototype de l'environnement OBIRS

Le prototype correspondant à l'approche *OBIRS* a été développé principalement en Java et correspond plus exactement en une partie métier (*couche métier*) implémentant le modèle de pertinence, une couche de données gérant l'ontologie de domaine et la collection, et une partie client correspondant à l'interface de visualisation.

La couche métier est concrètement implémentée sous forme de librairie Java englobée dans un composant EJB et déployée dans un conteneur J2EE (Jboss par exemple) sous forme de service Web accessible. L'avantage d'un déploiement est que les données sont centralisées et aussi bien l'ontologie que les index sont chargés une seule fois. Le composant ne fait dès lors que répondre à des requêtes qui lui sont adressées via le Web. Il est aussi possible d'utiliser *OBIRS* comme une librairie pour pouvoir mettre en œuvre d'autres mesures de similarité sémantiques et d'autres modèles d'agrégation.

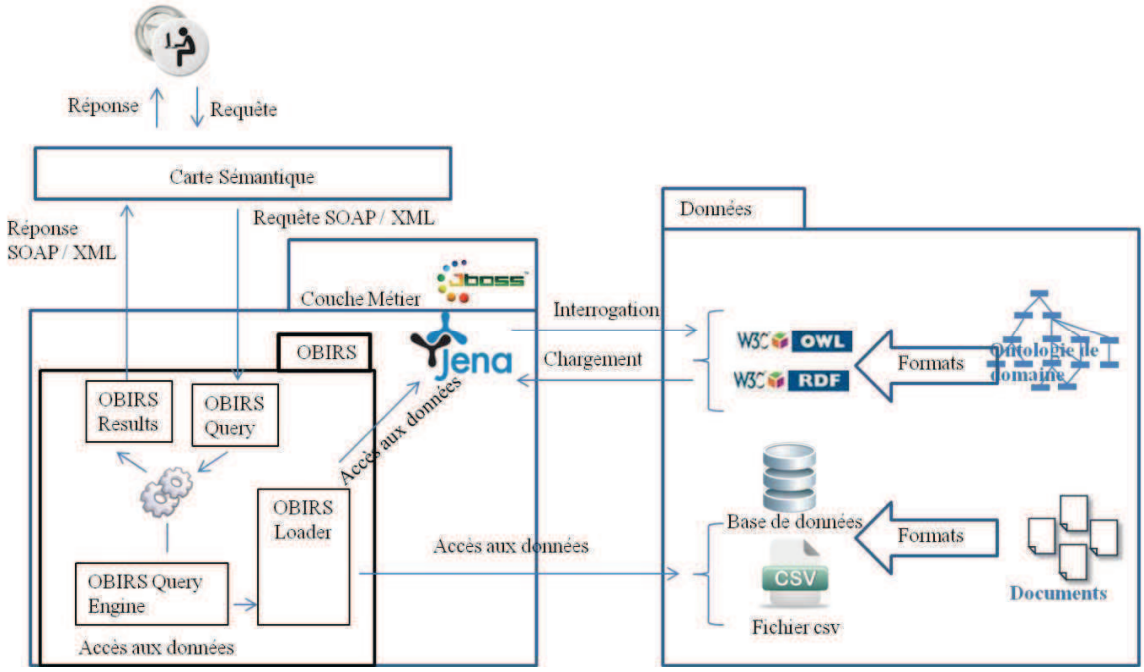


Figure 3.10 : Architecture globale du prototype *OBIRS*

OBIRS accède à l'ontologie via le module *OBIRS Loader* et l'outil open source Jena[26] qui permet de stocker et de gérer des données de types OWL et RDF. Par accès à une ontologie, nous entendons son chargement et le parcours de sa structure de graphe. Cependant, nous n'effectuons pas de requêtes SPARQL, étant entendu qu'un tel type de requêtes constitue de la recherche de données et non de la recherche d'information. Jena n'est qu'une interface pour pouvoir construire notre structure de graphe acyclique directe représentant la restriction que nous avons opérée dans la Définition 2.2.

Une matrice d'adjacence classique est utilisée pour stocker un tel graphe après avoir extrait de l'ontologie l'ensemble des relations de subsomption à l'aide de la propriété "*rdfs:subClassOf*". Cette dernière est la syntaxe standard de la relation de subsomption.
Les documents ainsi que les concepts qui les indexent peuvent être stockés dans des fichiers CSV ou dans une base de données relationnelle. Nous utilisons une structure de liste inversée pour les stocker.

La visualisation est effectuée à l'aide d'une interface Web réalisée avec la technologie *FLEX*. Nous avons déjà présenté les principales fonctionnalités de la visualisation que nous proposons.

La saisie des requêtes par des utilisateurs est détaillée dans les études de cas qui suivent. Nous avons opté pour deux types d'évaluation concernant l'approche *OBIRS*. Le premier type d'évaluation, traité dans ce chapitre, met à contribution des experts (des biologistes en l'occurrence) pour tester à la fois la pertinence des résultats que nous retrouvons et l'interface de visualisation que nous proposons. Dans ce cadre, nous décrivons deux études de cas concernant l'identification de gènes liés à une pathologie et une étude concernant une

[26] http://jena.apache.org/

recherche bibliographique concernant la prolifération des cellules que peut induire *BRCA1* (un gène impliqué dans le cancer du sein). Les ontologies utilisées dans cette première série d'évaluations sont le *MeSH* et la *Gene Ontology*. Le second type d'évaluation, quant à lui, est plus objectif et est basé sur la collection de documents biomédicaux *MuCHMORE* indexée par le *MeSH*. Ce même protocole étant également appliqué à l'évaluation de notre stratégie de reformulation de requête, l'ensemble des résultats obtenus sera présenté dans le Chapitre 4.

3.5.2. Cas d'études : application *d'OBIRS* à l'identification de gènes

Cette section décrit deux cas d'études pour tester la pertinence *d'OBIRS* dans la tâche d'identification de gènes. Ils ont été réalisés dans le cadre de notre partenariat avec des chercheurs de l'Inserm de l'ITMO IHP. Nous décrivons ici leurs retours concernant l'utilisation d'*OBIRS*. Durant le processus de génération des globules rouges (l'érythropoïèse), l'expression de plusieurs facteurs de transcription est requise dans les cellules souches pour induire leur différentiation. Nous avons utilisé *OBIRS* pour obtenir une liste de gènes codants pour des facteurs de transcription impliqués dans l'hématopoïèse humaine. Parmi les gènes impliqués, certains dont *GATA1*, *TAL1*, et *SP3* sont essentiels et connus ce qui permet d'évaluer la qualité des résultats obtenus.

Pour cette recherche, nous avons construit une requête formée de trois concepts : {"*erythrocyte development*", "*regulation of transcription*, *DNA dependent*", "*DNA binding*"}, utilisé la distance sémantique de Lin et restreint le nombre de résultats proposés aux 30 meilleurs gènes (c.f. Figure 3.11). La saisie de requêtes conceptuelles se fait à l'aide de termes spécifiques appartenant au vocabulaire de l'ontologie de domaine. Les labels des concepts de l'ontologie sont organisés en arbre binaire de recherche accélérant ainsi la recherche de concepts correspondant aux termes saisis par un utilisateur.

Les 30 premiers gènes trouvés (c.f. Figure 3.12) sont connus et ont donc chacun un symbole (code de quelques lettres) qui leur est attribué par le *HUGO Gene Nomenclature Committee*. Parmi cette liste de 30 gènes, 22 sont directement liés à l'érythropoïèse. Les gènes restant sont impliqués dans des processus connexes notamment liés à la leucémie et à des étapes embryonnaires de la formation du sang. De plus, les 15 premiers gènes (ayant les meilleurs *RSV*), parmi lesquels *SP3*, *GATA1* et *TAL1*, jouent un rôle majeur dans l'hématopoïèse. En dépit du grand nombre de gènes humains dans la base de données UniProt (~45000 gènes) et de concepts dans la Gene Ontology (~30000), le résultat de la requête traitée dans ce cas d'étude est obtenu en quelques secondes.

Le second cas d'étude concerne le *poisson-zèbre*, un organisme modèle utilisé en agronomie pour étudier la réponse immunitaire des poissons aux virus. Il est choisi comme organisme modèle notamment parce que son génome est simple, entièrement séquencé et très bien annoté. Au cours d'infections virales, plusieurs gènes sont impliqués dans la réponse antivirale. Parmi ceux-ci, nous pouvons noter les gènes responsables de l'inflammation. Cependant, l'inflammation peut aussi être induite par d'autres conditions telles que des maladies auto-immunes ou des cancers. Nous avons utilisé *OBIRS* afin d'obtenir la liste des gènes connus pour être impliqués dans cette réponse antivirale.

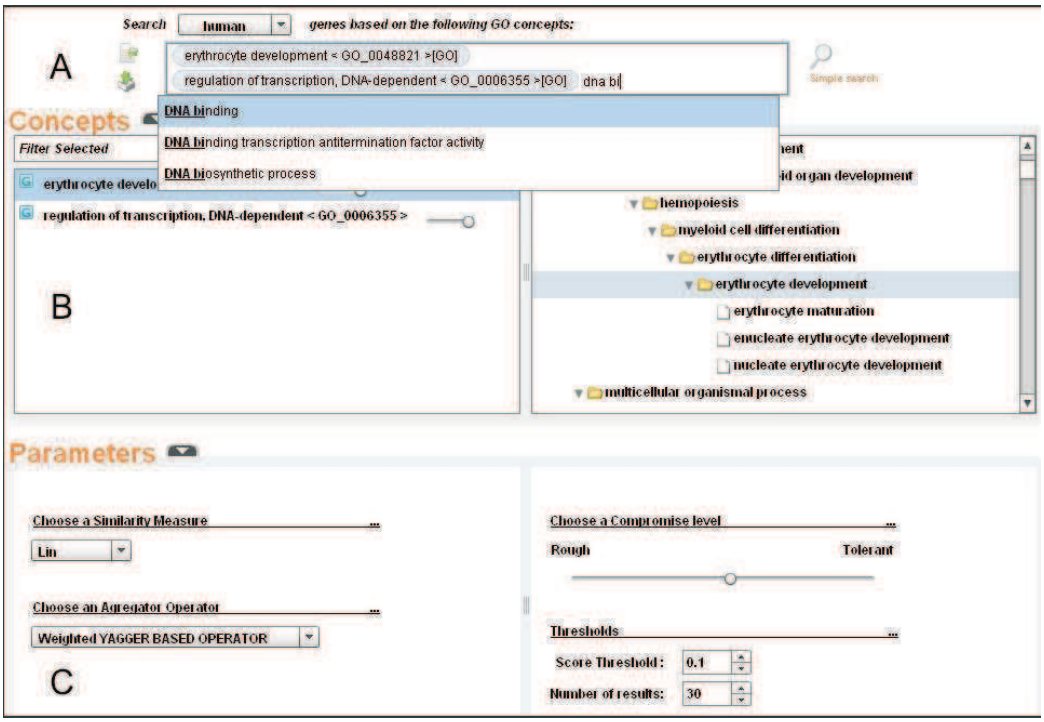


Figure 3.11 : Interface de saisie de requêtes conceptuelles *d'OBIRS* avec des fonctionnalités d'auto complétion (A) et de visualisation de la position de chaque concept de la requête dans le graphe induit par la taxonomie de l'ontologie (B). La partie (C) montre le panel de paramétrage de la requête.

Pour cela, nous avons construit une première requête formée de deux concepts non pondérés : {"*defense response to virus*", "*inflammatory response*"} tout en limitant le nombre de réponse à 20 gènes. La plupart des gènes retrouvés ont un grand intérêt mais certains, dont le gène *PXK*, ne sont pas directement liés à une réponse antivirale mais sont plutôt liés à un lupus. Le *lupus* est une maladie auto-immune induisant aussi une inflammation (Harley et al. 2008). Nous avons donc affiné la requête en donnant deux fois plus de poids au concept "*defense response to virus"* qu'au concept "*inflammatory response*". Comme prévu, le gène *PXK* n'est plus dans les premiers résultats. Les nouveaux résultats contiennent 19 gènes effectivement liés à une réponse virale en plus d'un locus (*LOC565099*) n'ayant pas de nom de gène officiel.

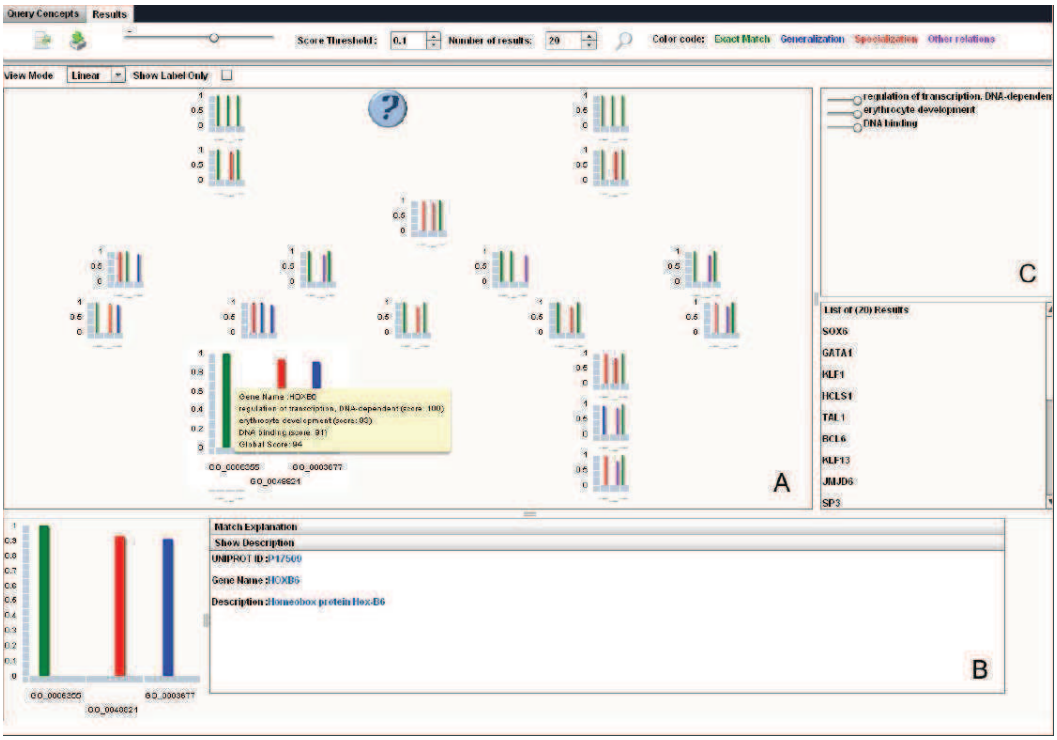


Figure 3.12 : Interface de visualisation des résultats *d'OBIRS*. Les gènes retrouvés, par la requête {"erythrocyte development", "regulation of transcription, DNA dependent", "DNA binding"}, sont présentés sur une carte sémantique (A) relativement à leur pertinence. Lorsqu'un gène est sélectionné (*HOXB6* dans cet exemple), des informations détaillées le concernant sont données (B). L'utilisateur peut modifier les poids relatifs des concepts de sa requête et voir la carte sémantique s'adapter en conséquence (C).

3.5.3. Cas d'études : recherche bibliographique autour de protéines limitant la prolifération de cellules que peut induire *BRCA1*

Nous décrivons, dans cette section, une étude de cas dans laquelle nous procédons à une étude bibliographique autour des protéines qui pourraient empêcher la prolifération de cellules que peut induire la protéine *BRCA1*. Dans cette étude, nous utilisons le système *CoLexIR* (c.f. 3.4.3) pour interroger un corpus contenant l'ensemble des publications (~2200) de *BMC Cancer* [27] entre l'année 2001 et l'année 2011 indexées par des concepts du *MeSH*. Notons que les publications de *BMC Cancer* sont disponibles sur le site Web de cette revue « open access » et que leur indexation est manuelle donc et est accessible via *PubMed*.

Nous avons formulé une première requête à l'aide de trois concepts du *MeSH* : {"*tumor suppressor proteins*", "*cell proliferation*", "*brca1 protein*"}. La présentation graphique des résultats (histogrammes et carte sémantique) permet de réaliser rapidement que la plupart de ces documents ne traitent pas de "*brca1 protein*" (du fait de leur faible score élémentaire pour ce concept). Un aperçu de quelques extraits d'articles retrouvés confirme ce premier constat selon lequel la requête formulée n'insiste pas suffisamment sur l'intérêt spécifique de notre étude de cas à savoir le gène *BRCA1*. Cette première formulation, utilisait la pondération par défaut qui attribue la même importance à tous les concepts de la requête. Nous l'avons donc reformulée en ajustant les pondérations des concepts : "*tumor suppressor proteins*" (0.25),

[27] http ://www.biomedcentral.com/bmccancer/

"*cell proliferation*“ (0.25) et “*brca1 protein*“ (0.5). Cette nouvelle pondération améliore la qualité des réponses et permet de retrouver plusieurs articles pertinents.

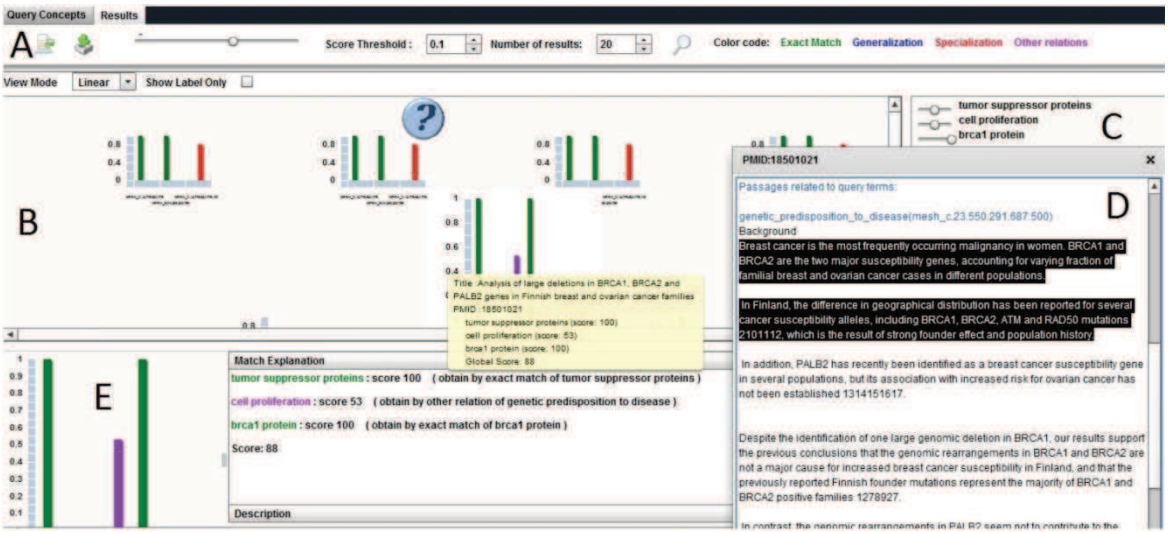


Figure 3.13 : Interface de visualisation du système *CoLexIR* couplant la visualisation par carte sémantique *d'OBIRS* et une stratégie de segmentation de textes pour mettre en exergue les passages traitant des concepts d'une requête.

Pour la plupart des articles sélectionnés, le processus de segmentation met en exergue des passages contenant des informations pertinentes qui parfois ne figurent pas dans le titre de l'article ni dans son résumé. Par exemple, dans (Pylkäs et al. 2008) le "*founder effect*" noté dans les études précédentes n'a pas été mentionné dans l'abstract mais il est retrouvé et mis en exergue par le processus de segmentation. Le même constat est valable concernant le fait qu'un réarrangement génomique entre *BRCA1* et *BRCA2* n'est pas un facteur déterminant de la prédisposition au cancer en Finlande. Une telle information, bien qu'omise du résumé, peut s'avérer fort utile pour quiconque s'intéresse à la distribution génomique des allèles du gène *BRCA* dans le cancer du sein. De même, dans (Friedenson 2007), plusieurs résultats importants concernant les gènes associés à la leucémie et au lymphome sont retrouvés par *CoLexIR* alors que ces informations ne sont pas dans le résumé, pourtant relativement long, de cet article sur le rôle du gène *BRCA1* dans des cancers autres que le cancer du sein. De même, un passage extrait par *CoLexIR* concernant l'interaction entre le gène *BRCA1* et les *protéines Fanconi* est particulièrement intéressant et pourrait fournir, aux chercheurs qui travaillent dans le domaine du cancer du sein et en immunologique, une base pour l'étude de cette interaction dans d'autres types de cancer. La Figure 3.13 montre les résultats renvoyés par *CoLexIR* pour cette requête et les passages mis en exergues pour l'article (Pylkäs et al. 2008).

3.6. Conclusion

L'approche de RI conceptuelle décrite dans ce chapitre constitue un pas important vers un SRI bénéficiant de la structuration des connaissances d'un domaine fournie par une ontologie, notamment dans sa dimension taxonomique, tout en restant d'un usage simple et intuitif. Nous avons proposé un calcul de pertinence des documents original mettant en œuvre un modèle d'agrégation des scores élémentaires de chaque concept de la requête. Ce calcul du *RSV* a

comme particularité d'intégrer les préférences de l'utilisateur. Les prototypes résultants, à savoir *OBIRS* et *CoLexIR*, permettent d'expliquer les raisons de la pertinence des résultats présentés à l'utilisateur suivant trois axes : i) décomposition du score de pertinence en scores élémentaires correspondant à la manière dont le document satisfait chacun des concepts de la requête ; ii) calcul de la contribution d'un concept au score global d'un document ; et iii) s'il s'agit de documents textuels, segmentation pour retrouver et mettre en exergue les passages traitant des concepts de la requête. Notre approche est, à notre connaissance, l'un des premiers SRI mettant en œuvre un tel niveau d'explication des résultats dans le but de favoriser l'interaction homme-SRI. Pour prendre en compte cette spécificité, nous avons évalué notre approche à travers des cas d'étude réels menés avec des biologistes.

Bien que notre système intègre l'utilisateur dans la *boucle de pertinence* à travers l'expression de ses préférences, nous pouvons aller encore plus loin dans cette prise en compte. Il s'agit notamment de lui permettre d'indiquer les documents qui l'intéressent et grâce à cela d'apprendre ses concepts d'intérêts pour enrichir et affiner la requête. Une telle stratégie constitue une reformulation de requête dans le sens où des sources d'évidences (issues de l'utilisateur ou non) sont exploitées en vue d'améliorer la compréhension du SRI sur le besoin en information de l'utilisateur. La reformulation de requête soulève d'importantes questions d'optimisation surtout lorsqu'elle est basée sur une ontologie de grande taille. La stratégie de justification mise en œuvre dans ce chapitre fournit des indications à des niveaux différents sur comment reformuler sa requête en vue d'obtenir de meilleurs résultats. Le Chapitre 4 présente une stratégie de reformulation de requête.

Chapitre 4 : Méthodes de réinjection de pertinence explicite utilisant une ontologie de domaine

Publications représentatives de ce travail

CORIA 2012 : Mohameth François Sy, Sylvie Ranwez, Jacky Montmain, Vincent Ranwez. "OBIRS-*feedback*, une méthode de reformulation utilisant une ontologie de domaine". CORIA 2012, 9e édition de la Conférence en Recherche d'Information et Applications, Bordeaux, France, 21-23 Mars 2012.

IEEE TKDE 2012 : Mohameth François Sy, Sylvie Ranwez, Jacky Montmain, Vincent Ranwez. "Efficient conceptual relevance feedback using connectivity of semantic neighborhood". IEEE Transactions on Knowledge and Data Engineering (soumis)

4.1. Introduction

Dans ce chapitre, une stratégie de reformulation de requêtes combinant une ontologie de domaine et des jugements utilisateurs est présentée. A partir d'une requête conceptuelle (c.f. Définition 3.2) et d'un ensemble de documents pour lesquels l'utilisateur a émis un jugement positif, notre approche consiste à rechercher une nouvelle requête qui optimise une fonction *objectif*. Cette dernière est définie de sorte à évaluer la capacité d'une requête reformulée à retrouver des documents jugés pertinents par l'utilisateur. Si les approches vectorielles utilisent de tels indicateurs depuis longtemps (Rocchio 1971), leur transposition à la reformulation conceptuelle n'est pas encore explorée à notre connaissance. L'état de l'art (c.f. section 2.5.3) montre, en particulier, que les solutions utilisant une ressource sémantique se contentent souvent de rajouter des concepts en relation (synonymie par exemple) avec ceux de la requête initiale.

Pour mettre en œuvre notre stratégie, nous utilisons la structure de *dag* (graphe acyclique directe) de la restriction d'une ontologie aux relations *is-a* (c.f. Définition 2.2) et nous mettons à jour deux propriétés simples et intuitives à partir desquels nous démontrons (c.f. 4.2.3.1) une propriété de convexité de la plupart des mesures de similarité sémantique respectant la Définition 2.5. Ensuite, nous présentons clairement des indicateurs de performance (c.f. 4.2.2) de requêtes conceptuelles au regard des documents d'intérêt de l'utilisateur et nous proposons des solutions heuristiques (c.f. 4.2.3.2) permettant une reformulation rapide même dans le cas d'ontologies contenant un très grand nombre de concepts.

Le modèle a été intégré à l'environnement *OBIRS* et les deux approches sont évaluées (c.f. 4.3) en utilisant la collection *MuCHMORE* et suivant le protocole TREC. Nous comparons notre stratégie au modèle PL2 (He & Ounis 2005).

4.2. OBIRS-*feedback* : un modèle conceptuel de réinjection de pertinence explicite

Dans cette section, nous présentons une approche originale de reformulation de requête conceptuelle, notée OBIRS-*feedback*, combinant une approche *locale* à travers une réinjection de pertinence explicite (*relevance feedback*), et une approche *globale* en exploitant une ontologie de domaine. Dans notre stratégie, la reformulation de requête conceptuelle est formalisée comme étant un problème d'optimisation conduisant à la recherche d'une nouvelle requête dite optimale, maximisant un indicateur bien défini. Nous définissons une famille d'indicateurs donnant un score aux requêtes conceptuelles candidates qui est d'autant plus grand que les requêtes sont sémantiquement proches des documents de D_p(documents jugés positifs) et éloignées de ceux de D_{np} (documents jugés négatifs). La composante "réinjection de pertinence" de notre approche adapte, dans le contexte conceptuel, la stratégie générale de Rocchio établie dans le modèle vectoriel, en exploitant la famille d'opérateurs d'agrégation de compromis de Yager (Yager 1979).

La dimension "*globale*" de l'approche OBIRS-*feedback*, quant à elle, permet de construire un espace de concepts d'intérêts dans lequel différentes requêtes conceptuelles candidates peuvent être construites et évaluées par rapport à l'indicateur défini. La construction d'un tel espace de concepts s'effectue en considérant les concepts indexant les documents positifs ainsi que leur voisinage sémantique. Dans les différentes approches globales de reformulation conceptuelle de la littérature, l'exploration du voisinage sémantique d'un concept d'intérêt pour l'utilisateur est considérée comme une tâche triviale et aucune indication précise n'est généralement fournie concernant sa mise en œuvre. Or définir et trouver tous les concepts similaires à un autre n'est pas une tâche si évidente, en particulier dans le cas d'une ontologie de grande taille. De plus, si l'on considère que le temps nécessaire pour mener une opération de reformulation est souvent un motif d'acceptation ou de refus d'une telle stratégie selon (Carpineto et al. 2012), il devient nécessaire de fournir des algorithmes ou des heuristiques à même de rendre cette recherche la plus rapide possible. Rappelons que le temps total nécessaire à la reformulation d'une requête est de deux ordres : i) le temps nécessaire pour générer des termes pour la reformulation (étape 2 sur la Figure 4.1) et ii) le temps nécessaire pour évaluer la requête reformulée (étape 4 sur la Figure 4.1). Nous proposons donc un

algorithme *(Algorithme 4.1*) rapide de construction du voisinage sémantique d'un concept d'une ontologie de domaine en formalisant et en exploitant les propriétés de connexité induites par certaines familles de mesures de similarité sémantique dans le graphe de l'ontologie.

Dans ce chapitre, nous supposons l'existence de requêtes conceptuelles en nous plaçant après les phases d'extraction de concepts d'une requête qui serait exprimée en langage naturel par exemple. De même, les documents des collections que nous traitons sont considérés comme étant indexés par des concepts issus d'une ontologie de domaine. La Figure 4.1 positionne les différentes stratégies mises en œuvre dans l'environnement OBIRS-*feedback* par rapport à la description générale de la reformulation qui a été présentée dans la Figure 2.15.

Le reste de la section est organisé comme suit. Nous commençons, dans la section 4.2.1, par introduire les notations, définitions et hypothèses que nous exploitons dans le cadre de l'approche OBIRS-*feedback*. Dans la section 4.2.2, nous présentons la formalisation de la reformulation d'une requête conceptuelle comme étant un problème de maximisation d'un indicateur que nous définissons. Pour identifier un sous-ensemble de concepts d'intérêts à partir duquel sont construites des requêtes conceptuelles candidates à la reformulation, nous introduisons un algorithme de détermination rapide du voisinage sémantique d'un concept d'une ontologie de domaine dans la section 4.2.3.1. Nous proposons un algorithme glouton permettant une construction incrémentale d'une requête approchant celle maximisant l'indicateur choisi dans la section 4.2.3.2.

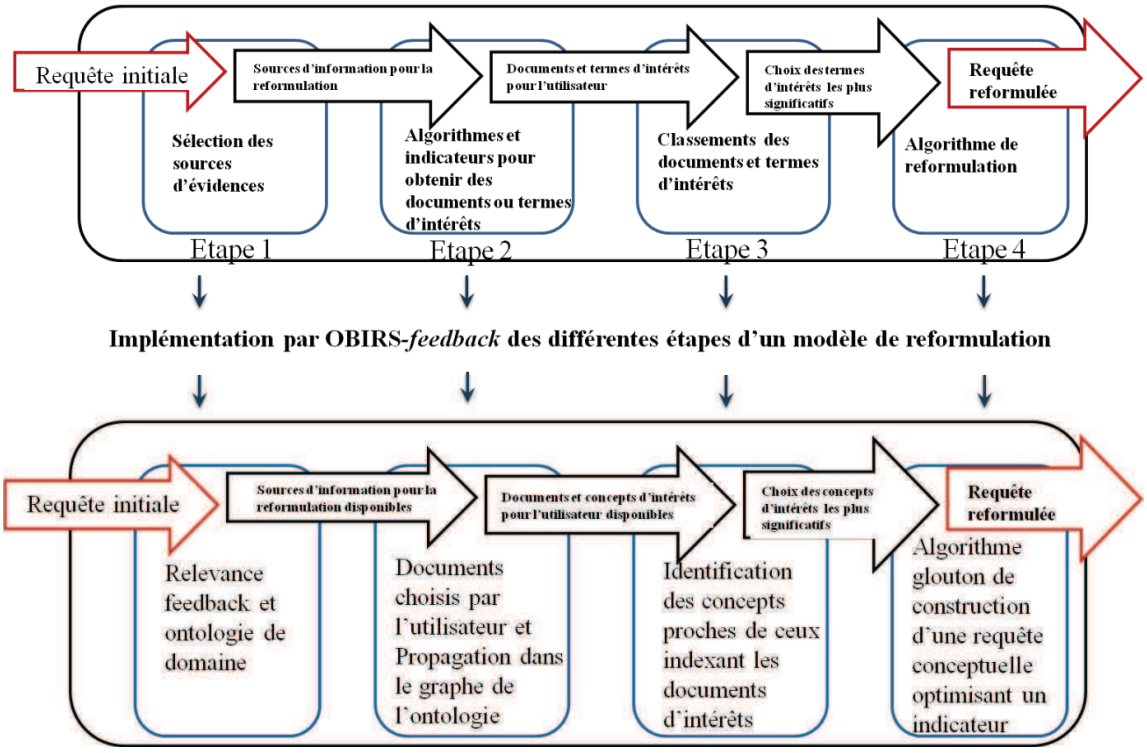


Figure 4.1 : Implémentation par OBIRS-*feedback* des différentes étapes d'un modèle de reformulation

4.2.1. Notations et définitions préliminaires

Dans cette section, nous introduisons les notations utiles pour la suite de ce chapitre. Nous réutilisons les notations déjà introduites dans les chapitres précédents concernant la définition de la restriction d'une ontologie aux relations de subsomption *is-a* (c.f. Définition 2.2 à la page 35), la définition de l'indexation conceptuelle d'un document ainsi que celle d'une collection de documents (c.f. Définition 3.1 à la page 71) et la définition d'une requête conceptuelle (c.f. Définition 3.2 à la page 71).

Une ontologie de domaine Θ restreinte aux seules relations *is-a*, que nous noterons $H_{isa} \subseteq H_C$, ainsi qu'à l'ensemble de ses concepts C peut être représentée par une structure de graphe acyclique direct (*dag*) ayant une racine et dont les nœuds sont les concepts de C et les arcs orientés sont les liens *is-a* dans H_{isa}. Notons un tel *dag* par $\Theta_{DAG} = (C, H_{isa})$. Un concept c_x est le *père* (respectivement le *fils*) d'un autre concept c_y si l'arc (c_y, c_x) (respectivement (c_x, c_y)) appartient à H_{isa}. Un concept c_x est un *ancêtre* (respectivement un *descendant*) d'un concept c_y s'il existe un chemin orienté $(c_y, c_x) - \text{path}$ (respectivement $(c_x, c_y) - path$) dans Θ_{DAG}. L'ensemble des concepts *père* de c_x (respectivement l'ensemble de ses *fils*) est noté $fathers(c_x)$ (respectivement $sons(c_x)$). Pour chaque nœud c_x, $Anc(c_x)$ (respectivement $Desc(c_x)$) représente l'ensemble de ses ancêtres (respectivement l'ensemble de ses descendants). Nous considérons par la suite que $c_x \in Anc(c_x)$ de même que $c_x \in Desc(c_x)$. En considérant la structure de *dag* hypothétique d'une ontologie de domaine représentée dans la Figure 4.2 (reprise ici pour la clarté des définitions présentées dans cette partie), $Anc(h) = \{h, c, d, a, Thing\}$ et $Desc(h) = \{h, i, j\}$. Rappelons que I_{c_x} est l'ensemble des instances du concept c_x. Nous pouvons alors remarquer que si $c_y \in Desc(c_x)$ alors $I_{c_y} \subseteq I_{c_x}$. De plus, si $c_z \in Desc(c_x) \cap Desc(c_y)$, alors $I_{c_z} \subseteq I_{c_y} \bigcap I_{c_x}$.

Tous les chemins entre deux concepts c_x et c_y dans Θ_{DAG} sont notés $paths_{\Theta_{DAG}}(c_x, c_y)$ et le plus court (en termes de nombre d'arcs) d'entre eux est noté $sp_{\Theta_{DAG}}(c_x, c_y)$ (pour *shortest path*).

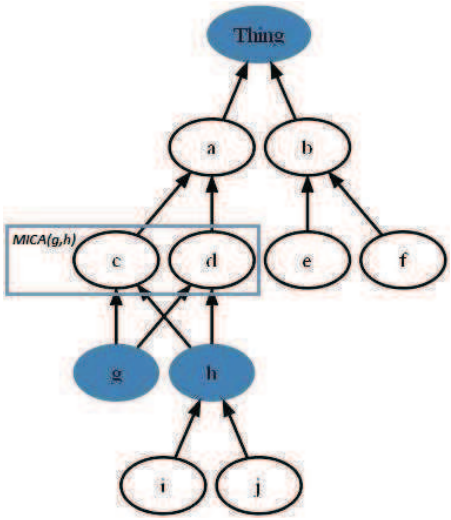


Figure 4.2 : un exemple hypothétique de la structure de dag d'une ontologie de domaine

4.2.2. Fonctions objectif pour la reformulation de requêtes conceptuelles

Dans cette partie, nous présentons un cadre formel pour la reformulation de requêtes conceptuelles.

Après la soumission d'une requête conceptuelle Q, le moteur de recherche retourne une liste de documents $D_{res} \subseteq D$. Parmi ces résultats, notons $D_{see} \subseteq D_{res}$ l'ensemble des documents que l'utilisateur a vus ou consultés, par $D_p \subseteq D_{see}$ les documents qu'il marque comme étant pertinents (documents positifs) et par $D_{np} \subseteq D_{see}$ les documents qu'il marque comme non pertinents (documents négatifs). Les deux ensembles D_p et D_{np} sont considérés comme étant disjoints ($D_p \cap D_{np} = \emptyset$).

Nous proposons de formaliser la réinjection de pertinence basée sur une ontologie de domaine comme un problème de recherche d'une requête conceptuelle Q_{max} optimisant un compromis entre sa proximité avec les documents positifs et son éloignement avec ceux négatifs. Un tel compromis est formalisé par un indicateur à valeurs réelles permettant d'évaluer la capacité d'une requête conceptuelle Q à respecter les jugements fournis par l'utilisateur. La réinjection de pertinence est, dès lors, vue comme un problème d'optimisation selon la définition suivante :

Définition 4.1 : Une requête reformulée optimale $Q_{max} \in \mathcal{P}(C)$ est une requête conceptuelle qui maximise une fonction objectif à valeur dans $\mathbb{R}$ $ind : \mathcal{P}(C) \times \mathcal{P}(D) \times \mathcal{P}(D) \rightarrow \mathbb{R}$:

$$Q_{max} = \underset{Q}{argmax} \left(ind(Q, D_p, D_{np}) \right) \qquad \textbf{(4.1)}$$

Le gain réel de la stratégie de réinjection de pertinence telle que formalisée ci-dessus dépend de la validité de la fonction indicateur ind choisie et de la possibilité de mettre en œuvre une méthodologie pour identifier la requête optimale Q_{max} correspondante, ou à défaut une requête conceptuelle qui lui est proche. Dans la suite, nous présentons et discutons des familles d'indicateurs ind ainsi qu'un algorithme permettant une identification rapide des concepts pertinents pour rechercher Q_{max}. Cette restriction de l'ensemble des concepts à tester se base sur deux propriétés simples des mesures de similarités sémantiques relatives à la connectivité du voisinage sémantique d'un concept donné.

Dès lors que l'objectif premier d'un modèle de reformulation est d'améliorer les performances d'un SRI en termes de précision et de rappel, une approche naturelle pour définir un indicateur $ind(Q, D_p, D_{np})$ est de baser le calcul de la pertinence des documents sur la mesure de ces critères de précision ou de rappel ou sur une combinaison des deux. Les valeurs globales (par rapport à toute la collection D) de rappel et de précision ne peuvent être calculées puisque l'ensemble des documents pertinents pour une requête donnée est inconnu. Cependant en considérant D_{see} comme un corpus dans lequel nous connaissons l'ensemble D_p des documents pertinents (au sens de l'utilisateur) et l'ensemble D_{np} des documents négatifs, les valeurs de rappel et de précision peuvent être estimées. L'ensemble D_{see} est vu comme une base d'apprentissage à travers laquelle la requête optimale Q_{max} peut être apprise.

En utilisant cette collection restreinte D_{see}, chaque requête candidate $Q \in \mathcal{P}(C)$ est évaluée en utilisant le modèle de pertinence du système de recherche d'information de base. Les documents de D_{see} sont réordonnés relativement à leurs scores (*RSV*) par rapport à la requête Q à évaluer. Si nbr est le nombre de documents pertinents (documents inclus dans D_p) effectivement retournés par une requête Q jusqu'à la $|D_p|$-ième position, alors l'indicateur est défini par :

$$ind(Q, D_p, D_{np}) = \frac{nbr}{|D_p|} \qquad \textbf{(4.2)}$$

Cet indicateur est maximal, c'est-à-dire vaut 1, si tous les documents classés jusqu'au $|D_p|$-ième document sont inclus dans D_p et sont donc des documents positifs. On estime ainsi la capacité d'une requête Q à classer en haut de la liste des résultats les documents que l'utilisateur a jugés pertinents. Cependant, cet indicateur permet une réinjection de pertinence ne tenant compte que des documents positifs.

Pour prendre en compte aussi bien les documents positifs que ceux négatifs, nous définissons l'indicateur suivant (Equation (4.3)) basé sur le schéma général de Rocchio :

$$\begin{aligned} ind(Q, D_p, D_{np}) = \quad & \alpha\, RSV(Q, Q_{init}) \; + \beta \underset{d_i \in D_p}{agregpos}\left(RSV(Q, d_i)\right) \\ & - \gamma \underset{d_j \in D_{np}}{agregneg}\left(RSV(Q, d_j)\right) \end{aligned} \qquad \textbf{(4.3)}$$

Q_{init} est la requête initiale et $(\alpha, \beta, \gamma) \in [0,1]^3$. $agregpos$ et $agregneg$ sont deux opérateurs d'agrégation, potentiellement différents et issus de la famille des opérateurs de compromis de Yager. Cet indicateur est général et nous permet de contrôler dans quelle mesure nous acceptons que la requête candidate Q dérive de la requête initiale Q_{init}. L'originalité de notre approche réside dans notre stratégie de sélection des concepts d'intérêts de l'utilisateur et de la solution heuristique que nous proposons pour l'exploiter afin d'optimiser l'indicateur choisi.

4.2.3. Algorithmes heuristiques pour la reformulation conceptuelle

Trouver la requête optimale Q_{max} nécessite de tester tous les sous-ensembles de concepts issus de l'ontologie de domaine Θ en tant que requêtes conceptuelles candidates. En considérant la restriction de l'ontologie à sa structure de *dag* $\Theta_{DAG} = (C, H_{isa})$, il y a $2^{|C|}$ requêtes candidates possibles et les évaluer toutes n'est pas réaliste. Il est donc nécessaire de trouver des heuristiques permettant de trouver une solution approchée de Q_{max} en un temps raisonnable étant donnée la nécessité de ne pas allonger les temps d'attente de l'utilisateur pour la prise en compte de ses retours d'information. Cette solution approchée est construite de manière gloutonne en retenant, à partir d'un ensemble de concepts de départ que nous noterons C_u ($C_u = C$ si nous considérons tous les concepts de l'ontologie) et de manière itérative, le concept qui maximise localement l'indicateur ind choisi.

Même si cette stratégie, explore un sous-ensemble des $2^{|C|}$ requêtes candidates, elle reste encore gourmande en temps de calcul, étant donné que $|C|$ concepts sont testés à chaque itération. Pourtant, la plupart des concepts de Θ_{DAG} peuvent être ignorés car ils sont totalement sans rapport avec les besoins de l'utilisateur exprimés à travers ses jugements. Il convient donc de bien choisir l'ensemble C_u des concepts parmi lesquels nous pouvons construire les requêtes candidates. Nous proposons de construire l'ensemble C_u à partir de chaque concept indexant un document positif appartenant à D_p auquel nous ajoutons les concepts qui sont dans son voisinage sémantique. Plus formellement, nous définissons le voisinage sémantique $S_\pi(c_x, t)$ d'un concept $c_x \in C$ et de rayon t comme suit :

Définition 4.2 : (*voisinage sémantique d'un concept*) Considérons la restriction d'une ontologie de domaine Θ à sa structure de *dag* $\Theta_{DAG} = (C, H_{isa})$, soit π une mesure de similarité sémantique, $c_x \in C$ un concept et $t \in [0,1]$ un réel. Le voisinage sémantique de c_x de rayon t est l'ensemble des concepts $S_\pi(c_x, t)$ tel que :

$$\begin{array}{ccc} S_\pi : C \times [0,1] & \rightarrow & \mathcal{P}(C) \\ S_\pi(c_x, t) & & \{c_j \in C / \pi(c_x, c_j) \geq t\} \end{array} \quad \textbf{(4.4)}$$

L'ensemble $S_\pi(c_x, t)$ regroupe tous les concepts de l'ontologie proches de c_x à hauteur de t. Etant donnés la fonction S_π, $C(d_i)$ et $C(d_i)$ l'index du document d_i, nous définissons par extension l'ensemble C_u par :

$$C_u = \bigcup_{c_x \in \bigcup_{d_i \in D_p} C(d_i)} S_\pi(c_x, t) \quad \textbf{(4.5)}$$

Le seuil t détermine le sous-ensemble C_u de concepts à considérer durant la recherche de la requête optimale Q_{max}. En effet, si $t = 0$ alors tous les concepts de l'ontologie sont pris en compte ($C_u = C$) tandis que si $t = 1$ seuls les concepts indexant les documents positifs de l'ensemble D_p sont considérés. La valeur de t définit donc un compromis entre une rapidité de mise en œuvre ($t = 1$) et une exhaustivité ($t = 0$).

Le gain en termes de temps dans l'évaluation des requêtes candidates à partir de l'ensemble C_u n'est effectif que si le voisinage $S_\pi(c_x, t)$ d'un concept c_x se détermine rapidement, c'est-à-dire sans avoir à calculer la similarité de c_x avec tous les concepts de C. Cette configuration peut être évitée si l'ensemble $S_\pi(c_x, t)$ induit un sous-graphe connexe dans Θ_{DAG}.

Dans la section 4.2.3.1, nous caractérisons une famille de mesures de similarité sémantique basées sur le contenu informationnel, en étudiant les propriétés de connexité des voisinages sémantiques qu'elles induisent pour un concept donné d'une ontologie de domaine. Un algorithme (*Algorithme 4.1)* est ensuite présenté pour obtenir un tel voisinage. La section 4.2.3.2 présente l'algorithme glouton de reformulation que nous proposons.

4.2.3.1. Etude de la connexité de voisinages sémantiques de concepts induits par des mesures de similarité

Il s'agit de donner une caractérisation des mesures de similarité sémantique basées sur le contenu informationnel permettant d'assurer une connexité des voisinages sémantiques qu'elles induisent.

Soient $(c_x, c_y) \in C^2$ deux concepts de l'ontologie Θ_{DAG} et $\pi : C \times C \rightarrow [0,1]$ une mesure de similarité sémantique basée sur le contenu informationnel telle que définie dans Définition 2.5. Nous rappelons ici les propriétés élémentaires de π :

- $\pi(c_x, c_y) \geq 0$ (positivité)
- $\pi(c_x, c_y) = \pi(c_y, c_x)$ (symétrie)
- $\pi(c_x, c_x) = 1$ (identité)

Aux précédentes propriétés, nous ajoutons les deux propriétés (4.6) et (4.7) suivantes basées sur l'intuition que plus un concept c_h s'éloigne d'un concept c_x (sur lequel est centrée la mesure de similarité) sur un chemin non nécessairement orienté (le chemin $(c_x, c_y) - path$ dans la Figure 4.3 par exemple), plus la similarité $\pi(c_x, c_h)$ décroit :

$$\pi(c_x, c_y) \geq \pi(c_x, c_{y'}) \; si \begin{cases} c_y \in Anc(c_x) \; et \\ c_{y'} \in Anc(c_y) \end{cases} \quad \textbf{(4.6)}$$

$$\pi(c_x, c_y) \leq \pi(c_x, c_{y'}) \quad si \begin{cases} c_{y'} \in Anc(c_y) \; et \\ MICA(c_x, c_y) \cap MICA(c_x, c_{y'}) \neq \emptyset \end{cases} \quad \textbf{(4.7)}$$

Avec $MICA(c_x, c_y)$ les ancêtres communs de c_x et c_y les plus informatifs. Cette notion est déjà introduite dans Définition 2.4 mais rappelons sa définition :

$$MICA(c_x, c_y) = \left\{ \underset{c_j \in Anc(c_x) \cap Anc(c_y)}{\text{argmax}} (IC_j) \right\} \quad \textbf{(4.8)}$$

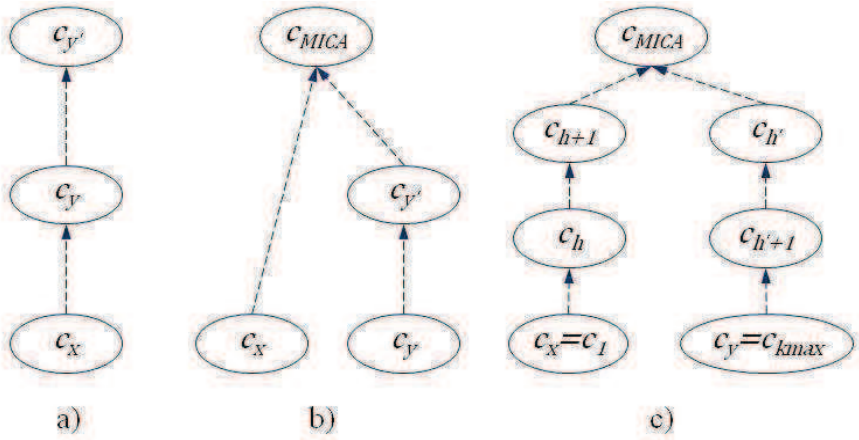


Figure 4.3 : Illustration des positions des concepts c_x

Il est facile de montrer que les principales mesures de similarité sémantique basées sur le contenu informationnel vérifient les propriétés précédentes. La mesure de Lin (D. Lin 1998) π_{Lin} dont nous rappelons la définition dans l'équation (4.9) respecte les propriétés définies dans les équations (4.7) et (4.8). En effet, la fonction $f(y) = (2 * y)/(a + y)$, a représentant IC_x (une constante donc), est une fonction croissante de y et donc induit la propriété (4.6). De manière similaire, la propriété (4.8) est induite par le fait que la fonction $f(y) = (2 * b)/(a + y)$, a étant toujours IC_x et constant et $b = IC_{MICA(c_x,c_y)} = IC_{MICA(c_x,c_y')}$.

$$\pi_{Lin}(c_x, c_y) = \frac{2 * IC_{MICA(c_x,c_y)}}{IC_x + IC_y} \qquad \textbf{(4.9)}$$

Jiang et Conrath (Jiang et al. 1997) ont proposé une distance sémantique entre deux concepts c_x et c_y comme étant $IC_x + IC_y - 2IC_{MICA(c_x,c_y)}$. Une manière usuelle de transformer cette distance en mesure de similarité sémantique est donnée par :

$$\pi_{j\&c}(c_x, c_y) = 1 - \frac{IC_x + IC_y - 2IC_{MICA(c_x,c_y)}}{2} \qquad \textbf{(4.10)}$$

Il est trivial de vérifier que la mesure de Jiang respecte les propriétés (4.6) et (4.7). Notons que les propriétés (4.6) et (4.7) sont assez générales et peuvent être satisfaites par des mesures de similarité sémantique non basées sur le contenu informationnel. Bien qu'elles soient simples et assez intuitives, il n'en demeure pas moins qu'elles sont suffisantes pour garantir la connexité du voisinage sémantique d'un concept étant donné un rayon fourni (*Proposition 4.1)*.

Lemme 4.1 : Soient $\Theta_{DAG} = (C, H_{isa})$ la restriction d'une ontologie de domaine à sa structure de *dag*, π une mesure de similarité sémantique basée sur le contenu informationnel définie sur Θ_{DAG} et $(c_x, c_y) \in C^2$ deux concepts de Θ_{DAG}. Soit $c_{MICA} \in MICA(c_x, c_y)$ un de leurs ancêtres communs les plus informatifs. En considérant le chemin $(c_x, c_y) - path$ dans Θ_{DAG}, non nécessairement orienté, et formé par l'union des deux chemins p_x et p_y orientés dans Θ_{DAG} et définis comme suit :

$$p_x = (c_x, c_{MICA}) - path = \{c_x = c_1, \dots, c_h, \dots, c_{h_{max}} = c_{MICA}\},$$
$$c_{h+1} \in fathers(c_h), 1 \leq h < h_{max}$$

$$p_y = (c_{MICA}, c_y) - path = \{c_{h_{max}} = c_{mica}, \dots, c_k \dots, c_{k_{max}} = c_y\},$$

$$c_{k+1} \in sons(c_k), h_{max} \leq k < k_{max}$$

alors $\forall 1 \leq h$ et $h + l < k_{max}$, $\pi(c_x, c_h) \geq \pi(c_x, c_{h+l})$.

Notons que cela implique que $\pi(c_x, c_h) \geq \pi(c_x, c_y)$.

Preuve : Pour prouver le lemme, il suffit de prouver le cas où $l = 1$. Cela signifie qu'il suffit de prouver que $\pi(c_x, c_h) \geq \pi(c_x, c_{h+1}), \forall 1 \leq h < k_{max}$ dès lors la propriété générale peut s'établir par déduction. Deux configurations sont envisageables suivant que l'arc (c_h, c_{h+1}) appartient à p_x où à p_y(Figure 4.3.c) :

$(c_h, c_{h+1}) \subseteq p_x$: dans ce cas de figure $c_h \in Anc(c_x)$ et $c_{h+1} \in Anc(c_h)$. Ce cas de figure est celui traité par la propriété donnée par l'équation (4.6 **)**. D'après cette propriété :

$$\pi(c_x, c_h) \geq \pi(c_x, c_{h+1})$$

$(c_{h'}, c_{h'+1}) \subseteq p_y$: donc $c_{MICA} \in MICA(c_x, c_{h'}) \cap MICA(c_x, c_{h'+1})$ et $c_{h'} \subseteq Anc(c_{h'+1})$. Cette configuration est celle décrite par l'équation (4.7 **)** et implique donc que :

$$\pi(c_x, c_{h'}) \geq \pi(c_x, c_{h'+1}) \square$$

Nous allons monter, en utilisant le *Lemme 4.1*, que si π est une mesure de similarité sémantique respectant les propriétés (4.6 **)** et (4.7 **)** alors le voisinage sémantique $S_\pi(c_x, t)$ centré sur le concept c_x et de rayon t induit un sous-graphe connexe dans Θ_{DAG}.

***Proposition 4.1* :** Soient $\Theta_{DAG} = (C, H_{isa})$ la restriction d'une ontologie de domaine à sa structure de *dag*, π une mesure de similarité sémantique satisfaisant les propriétés (4.6 **)** et (4.7 **)**, $c_x \in C$ un concept de Θ_{DAG} et $t \in [0,1]$ une valeur de similarité seuil. L'ensemble des concepts $S_\pi(c_x, t) \subseteq C$ est non vide et le sous-graphe $\theta_{DAG}[S_\pi(c_x, t)]$ induit par $S_\pi(c_x, t)$ est connexe.

Preuve : Il découle de la propriété d'identité que $S_\pi(c_x, t)$ est non vide puisque $\pi(c_x, c_x) = 1, \forall c_x \in C$ et donc $c_x \in S_\pi(c_x, t), \forall t \in [0,1]$.

Pour montrer que $\theta_{DAG}[S_\pi(c_x, t)]$ est connexe, nous devons montrer que $\forall c_y \in S_\pi(c_x, t)$, c'est-à-dire $\pi(c_x, c_y) \geq t$ et distinct de c_x, il existe un chemin $(c_x, c_y) - path$ non nécessairement orienté dans $\theta_{DAG}[S_\pi(c_x, t)]$.

Pour cela, considérons $c_{MICA} \in MICA(c_x, c_y)$ comme étant un des ancêtres communs les plus informatifs de c_x et c_y. Selon le *Lemme 4.1*, la concaténation des plus courts chemins $sp_{\Theta_{DAG}}(c_x, c_{mica})$ et $sp_{\Theta_{DAG}}(c_{mica}, c_y)$ est telle que $\forall c_h$ appartenant à cette concaténation :

$\pi(c_x, c_h) \geq \pi(c_x, c_y) \geq t \Longrightarrow c_h \in S_\pi(c_x, t) \square\blacksquare$

Une fois que nous nous sommes assurés de la connexité du voisinage sémantique d'un concept étant donnée une mesure de similarité sémantique $\boldsymbol{\pi}$ respectant les propriétés (4.6 **)** et (4.7 **)**, nous montrons l'exploitation de cette connexité dans la construction d'un tel voisinage.

Nous proposons l'algorithme *VoisinageSemantique* (*Algorithme 4.1*) de recherche de graphe pour construire l'ensemble C_u(voir l'équation (4.5 **)**) constituant l'espace de concepts à partir duquel les requêtes candidates à la reformulation sont construites. L'algorithme *VoisinageSemantique* recherche dans le graphe θ_{DAG} à partir d'un concept donné c_x indexant un document positif, i.e. appartenant à l'ensemble D_p. Chacun des concepts c_y adjacents à c_x dans θ_{DAG} (ses pères et fils en l'occurrence) est exploré (ligne 3) et sa similarité avec c_x est évaluée (ligne 4). Si cette valeur de similarité sémantique est supérieure au seuil t, alors le concept adjacent c_y est ajouté à C_u et l'exploration continue avec les concepts qui sont adjacents à c_y. Au contraire si la similarité entre c_x et c_y est inférieure au seuil t, alors le concept $c_x$est une impasse dans la traversée du graphe θ_{DAG} en vertu de la *Proposition 4.1*.

Nom : VoisinageSemantique
Entrée : Θ_{DAG}// *dag* de l'ontologie
t // une similarité limite appartenant à [0,1]
D_p // documents positifs
Sortie : C_u // un ensemble de concepts d'intérêts

$C_u \leftarrow \emptyset$ //queue de concepts
$S_{interest} \leftarrow \bigcup_{d_i \in D_u} C(d_i)$// concepts des documents de D_p
Pour chaque $c_x \in S_{interest}$
$S_{queue} \leftarrow \{c_x\}$ //le concept à explorer
$S_{visited} \leftarrow \emptyset$
1. $S_\pi(c_x, t) \leftarrow \{c_x\}$
2. **Tant que** $(S_{queue} \neq \emptyset)$ **Faire**
$c_{current} \leftarrow S_{queue}.pop()$
3. $S_{related} = fathers(c_{current}) \cup sons(c_{current})$
Pour chaque $c_{related} \in S_{related} \backslash S_{visited}$
Si $(\pi(c_x, c_{related}) \geq t)$ **Alors**
$S_\pi(c_x, t) \leftarrow S_\pi(c_x, t) \cup \{c_{related}\}$
$S_{queue} \leftarrow S_{queue} \cup \{c_{related}\}$
Fin Si
Fin Pour
$S_{visited} \leftarrow S_{visited} \cup \{c_{current}\}$
Fin Tant que
$C_u \leftarrow C_u \cup S_\pi(c_x, t)$
Fin Pour
return C_u

Algorithme 4.1 : Construction du voisinage sémantique d'un ensemble de concepts C_u

4.2.3.2. Algorithme de recherche heuristique d'une requête conceptuelle maximisant une fonction objectif

Nous présentons maintenant un algorithme glouton de recherche d'une requête conceptuelle (voir *Définition 4.1*) maximisant une fonction objectif telle que définie dans l'équation (4.3). La recherche commence par considérer une requête utilisateur conceptuelle Q potentiellement vide comme point de départ. Ensuite, il l'enrichit de manière itérative et à partir des concepts de l'ensemble C_u (4.3), construit un voisinage sémantique (ligne 1) en utilisant l'algorithme *Algorithme 4.1*. Ce voisinage permet une plus grande amélioration de la fonction objectif ind. Ce processus continue aussi longtemps qu'il subsiste un concept de C_u à même d'améliorer l'indicateur. La requête Q_{max} obtenue au final est une approximation de la requête optimale. Cet algorithme glouton a l'avantage d'être rapide mais il n'est pas complet car ne garantissant pas que la requête retournée Q_{max} est l'optimal globale relativement à ind.

Nom : RechercheRequetes
Entrée : D_{np}// ensemble des documents négatifs
D_p // ensemble des documents positifs
Q // la requête initiale
ind // fonction objectif à maximiser
θ_{DAG} // dag de l'ontologie
t // une similarité limite appartenant à $[0,1]$

Sortie : Q_{max} // la requête reformulée

---

1. $C_u \leftarrow VoisinageSemantique(\theta_{DAG}, t, D_u)$
2. $bestInd \leftarrow ind(Q, D_u, D_{see})$
3. $Q_{max} \leftarrow Q$

Faire
$improved \leftarrow false$
4. **Pour Chaque** $c_x \in C_u$
5. **Si** $(ind(Q_{max} \cup \{c_x\}, D_u, D_{see}) > bestInd)$ **Alors**
6. $bestC \leftarrow c_x$
7. $bestInd \leftarrow ind(Q_{max} \cup \{c_x\}, D_u, D_{see})$;
8. $improved \leftarrow true$
Fin Si
Fin Pour
9. **Si** ($improved$) **Alors**
10. $Q_{max} = Q_{max} \cup \{bestC\}$
Fin Si
Tant que ($improved$)
return Q_{max}

Algorithme 4.2 : Recherche heuristique d'une requête conceptuelle optimisant un indicateur choisi

Ces deux algorithmes ont été implémentés comme module de l'environnement *OBIRS*. Cependant, il n'est pas encore disponible dans l'interface de visualisation disponible à l'adresse : *http://www.ontotoolkit.mines-ales.fr/ObirsClient/*.

4.3. Expérimentation

4.3.1. Données expérimentales

L'évaluation de notre modèle a été réalisée en utilisant *MuCHMORE*, une collection de résumés d'articles scientifiques du domaine biomédical. Dans cette collection, un ensemble de requêtes a été constitué et pour chacune d'elles, des experts du domaine ont confectionné une liste de documents pertinents. Les résumés de la collection *MuCHMORE* sont des textes en anglais ayant une taille homogène indexés par des concepts de la version 2001 du *MeSH* (*Medical Subject Headings*, c.f. section Figure 2.8). Cette version du *MeSH* n'étant plus disponible, les indexations des documents ont été mises à jour sur la base de la version 2010 du *MeSH*. Nous avons établi avec succès la correspondance de 80% des concepts trouvés dans les indexations originelles des documents avec des concepts de la version 2010 du *MeSH*. Le processus d'indexation conceptuelle de ces documents sort du cadre de cette thèse ; le lecteur intéressé pourra trouver plus de détails sur cette étape dans la documentation disponible gratuitement sur le site Web de *MuCHMORE*[28]. Le Tableau 4.1 présente les données statistiques pertinentes concernant la collection.

Corpus MuCHMORE (Springer)	
Type de jugement	Par des experts
Taille du corpus $\|D\|$	7823
Nombres de requêtes	23
Taille moyenne :	
de vl'indexation d'un abstract $\|C(d_i)\|$	12
de la longueur d'une requête $\|Q\|$	2.3
du nombre moyen de documents pertinents $\|D^Q\|$	13.4

Tableau 4.1 : Statistiques descriptives de la collection *MuCHMORE*

Nous allons montrer (c.f. Tableau 4.2) un extrait de l'indexation du document "*Arthroskopie.00130003.eng.abstr"* de la collection. Dans cet extrait, la phrase *"The posterior cruciate ligament (PCL) is the strongest ligament of the human knee joint.*" est reliée à des concepts (balises <*concept*> et <*msh*>) du *MeSH* au travers des termes (balises <*token*>) qui le constituent. Par exemple, le terme "*posterior"* (*token* d'identifiant "*w2*") est associé au concept "*Pituitary Gland, Posterior*".

[28] http ://muchmore.dfki.de/pubs/D4.1.pdf : consulté le 31 août 2012

```
<document id="Arthroskopie.00130003.eng.abstr" type ="abstract" …>
        <sentence id="s1" corresp="s1">
         <umlsterms>
          <umlsterm id="t1" from="w2" to="w2" >
          <concept cui="C0032009" preferred="Pituitary Gland, Posterior" tui="T023">
                  <msh code="A6.407.747.734"/>
                  <msh code="A8.186.211.730.385.357.352.600.734"/>
          </concept>
          </umlsterm>
          …
          </umlsterms>
          <text>
              <token id="w1" pos="DT" lemma="the">The</token>
              <token id="w2" pos="JJ" lemma="posterior">posterior</token>
              <token id="w3" pos="JJ" lemma="cruciate">cruciate</token>
              <token id="w4" pos="NN" lemma="ligament">ligament</token>
             <token id="w5" pos="PUNCT">(</token>
              <token id="w6" pos="NN">PCL</token>
              <token id="w7" pos="PUNCT">)</token>
             <token id="w8" pos="VBZ" lemma="be">is</token>
              <token id="w9" pos="DT" lemma="the">the</token>
              <token id="w10" pos="JJS" lemma="strong">strongest</token>
              <token id="w11" pos="NN" lemma="ligament">ligament</token>
              <token id="w12" pos="IN" lemma="of">of</token>
              <token id="w13" pos="DT" lemma="the">the</token>
              <token id="w14" pos="JJ" lemma="human">human</token>…
          </text>
          …
          </sentence>…
</document>
```

Tableau 4.2 : Extrait de l'indexation du document "Arthroskopie.00130003.eng.abstr" de la collection *MuCHMORE*

Utilisation d'un modèle probabiliste DFR (Divergence From Randomness) comme SRI classique de base

Afin de comparer *OBIRS* à des approches de RI classiques, nous avons généré une indexation par termes des documents de *MuCHMORE*. Pour identifier les termes (mots ou phrases nominales) pertinents pour cette indexation, nous avons utilisé la plate-forme *Terrier*[29] disponible gratuitement en téléchargement. Il s'agit d'une plateforme standard d'évaluation et de développement en RI. Notons que *Terrier* suit le protocole *TREC* (Voorhees et al. 1997) quant à son modèle d'évaluation. Nous l'utilisons pour extraire des termes à partir des documents en texte brut de *MuCHMORE* à l'aide des outils classiques de suppression de mots vides (*stop words*) et de lemmatisation. Après extraction des termes indexant un document, chacun d'eux est pondéré à l'aide d'un des modèles *DFR* les plus populaires à savoir *PL2* (He et al. 2005). Nous donnons un aperçu de ce modèle avec notamment son estimation du *RSV* d'un document d_j indexé par des termes, étant donnée une requête Q constituée d'un seul terme t_r :

[29] http ://www.terrier.org/

$$RSV(Q,d_j)=\frac{1}{tfn+1}\sum_{t_r\in Q}\left[\left(tfn*log_2\frac{tfn}{\lambda}\right)+log_2e*\left(\lambda+\frac{1}{12*\ tfn}-tfn\right)+0.5*log_2(2\pi*\ tfn)\right] \quad \textbf{(4.11)}$$

où λ est la fréquence normalisée du terme t_r dans la collection; tfn est la fréquence normalisée du terme t_r dans le document d_j en fonction de la fréquence d'origine $tf\left(tf_{t_r}^{d_j}\right)$ de ce terme. Plus précisément, $tfn=tf_{t_r}^{d_j}*log_2\left(1+c*\frac{dlAv}{dl_j}\right)$, avec dl_j la longueur du document d_j ; $dlAv$, la moyenne des longueurs des documents dans toute la collection et $c>0$ un paramètre réel libre. Selon les auteurs de (Clinchant et al. 2009), le paramétrage $c=5$ donne les meilleurs résultats dans les collections *TREC*, en particulier pour les requêtes courtes. Notre jeu de tests basé sur *MuCHMORE* ayant des requêtes dont la longueur moyenne est 2,3 (c.f. Tableau 4.1), nous avons donc utilisé ce paramétrage pour *PL2* dans nos tests.

4.3.2. Protocole expérimental de validation

Le scénario suivant, connu sous le nom de "*simulated-feedback technique"* (Efthimiadis 1996), a été adopté afin de simuler une activité de recherche en considérant que si : 1) $D_{see}^{Q}\subseteq D_{res}^{Q}$ est l'ensemble des documents renvoyés par le SRI et vus par l'utilisateur après une requête Q et 2) D_E^Q est l'ensemble des documents que les experts ont jugés pertinents pour la même requête alors, les documents de $D_{see}^{Q}\cap D_E^Q$ peuvent être utilisés comme une approximation de l'ensemble D_u^Q des documents pertinents qu'un utilisateur aurait pu sélectionner. Ainsi, l'ensemble D_{see}^{Q}, où des jugements d'utilisateurs sur la pertinence des documents sont disponibles, est utilisé comme collection de base pour l'apprentissage de la requête reformulée. Lorsque la requête est reformulée (c.f. Algorithme 4.2), sa performance est évaluée en utilisant la globalité de la collection D.

Nous allons évaluer l'approche *OBIRS* de même que OBIRS-*feedback* en utilisant la mesure de similarité sémantique de Lin (D. Lin 1998) tant pour calculer le *RSV* d'un document – c.f. l'équation (3.7), que pour définir l'indicateur de qualité d'une reformulation – c.f. l'équation (4.3). Cette mesure de similarité est donc également celle utilisée pour construire l'ensemble des concepts d'intérêt (c.f. Algorithme 4.1). Le paramètre q de notre modèle de *RSV* est fixé à 2 pour *OBIRS* ainsi que pour l'évaluation des indicateurs dans OBIRS-*feedback*. Lors de notre reformulation nous ne faisons que compléter la requête initiale ce qui assure que l'on ne pourra pas trop s'en éloigner et rend le paramètre α quasiment inutile. En fixant α à 0, on prend donc peu de risque tout en simplifiant nettement le problème d'optimisation puisqu'il ne reste alors qu'un seul paramètre libre. Ainsi, dans nos tests, les fonctions indicatrices – c.f. l'équation (4.3), étudiées ont un seul paramètre libre γ , α étant fixé à 0 et β à 1.

Nos analyses suivent le protocole *TREC* et impliquent la recherche de documents *ad hoc*.

Nous considérons les 1000 premiers résultats de chaque requête et mesurons la précision moyenne – c.f l'équation.(2.41), sur toutes les requêtes ainsi que la R-précision – c.f l'équation (2.40), dans la collection *MuCHMORE*. Nous fournissons également des valeurs de rappel, les courbes de précision et de rappel (Manning et al. 2008) (en utilisant onze points de rappel) afin d'avoir une vision globale de l'évolution de la précision du système.

4.3.3. Résultats

Dans cette section, nous évaluons le gain en performance (précision moyenne, R-précision et rappel) obtenu lorsque l'on ajoute l'étape de reformulation au système *OBIRS* via la phase de réinjection de pertinence basée sur les fonctions objectif *ind* de l'équation (4.3). Nous examinons d'abord l'impact des documents négatifs ($D_{see}^{Q} \backslash D_{u}^{Q}$) sur les performances globales de OBIRS-*feedback* en faisant varier le paramètre γ. Ensuite, nous étudions l'impact du seuil de similarité t, limitant l'ensemble C_u des concepts pris en compte lors de la recherche d'une requête reformulée – c.f. l'équation (4.5). Nous finissons en comparant les résultats d'OBIRS-*feedback* avec ceux d'*OBIRS* et de *PL2*.

Impact des documents négatifs dans la reformulation

La Figure 4.4 montre l'évolution de la précision moyenne, de la R-précision et du rappel d'OBIRS-*feedback* selon le poids γ donné aux documents négatifs. Onze valeurs de γ sont considérées et pour chacune d'elles tous les concepts de l'ontologie sont testés lors de la recherche de Q_{max} (dans ce cas t = 0 et, donc, $C_u = C$).

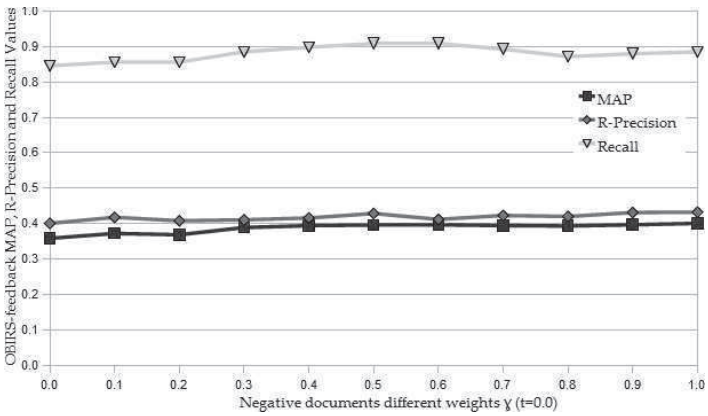


Figure 4.4 : Evolution de la précision moyenne (MAP), de la R-précision et du rappel suivant différents seuils pour le paramètre γ contrôlant les documents négatifs (sous la condition t = 0)

Les valeurs de précision moyenne (respectivement de rappel) dans OBIRS-*feedback* commencent à se dégrader lorsque le poids γ affecté à l'importance des documents négatifs est supérieur à 0,7 (respectivement $\gamma \geq 0{,}6$) alors que les valeurs de la R-précision varient très peu. Avec des valeurs de γ aussi élevées, le score de l'indicateur – c.f. l'équation (4.3), est trop fortement influencé par les documents que l'utilisateur a jugés non pertinents, ce qui introduit du bruit. Le paramétrage de γ à 0,5 conduit au meilleur rappel (0,9083), à une bonne précision moyenne (0,3968) et à la meilleure R-précision (0.4293). Cette valeur de $\gamma = 0{,}5$ a donc été retenue pour toutes les autres expériences.

Impact de la taille du voisinage exploré ($S_\pi(c_x, t)$)

Dans cette expérience, nous étudions l'impact lié au fait de ne tester, lors de la reformulation, que les concepts sémantiquement proches de ceux indexant les documents jugés pertinents. Le degré de proximité nécessaire pour qu'un concept soit testé est contrôlé par le seuil de similarité t. De faibles valeurs de t permettent de s'éloigner des concepts que l'utilisateur a jugés pertinents et donc d'augmenter la taille de C_u. L'ensemble C_u de concepts, à partir duquel, une requête reformulée Q_{max} varie d'un sous-ensemble limité aux seuls concepts présents dans l'indexation des documents pertinents D_u ($t = 1$) jusqu'à la prise en compte de l'ensemble de l'ontologie ($t = 0$). Nous avons testé ces deux extrêmes ainsi que neuf valeurs intermédiaires. Une courbe de précision-rappel est proposée pour le seuil t variant de 0 à 1 par pas de 0,1 (c.f. Figure 4.5).

Il n'y a presque pas de différence entre les valeurs de précision (interpolées) pour les différents seuils de similarité t comme en témoignent les courbes de la Figure 4.5 qui se chevauchent ainsi que les valeurs de précision dans le Tableau 4.3. L'information importante est le chevauchement des courbes. Les concepts sémantiquement éloignés de tous ceux indexant les documents pertinents (c'est-à-dire ceux considérés seulement avec de faibles valeurs de t n'améliorent pas la pertinence d'OBIRS-*feedback*. Notons cependant, que le bruit qu'ils pourraient introduire est largement neutralisé par la recherche de concepts qui optimisent l'indicateur de reformulation. Pour tous les seuils de similarité, la précision d'OBIRS-*feedback* est meilleure que celle du système de base *OBIRS*.

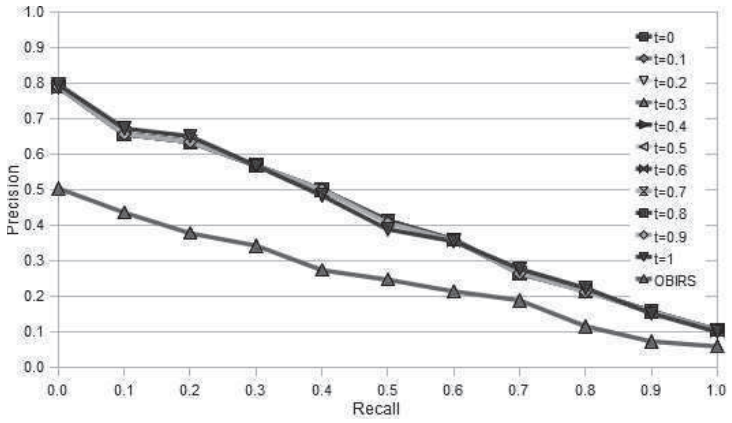


Figure 4.5 : Courbes de rappel précision d'OBIRS-*feedback* pour 11 valeurs de seuil t de similarité (sous la condition $\gamma = 0.5$)

La Figure 4.6 montre les différentes valeurs de MAP, de R-Précision et de Rappel obtenues par OBIRS-*feedback*, en fonction de différentes valeurs du seuil (t). Les meilleurs résultats sont obtenus pour $t = 0{,}8$ (MAP=0,3968, R-Précision=0,4293 et Rappel = 0,9083) et pour $t = 0{,}9$ (MAP=0,3963, R-Précision=0,4300 and Rappel = 0,8900).

	11 points de Rappel										
*Seuil (**t**)*											
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.1	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.2	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.3	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.4	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.5	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.6	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.7	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.8	0.789	0.656	0.634	0.568	0.500	0.413	0.360	0.265	0.216	0.158	0.106
0.9	0.789	0.656	0.634	0.569	0.500	0.407	0.358	0.265	0.213	0.158	0.106
1	0.795	0.672	0.650	0.568	0.484	0.389	0.355	0.278	0.224	0.152	0.101
OBIRS	0.504	0.436	0.378	0.343	0.274	0.247	0.214	0.189	0.117	0.074	0.060

Tableau 4.3 : Valeurs de précision aux 11 points de Rappel pour chaque seuil de similarité t.

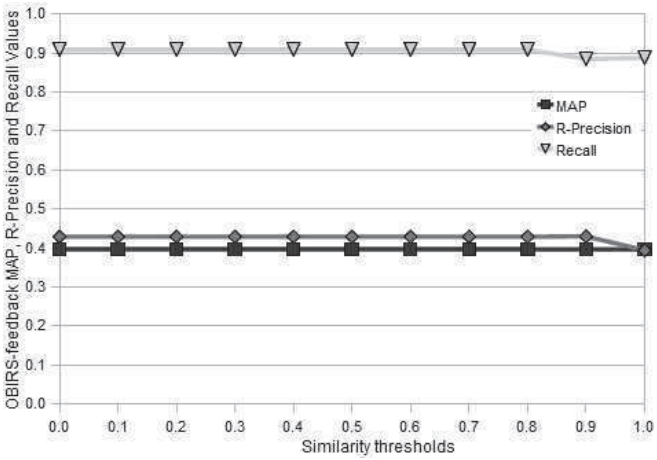


Figure 4.6 : Evolution de la précision moyenne (MAP), de la R-précision et du rappel en fonction du seuil de similarité utilisé pour définir l'ensemble de concepts testés par OBIRS-*feedback* (sous la condition $\gamma = 0.5$)

La Figure 4.7 montre la variation, en fonction de la variation du seuil de similarité (t), du nombre de concepts testés lors de la reformulation par OBIRS-*feedback* ($|C_u|$). Quand $t = 1$, le nombre moyen de concepts dans C_u est 104 (~273 quand $t = 0{,}9$; ~723 quand $t = 0{,}8$ et 25 603 quand $t = 0$). Ce paramètre influence donc grandement les temps d'exécution. En effet, la durée moyenne de l'exécution d'OBIRS-*feedback* va de 1.654s (quand $t = 1$) à 395.832s (quand $t = 0$) quand tous les concepts de l'ontologie sont considérés. Ce temps d'exécution comprend la construction de l'ensemble des concepts d'intérêt C_u (c.f. Figure 4.8), ainsi que la recherche de la requête optimale Q_{max}.

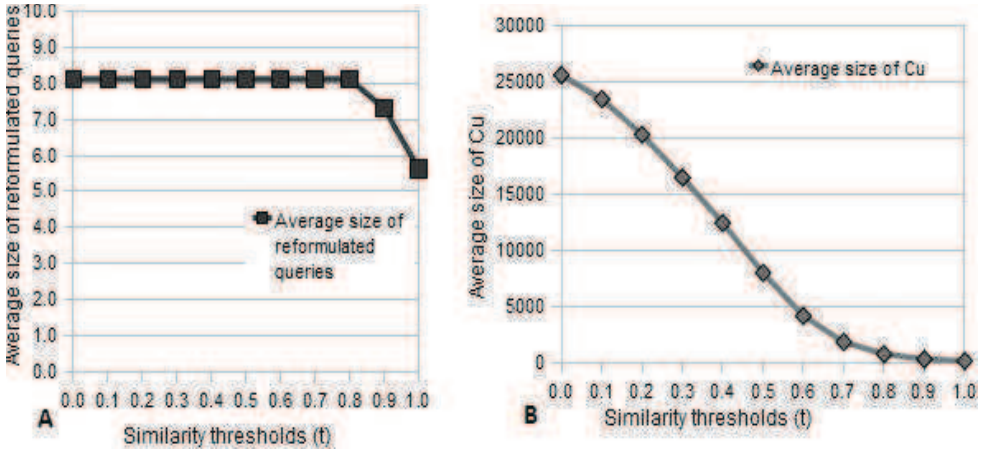


Figure 4.7 : Taille moyenne des requêtes reformulées (A) et de l'ensemble des concepts d'intérêt C_u pour chaque seuil de similarité t (sous la condition $\gamma = 0.5$)

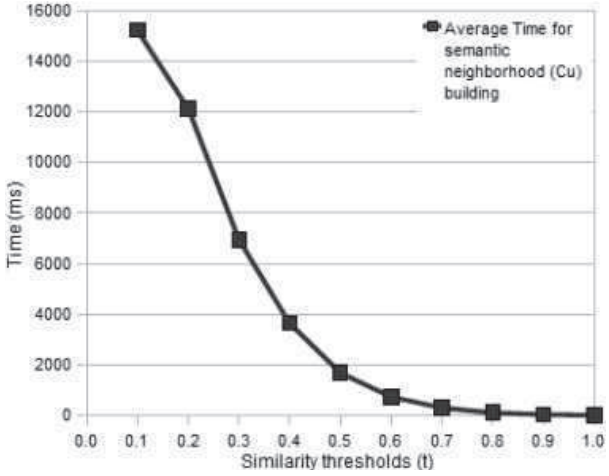


Figure 4.8 : Evolution du temps d'exécution d'OBIRS-*feedback* en fonction du seuil de similarité t

Comparaison des performances d'OBIRS-*feedback* avec celles d'*OBIRS* et de PL2

Etudions les performances d'OBIRS-*feedback* lorsque celui-ci est paramétré par le poids des documents négatifs et le seuil de similarité obtenu dans les expérimentations précédentes (i.e. $t = 0{,}8$ ou $0{,}9$ et $\gamma = 0{,}5$).
OBIRS-*feedback* obtient une meilleure valeur de R-précision et de rappel que le système PL2 (avec $c = 5$) tandis que ce dernier obtient la meilleure précision moyenne.

Modèles de RI	**MAP**	**R-précision**	**Rappel**
OBIRS $q = 2{,}0$	0,2401	0,2592	0,7696
PL2 (DFR) $c = 5$	**0,4332**	0,4293	0,8900
OBIRS-feedback $\gamma = 0{,}5,\ t = 0{,}8,\ q = 2{,}0$	0,3968	0,4293	**0,9083**
OBIRS-feedback $\boldsymbol{\gamma = 0{,}5,\ t = 0{,}9,\ q = 2{,}0}$	0,3963	**0,4300**	0,8900

Tableau 4.4 : Comparaison d'OBIRS-*feedback* avec le système de base *OBIRS* et le modèle de RI classique PL2 (c=5). Les meilleures valeurs de précision moyenne, R-précision et de rappel sont mis en gras.

4.4. Conclusion

Dans ce chapitre, nous avons présenté une méthode de reformulation de requêtes conceptuelles à la fois *globale*, utilisant une ontologie de domaine, et *locale*, i.e. s'appuyant sur des jugements utilisateurs. Suivant le schéma de Rocchio, nous formalisons le problème de reformulation de requêtes comme la recherche d'un sous-ensemble de concepts améliorant une fonction objectif qui reflète la proximité sémantique de la requête reformulée par rapport aux documents d'intérêts pour l'utilisateur. Cette fonction objectif peut aussi être vue comme un indicateur de la qualité d'un processus d'apprentissage utilisant les documents présentés à l'utilisateur comme un corpus d'apprentissage. A notre connaissance, c'est la première fois que la reformulation de requêtes conceptuelles est formellement exprimée sous la forme d'un problème d'optimisation combinatoire.

Les résultats obtenus par l'intégration de cette approche dans l'environnement *OBIRS* montrent que la précision et le rappel de *OBIRS* sont nettement améliorés tout en conservant des temps de réponses courts. Ce résultat reste valable même si l'ontologie est de grande taille. Cette efficacité est obtenue grâce à une heuristique qui réduit significativement l'espace des requêtes conceptuelles dans lequel la requête optimale est recherchée. Cette réduction de l'espace de recherche est fondée sur deux propriétés simples et intuitives de nombreuses mesures de similarité sémantique. En effet, nous avons prouvé que ces deux propriétés garantissent la connexité du voisinage sémantique d'un concept et cela quel qu'en soit le rayon. Notre stratégie obtient de meilleures valeurs de précision aux premières positions des résultats retrouvés de même qu'un meilleur rappel que le système PL2 (c=5). Le modèle PL2 existe de longue date et bénéficie d'années de développement et d'amélioration. Atteindre des performances comparables est donc très prometteur d'autant plus qu'il y a encore matière à améliorer ces résultats en intégrant, par exemple, une pondération des concepts dans OBIRS-*feedback*.

Notre stratégie de reformulation de requêtes explore de nouvelles voies dans le cadre des SRI conceptuels. Elle fournit un cadre général pour la mise en œuvre d'une stratégie de reformulation de requêtes dès lors qu'une ontologie est utilisée aussi bien pour l'indexation que pour l'appariement dans un SRI. En outre, elle définit une famille de stratégies de reformulation de requêtes conceptuelles grâce à l'intégration d'un modèle des préférences de l'utilisateur, autorisant différents comportements de choix à travers les opérateurs d'agrégation retenus et la pondération paramétrable de la contribution des documents négatifs (ceux jugés non pertinents par l'utilisateur). La construction rapide d'un ensemble de concepts d'intérêts pour un utilisateur donné, peut avoir d'importantes applications dans le domaine de la personnalisation en RI surtout dans la phase d'apprentissage de profils utilisateurs.

Chapitre 5 : Conclusion générale

5.1. Synthèse des contributions

Historiquement, les principaux Systèmes de Recherche d'Information (SRI) se sont basés sur une représentation des documents et des requêtes, proche du langage naturel, sous forme de *sac de termes*. Dans ces représentations, la statistique d'occurrence des termes dans les documents est utilisée pour évaluer leur importance. Or ces approches, dites classiques, souffrent de différentes limites. La première concerne l'ambiguïté intrinsèque au langage naturel, dès lors qu'un terme peut couvrir plusieurs sens selon le contexte dans lequel on l'utilise. Une telle limite a de sérieuses répercussions dans les principaux processus de la RI. De plus, de telles approches ignorent les possibles relations entre les termes considérés. Enfin, souvent ces approches offrent une justification et une interaction limitées.

En réponse à ces différents écueils, plusieurs pistes ont été explorées, dont la plus prometteuse est celle des approches de RI dites conceptuelles. Dans celles-ci, les termes sont remplacés par des concepts qui forment des unités de sens. L'utilisation de concepts issus d'une ontologie de domaine permet de s'affranchir des possibles ambiguïtés (un concept est défini de façon unique dans une ontologie) et permet de trouver des liens possibles entre les concepts exprimés dans la requête. Les ontologies, en tant que formalisation explicite et partagée des notions et sens d'un domaine, ainsi que de leurs relations, offrent à la RI l'opportunité d'améliorer son processus d'indexation et par là, ses processus d'appariement et de reformulation. Ces ressources conceptuelles sont devenues très nombreuses et diversifiées, dans le domaine biomédical notamment, autorisant leur utilisation dans un large spectre d'applications.

Les travaux présentés dans cette thèse entrent dans le sillage des approches conceptuelles de la RI, notamment dans ses processus d'appariement et de reformulation. Nous avons proposé un modèle de pertinence utilisant un opérateur d'agrégation pour combiner les proximités sémantiques, évaluées à partir d'une ontologie de domaine, entre un document et l'ensemble des concepts de la requête. Ce modèle de pertinence permet de justifier la sélection des documents pertinents, en indiquant les contributions respectives des différents concepts de la requête au score global de chaque document. Une restitution visuelle des résultats est proposée. Ensuite, un modèle de reformulation de requêtes conceptuelles a été mis en œuvre. L'approche originale retenue implique l'utilisateur dans la boucle de pertinence pour lui

fournir des résultats au plus proche de ses attentes. Ces deux contributions ont été implémentées et mises à la disposition de nos partenaires dans le cadre des plateformes collaboratives de gestion de ressources. Ces ressources peuvent, par exemple, être des publications scientifiques partagées pour les besoins d'un collectif de chercheurs (comme c'est le cas dans *CoLexIR*) ou des gènes indexés par des concepts de la *Gene Ontology* (comme c'est le cas dans *OBIRS-feedback*).

La première contribution de cette thèse est relative à un modèle de pertinence utilisant un modèle d'agrégation à trois niveaux et qui suppose, en pré requis, l'existence d'une indexation conceptuelle des ressources sur lesquelles porte la recherche. Le premier niveau concerne le calcul de la similarité sémantique entre deux concepts, l'un indexant la requête de l'utilisateur et l'autre appartenant à l'indexation d'un document à évaluer. Le second niveau propose une agrégation de ces similarités sémantiques élémentaires pour calculer la pertinence d'un document relativement à un concept de la requête. Le troisième et dernier niveau permet d'agréger les scores d'un document relatifs aux concepts de la requête pour lui associer un score de pertinence global (*RSV*). Les préférences de l'utilisateur peuvent être prises en compte au niveau des opérateurs d'agrégation et de la pondération des concepts de la requête. Le *RSV* permet par la suite d'ordonner l'ensemble des documents suivant leur degré de pertinence. Dans nos modèles d'agrégation, il n'y a pas de problème quant à la nature des entités à agréger puisqu'il ne s'agit que de similarités sémantiques : elles sont assimilées à des utilités qui mesurent l'adéquation d'un concept avec un autre, d'un concept avec un document, etc. La commensurabilité est garantie dès lors que ces scores sont calculés à partir de mesures de similarité sémantique définies dans un même cadre : une ontologie de domaine. Cet aspect répond à notre premier objectif de **mise en place d'un modèle de pertinence utilisant une ontologie de domaine** (c.f. section 1.3, page 8).

Par ailleurs, nous avons montré (c.f. section 3.4.1) que la famille d'opérateurs de compromis de Yager, que nous avons utilisée pour le dernier niveau d'agrégation, permet de justifier le score global obtenu en termes de contribution des différents concepts d'une requête. Une analyse de sensibilité de l'opérateur d'agrégation qui synthétise le système de préférences de l'utilisateur sous une forme analytique permet de calculer simplement ces contributions. Avec notre modèle, il est ainsi possible d'identifier et de mettre en évidence les concepts sur lesquels l'utilisateur peut intervenir (en les modifiant ou en modifiant leur pondération) pour améliorer sa requête. Nous avons fortement tiré parti de cette justification en présentant les scores élémentaires de chaque document au travers d'une icône le représentant dans l'espace de visualisation. Ainsi l'utilisateur peut apprécier en un coup d'œil en quoi le document correspond à sa requête. Les documents sont positionnés sur une carte sémantique où leur distance géométrique à une *sonde* représentant la requête est proportionnelle à leur score de pertinence (*RSV*). Cet effort de visualisation pour une meilleure présentation des résultats du SRI et une meilleure justification de leur score constitue une contribution majeure de cette thèse et répond aux deuxième et troisième objectifs que nous nous étions donné concernant la **mise en place d'un modèle de justification et de diagnostic des résultats** pour l'utilisateur et d'une **stratégie de visualisation de ces résultats** (c.f. section 1.3, page 8).

Pour finir, nous avons proposé une stratégie de reformulation de requêtes conceptuelles basée sur deux propriétés simples et intuitives de la plupart des mesures de similarité sémantique, dont celles basées sur le contenu informationnel. Nous avons formalisé ces deux propriétés et démontré qu'elles conduisent à la connexité du voisinage conceptuel d'un concept et cela quel que soit le rayon choisi. Nous avons utilisé cette connexité pour proposer deux heuristiques efficaces de construction d'une requête conceptuelle optimisant une fonction objectif. Notre stratégie de reformulation peut alors être vue comme une phase d'apprentissage de la requête sur les documents initialement retournés à l'utilisateur. Cette contribution répond à notre dernier objectif de **mise en place d'une stratégie de reformulation de requêtes conceptuelles rapide et efficiente**.

Dans le calcul du *RSV*, nous avons clairement dissocié dans notre approche ce qui relève du modèle structurel des connaissances et des distances qu'on peut lui associer, de ce qui relève de l'expression des préférences ou des objectifs de l'utilisateur. Cette association est une autre contribution importante de nos travaux. En effet, la décomposition du calcul du *RSV*, les paramétrages que nous avons proposés permettent d'envisager des outils de RI plus faciles à contrôler et de plus personnalisables.

5.2. Valorisation des contributions

Outre les différentes publications qui ont été mentionnées dans le mémoire, il est à noter que les solutions proposées vont être mises à la disposition de nos partenaires industriels de différentes manières. La version de l'environnement *OBIRS* développée dans le cadre de cette thèse et disponible en ligne, n'est qu'un prototype. Un transfert est en cours pour proposer une version plus aboutie et intégrée aux différentes plateformes collaboratives de nos partenaires (principalement ITMO IHP, ITMO Cancer et CEA). Mais *OBIRS* a également suscité l'intérêt des membres du projet ANR PhylARIANE avec lesquels nous collaborons pour qu'il soit intégré à leurs applications et utilisé pour l'exploitation des corpus de gènes constitués dans le cadre de ce projet (http://www.lirmm.fr/phylariane/resources.php).

5.3. Perspectives

Les travaux que nous avons détaillés dans cette thèse débouchent sur plusieurs perspectives que cette section se propose de détailler. Ces dernières peuvent être regroupées en deux groupes selon qu'elles se rapportent à l'une ou l'autre des deux contributions principales de cette thèse : les approches *OBIRS* et OBIRS-*feedback*.

Utilisation de plusieurs ressources conceptuelles

La première perspective concerne l'utilisation de plusieurs modèles de connaissances dans notre approche de RI conceptuelle. En effet, plusieurs concepts issus d'ontologies différentes peuvent être nécessaires pour décrire complètement le contenu d'un document ou d'un corpus. Cela est d'autant plus vrai qu'une ontologie même généraliste ne couvre souvent pas tous les aspects d'un domaine ou d'une discipline. Dans ce cas les mesures de similarité sémantique à mettre en œuvre doivent être repensées, ainsi que des ajustements concernant le calcul de

pertinence des documents. En effet, le problème de la commensurabilité dans l'agrégation se posera alors puisque il s'agira de fusionner des similarités induites par plusieurs ontologies de domaine.

Diversification

La majeure partie des stratégies de reformulation locales considèrent que la totalité (*pseudo relevance feedback*) ou une partie (*relevance feedback*) des *k* premiers documents présentés à l'utilisateur après une requête, sont pertinents. Derrière cette considération, se cachent deux hypothèses : i) ces méthodes de reformulation considèrent que l'utilisateur ne consulte que rarement les documents classés au-delà d'un certain seuil relativement petit (∼20 résultats correspondant généralement à deux pages de résultats sur Google par exemple) et ii) les *k* premiers documents sont représentatifs de l'ensemble des résultats en termes de contenu et qu'ils sont donc suffisants pour fournir des éclairages assez divers sur la requête. La première hypothèse correspond généralement au comportement d'un utilisateur qui s'attend à être satisfait dès les premiers résultats et "rechignant" à aller au-delà. Cette hypothèse amène la plupart des SRI à améliorer leur précision aux premiers documents. La seconde hypothèse n'est souvent pas satisfaite en pratique : les premiers documents retournés par la majeure partie des SRI sont proches en termes de contenu, ce qui pénalise la reformulation. Cette faible diversité est due au fait que la plupart des modèles de pertinence qu'ils mettent en œuvre sont basés sur des calculs de proximité (vectoriel, conceptuel). Le gain cumulé en termes d'information, obtenu en considérant les *k* premiers documents, n'est pas très différent de celui obtenu en considérant un seul d'entre eux. Lorsqu'un module de reformulation est activé dans un SRI, un compromis doit donc être trouvé entre un classement par pertinence (dépendant du *RSV*) et un classement par diversité (dépendant de la couverture des thèmes de la requête).

Personnalisation

Une autre perspective intéressante est celle relative à l'évolution de notre stratégie de reformulation pour tendre vers une stratégie de personnalisation. En effet, nous nous sommes limités à une seule session de recherche dans notre approche de reformulation. Une session de recherche regroupe l'ensemble des interactions entre l'utilisateur et le SRI pour aboutir à la satisfaction d'un besoin en information. Une fois que l'on considère plusieurs sessions, la construction et la maintenance dans le temps de l'ensemble des concepts d'intérêts de l'utilisateur deviennent problématiques. Cet ensemble de concepts d'intérêts est appelé *profil utilisateur*. Des questions majeures en termes de structure de données à utiliser pour le représenter, d'algorithmes pour le construire et de stratégies pour le maintenir se posent, ainsi que des questions relatives au stockage et à la diffusion de *profil* (vie privée). Pour illustrer la pertinence de la personnalisation, reprenons le dialogue entre le professeur et son élève :

- Le professeur : "Quel est le dernier livre que tu as lu ?"
- Elève : Peut-on considérer un essai comme étant un livre ?
- Professeur : Oui
- L'élève : "J'ai lu le dernier essai de Léopold Sedar Senghor"

Nous avons indiqué, dans la section 2.2.2, que l'élève n'a pu répondre à cette question que parce qu'il a effectué un alignement de connaissance avec le professeur à travers la deuxième question. Il a aussi pu répondre à cette question parce qu'il a mobilisé sa connaissance antérieure sur ce qu'est un livre (définition qu'on lui aura fournie lors de précédentes questions par exemple). Enfin, on peut noter que même si le dernier livre que l'élève a effectivement lu est un roman policier, il sait que cette réponse n'est pas pertinente dans le cadre d'une question posée par son professeur. Il s'agit donc de connaissances générales et de connaissances relatives aux attentes de son professeur que l'élève a capitalisées sur le long terme en vue de répondre à une question sur le court terme. C'est tout l'objet de la personnalisation en RI : prendre en compte l'historique de recherche pour déterminer les centres d'intérêt de l'utilisateur à même de préciser son besoin actuel en information. La séparation des préférences de l'utilisateur et de la distance associée à la structure des connaissances dans notre modèle de RI et de calcul du *RSV* nous donne une grande flexibilité pour aborder cette problématique de la personnalisation.

Indexation par propagation

Cette dernière perspective concerne l'application de notre stratégie de reformulation dans le contexte d'une indexation par propagation. Le principe est le suivant. On dispose d'un ensemble de documents indexés par des concepts et représentés sur une carte sémantique. Lorsque l'utilisateur souhaite indexer un nouveau document, il lui suffit de cliquer sur la zone de la carte sémantique où il pense que ce document devrait être. Le système d'assistance à l'indexation peut alors récupérer les documents correspondant aux k plus proches voisins de ce point (qui vont constituer l'ensemble des documents D_u d'intérêt de l'utilisateur) tandis que les k suivants constitueront l'ensemble D_{see}. L'approche OBIRS-*feedback* permet alors d'obtenir ainsi automatiquement une première indexation du nouveau document, que l'utilisateur n'aura plus qu'à amender.

Les travaux menés durant ces trois années ont permis tout à la fois de proposer de nouveaux modèles en recherche d'information et de répondre à certains besoins identifiés chez nos partenaires, futurs utilisateurs des outils que nous avons développés. Mais au-delà de la satisfaction de voir aboutir ces travaux de recherche, les nombreuses perspectives qu'ils ouvrent me confortent dans le choix que j'ai fait en m'engageant dans cette thèse et dans l'activité de chercheur.

Références bibliographiques

Alexopoulou, Dimitra, Bill Andreopoulos, Heiko Dietze, et al. 2009. 'Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy.' *BMC Bioinformatics* 10(1), pp. 28.

Amati, Gianni, and Cornelis Joost Van Rijsbergen. 2002. 'Probabilistic models of information retrieval based on measuring the divergence from randomness.' *ACM Trans. Inf. Syst.* 20(4), pp. 357–389.

Aussenac-Gilles, Nathalie. 2008. 'Le web sémantique, quel renouvellement pour la recherche d'information ?'. In *Recherche d'information : état des lieux et perspectives*, Recherche d'information et web, Mohand boughanem, Jacques Savoy pp. 231–266.

Badra, Fadi, Sylvie Despres, and Rim Djedidi. 2011. 'Ontology and Lexicon: The Missing Link'. In *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, eds. Monique Slodzian, Mathieu Valette, Nathalie Aussenac-Gilles, et al. Paris, France: INALCO pp. 16–18.

Bannour, Ines, and Haïfa Zargayouna. 2012. 'Une plate-forme open-source de recherche d'information sémantique'. In *CORIA*, pp. 167–178.

Baziz, Mustapha, Nathalie Aussenac-Gilles, and Mohand Boughanem. 2003. 'Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information'. In *Inforsid*, pp. 121–134.

Baziz, Mustapha, Mohand Boughanem, Nathalie Aussenac-Gilles, et al. 2005. 'Semantic cores for representing documents in IR'. In *Proceedings of the 2005 ACM symposium on Applied computing - SAC '05*, Santa Fe, New Mexico pp. 1011.

Baziz, Mustapha, Mohand Boughanem, Gabriella Pasi, et al. 2007. 'An information retrieval driven by ontology from query to document expansion'. In *Large Scale Semantic Access to Content (Text, Image, Video, and Sound)*, RIAO '07, Paris, France, France: Le Centre de Hautes Etudes Internationales d'Informatique Documentaire pp. 301–313.

Belkin, Nicholas J., Peter Ingwersen, and Annelise Mark Pejtersen, eds. 1992. '*Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*.' ACM.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. 'The Semantic Web: Scientific American.' *Scientific American*.

Bhagdev, Ravish, Sam Chapman, Fabio Ciravegna, et al. 2008. 'Hybrid Search: Effectively Combining Keywords and Ontology-based Searches'. In *Proceedings of the 5th European Semantic Web Conference*, Springer Verlag.

Bhogal, Jagdev, Andy Macfarlane, and Peter Smith. 2007. 'A review of ontology based query expansion.' *Information Processing & Management* 43(4), pp. 866–886.

Borlund, Pia. 2003. 'The concept of relevance in IR.' *J. Am. Soc. Inf. Sci. Technol.* 54(10), pp. 913–925.

Borst, Willem Nico. 1997. 'Construction of Engineering Ontologies For Knowledge Sharing ans Reuse.'

Bossy, Robert, Alain Kotoujansky, Sophie Aubin, et al. 2008. 'Close Integration of ML and NLP Tools in BioAlvis for Semantic Search in Bacteriology.'. In *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Sciences*, eds. Albert Burger, Adrian Paschke, Paolo Romano, et al. Edinburgh, United Kingdom.

Boughanem, Mohand. 2008. 'Introduction à la recherche d'information'. In *Recherche d'information : état des lieux et perspectives*, Recherche d'information et web, J. Savoy, M. Boughanem pp. 19–44.

Buitelaar, Paul, Philip Cimiano, John McCrae, et al. 2011. 'Ontology Lexicalisation: The lemon Perspective'. In *Workshop Proceedings of the 9th International Conference on Terminology and Artificial Intelligence*, eds. Monique Slodzian, Mathieu Valette, Nathalie Aussenac-Gilles, et al. Paris, France: INALCO pp. 33–36.

Le Capitaine, Hoel. 2009. 'Opérateurs d'agrégation pour la mesure de similarité. Application à l'ambiguïté en reconnaissance de formes.' Thèse de doctorat en Automatique, Image et Signal, Université de La Rochelle.

Carpineto, Claudio, and Giovanni Romano. 2012. 'A Survey of Automatic Query Expansion in Information Retrieval.' *ACM Comput. Surv.* 44(1), pp. 1:1–1:50.

Cena, Federica, Silvia Likavec, and Francesco Osborne. 2011. 'Propagating user interests in ontology-based user model'. In *Proceedings of the 12th international conference on Artificial intelligence around man and beyond*, AI*IA'11, Berlin, Heidelberg: Springer-Verlag pp. 299–311.

Chu, Wesley W., Zhenyu Liu, and Wenlei Mao. 2002. '*Textual Document Indexing and Retrieval via Knowledge Sources and Data Mining*.'

Cimiano, Philip, Paul Buitelaar, John McCrae, et al. 2011. 'LexInfo: A Declarative Model for the Lexicon-Ontology Interface.' *Web Semantics: Science, Services and Agents on the World Wide Web* 9(1), pp. 29–51.

Clinchant, Stéphane, and Eric Gaussier. 2009. 'Bridging Language Modeling and Divergence from Randomness Models: A Log-Logistic Model for IR'. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, Berlin, Heidelberg: Springer-Verlag pp. 54–65.

Clinchant, Stéphane, and Eric Gaussier. 2010. 'Information-based models for ad hoc IR'. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, New York, NY, USA: ACM pp. 234–241.

Cockburn, Andy, and Bruce McKenzie. 2001. '3D or not 3D?'. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01*, Seattle, Washington, United States pp. 434–441.

Collins, Allan, and Elizabeth Loftus. 1975. 'A spreading-activation theory of semantic processing.' *Psychological Review* 82(6), pp. 407–428.

Crampes, Michel, and Jeremy De Oliveira-Kumar. 2010. 'La sonde sémantique pour la fouille sémantique visuelle'. In *Actes de la conférence Ingénierie des Connaissances 2010*, Nimes, France pp. 259–270.

Daoud, Mariam, Lynda Tamine-Lechani, Mohand Boughanem, et al. 2009. 'A session based personalized search using an ontological user profile'. In *Proceedings of the 2009 ACM symposium on Applied Computing*, SAC '09, New York, NY, USA: ACM pp. 1732–1736.

Dinh, Duy, and Lynda Tamine. 2011. 'Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval'. In *Advances in Information Retrieval*, Springer Berlin / Heidelberg pp. 375–386.

Dinh, Duy, and Lynda Tamine. 2012. 'Towards a context sensitive approach to searching information based on domain specific knowledge sources.' *Web Semantics: Science, Services and Agents on the World Wide Web* 12, pp. 41–52.

Dominich, Sandor 2008. 'Introduction'. In *The Modern Algebra of Information Retrieval*, The Information Retrieval Series, Springer Berlin Heidelberg pp. 1–26.

Doms, Andreas, and Michael Schroeder. 2005. 'GoPubMed: exploring PubMed with the Gene Ontology.' *Nucleic Acids Research* 33(Web Server), pp. W783–W786.

Dragoni, Mauro, Célia da Costa Pereira, and Andrea Tettamanzi. 2012. 'A conceptual representation of documents and queries for information retrieval systems by using light ontologies.' *Expert Systems with Applications* 39(12), pp. 10376 – 10388.

Dubois, Didier, and Henri Prade. 1985. 'A review of fuzzy set aggregation connectives.' *Information Sciences* 36(1–2), pp. 85 – 121.

Dubois, Didier, and Henri Prade. 2004. 'On the use of aggregation operations in information fusion processes.' *Fuzzy Sets and Systems* 142(1), pp. 143 – 161.

Duthil, Benjamin, François Trousset, Mathieu Roche, et al. 2011. 'Towards an automatic characterization of criteria'. In *Proceedings of the 22nd international conference on Database and expert systems applications - Volume Part I*, DEXA'11, Berlin, Heidelberg: Springer-Verlag pp. 457–465.

Efthimiadis, Efthimis N. 1996. 'Query expansion.' *Annual review of information science and technology* 31, pp. 121–187.

Farah, Mohamed, and Daniel Vanderpooten. 2006. 'A Multiple Criteria Approach for Information Retrieval'. In *SPIRE*, pp. 242–254.

Farah, Mohamed, and Daniel Vanderpooten. 2007. 'L'Agrégation en Recherche d'Information'. In *CORIA*, pp. 125–136.

Friedenson, Bernard. 2007. 'The BRCA1/2 pathway prevents hematologic cancers in addition to breast and ovarian cancers.' *BMC Cancer* 7(1), pp. 152.

Gandon, Fabien. 2008. 'Graphes RDF et leur Manipulation pour la Gestion de Connaissances.'. HDR de l'Université de Nice.

Gandon, Fabien. 2002. '*Ontology Engineering: a Survey and a Return on Experience*.' Rapport de recherche N 4396, INRIA.

Gandon, Fabien, Moussa Lo, and Cheikh Niang. 2008. 'Un modèle d'index pour la résolution distribuée de requêtes sur un nombre restreint de bases d'annotations RDF'. In *Actes d'IC*, pp. 25–35.

Giunchiglia, Fausto, Uladzimir Kharkevich, and Ilya Zaihrayeu. 2009. 'Concept Search'. In *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, eds. Lora Aroyo, Paolo Traverso, Fabio Ciravegna, et al. Springer Berlin / Heidelberg pp. 429–444.

Grabisch, Michel, Toshiaki Murofushi, Michio Sugeno, et al. 2000. '*Fuzzy Measures and Integrals. Theory and Applications*.' Physica Verlag, Berlin.

Grabisch, Michel, Sergei A. Orlovski, and Ronald R. Yager. 1998. 'Fuzzy sets in decision analysis, operations research and statistics'. In ed. Roman Słowiński. Norwell, MA, USA: Kluwer Academic Publishers pp. 31–68.

Gruber, Thomas R. 1993. 'A translation approach to portable ontology specifications.' *Knowl. Acquis.* 5(2), pp. 199–220.

Guarino, Nichola, Claudio Masolo, and Guido Vetere. 1999. 'OntoSeek: content-based access to the Web.' *Intelligent Systems and their Applications, IEEE* 14(3), pp. 70 –80.

Guarino, Nichola, Daniel Oberle, and Steffen Staab. 2009. 'What Is an Ontology?'. In *Handbook on Ontologies*, International Handbooks on Information Systems, eds. Steffen Staab and Dr. Rudi Studer. Springer Berlin Heidelberg pp. 1–17.

Guarino, Nicola, and Christopher A. Welty. 2000. 'A Formal Ontology of Properties'. In *Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management*, EKAW '00, London, UK, UK: Springer-Verlag pp. 97–112.

Haav, Hele-Mai, and Tanel-Lauri Lubi. 2001. 'A Survey of Concept-based Information Retrieval Tools on the Web'. In *5th East-European Conference, ADBIS 2001*, Vilnius, Lithuania.

Harley, John B, Marta E Alarcón-Riquelme, Lindsey A Criswell, et al. 2008. 'Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci.' *Nature Genetics* 40(2), pp. 204–210.

He, Ben, and Iadh Ounis. 2005. 'Term Frequency Normalisation Tuning for BM25 and DFR Models'. In *ECIR*, pp. 200–214.

Hersh, William. 2005. 'Evaluation of biomedical text-mining systems: Lessons learned from information retrieval.' *Briefings in Bioinformatics* 6(4), pp. 344–356.

Hirst, Graeme, and David St Onge. 1998. 'Lexical Chains as representation of context for the detection and correction malapropisms'. In *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, The MIT Press.

Hliaoutakis, Angelos, Giannis Varelas, Epimeneidis Voutsakis, et al. 2006. 'Information retrieval by semantic similarity.' *International Journal on Semantic Web and Information Systems* 2(3), pp. 55–73.

Huang, Chu-ren, Nicoletta Calzolari, Aldo Gangemi, et al. 2010. 'Interfacing ontologies and lexical resources'. In *Ontology and the Lexicon*, Studies in Natural Language Processing, Cambridge University Press pp. 185,200.

Jalabert, Fabien. 2007. 'Cartographie des connaissances : l'intégration et la visualisation au service de la biologie Application à l'ingénierie des connaissances et à l'analyse de données d'expression de gènes.' Thèse de doctorat en Informatique, Université Montpellier 2.

Jansen, Bernard 2000. 'Real life, real users, and real needs: a study and analysis of user queries on the web.' *Information Processing & Management* 36(2), pp. 207–227.

Jansen, Bernard 2000. 'The effect of query complexity on Web searching results.' *Information Research* 6(1), pp. Paper87.

Jiang, Jay J., and David W. Conrath. 1997. 'Semantic similarity based on corpus statistics and lexical taxonomy'. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pp. 19–33.

Joho, Hideo, Mark Sanderson, and Micheline Beaulieu. 2004. 'A Study of User Interaction with a Concept-Based Interactive Query Expansion Support Tool'. In *ECIR*, pp. 42–56.

Khoo, Christopher S. G., and Jin-Cheon Na. 2007. 'Semantic relations in information science.' *Annual Review of Information Science and Technology* 40(1), pp. 157–228.

Kleiber, Georges. 1996. 'Noms propres et noms communs : un problème de dénomination.' *Meta*, pp. 567–589.

Krantz, David H., R. David Luce, Patrick Suppes, et al. 1971. '*Foundations of measurement*.' Academic Press, New York.

Lassila, Ora, and Deborah L. McGuinness. 2001. '*The Role of Frame-Based Representation on the Semantic Web*.' Stanford: Stanford University.

Lin, Dekang 1998. 'An Information-Theoretic Definition of Similarity'. In *ICML*, pp. 296–304.

Lin, Hsien-Tang, Nai-Wen Chi, and Shang-Hsien Hsieh. 2012. 'A concept-based information retrieval approach for engineering domain-specific technical documents.' *Advanced Engineering Informatics*, pp. 349–360.

Lucas, Wendy, and Heikki Topi. 2004. 'Training for Web search: Will it get you in shape?' *Journal of the American Society for Information Science and Technology* 55(13), pp. 1183–1198.

Luhn, H. P. 1957. 'A statistical approach to mechanized encoding and searching of literary information.' *IBM J. Res. Dev.* 1(4), pp. 309–317.

Maedche, Alexander, and Steffen Staab. 2001. 'Ontology learning for the Semantic Web.' *IEEE Intelligent Systems* 16(2), pp. 72–79.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. '*Introduction to information retrieval*.' Cambridge University Press.

Maron, M. E., and J. L. Kuhns. 1960. 'On Relevance, Probabilistic Indexing and Information Retrieval.' *Journal of the ACM* 7(3), pp. 216–244.

Mendes, Pablo N., Max Jakob, and Christian Bizer. 2012. 'DBpedia for NLP: A Multilingual Cross-domain Knowledge Base'. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey.

Mizzaro, Stefano. 1998. 'How many relevances in information retrieval?' *Interacting with Computers* 10(3), pp. 303–320.

Mizzaro, Stefano. 1997. 'Relevance: The whole history.' *Journal of the American Society for Information Science* 48(9), pp. 810–832.

Modave, François, and Michel Grabisch. 1998. 'Preference representation by a Choquet integral: Commensurability hypothesis'. In *7th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'98)*, Paris, France: Editions EDK, Paris pp. 164–171.

Nelsen, Roger B. 1998. '*An Introduction to Copulas (Lecture Notes in Statistics)*.' Springer.

Nenad, Stojanovic. 2005. 'On the query refinement in the ontology-based searching for information.' *Information Systems* 30(7), pp. 543–563.

Oliver, Helen, Gayo Diallo, Ed de Quincey, et al. 2009. 'A user-centred evaluation framework for the Sealife semantic web browsers.' *BMC Bioinformatics* 10(S-10), pp. 14.

Park, Laurence A. F., and Kotagiri Ramamohanarao. 2007. 'Query expansion using a collection dependent probabilistic latent semantic thesaurus'. In *Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining*, PAKDD'07, Berlin, Heidelberg: Springer-Verlag pp. 224–235.

Peat, Helen J., and Peter Willett. 1991. 'The limitations of term co-occurrence data for query expansion in document retrieval systems.' *Journal of the American Society for Information Science* 42, pp. 378–383.

Pedersen, Ted, Serguei V.S. Pakhomov, Siddharth Patwardhan, et al. 2007. 'Measures of semantic similarity and relatedness in the biomedical domain.' *Journal of Biomedical Informatics* 40(3), pp. 288–299.

Pereira, Célia da Costa, Mauro Dragoni, and Gabriella Pasi. 2012. 'Multidimensional relevance: Prioritized aggregation in a personalized Information Retrieval setting.' *Inf. Process. Manage.* 48(2), pp. 340–357.

Pereira, Célia da Costa, Mauro Dragoni, and Gabriella Pasi. 2009. 'Multidimensional Relevance: A New Aggregation Criterion'. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, eds. Mohand Boughanem, Catherine Berrut, Josiane Mothe, et al. Springer Berlin / Heidelberg pp. 264–275.

Pereira, Célia da Costa, and Andrea Tettamanzi. 2006. 'An Ontology-Based Method for User Model Acquisition'. In *Soft Computing in Ontologies and Semantic Web*, Studies in Fuzziness and Soft Computing, ed. Zongmin Ma. Springer Berlin / Heidelberg pp. 211–229.

Pesquita, Catia, Daniel Faria, André O. Falcão, et al. 2009. 'Semantic Similarity in Biomedical Ontologies.' *PLoS Comput Biol* 5(7), pp. e1000443.

Pirró, Giuseppe. 2009. 'A semantic similarity metric combining features and intrinsic information content.' *Data & Knowledge Engineering* 68(11), pp. 1289–1308.

Pirro, Giuseppe, and Jérôme Euzenat. 2010. 'A Feature and Information Theoretic Framework for Semantic Similarity and Relatedness'. In *9th International Semantic Web Conference (ISWC2010)*,

Ponte, Jay M., and W. Bruce Croft. 1998. 'A language modeling approach to information retrieval'. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, Melbourne, Australia pp. 275–281.

Pylkäs, Katri, Hannele Erkko, Jenni Nikkilä, et al. 2008. 'Analysis of large deletions in BRCA1, BRCA2 and PALB2 genes in Finnish breast and ovarian cancer families.' *BMC Cancer* 8(1), pp. 1–5.

Qiu, Yonggang, and Hans-Peter Frei. 1993. 'Concept based query expansion'. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '93, New York, NY, USA: ACM pp. 160–169.

Quillian, Ross. 1968. 'Semantic Memory'. In *Semantic Information Processing*, MIT Press pp. 216–270.

Rada, Roy, Judith Barlow, Jan Potharst, et al. 1991. 'Document Ranking Using An Enriched Thesaurus.' *Journal of Documentation* 47(3), pp. 240–253.

Rada, Roy, H. Mili, E. Bicknell, et al. 1989. 'Development and application of a metric on semantic nets.' *IEEE Transactions on Systems, Man, and Cybernetics* 19(1), pp. 17–30.

Ranwez, Sylvie, Benjamin Duthil, Mohameth François Sy, et al. 2012. 'How Ontology Based Information Retrieval Systems may Benefit from Lexical Text Analysis'. In *New Trends of Research in Ontologies and Lexical Resources*, Theory and Applications of Natural Language Processing, eds. Alessandro Oltramari, P. Vossen, L. Qin, et al. Springer Verlag.

Ranwez, Sylvie, Vincent Ranwez, Jean Villerd, et al. 2006. 'Ontological Distance Measures for Information Visualisation on Conceptual Maps'. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, LNCS, eds. Robert Meersman, Zahir Tari, and Pilar Herrero. Springer pp. 1050–1061.

Resnik, Philip. 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, pp. 95–130.

Reymonet, Axel, Jérôme Thomas, and Nathalie Aussenac-Gilles. 2007. 'Modélisation de ressources termino-ontologiques en OWL'. In *Journées Francophones d'Ingénierie des Connaissances (IC), Grenoble (F), 04/07/2007-06/07/2007*, ed. Francky Trichet. http://www.cepadues.com/: Cépaduès Editions pp. 169–180.

Rijsbergen, C. J. van. 1979. '*Information Retrieval*.' 2nd ed. London: Butterworths.

Robertson, Stephen E., Steve Walker, Susan Jones, et al. 1994. 'Okapi at TREC-3'. In pp. 109–126.

Robertson, Stephen, and Susan Walker. 1994. 'Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval'. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, New York, NY, USA: Springer-Verlag New York, Inc. pp. 232–241.

Robertson, Stephen E. 1977. 'The probability Ranking Principle In IR.' *Journal of Documentation* 33(4), pp. 294–304.

Rocchio, J. J. 1971. 'Relevance feedback in information retrieval'. In *The Smart retrieval system - experiments in automatic document processing*, ed. G. Salton. Englewood Cliffs, NJ: Prentice-Hall pp. 313–323.

Roy, Bernard. 1991. 'The outranking approach and the foundations of electre methods.' *Theory and Decision* 31(1), pp. 49–73.

Salton, Gerard 1969. 'A comparison between manual and automatic indexing methods.' *American Documentation* 20(1), pp. 61–71.

Salton, Gerard 1968. '*Automatic Information Organization and Retrieval*.' McGraw Hill Text.

Salton, Gerard 1986. 'On the use of term associations in automatic information retrieval'. In *Proceedings of the 11th conference on Computational linguistics*, COLING '86, Stroudsburg, PA, USA: Association for Computational Linguistics pp. 380–386.

Salton, Gerard, and Christopher Buckley. 1988. 'Term-weighting approaches in automatic text retrieval.' *Information Processing And Management: an International Journal* 24(5), pp. 513–523.

Salton, Gerard, Edward A. Fox, and Harry Wu. 1983. 'Extended Boolean information retrieval.' *Communications of the ACM* 26(11), pp. 1022–1036.

Salton, G., A. Wong, and C. S. Yang. 1975. 'A vector space model for automatic indexing.' *Commun. ACM* 18(11), pp. 613–620.

Schmid, Helmut. 1994. 'TreeTagger.' *In TC project at the institute for Computational Linguistics of the University of Stuttgart*.

Seco, Nuno, Tony Veale, and Jer Hayes. 2004. 'An Intrinsic Information Content Metric for Semantic Similarity in WordNet'. In *ECAI*, pp. 1089–1090.

Sowa, John F. 2000. 'Ontology, Metadata, and Semiotics'. In *ICCS*, Lecture Notes in Computer Science, Darmstadt, Germany: Ganter Bernhard, Mineau Guy W. pp. 55–81.

Stokoe, Christopher, Michael P. Oakes, and John Tait. 2003. 'Word sense disambiguation in information retrieval revisited'. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, Toronto, Canada pp. 159.

Studer, Rudi, V. Richard Benjamins, and Dieter Fensel. 1998. 'Knowledge Engineering: Principles and Methods.' *Data Knowl. Eng.* 25(1-2), pp. 161–197.

Supekar, Kaustubh, Christopher G Chute, and Harold Solbrig. 2005. 'Representing Lexical Components of Medical Terminologies in OWL.' 2005, pp. 719–723.

Tversky, Amos 1977. 'Features of Similarity'. In *Psychological Review*, pp. 327–352.

Vallet, David, Miriam Fernandez, and Pablo Castells. 2005. 'An Ontology-Based Information Retrieval Model'. In *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pp. 455–470.

Ventresque, Anthony, Sylvie Cazalens, Philippe Lamarre, et al. 2008. 'Enrichissement sémantique de requête utilisant un ordre sur les concepts'. In *EGC 2008*,

Villerd, Jean. 2008. 'Représentations visuelles adaptatives de connaissances associant projection multidimensionnelle (MDS) et analyse de concepts formels (FCA)'. Thèse en Informatique, Ecole des Mines de Paris.

Voorhees, Ellen M. 1994. 'Query Expansion Using Lexical-Semantic Relations'. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, New York, NY, USA: Springer-Verlag New York, Inc. pp. 61–69.

Voorhees, Ellen M., and Donna Harman. 1997. 'Overview of the Sixth Text REtrieval Conference (TREC-6)'. In *TREC*, pp. 1–24.

Wiss, Ulrika, and David Carr. 1998. '*A Cognitive Classification Framework for 3-Dimensional Information Visualization*.'

Wong, S., P. Bollmann, and Y. Yao. 1991. 'Information Retrieval Based on Axiomatic Decision Theory.' *International Journal of General Systems* 19(2), pp. 107–117.

Wu, Zhibiao, and Martha Palmer. 1994. 'Verbs semantics and lexical selection'. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* -, Las Cruces, New Mexico pp. 133–138.

Yager, Ronald R. 2007. 'Aggregation of ordinal information.' *Fuzzy Optimization and Decision Making* 6(3), pp. 199–219.

Yager, Ronald R. 1988. 'On ordered weighted averaging aggregation operators in multicriteria decisionmaking.' *IEEE Trans. Syst. Man Cybern.* 18(1), pp. 183–190.

Yager, Ronald R. 1979. 'Possibilistic decision making.' *IEEE Trans on Systems, Man and Cybernetics* 9, pp. 388–392.

Zargayouna, Haïfa, and Sylvie Salotti. 2004. 'Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML.' Available at: http://hal.archives-ouvertes.fr/hal-00380573/fr/ [Accessed November 25, 2010].

Zipf, George 1949. 'Human Behaviour and the Principle of Least-Effort'. In Cambridge, MA: Addison-Wesley.