OBIRS-feedback, une méthode de reformulation utilisant une ontologie de domaine

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

OBIRS-*feedback*, une méthode de reformulation utilisant une ontologie de domaine

Mohameth François Sy* — Sylvie Ranwez* — Jacky Montmain* — Vincent Ranwez**

** Laboratoire de Génie Informatique et d'Ingénierie de Production*
EMA - Site EERIE, Parc Scientifique Georges Besse
F-30 035 Nîmes cedex 1
prenom.nom@mines-ales.fr

***Institut des Sciences de l'Evolution de Montpellier (ISE-M), UMR 5554 CNRS*
Université Montpellier II, place E. Bataillon, CC 064, F-34 095 Montpellier cedex 5
vincent.ranwez@univ-montp2.fr

RÉSUMÉ. *Les performances d'un système de recherche d'information (SRI) peuvent être dégradées en termes de précision du fait de la difficulté pour des utilisateurs à formuler précisément leurs besoins en information. La reformulation ou l'expansion de requêtes constitue une des réponses à ce problème dans le cadre des SRI. Dans cet article, nous proposons une nouvelle méthode de reformulation de requêtes conceptuelles qui, à partir de documents jugés pertinents par l'utilisateur et d'une ontologie de domaine, cherche un ensemble de concepts maximisant les performances du SRI. Celles-ci sont évaluées, de manière originale, à l'aide d'indicateurs dont une formalisation est proposée. Cette méthode a été évaluée en utilisant notre moteur OBIRS, l'ontologie de domaine MeSH et la collection de tests MuCHMORE.*

ABSTRACT. *The lack of accuracy of an information retrieval system (IRS) may be due to an inadequate formulation of user's queries. Reformulation or query expansion is a possible solution to this problem. In this paper, we introduce a reformulation method based upon a domain ontology. This conceptual relevance feedback method uses a set of documents a user has deemed relevant to search a set of concepts that maximizes IRS performances. Those performances are assessed in an original way, using indicators that are formalized. This method has been evaluated using our environment OBIRS (Ontology Based Information Retrieval System) as base system, MeSH as domain ontology and MuCHMORE as a test collection.*

MOTS-CLÉS: *Reformulation de requête, ontologie, Systèmes de Recherche d'Information, requêtes conceptuelles, expansion de requêtes.*

KEYWORDS: *Relevance feedback, ontology based Information Retrieval Systems, concept-based information retrieval, query expansion.*

1. Introduction

Avec l'augmentation des informations accessibles et l'essor de nouvelles techniques de communication, la recherche d'information constitue un enjeu majeur dans différents secteurs industriels et académiques, en particulier pour les acteurs impliqués dans la recherche scientifique, la veille technologique, l'innovation industrielle, la prise de décision ou la gouvernance d'instituts de recherche. Leur performance et leur capacité d'innovation dépendent fortement de leur capacité à trouver la bonne information, le plus rapidement possible. Durant les dernières décennies, les technologies du Web sémantique et les ontologies de domaine ont confirmé leur efficacité dans les processus de recherche d'information.

Mises en œuvre au sein des Systèmes de Recherche d'Information (SRI), les ontologies de domaine peuvent intervenir dans trois phases de leur processus : i) elles améliorent l'indexation en permettant d'associer des concepts à un document, levant par là l'ambigüité liée à une indexation par mots clés ; ii) elles fournissent un espace conceptuel dans lequel des mesures de similarités ou de proximités sémantiques sont envisageables pour pouvoir mettre en œuvre un processus d'appariement entre requêtes utilisateurs et objets du corpus ; iii) elles permettent de décrire et d'exploiter les relations sémantiques entre concepts pour mieux comprendre et désambigüiser les requêtes des utilisateurs en les reformulant. Nous avons tiré parti des avantages procurés par les ontologies de domaine dans l'environnement OBIRS – *Ontological Based Information Retrieval System* (Ranwez et al., 2010). Dans ce dernier, une agrégation à trois niveaux est proposée pour calculer une mesure de pertinence globale d'un document en fonction de la proximité sémantique des documents du corpus avec chaque concept de la requête d'une part et des préférences de l'utilisateur, d'autre part. Une carte sémantique qui reflète la pertinence des documents sélectionnés est ensuite présentée à l'utilisateur, qui explicite l'adéquation des résultats avec la requête.

Le travail présenté dans cet article s'inscrit dans le cadre de la reformulation de requêtes utilisateurs et vient enrichir l'environnement OBIRS. A partir d'une requête conceptuelle (liste de concepts) initiale, d'un ensemble de documents jugés pertinents et d'une ontologie de domaine, la reformulation est, ici, envisagée comme la recherche d'une nouvelle requête maximisant un indicateur de performance. 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 (section 2) montre, en particulier, que les solutions utilisant une ressource sémantique se contentent de rajouter des concepts en relation (synonymie par exemple) avec ceux de la requête initiale. Après avoir présenté et clairement défini des indicateurs de performance, nous proposons des solutions heuristiques permettant une reformulation rapide même dans le cas d'ontologies contenant un très grand nombre de concepts (section 3). Le modèle a été intégré à l'environnement OBIRS et est évalué (section 4) en utilisant la

collection de résumés scientifiques *MuCHMORE*[1] et suivant le protocole *TREC* (Text REtrieval Conference). Ces travaux ouvrent de nombreuses perspectives qui sont discutées en conclusion de cette contribution.

2. Problématique et état de l'art

L'ambigüité intrinsèque au langage naturel ainsi que la difficulté pour un utilisateur, fut-il un expert du domaine, à exprimer ses besoins en information par une requête qui se résume bien souvent à quelques termes, conduisent la plupart des SRI à mettre en œuvre des mécanismes pour en préciser le sens. La prise en compte du contexte des requêtes des utilisateurs constitue une des solutions proposées (Bhogal et al., 2007) (Hernandez et al., 2007). Plusieurs pistes, dont l'exploitation de l'historique de recherche à long terme au travers de profils des utilisateurs, i.e. la personnalisation (Farah, 2009) (Leung et al., 2010), ainsi que la reformulation automatique ou manuelle de la requête initiale (Hliaoutakis et al., 2006) (Baziz et al., 2007), ont été explorées pour mieux préciser les besoins des utilisateurs.

L'utilisation de ressources sémantiques plus ou moins formalisées (thésaurus et ontologies) dans les SRI permet de lever, jusqu'à un certain niveau, l'ambigüité des requêtes qui, d'une expression en langue naturelle, sont transformées en concepts organisés de différentes manières : listes, réseaux sémantiques, etc. (Giunchiglia et al., 2009). De nombreux travaux utilisent WordNet (Miller, 1995) dans ce sens, notamment ceux de (Voorhees, 1994) et (Baziz et al., 2003). Dans ces méthodes, les concepts correspondant aux mots ou phrases d'une requête sont extraits et celle-ci est étendue en ajoutant des termes en relation sémantique avec les concepts initiaux. D'autres travaux autorisent l'utilisateur à exprimer sa requête directement en utilisant les concepts d'une ontologie (Hliaoutakis et al., 2006) (Ranwez et al., 2010), le corpus étant aussi indexé par des concepts de la même ontologie.

Dans la littérature plusieurs dénominations existent pour désigner le fait de reformuler une requête. Dans (Manning et al., 2008), une catégorisation des méthodes de reformulation est proposée. Un premier type de reformulation est constitué des méthodes globales permettant d'étendre une requête indépendamment des résultats retournés. Elles sont désignées comme des méthodes d'expansion de requêtes et opèrent en amont du processus d'appariement dans les SRI (Baziz et al., 2007). Par opposition aux précédentes, les méthodes locales ajustent la requête initiale à partir des premiers documents ou termes résultats. Il s'agit principalement des méthodes de retour de pertinence (*relevance feedback*) lorsque la reformulation a pour base les documents manuellement jugés pertinents (ou non pertinents) par l'utilisateur ainsi que des méthodes de retour de pertinence aveugle ou implicite (*pseudo relevance feedback*) lorsque les n premiers documents des résultats sont automatiquement supposés pertinents. (Bhogal et al., 2007) présentent un état de

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

l'art des méthodes d'expansion et de retour de pertinence mettant en œuvre des ontologies qu'elles soient de domaine ou génériques.

Notre contribution concerne les méthodes de *relevance feedback* et suit le scénario suivant :

– l'utilisateur soumet une requête initiale simple ;

– le SRI retourne une liste de documents ordonnés selon l'estimation de leur pertinence par rapport à la requête ;

– l'utilisateur marque certains documents comme étant pertinents ;

– la requête initiale est alors reformulée automatiquement pour constituer une nouvelle requête qui va être soumise au SRI ;

– le SRI retourne une nouvelle liste de documents plus en adéquation avec les attentes de l'utilisateur que la première liste.

Bien que l'expression de requêtes en utilisant des concepts d'une ontologie de domaine permette aux utilisateurs d'indiquer à un SRI leurs besoins en information en évitant les problèmes de polysémie et de synonymie, il n'en demeure pas moins qu'une certaine ambigüité peut subsister. (Nenad, 2005) identifie certains facteurs conduisant à l'ambigüité de requêtes conceptuelles. Selon l'auteur, une requête trop générale, i.e. composée par des concepts de haut niveau dans la hiérarchie de l'ontologie, conduit à des résultats bruités. Le problème pointé ici est clairement celui du manque de contenu informationnel des concepts trop génériques. Un autre facteur qui peut conduire à une mauvaise interprétation est la redondance due au fait qu'une partie d'une requête conceptuelle peut être déduite (au sens ontologique) à partir d'une autre partie. Typiquement, il s'agit de cas où l'utilisateur indique, de manière involontaire, des concepts liés par des relations de subsomption par exemple. Ceci montre que, même si les SRI conceptuels permettent de surmonter l'ambigüité du langage naturel, il subsiste des risques de mauvaises interprétations.

Plusieurs méthodes de reformulation par retour de pertinence (*relevance feedback*) ont été proposées dans la littérature et leur efficacité pour améliorer les performances des SRI en termes de précision a été largement étudiée (Abdelali et al., 2007). L'idée de trouver une requête Q' qui, à partir d'un ensemble D_s de documents renvoyés par un SRI en réponse à une requête Q, se rapproche des documents $D_u \subset D_s$ jugés pertinents par l'utilisateur et s'éloigne des documents $D_s \backslash D_u$ supposés non pertinents, date de (Rocchio, 1971) :

$$Q' = \underset{Q}{argmax}\left(sim(Q, D_u) - sim(Q, D_s \backslash D_u)\right) \quad [1]$$

Avec *sim* une mesure de similarité entre une requête Q et un ensemble de documents. Le choix de cette mesure de similarité est fortement dépendant de la collection de documents sous jacente. Dans le cas du modèle vectoriel, avec l'ensemble des termes d'indexation comme base, le cosinus est souvent utilisé comme mesure de similarité. Dès lors, l'équation 1 peut s'écrire :

$$\overrightarrow{Q'} = \alpha\vec{Q} + \beta\frac{1}{|D_u|}\sum_{\vec{D}_i \in D_u}\vec{D}_i - \gamma\frac{1}{|D_s \backslash D_u|}\sum_{\vec{D}_j \in D_s \backslash D_u}\vec{D}_j \quad [2]$$

Les paramètres α, β, et $\gamma \in [0,1]$ permettent de donner plus ou moins de poids à l'un ou l'autre terme suivant que l'on veut mettre l'accent sur la requête initiale (α grand), l'apport des résultats positifs (β grand) ou celui des résultats négatifs (γ grand). Notre approche reprend cette idée, dans le cadre d'un modèle conceptuel de recherche d'information et se donne pour objectif de trouver une requête conceptuelle (un ensemble de concepts) qui maximise sa proximité avec les documents jugés pertinents et minimise sa proximité avec les documents supposés non pertinents.

Dans (Wang et al., 2008), les auteurs proposent une stratégie de reformulation qui consiste à réordonner l'ensemble N des documents non visualisés par l'utilisateur en utilisant l'ensemble D_{np} des documents jugés non pertinents. Le score $RSV(Q, D_i)$ (*Retrieval Status Value*) d'un document $D_i \in N$ par rapport à une requête Q est mis à jour suivant qu'il est plus ou moins proche des documents de D_{np} :

$$\forall\, D_i \in N, S_{combined}(Q, D_i) = RSV(Q, D_i) - \gamma Score\left(Q_{neg}, D_i\right) avec\ \gamma \in [0,1] \quad [3]$$

Q_{neg} est une représentation d'une requête négative construite à partir des documents de D_{np}. Elle est utilisée ici pour évaluer $Score\left(Q_{neg}, D_i\right)$ qui correspond à une estimation de l'adéquation entre les exemples négatifs et un document D_i. Le paramètre γ permet de plus ou moins mettre l'accent sur l'influence des documents non pertinents. Dans le cadre du modèle vectoriel (avec l'ensemble des termes d'indexation comme base), les auteurs considèrent chaque document de D_{np} comme une requête négative (Q_{neg}) dont chaque document $D_i \in N$ doit s'éloigner :

$$S\left(Q_{neg}, D_i\right) = max\left(\cup_{Q \in D_{np}} \{\vec{Q}.\vec{D_i}\}\right) \quad [4]$$

L'utilisation du *max* permet de pénaliser tout document D_i proche d'au moins un document de D_{np} jugé non pertinent. L'objectif est de mettre l'accent sur les exemples négatifs plutôt que sur les exemples positifs dans le cas de requêtes difficiles (Voorhees, 2006).

Notre méthode n'adresse pas les corpus difficiles (requêtes difficiles) et se veut d'abord une approche de reformulation positive même si l'inclusion d'exemples négatifs est prévue. Dans notre stratégie de reformulation, il s'agit de construire une nouvelle requête à soumettre au SRI de base pour trouver des résultats plus pertinents à présenter à l'utilisateur. Ainsi, notre stratégie ne se limite pas à réordonner la liste des documents non présentés à l'utilisateur.

3. Reformulation de requêtes conceptuelles

Dans cette section, nous présentons un cadre formel pour la reformulation de requêtes conceptuelles. Concrètement, nous définissons des indicateurs de performances à améliorer ainsi que des heuristiques permettant de les évaluer rapidement. Nous montrons comment cette approche peut être intégrée dans un SRI existant, à savoir OBIRS, et nous étudions son impact sur les résultats.

NOTE. — Dans la suite, nous appelons D_i les entités indexées que l'on recherche, même s'il ne s'agit pas forcément de documents textuels. Il peut s'agir, d'articles, de gènes, d'images, ou de tout autre élément qui aura été préalablement indexé par des concepts d'une ontologie de domaine.

3.1. *Reformulation de requête : notre proposition*

Comme indiqué dans l'état de l'art, la méthode de reformulation proposée dans cet article s'appuie sur un retour de pertinence de l'utilisateur à travers les documents qu'il juge pertinents parmi ceux qui lui sont présentés.

Considérons les définitions suivantes :

– Θ : une ontologie de domaine,

– $C(\Theta)$: un ensemble de concepts de Θ

– D_S : la liste des documents présentés à l'utilisateur par le SRI,

– D_u : l'ensemble des documents jugés pertinents par l'utilisateur $(D_u \subset D_s)$,

– $C(D_i)$: l'ensemble des concepts indexant un document D_i

– Q_h : une requête quelconque formée d'un ensemble de concepts de Θ,

– Q : une requête initiale,

– Q' : une nouvelle requête obtenue après reformulation,

– $ind(Q_h, D_u, D_S)$: un indicateur sur la qualité de Q_h,

Notre méthode ramène le problème de la reformulation à la recherche d'un ensemble de concepts Q' offrant les meilleures performances par rapport à un indicateur $ind(Q_h, D_u, \mathrm{D}_s)$ qu'il convient de bien choisir :

$$Q' = \underset{Q_h}{argmax}\big(ind(Q_h, D_u, D_s)\big) \qquad [5]$$

Par construction, cette approche est censée fournir l'ensemble des concepts permettant les meilleures performances au sens de l'indicateur défini. Cependant, le gain réel de la reformulation dépend de la pertinence de l'indicateur choisi et de la possibilité d'une mise en œuvre permettant d'identifier une requête optimale ou qui s'en approche. Dans la suite plusieurs indicateurs sont discutés.

3.1.1. *Différents indicateurs de performances*

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 $ind(Q_h, D_u, D_s)$ 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. Ces mesures sont calculées en considérant D_s comme corpus dans lequel nous connaissons l'ensemble D_u des documents pertinents (au sens de l'utilisateur). La mesure de *R-precision* (Equation 6) peut constituer un bon

compromis entre *précision* et *rappel*. En effet, il s'agit de la *précision* au $|D_u|$-ième document, point pour lequel la *précision* équivaut au *rappel* (avec D_s comme corpus). Cet indicateur sera maximal si tous les documents de D_u sont classés en tête de liste des résultats obtenus après reformulation. Si nbr est le nombre de documents pertinents (documents inclus dans D_u) effectivement retournés par une requête Q_h jusqu'à la $|D_u|$-ième position :

$$ind(Q_h, D_u, D_s) = \frac{nbr}{|D_u|} \quad [6]$$

Avec la définition d'indicateur donnée précédemment, une reformulation strictement positive est opérée. On cherche à se rapprocher des documents de D_u et aucun cas n'est fait des documents non pertinents potentiels.

Pour intégrer une reformulation négative, on peut combiner deux indicateurs élémentaires, l'un sur les documents positifs et l'autre sur ceux négatifs. L'idée est de trouver une requête Q_h dont les résultats s'approchent des documents de D_u et s'éloignent de ceux de $D_s \backslash D_u$:

$$ind(Q_h, D_u, D_s) = \underset{D_i \in D_u}{agreg}\left(RSV(Q_h, D_i)\right) - \gamma \underset{D_j \in D_s \backslash D_u}{agreg'}\left(RSV(Q_h, D_j)\right) \quad [7]$$

Avec $\gamma \in [0,1]$ et $RSV(Q_h, D_i)$ un score de pertinence du document D_i par rapport à la requête Q_h. Cette solution reprend le schéma classique de l'équation 1. Il est également à noter que les deux agrégations peuvent ne pas être de même nature. En pratique, soit l'utilisateur indique explicitement les documents qui ne correspondent pas à son besoin, soit on peut faire l'hypothèse que les documents de $D_s \backslash D_u$ qu'il a vus et pour lesquels il n'a pas donné d'avis favorable ne sont pas pertinents. Le paramètre γ permet de pondérer l'apport des documents non pertinents dans l'évaluation de l'indicateur. La valeur de l'indicateur ainsi défini est élevée lorsque $RSV(Q_h, D_i)$ est élevée pour tout document de D_u et faible pour tout document de $D_s \backslash D_u$.

$RSV(Q_h, D_i)$ dépend du système de base dans lequel la méthode de reformulation s'intègre. La section 4 détaille le modèle d'agrégation utilisé comme modèle de pertinence pour évaluer l'adéquation entre un document et une requête.

3.1.2. *Sélection des concepts de l'ontologie utilisés pour la reformulation*

Dans la méthode proposée, le nombre de reformulations possibles est exponentiel : typiquement $2^{|\theta|}$. Il est évident que tester toutes les combinaisons de concepts n'est pas une solution réaliste. Nous proposons donc une heuristique pour trouver une solution approchée en un temps raisonnable. Il s'agit d'une approche gloutonne qui à partir d'un ensemble de concepts $C(\Theta)$ et d'une requête initiale Q, enrichit de manière itérative Q' tant que cela permet d'améliorer l'indicateur de performance choisi. L'algorithme 1 détaille la stratégie de recherche de Q'.

Dans l'algorithme 1, si $C(\Theta)$ correspond à l'ensemble des concepts de l'ontologie Θ (i.e. $|C(\Theta)| = |\Theta|$), le temps de calcul est trop élevé pour une ontologie de grande taille. Dans le cadre d'une approche heuristique, il est raisonnable de ne tester que des concepts au voisinage sémantique des concepts indexant les documents de D_u. Si $\pi(C_x, C_y)$ est une mesure de proximité sémantique entre C_x et C_y deux concepts de Θ, nous construisons $C(\Theta)$ de la manière suivante :

$$C(\Theta) = \left\{C_i \in \Theta \,/\exists C_j \in \bigcup_{D_i \in D_u} C(D_i), \pi\left(C_j, C_i\right) > \varepsilon\right\} \quad [8]$$

L'équation précédente, bien que réduisant l'espace de concepts, transpose le problème dans la construction de $C(\Theta)$. En effet, la solution naïve pour construire $C(\Theta)$ nécessite de parcourir toute l'ontologie Θ. Cependant, il est possible de construire $C(\Theta)$ en identifiant pour chaque concept indexant un document de D_u les concepts de son voisinage qui doivent être considérés. Il est alors possible, pour certaines mesures de proximités sémantiques, de construire $C(\Theta)$ en parcourant son seul voisinage et non la totalité de l'ontologie. Cette amélioration utilisée dans OBIRS-*feedback* avec la mesure de similarité de Lin (Lin, 1998) sort du cadre de cet article.

Données : $\boldsymbol{D_u}$ *// l'ensemble des documents jugés pertinents*
$\boldsymbol{D_s}$ *// l'ensemble des documents présentés à l'utilisateur*
$\boldsymbol{C(\Theta)}$ *// un ensemble de concepts de l'ontologie* $\boldsymbol{\Theta}$
$\boldsymbol{Q}$ *// la requête initiale*
Résultat : $\boldsymbol{Q'}$ *// requête maximisant l'indicateur de performance choisi*

$\boldsymbol{Q' = Q}$;
$\boldsymbol{ind = ind(Q, D_u, D_s)}$;
Faire
  $\boldsymbol{bestInd = ind}$;
  improve = False ;
  Pour chaque concept $\boldsymbol{C \in C(\Theta)}$ **Faire**
    Si $(\boldsymbol{ind(Q' \cup \{C\}, D_u, D_s) > bestInd})$ **Alors**
      $\boldsymbol{bestC = C}$;
      $\boldsymbol{bestInd = ind(Q' \cup \{C\}, D_u, D_s)}$ **;**
    Fin Si
  Fin Pour
  Si ($\boldsymbol{bestInd}$ > *ind*) **Alors**
    $\boldsymbol{ind = bestInd;}$
    $\boldsymbol{Q' = Q' \cup \{bestC\}}$;
    improve = True ;
  Fin Si
Tant Que (*improve*)
renvoyer ($\boldsymbol{Q'}$) ;

Algorithme 1 : *Recherche d'une requête* Q' *maximisant* $ind(Q_h, D_u, D_s)$

3.2. *Intégration dans l'environnement OBIRS*

La méthode que nous proposons s'appuie sur le système OBIRS[2] – *Ontology based Information Retrieval System* (Ranwez et al., 2010) dans lequel elle introduit un processus de reformulation de requêtes.

OBIRS est un SRI permettant la formulation assistée de requêtes à base de concepts d'une ontologie de domaine et mettant en œuvre un modèle de pertinence utilisant des proximités sémantiques.

Dans un premier temps, la proximité sémantique π entre un concept C_x d'une requête Q_h et un document D_i est définie comme une agrégation des proximités entre C_x et les concepts de $C(D_i)$:

$$\pi(C_x, D_i) = \underset{C_y \in C(D_i)}{agreg}\left(\pi(C_x, C_y)\right) \quad [9]$$

La mesure de proximité sémantique $\pi(C_x, C_y)$ peut être basée sur un calcul du plus court chemin, sur des mesures utilisant l'*Information Content* (IC) (Resnik, 1999) (Lin, 1998) ou encore sur des mesures ensemblistes telles que celle exposée dans (Ranwez et al., 2006).

Une stratégie de *best match* peut être mise en œuvre pour établir la proximité $\pi(C_x, D_i)$ en choisissant l'opérateur *max*. On aura alors :

$$\pi_{max}(C_x, D_i) = \max_{C_y \in C(D_i)} \pi(C_x, C_y) \quad [10]$$

L'adéquation entre un document et une requête est évaluée, dans un deuxième temps, en utilisant la famille d'opérateurs d'agrégation proposée par Yager (Yager, 1979). Chaque concept C_x d'une requête Q étant considéré comme un critère, il s'agit de considérer les documents du corpus comme des alternatives pour lesquelles l'évaluation par rapport aux critères est donnée par $\pi_{max}(C_x, D_i)$:

$$RSV(Q, D_i) = \underset{C_x \in C(Q)}{agreg}\left(\pi(C_x, D_i)\right) = \left(\frac{\left(\sum_{x=1}^{|Q|} \pi_{max}(C_x, D_i)^q\right)}{|Q|}\right)^{1/q}, q \in \mathbb{R} \quad [11]$$

Ce modèle d'agrégation permet d'adjoindre un modèle des préférences de l'utilisateur à la notion de proximité sémantique pour définir la pertinence globale d'un document vis-à-vis d'une requête. La préférence est ici exprimée comme l'exigence plus ou moins prononcée de la satisfaction simultanée de tous les critères, i.e. la contrainte plus ou moins nécessaire que tous les concepts de la requête soient sémantiquement proches de D_i. En effet, si dans l'équation 11, q tend vers $-\infty$ alors la requête tend à être conjonctive (l'agrégation tend vers le *min*), lorsque q tend vers $+\infty$ elle tend à être disjonctive (l'agrégation tend vers le *max*).

[2] http://www.ontotoolkit.mines-ales.fr/ObirsClient/

Ce score nous permet d'ordonner les documents de façon à ne proposer à l'utilisateur que ceux qui ont le meilleur score. Il permet aussi d'expliquer le choix effectué par le système à travers une visualisation par carte sémantique aussi bien des documents que de la mesure de l'adéquation de chaque concept de la requête à chaque document.

Le système OBIRS est utilisé comme système de base dans la section suivante pour évaluer notre approche de reformulation.

4. Evaluations et tests

4.1. *Protocole expérimental*

L'évaluation de notre modèle a été réalisée en utilisant *MuCHMORE*, un corpus regroupant des résumés de documents scientifiques dans le domaine médical. Chaque résumé est indexé par des concepts de l'ontologie MeSH (*Medical Subject Headings*) comportant environ 25603 concepts. Des requêtes, exprimées par des listes de concepts de la même ontologie, sont disponibles ainsi que les documents pertinents leur correspondant. Le tableau 1 indique les détails de la collection.

***MuCHMORE* Springer Corpus English (2001)**				
Nombre de documents	Nombre de requêtes	Nombre de concepts de la collection	Nombre moyen de concepts indexant un document	Jugements
7823	**23**	**8215**	**12**	**Expert**

Tableau 1. *Description du corpus MuchMore*

Le scénario suivant est adopté pour simuler une activité de recherche : si D_s est l'ensemble des documents présentés à l'utilisateur après une requête Q et D_e les documents que les experts ont jugés pertinents pour la même requête lors de la constitution de la collection *MuCHMORE*, alors les documents de $D_s \cap D_e$ constituent l'ensemble D_u qu'aurait pu sélectionner un utilisateur.

Les performances et l'apport de OBIRS ont déjà fait l'objet d'une évaluation (Ranwez et al., 2010), l'objectif est ici de mesurer l'apport de la stratégie de feedback présenté dans cet article. Néanmoins, dans le cadre de cette comparaison entre OBIRS-*feedback* et OBIRS, nous fournissons également à titre indicatif les performances de deux stratégies booléennes simple : AND (tous les concepts de la requête doivent être présents dans l'indexation des documents jugés pertinents) et OR (au moins un des concepts doit être présent). Pour chaque requête, nous avons considérés les 100 premiers résultats. Ainsi les performances sont mesurées en utilisant la précision moyenne sur toutes les requêtes de la collection (*MAP : Mean*

Average Precision) et la courbe de *précision rappel*. La mesure de similarité conceptuelle utilisée dans ces expérimentations est celle de Lin : sim_{Lin} (Lin, 1998), avec une évaluation du contenu informationnel (IC) correspondant à celle de (Seco et al., 2004). L'IC d'un concept C_x est donné par la probabilité $P(C_x)$ de présence de ce concept ou de ses descendants dans un corpus $\left(IC(C_x) = -log\left(P(C_x)\right)\right)$. L'estimation de Seco est basée sur la structure de l'ontologie et considère que plus un concept C_x a d'hyponymes $\left(hypo(C_x)\right)$, plus grande est cette probabilité. Si $MICA\left(C_x, C_y\right)$ des concepts C_x et C_y est leur ancêtre commun ayant le plus grand IC, alors :

$$sim_{Lin}(C_x, C_y) = \frac{2*IC\left(MICA(C_x,C_y)\right)}{IC(C_x)+IC(C_y)} \ , IC(C_x) = 1 - \frac{log(hypo(C_x)+1)}{log(|\Theta|)} \qquad [12]$$

La valeur du paramètre q de la famille d'opérateurs d'agrégation utilisée (Equation 11) est fixée à 2.0.

4.2. *Résultats*

Nous avons comparé les résultats obtenus par OBIRS et une stratégie booléenne (systèmes de base) avec ceux obtenus après la phase de reformulation notamment en mettant en œuvre l'indicateur défini par l'équation 7. Cette section étudie l'impact du paramètre de pondération (γ) des documents négatifs et celui de la restriction de l'espace de recherche $C(\Theta)$.

4.2.1. Impact du paramètre de combinaison (γ) sur la *précision* et le *rappel*

Dans cette section, nous montrons l'évolution de la *précision* moyenne (MAP : *Mean Average precision*) et du *rappel* (Figure 1) en fonction du paramètre de pondération γ des documents négatifs (Equation 7). Onze valeurs sont considérées pour ce paramètre. Pour cela, tous les concepts de l'ontologie sont considérés (i.e. $|C(\Theta)| = |\Theta|$ et $\varepsilon = 0$).

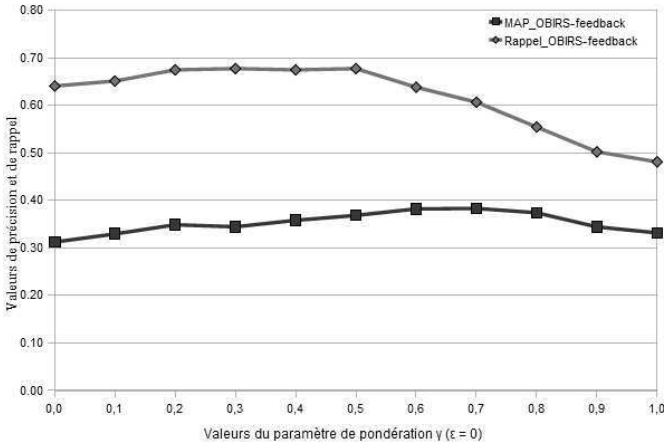

Figure 1. *Evolution du rappel et de la précision moyenne en fonction de* γ

Quand la valeur du paramètre de pondération γ est supérieure à 0.8, le *rappel* est dégradé (Figure 1). Pour ces valeurs extrêmes, la méthode de reformulation donne une grande importance aux documents négatifs. Pour $\gamma = 0.5$, nous obtenons le meilleur *rappel* ainsi qu'une bonne précision. C'est pourquoi dans la suite, nous utilisons cette valeur.

4.2.2. Impact des concepts de l'espace de recherche $C(\Theta)$

Dans cette partie, nous montrons l'impact (Figure 2) du nombre de concepts de $C(\Theta)$ (Equation 8). Ce nombre dépend du seuil ε et varie entre $|\Theta|$ (le nombre de concepts de l'ontologie) et $\left|\bigcup_{D_i \in D_u} C(D_i)\right|$ (la taille de l'union des concepts indexant les documents jugés pertinents). Les cas extrêmes correspondent respectivement à $\varepsilon = 0$ et $\varepsilon = 1$. Nous avons fait varier le seuil ε de 0 à 1 par pas de 0.1 et pour chacune de ces valeurs une courbe de *précision rappel* est produite.

Il n'y a presque pas de différence entre les valeurs de *précision* pour les divers seuils de ε comme en témoigne le chevauchement des courbes (Figure 2). Quand ε vaut 1, i.e. $C(\Theta)$ correspond à l'union des concepts indexant les documents jugés pertinents, la *précision* est meilleure sur quelques points de *rappel*. Les concepts assez éloignés de ceux indexant les documents de D_u (lorsque $\varepsilon \neq 1$) n'apportent pas de la pertinence, mais le bruit qu'ils introduisent est neutralisé par le fait que nous cherchons les concepts qui maximisent notre indicateur. Pour toutes valeurs de ε, la *précision* est meilleure que celle du système de base (OBIRS).

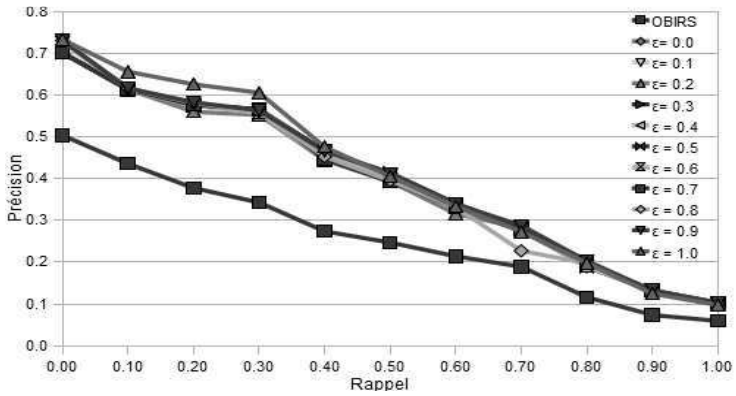

Figure 2. *Courbe de rappel précision en fonction de ε ($\gamma = 0.5$)*

4.2.3. Nombre de concepts moyen dans les requêtes reformulées

Il est intéressant de voir la taille des requêtes reformulées suivant le seuil ε fixé. Ce nombre varie peu même si on s'éloigne du groupe de concepts très proches de ceux qui indexent les documents jugés pertinents. Cela signifie que la taille de $C(\Theta)$ a peu d'influence sur la requête reformulée aussi bien sur sa qualité (comme montré dans la section précédente) que sur sa taille.

ε	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Taille moyenne des requêtes reformulées	8.04	8.13	8.04	8.13	7.9	7.77	7.22	5.22
Taille moyenne de $C(\Theta)$	11326	7719	4855	2512	1198	544	219	84.1

Tableau 2. *Taille moyenne des requêtes reformulées*

Le tableau 2 ainsi que la figure 2 montrent que des valeurs plus ou moins grandes de ε n'impliquent pas de perte de qualité sur la *précision*. Dès lors qu'il n'est pas nécessaire de trop s'éloigner (ε faible) des concepts de $\left|\bigcup_{D_i \in D_u} C(D_i)\right|$, nous choisissons pour la suite une valeur de 0.5 pour ε. Cela nous permet de gagner en temps de calcul du fait de la taille réduite de $C(\Theta)$ sans dégrader les résultats obtenus.

4.2.4. Apports de OBIRS-*feedback*

La figure 3 compare les performances de OBIRS et celles de OBIRS-*feedback* en termes de *précision*. Les courbes de *précision rappel* des deux stratégies de recherche booléennes simples AND et OR sont données à titre indicatif. L'approche OBIRS-*feedback* est testée en utilisant $\gamma = 0.5$ pour la pondération des exemples négatifs et $\varepsilon = 0.5$ pour la construction de l'espace de recherche $C(\Theta)$. L'écart de performance est significatif entre OBIRS et OBIRS-*feedback* en termes de *précision* moyenne (0.3708 pour OBIRS-*feedback* et 0.2318 pour OBIRS, soit 13,9 % d'amélioration) et de *rappel* (0,67277 pour OBIRS-*feedback* et 0,55236 pour OBIRS, soit 12,041 % d'amélioration).

Il est intéressant de souligner que le temps moyen d'exécution de OBIRS-*feedback* sur un ordinateur doté d'un système Linux (*Debian*, 8 Go de mémoire vive et processeur double cœur de 2.7 GHZ 64 bits) varie de 1,654 s (pour ε = 1) à 395,832 s (pour ε = 0). Ce temps inclus la construction de $C(\Theta)$ et la recherche de Q' dont la taille moyenne est de 8.04 concepts comme indiquée dans le tableau 2.

La collection sur laquelle repose notre évaluation est annotée par une ancienne version du MeSH et un travail de mise à niveau de cette annotation a été effectué jusqu'à obtenir un taux de couverture de 80 % avec une version de l'ontologie plus récente (2010). Par ailleurs, la taille relativement réduite de ce corpus peut biaiser nos résultats. Un test sur le corpus *TREC Genomics Tract 2006* contenant 162 259 documents est envisagé.

La collection MuCHMORE est constituée de résumés de documents scientifiques (biomédicaux). Ces derniers sont donc de taille homogène et l'ontologie de domaine MeSH permet de les représenter de manière satisfaisante. Notre approche concerne la recherche *ad hoc* de documents (différente de la

recherche de passages) et nécessite une collection de documents indexés par des concepts issus d'une ontologie de domaine (la granule est ici le document). Le travail d'indexation en soit (extraction de concepts, pondération) n'entre pas dans le cadre de cet article.

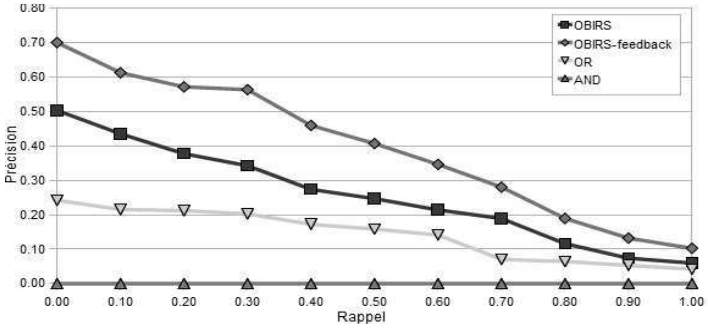

Figure 3. *Courbes de précision rappel pour les approches OBIRS, AND, OR et OBIRS-feedback ($\varepsilon = \gamma = 0.5$)*

5. Conclusion et perspectives

Nous avons développé dans cet article une méthode de reformulation de requêtes utilisant une ontologie de domaine et mettant à contribution l'utilisateur par le biais de ses jugements sur les documents qui lui sont présentés. Le problème de la reformulation est formalisé comme étant la recherche d'un sous-ensemble de concepts maximisant un indicateur de performance. A notre connaissance, il s'agit de la première application, dans le domaine conceptuel, d'une technique éprouvée dans le domaine vectoriel. Les résultats obtenus, en intégrant cette approche de reformulation à notre environnement OBIRS, montrent que la *précision* et le *rappel* de celui-ci sont sensiblement améliorés et cela en un temps court (quelques secondes) même si l'ontologie comporte un grand nombre de concepts. Ce gain de temps est obtenu grâce à la mise en œuvre d'heuristiques nous permettant de réduire l'espace des concepts dans lequel chercher la requête optimale. De nouveaux documents pertinents sont ainsi retrouvés après reformulation et sont bien classés.

L'approche présentée dans cet article explore de nouvelles voies concernant la reformulation conceptuelle. Elle fournit un cadre général pour mettre en œuvre la reformulation dans la plupart des SRI, du moment qu'ils utilisent une ontologie de domaine dans leur processus de pertinence. Aussi, elle définit une famille de méthodes de reformulation grâce d'une part à l'intégration d'un modèle de préférences des utilisateurs permettant différents types d'agrégations et d'autre part à la pondération de l'apport des documents négatifs. Les documents jugés pertinents par l'utilisateur peuvent couvrir plusieurs thèmes n'étant pas forcément indiqués dans sa requête initiale. Nous envisageons de partir d'une requête vide pour prendre en compte l'essentiel des thèmes que recouvrent les documents pertinents sans être

contraints par la requête initiale. Par ailleurs, cette construction rapide d'un ensemble de concepts d'intérêts peut avoir des applications importantes en personnalisation durant la phase d'apprentissage des profils des utilisateurs.

Remerciements

Cette publication est le résultat de la collaboration entre le LGI2P/Ecole des Mines d'Alès et l'ISEM/Université Montpellier 2.

6. Bibliographie :

Abdelali A., Cowie J., Soliman H. « Improving query precision using semantic expansion ». *Information Processing & Management*, vol. 43, n°3, 2007, p. 705-716.

Baziz M., Aussenac-Gilles N., Boughanem M. « Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information ». *Actes du 21e Congrès INFORSID*, Nancy, 24-27 Mai 2003, France, p. 121-134.

Baziz M., Boughanem M., Pasi G., Prade H. « An information retrieval driven by ontology from query to document expansion ». In : *Proceedings of the Large Scale Semantic Access to Content (Text, Image, Video, and Sound)* RIAO'2007, Pittsburgh, 30-01 June, USA, CID : Paris, France, p. 301–313.

Bhogal J., Macfarlane A., Smith P. « A review of ontology based query expansion », *Information Processing & Management*, vol. 43, n°4, 2007, p. 866-886.

Farah M. « Ordinal Regression Based Model for Personalized Information Retrieval », In : *Advances in Information Retrieval Theory*, 2009, Springer Berlin / Heidelberg, p. 66-78.

Giunchiglia F., Kharkevich U., Zaihrayeu I. « Concept Search ». In : *The Semantic Web: Research and Applications*, 2009, Springer Berlin / Heidelberg, p. 429-444.

Hernandez N., Mothe J., Chrisment C., Egret D., « Modeling context through domain ontologies ». Information Retrieval, vol. 10, n°2, p. 143-172.

Hliaoutakis A., Varelas G., Voutsakis E., Petrakis E. G., Milios E. « Information retrieval by semantic similarity », International Journal on Semantic Web and Information Systems, vol. 2, n°3, 2006, p. 55–73.

Leung K. W., Lee D. L. « Deriving Concept-Based User Profiles from Search Engine Logs », *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, n°7, 2010, p. 969-982.

Lin D. « An Information-Theoretic Definition of Similarity ». In : *Proceedings of the 15th International Conference on Machine Learning ICML'1998*, Madison,

Wisconsin, 24-27 July 1998, USA, Morgan Kaufmann, p. 296-304.

Manning C. D., Raghavan P., Schütze H. *Introduction to information retrieval*, Cambridge University Press, 2008.

Miller G. A. « WordNet: a lexical database for English », *Communications of the ACM*, vol. 38, n°11, 1995, p. 39-41.

Nenad S. « On the query refinement in the ontology-based searching for information », *Information Systems*, vol. 30, n°7, 2005, p. 543-563.

Ranwez S., Ranwez V., Sy M., Montmain J., Crampes M. « User Centered and Ontology Based Information Retrieval System for Life Sciences », In : *Proceedings of the Workshop on Semantic Web Applications and Tools for Life Science, Berlin, 8-10 December 2010, Germany*.

Ranwez S., Ranwez V., Villerd J., Crampes M. « Ontological Distance Measures for Information Visualisation on Conceptual Maps », *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, vol. 4278, 2006, Springer, p. 1050–1061.

Resnik P. « Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language », *Journal of Artificial Intelligence Research*, vol. 11, 1999, p. 95-130.

Rocchio J. J. « Relevance feedback in information retrieval », In : *The Smart retrieval system - experiments in automatic document processing*, 1971, Englewood Cliffs, NJ: Prentice-Hall, p. 313–323.

Seco N., Veale T., Hayes J. « An Intrinsic Information Content Metric for Semantic Similarity in WordNet », In : *Proceedings of the 16th Eureopean Conference on Artificial Intelligence ECAI'2004*, Valencia, 2004, Spain, IOS Press, p. 1089-1090.

Voorhees E. « Query expansion using lexical-semantic relations », In : *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, Dublin, 1994, Ireland, Springer-Verlag New York, Inc., p. 61-69.

Voorhees E. M. « The TREC 2005 robust track », *ACM SIGIR Forum*, vol. 40, n°1, 2006, p. 41.

Wang X., Fang H., Zhai C. « A study of methods for negative relevance feedback », In : *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, New York, 2008, USA, ACM, p. 219–226.

Yager RR. « Possibilistic decision making », *IEEE Trans on Systems, Man and Cybernetics*, vol. 9, 979, p. 388-392.