Choix d'une mesure d'association pour une extension de requête contrôlée : la question de l'orientation de la mesure

Choix d'une mesure d'association

Christophe Brouard

Equipe MRIM, Laboratoire d'Informatique de Grenoble
Bâtiment IMAG B
385 avenue de la Bibliothèque
38400 Saint-Martin d'Hères
Christophe.Brouard@imag.fr

RÉSUMÉ. *Cet article présente une étude comparative de mesures d'association dans le contexte de la construction automatique de thésaurus. L'étude porte plus particulièrement sur la question de l'orientation de la mesure d'association. Différentes solutions sont distinguées et testées dans le cadre d'une tâche de filtrage adaptatif dans laquelle le thésaurus est utilisé pour sélectionner des termes d'indexation à ajouter au cours de l'apprentissage. Les résultats obtenus sur le corpus OSHUMED montrent une forte influence de l'orientation considérée.*

ABSTRACT. *This paper describes a comparative study of association measures in the context of thesaurus construction. The study deals more particularly with the orientation of the association measure. Several solutions are identified and tested in an adaptive filtering task. In this task, the thesaurus is used to select the terms to take into account in the learning step. The results obtained on the OSHUMED corpus show the important role played by the orientation of the association measures.*

MOTS-CLÉS : *thésaurus, fouille de texte, mesures d'association, filtrage adaptatif.*

KEYWORDS: *thesaurus, text mining, association measures, adaptive filtering.*

1. Introduction

Un thésaurus considère un ensemble de termes et recense les différentes relations sémantiques par lesquelles ces termes sont liés. De très nombreux types de relations sémantiques peuvent être répertoriés. Il peut s'agir de relation de synonymie, ou quasi-synonymie, d'hyponymie (est un type de), de méronymie (partie de), etc... L'utilisation de thésaurus dans le cadre de tâches de recherche d'information est courante. On peut notamment citer le cas d'utilisation le plus fréquent consistant en l'extension (ou expansion) de requête. L'extension de requête consiste à ajouter à une requête à base de mots-clefs, les mots liés sémantiquement (le plus souvent les synonymes et les hyponymes) permettant ainsi de retrouver des documents pertinents contenant par exemple des synonymes mais pas nécessairement les mots-clefs initialement présents dans la requête.

Un thésaurus peut être construit manuellement. Le thésaurus Wordnet (Miller, 1995) en est l'exemple le plus populaire. Cependant, la nécessité de construire des thésaurus spécifiques à des domaines ainsi qu'à des modes d'utilisation particuliers ainsi que le coût de construction et de maintenance de tels thésaurus rendent souhaitable la définition de méthodes de construction automatique. Une assez grande diversité de méthodes de construction automatique ont été définies[1]. La méthode la plus répandue est de nature statistique. Elle consiste à établir une relation entre deux termes sur la base directe de leur présence dans les mêmes textes ou les mêmes passages de texte. Cette méthode de construction repose sur l'hypothèse que plus les termes surviennent ensemble, plus ils ont de chance d'être liés sémantiquement. D'autres méthodes procèdent de façon plus indirecte en comparant le contexte des termes, c'est-à-dire les termes qui co-occurrent respectivement avec les deux termes dont on étudie la relation. Il sera par exemple possible de déduire une relation de synonymie entre deux termes A et B si les termes qui accompagnent A et B sont les mêmes. Enfin, sans être exhaustif, on peut ajouter l'utilisation possible de la connaissance portant sur les catégories syntaxiques des termes étudiés (Grefenstette, 1994).

Dans la suite, nous considérons le premier type de construction mentionné ci-dessus se basant sur la présence des termes dans les mêmes documents. Dans ce cas, une mesure d'association entre deux termes rendant compte de la présence simultanée des deux termes est requise. La valeur de cette mesure doit croître avec le nombre de co-occurrences des deux termes. Pour le reste, il est possible d'imaginer différentes mesures d'association. Différentes mesures ont été proposées et on trouve quelques études comparatives empiriques (Chung et Lee 2001; Kim et Choi, 1999).

Les études comparatives considèrent des mesures d'association symétriques, c'est-à-dire des mesures M telles M(A,B)=M(B,A). Nous proposons dans la suite de cet article de considérer des mesures asymétriques et nous posons la question du choix de l'orientation de la mesure qui en découle. Nous testons les mesures par

[1] On pourra se reporter à (Bruandet et Chevallet 2003) pour une synthèse.

l'intermédiaire d'un test classique d'extension de requête, chaque mesure d'association donnant lieu à une extension différente. Cependant, nous considérons une forme d'extension contrôlée dans laquelle l'utilisation du thésaurus n'est pas aveugle dans le sens où elle peut s'appuyer sur les retours de pertinence donnés par l'utilisateur. On peut ainsi, s'attendre à ce que le contexte de test proposé supporte mieux le bruit inhérent à la nature statistique de la méthode de construction utilisée. Les performances n'étant alors plus aussi sensibles au bruit, on peut espérer ainsi mesurer plus finement la qualité du thésaurus construit en évitant la situation dans laquelle aucune amélioration n'est constatée quelle que soit la mesure d'association considérée.

La section suivante pose la question du choix de la mesure et en particulier celui de l'orientation de la mesure et décrit les motivations de notre étude. La section 3 présente les systèmes de filtrage adaptatif et l'utilisation du thésaurus d'association que nous proposons. La section 4 distingue plusieurs possibilités d'orientation de la mesure d'association et les teste expérimentalement sur le corpus OHSUMED. Enfin, nous concluons par un bilan en listant les éléments de réponses apportés par cette étude et en dégageant les perspectives de ce travail.

2. Thésaurus d'association

2.1. *Les mesures d'association les plus utilisées*

Dans la suite de cet article, par "thésaurus d'association", nous désignons un graphe dont les nœuds représentent des termes (éventuellement lemmatisés) et dont la valeur des arcs non typés entre les nœuds croît avec le nombre de documents ou passages de texte dans lesquels les deux termes (correspondant aux nœuds reliés) co-occurrent. En introduction, nous avons succinctement présenté la notion de thésaurus construit automatiquement sur la base de calcul statistique s'appuyant sur la co-occurrence des termes. Différentes mesures peuvent être proposées pour valuer ces liens et on peut s'interroger sur l'efficacité et les propriétés souhaitables de telles mesures. Parmi les mesures les plus souvent utilisées on retrouve essentiellement des mesures symétriques correspondant à une fraction rapportant la grandeur du cardinal de l'ensemble des documents contenant les 2 termes à la grandeur des cardinaux des ensembles de documents contenant respectivement chacun des deux termes. Les mesures Jaccard utilisé par Miyamoto (1990), Dice et Cosinus utilisés dans (Frakes et Yates, 1992) sont de ce type. On trouve assez rarement l'utilisation de mesures asymétriques. Néanmoins, on peut citer les travaux de Haddad (2002) sur l'utilisation de règles d'association. Dans cette étude, la mesure d'association considérée pour un terme t1 est la fréquence relative de t2 sachant t1 aussi appelée confiance de la règle $t1 \rightarrow t2$. A notre connaissance, aucune étude ne traite du problème de l'orientation de la mesure.

2.2. *La question de l'orientation de la mesure*

Notre questionnement sur l'orientation de la mesure a deux raisons. La première part du constat que lorsque l'on utilise un thésaurus construit manuellement l'extension est souvent réalisée par ajout de synonymes et d'hyponymes. C'est par exemple le cas du mécanisme d'extension intégré au système de recherche associé à la base de documents MEDLINE[2]. On ajoute donc des termes plus spécifiques que les termes présents dans la requête et non pas plus génériques. En effet, si une requête porte sur la culture des pommes, on peut légitimement penser que l'utilisateur peut s'intéresser à un document traitant de la culture de la golden. Par contre, son intérêt pour des documents traitant de la culture des fruits est moins évident. Un sens est donc généralement privilégié et quoi qu'il en soit, la question de l'orientation des relations est un sujet d'interrogation. On peut donc aussi se demander si dans le cas d'une construction automatique l'orientation n'est pas aussi un aspect important de la mesure à prendre en compte.

La seconde raison de notre questionnement au sujet de l'orientation part d'une décomposition de la mesure du cosinus. Comme nous le verrons au début de la section 4, la mesure du cosinus fréquemment utilisée correspond au produit de la racine carrée de la fréquence de t1 sachant t2 (c'est-à-dire la confiance de $t2 \rightarrow t1$) avec celle de t2 sachant t1 (c'est-à-dire la confiance de $t1 \rightarrow t2$). Si on ne s'attache qu'à établir un ordre entre les termes on peut se ramener (en oubliant les racines carrées) au produit de la confiance des règles $t1 \rightarrow t2$ et $t2 \rightarrow t1$. Or, si l'on souhaite étendre t1, l'utilisation de la mesure de confiance de la règle $t1 \rightarrow t2$ nous semble comporter un problème majeur : Cette mesure est étroitement liée à la fréquence marginale de t2. On peut donc s'interroger sur sa capacité à révéler une relation sémantique entre t1 et t2. Par contre, la règle inverse $t2 \rightarrow t1$ ne comporte pas le même problème. Elle est certes aussi étroitement liée à la fréquence marginale de t1, mais s'il s'agit d'étendre t1, cette fréquence est une constante pour tous les termes mis en concurrence. Il nous semble donc qu'un sens de l'association soit plus à même de rendre compte d'une véritable relation sémantique entre les termes. La tendance des mesures comme le cosinus à extraire des termes à fréquence marginale élevée est un problème connu (Peat & Willet, 1991). Nous proposons de gérer ce problème en évitant la partie de la mesure qui nous semble être responsable de ce phénomène. Nous souhaitons ainsi expérimenter l'idée qu'une orientation doit comme dans le cas des thésaurus construit manuellement être privilégiée.

[2] Le système de recherche associé à MEDLINE est accessible en ligne à l'url : www.ncbi.nlm.nih.gov/sites/entrez

3. Contexte d'utilisation du thésaurus

3.1. *Les systèmes de filtrage adaptatif*

Par système de filtrage, nous désignons ici un système dont la tâche consiste à sélectionner automatiquement, dans un flux entrant de documents, les documents pertinents (pour un besoin d'information donné d'un utilisateur donné). Par système de filtrage adaptatif, nous désignons un système de filtrage qui outre le flux de documents prend aussi en entrée les retours de pertinence sur les documents sélectionnés et qui exploite ces retours de pertinence (ce document est pertinent, ce document n'est pas pertinent) pour s'adapter à la demande des utilisateurs.

La conférence TREC a proposé pendant plusieurs années d'expérimenter différents systèmes sur ce type de tâche (Robertson et Hull, 2001). De ces différentes campagnes d'évaluation, il est possible d'extraire l'architecture utilisée par une majorité de systèmes. La plupart de ces systèmes représentent le besoin d'information par un ensemble de mots-clés pondérés (on parle de profil). Dans la plupart de ces systèmes, le retour de pertinence d'un utilisateur se traduit par l'ajout d'éventuels nouveaux mots-clés au profil et par une mise à jour des poids associés aux différents mots constituant le profil. Schématiquement, un retour positif augmente le poids des mots présents dans le document et diminue le poids des mots absents et un retour négatif a l'action inverse. Dans tous ces systèmes, l'évaluation de la pertinence d'un nouveau document repose essentiellement sur la somme des poids des mots du profil présent dans le document à évaluer. Enfin, ces systèmes comprennent aussi un module gérant le seuil de sélection d'un document (à partir de quel score décide-t-on de sélectionner le document ?). Ce seuil est variable et peut être adapté au cours de l'apprentissage. Il aura par exemple tendance à augmenter si le système génère trop de fausses alarmes (documents sélectionnés non pertinents) ou au contraire à diminuer si le système reste silencieux. Outre l'idée de profil, ces systèmes comprennent donc trois ingrédients principaux :

1. Une règle de mise à jour des poids
2. Une fonction d'évaluation générant un score de pertinence
3. Une méthode de mise à jour du seuil de sélection

La règle de mise à jour des poids la plus connue est Rocchio (Rocchio, 1971). Mais différents systèmes, comme le système Okapi (Robertson et Walker, 2001), proposent leurs propres règles. La fonction d'évaluation la plus utilisée est le cosinus du modèle vectoriel. Mais différentes autres fonctions existent aussi. Enfin, la méthode de mise à jour du seuil est souvent empirique et s'écrit comme un ensemble de règles (Wu *et al.*, 2001) et/ou repose sur un calcul probabiliste lié à la fonction que le système cherche à optimiser (Arampatzis *et al.*, 2001).

3.2. *Sélection des mots à ajouter au profil*

Dans ces systèmes, un autre élément important qui n'est pas toujours mis en évidence dans la description des systèmes concerne la sélection des mots à ajouter au profil. Certes, les poids des mots du profil sont mis à jour à chaque retour de pertinence, mais on peut s'interroger sur le moment et la façon dont ces mots sont ajoutés au profil. Si l'on se fie à la description des systèmes participant à la conférence TREC, on s'aperçoit que, dans la plupart des cas, les mots de la requête formulée initialement sont ajoutés automatiquement et que d'autres sont sélectionnés parmi l'ensemble des mots survenant dans les nouveaux documents pertinents trouvés par le système. On peut s'étonner de la présence de cette étape de sélection puisque l'on se prive de données d'apprentissage par cette sélection. La première raison invoquée pour cette sélection concerne l'accélération des traitements. Néanmoins on peut aussi penser qu'elle limite le bruit que représente la présence fortuite de mots dans des documents pertinents. Car bien que la mise à jour des poids puisse jouer son rôle par la suite, cet apprentissage réalisé sur un nombre limité de documents peut ne pas être suffisant compte tenu du nombre important de termes (Robertson, 1990). Cette sélection consiste généralement à garder les k meilleurs (nouveaux ou pas) termes présents dans le document sur la base d'un calcul qui varie selon les systèmes et qui ne correspond généralement pas au calcul des poids des termes dans le profil. Une autre possibilité permettant de compléter la sélection consisterait à exploiter un thésaurus permettant de sélectionner parmi les mots présents dans les documents pertinents ceux qui sont liés sémantiquement aux mots de la requête. C'est le mode d'utilisation du thésaurus que l'on souhaite proposer ici. Il s'agit d'une forme d'extension contrôlée.

4. Choix d'une mesure d'association

4.1. Trois orientations possibles

Nous proposons d'aborder le problème de l'orientation de la mesure d'association en comparant trois mesures. Chacune de ces mesures sera utilisée par la suite, pour, étant donné un terme T, sélectionner l'ensemble des k termes S les plus fortement liés à T. La première mesure $M1(T,S)$ mesure l'association d'un terme T à un terme S par la fréquence relative de S sachant T. Elle correspond à la confiance de la règle $T \rightarrow S$ (équation 1).

$$M1(T,S) = \frac{|T \cap S|}{|T|} \qquad [1]$$

La deuxième mesure $M2(T,S)$ que nous nous proposons d'étudier, mesure l'association d'un terme T à un terme S par la fréquence relative de T sachant S. Elle correspond à la confiance de la règle $S \rightarrow T$ (équation 2).

$$M2(T,S) = \frac{|T \cap S|}{|S|} \quad [2]$$

La troisième mesure que nous proposons d'intégrer dans nos comparaisons est la mesure $M3(T,S)$ correspondant au cosinus. Cette mesure est symétrique (équation 3). Comme on peut le remarquer (équation 4), elle correspond au produit de la racine carrée des deux premières mesures et donc correspond à la prise en compte simultanée des deux premières mesures et des deux orientations. Ces trois mesures correspondent chacune à un choix d'orientation résumé dans la figure 1.

$$M3(T,S) = \frac{|T \cap S|}{\sqrt{|T|}.\sqrt{|S|}} \quad [3]$$

$$M3(T,S) = \frac{|T \cap S|}{\sqrt{|T|}.\sqrt{|S|}} = \frac{\sqrt{|T \cap S|}}{\sqrt{|T|}}.\frac{\sqrt{|T \cap S|}}{\sqrt{|S|}} \quad [4]$$
$$= \sqrt{M1(T,S)}.\sqrt{M2(T,S)} = \sqrt{M1(T,S).M2(T,S)}$$

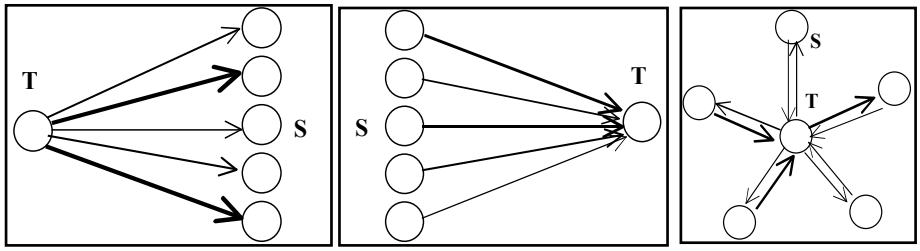

Figure 1. *La mesure d'association M1 (à gauche) permet de sélectionner les termes S tels que la l'association T→S est la plus forte. La mesure d'association M2 (au centre) permet de sélectionner les termes S tels que l'association S→T est la plus forte. La mesure d'association M3 (à droite) permet de sélectionner les termes S tels que l'association S→T ET l'association T→S (on fait le produit) sont les plus fortes.*

4.2 *Expérimentations*

4.2.1. Le corpus

Le corpus utilisé est une légère adaptation du corpus OHSUMED pour la tâche de filtrage de la conférence TREC9 (Robertson et Hull, 2001). Ce corpus contient des résumés d'articles de recherche dans le domaine biomédical qui correspondent aux documents de la base MEDLINE pour les années 1987-1991, soit

350 000 documents. Ce corpus contient soixante-trois requêtes et jugements de pertinence associés. La partie du corpus correspondant à l'année 1987 est réservée à l'apprentissage (en particulier pour l'apprentissage du thésaurus) soit environ 55 000 documents. A noter cependant que dans le cas du filtrage adaptatif, seule la connaissance de la pertinence de 2 documents de 1987 pour les différentes requêtes était donnée.

4.2.2. Construction préliminaire d'un graphe

Dans un premier temps, nous avons construit automatiquement, en utilisant la base d'apprentissage, le graphe dont les nœuds correspondent aux différents termes présents dans les 55 000 documents et dont les liens orientés sont valués par les fréquences relatives correspondantes. Ainsi le lien allant du nœud correspondant au terme t1 au nœud correspondant au terme t2 est valué par la fréquence relative de t2 sachant t1. Ce graphe a ensuite été simplifié. Nous avons supprimé les nœuds et connexions associées présents dans moins de 10 documents parmi les 55 000 considérant que la fréquence t1 sachant t2 n'est crédible qu'à partir d'un certain nombre d'observations de t2 (problème de taille d'échantillon). Pour accélérer les traitements, nous avons aussi supprimé les connexions dont la valeur est strictement inférieure à 0,01. L'impact de cette seconde simplification sur les extensions n'est cependant pas important puisqu'il s'agit de connexions qui ne seront que peu utilisées dans la construction des extensions qui s'appuie sur les meilleures connexions (la valeur de 0.01 peut être considérée comme faible). Nous utilisons ensuite ce graphe pour une extension des termes de la requête.

4.2.3. Une extension contrôlée

Nous souhaitons expérimenter la capacité des 3 mesures d'association (décrites ci-dessus) à sélectionner les termes liés sémantiquement à la requête. Cette extension considère tous les termes de la requête et favorise les termes qui apparaissent liés (selon la mesure d'association considérée) au nombre maximal de termes de la requête dans le même esprit de Qiu et Frei (1993). De façon à éviter le problème du choix d'un seuil commun pour toutes les mesures, on considère qu'un terme est plus ou moins lié à un terme de la requête, en fonction de son classement pour la mesure d'association considérée. La règle retenue attribue une valeur entre 0 et 1, plus ou moins grande selon que le terme fait partie des 5, 30 ou 100 meilleurs termes associés au terme étendu. Un poids est aussi attribué fonction de la fréquence du terme dans le corpus de façon à pénaliser l'extension des termes trop fréquents. Cette extension est résumée dans l'équation 6.

Comme nous l'avons expliqué, nous n'utilisons pas cette extension directement mais comme une aide à la sélection des mots présents dans les documents pertinents ramenés par un système de filtrage adaptatif. Lorsqu'un document pertinent est trouvé par le système, le système sélectionne d'abord tous les termes déjà présents dans le profil (c'est-à-dire présents dans la requête initiale ou ajoutés lors de

précédentes découvertes de documents pertinents) et il ajoute un certain nombre de nouveaux termes.

$$\text{Score}(t) = \sum_{t_i \in Q} \text{Asso}(t_i, t).\text{idf}(t_i) \quad [6]$$

Si t tel que $M(t_i, t)$ dans les 5 premiers Alors $\text{Asso}(t_i, t) = 1$

Si t tel que $M(t_i, t)$ dans les 30 premiers Alors $\text{Asso}(t_i, t) = 0.5$

Si t tel que $M(t_i, t)$ dans les 100 premiers Alors $\text{Asso}(t_i, t) = 0.2$

M est la mesure d'association considérée

$$\text{idf}(t_i) = \text{Log}\left(\frac{N}{\max(\text{df}(t_i), 10)}\right)$$

$\text{df}(t_i)$ est le nombre de documents contenant t_i

N est le nombre de documents dans le corpus d'apprentissage

4.2.4. Le système de filtrage utilisé

Le système de filtrage adaptatif utilisé est le système que nous avons conçu pour les participations à TREC9 et TREC11 décrit dans (Brouard, 2002). Il a l'avantage de considérer une mise à jour des poids extrêmement simple. De plus, il n'intègre, par exemple, pas de poids liés à la fréquence des termes dans le corpus (de type idf). Le poids des termes dépend uniquement de leur présence/absence dans les documents pertinents/non pertinents. Cet aspect est important pour nos tests car nous souhaitons tester exclusivement l'extension. En effet, la prise en compte d'autres informations (particulièrement l'idf) pourrait permettre de compenser les lacunes de certaines mesures d'association. Un terme très fréquent n'ayant pas de relation avec un terme de la requête extrait par la mesure M1 se retrouverait très peu pris en compte et la mesure M1, ne serait que peu pénalisée malgré la mauvaise qualité de l'extension. Nous avons ainsi, pour simplifier l'analyse des résultats, gardé le calcul du poids des termes le plus simple possible. Il correspond en fait au produit de la fréquence relative des documents pertinents sachant le terme avec la fréquence inverse ce qui correspond à la faculté du terme à être dans les documents pertinents et exclusivement ceux-là (compromis rappel/précision). De même, le calcul du score ne fait intervenir aucune statistique sur la fréquence des termes dans le corpus. Il correspond au rapport de la somme des poids des 30 meilleurs termes trouvés dans le document sur les 30 meilleurs termes (présents ou pas). Dans ce système, les termes des documents et des requêtes ont été lemmatisés et les mots vides supprimés. Le critère que nous utilisons pour juger de la qualité de notre extension est le score du système tel qu'il avait été défini pour la conférence TREC9. Ce Score est T9U = 2*R -N où R est le nombre de documents pertinents et N est le nombre de documents non pertinents. Le système optimise le score en sélectionnant les documents tels que la probabilité de pertinence est supérieure à 1/3. Ce score est systématiquement calculé pour les soixante-trois requêtes et le jugement de 275 000 documents.

4.2.5. Expérience préliminaire

Nous avons tout d'abord mené des expériences préliminaires en sélectionnant au hasard, puis les premiers termes du document pertinent. Nous avons fait varier le nombre de termes ajoutés. Ces expériences ont confirmé l'utilité de cette sélection dans notre contexte (figure 2). Lorsque aucun terme n'est ajouté ou lorsque tous les termes sont ajoutés, les scores sont particulièrement bas. Le score maximal est obtenu lorsque environ sept termes sont sélectionnés. Les scores pour une sélection au hasard sont nettement inférieurs aux scores obtenus pour une sélection des premiers termes. Cette dernière observation montre que les premiers termes des documents pertinents sont dans le cas particulier de ce corpus, un bon choix pour l'extension de la requête (ces termes correspondent souvent au titre du document ou au début du résumé).

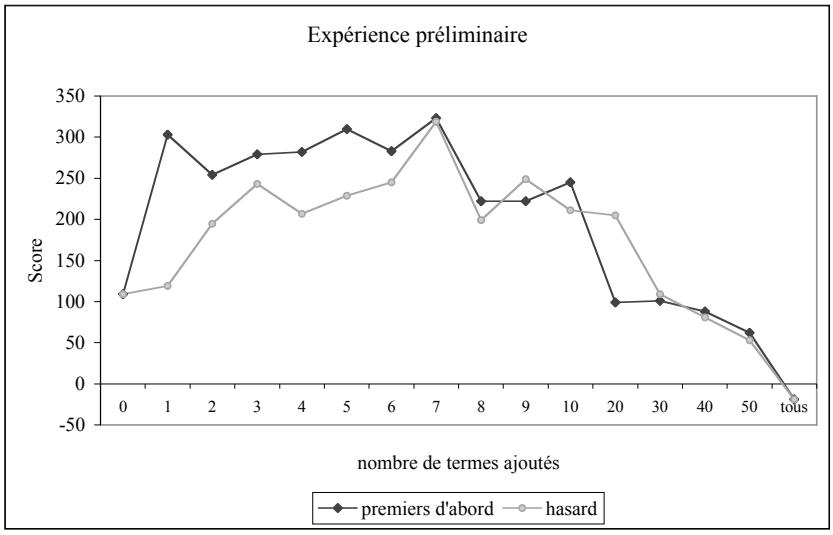

Figure 2. *Résultats de l'expérience préliminaire. Impact du nombre de termes ajoutés sur le score du système de filtrage. Le score correspond au nombre de documents pertinents sélectionnés par le système multiplié par 2 auquel on retranche le nombre de documents non pertinents sélectionnés.*

4.2.6. Comparaison préalable des extensions

Avant d'utiliser les extensions dans le contexte de la tâche de filtrage adaptatif, nous avons procédé à une comparaison des extensions indépendante du système de recherche d'information utilisé et de la tâche effectuée. Cette comparaison a consisté dans un premier temps à vérifier si les termes des différentes extensions avaient bien un lien avec la requête. Il en ressort clairement que les termes correspondant à une extension par M1 (qui correspond à l'orientation T →S, S étant le terme sélectionné) sont des termes à fréquence marginale élevée (environ 40 000 occurrences pour les premiers termes) sans nécessairement un lien évident avec les termes de la requête.

Les termes présents dans l'extension correspondant à la mesure M2 (qui correspond à l'orientation S→T, S étant le terme sélectionné) semblent majoritairement bien liés à la requête (pour les premiers au moins) et ont une fréquence marginale assez faible (environ 200 occurrences pour les premiers termes). Enfin, on observe que les termes correspondant à une extension par M3 (le cosinus) sont un mélange de termes à fréquence marginale élevée (15 000 occurrences en moyenne pour les premiers) et de termes effectivement liés sémantiquement à la requête. On retrouve donc bien les deux critères exprimés par le produit. Un exemple de requête avec les trois extensions associées est donné dans la figure 3.

La requête : DIABETIC GASTROPARESIS TREATMENT

Les 10 premiers termes de l'extension avec M1 :
patient study control effect *insulin diabetes* ***group compare use*** *treat*

Les 10 premiers termes de l'extension avec M2 :
csii insulin-dependent insulin-treated ketoacidosis aldose retinopathy diabetes nondiabetic mellitus non-diabetic

Les 10 premiers termes de l'extension avec M3 :
retinopathy insulin insulin-dependent diabetes mellitus ***patient study*** *treat therapy* ***trial***

Figure 3. *Les dix premiers termes de chaque extension pour une requête particulière. En gras, les termes à fréquence marginale élevée et sans lien particulier avec la requête (le corpus ne contient que des textes dans le domaine biomédical).*

La capacité des termes des extensions à prédire la pertinence des documents pour une requête donnée est un autre résultat indépendant de la tâche et du système utilisé qui constitue une nouvelle indication du lien sémantique de ces termes avec ceux de la requête. En calculant la précision moyenne pour le premier terme des différentes extensions pour les 63 requêtes, on obtient une précision de 4,5% pour M2. Cela signifie qu'en moyenne lorsque le premier terme de l'extension par M2 est présent dans un document, ce document est pertinent dans 4,5% des cas. Dans le cas de M1, la précision est de 0,03% et de 1,9% pour M3. La précision de M2 est globalement supérieure à celle de M3 pour les 10 premiers termes. L'écart se réduit ensuite (figure 4). Pour ce qui est du rappel, l'ordre est inversé. Le rappel moyen pour le premier terme de l'extension par M2 est 12,6% (le premier terme de M2 est présent dans 12,6% des documents pertinents) contre 42,7% pour M3 et 57,5% pour M1. L'ordre des mesures pour le rappel s'explique largement par la fréquence marginale des termes.

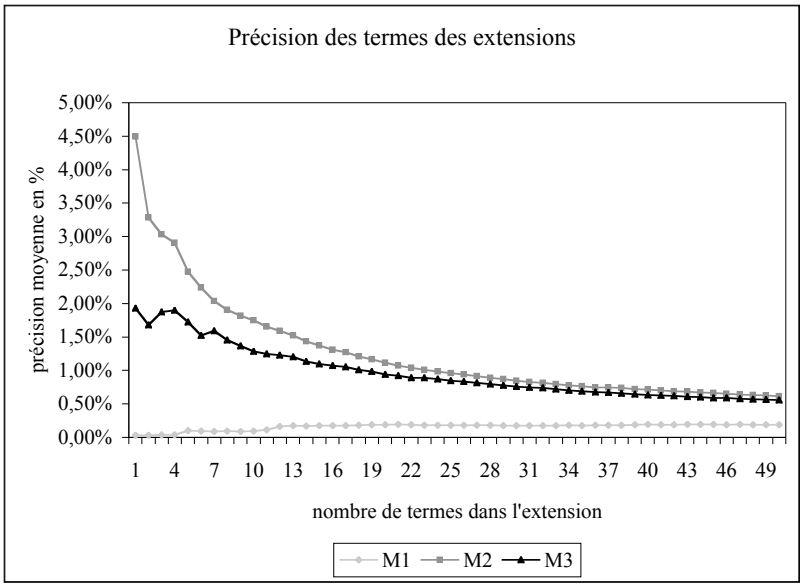

Figure 4. *Précision moyenne des n premiers termes des extensions. La précision avec M2 est meilleure que celle avec M3 pour les premiers termes.*

4.2.7. Comparaison des extensions sur la tâche de filtrage

Nous avons ensuite utilisé les extensions dans le cadre de la tâche de filtrage adaptatif. Comme décrit précédemment, nous avons sélectionné les termes des extensions présents dans les documents pertinents trouvés par le système en testant les différentes extensions et en faisant varier le nombre de termes dans l'extension. Les résultats montrent un impact très clair de la mesure d'association considérée et confirment nos hypothèses. Globalement, la mesure qui donne les meilleurs résultats est M2, suivie de M3, suivie de M1 (figure 5). Cependant, pour une extension de 10 termes, M3 et M2 obtiennent des résultats similaires. Dans ce cas, les termes de l'extension par M2 dont la fréquence marginale est nettement plus faible que ceux de M3 est défavorisée. En effet, le très petit nombre de termes (proche de 0) ajoutés nous ramène au cas de l'expérience préliminaire où aucun terme n'était ajouté. On peut aussi noter une baisse plus brutale des performances pour M2 que pour M3 lorsque l'on considère un plus grand nombre de termes.

Contrairement à ce que l'on pourrait attendre, la différence des résultats du système de filtrage ne se fait pas sur la précision. La précision du système de filtrage avec M2 qui est de 27,5% est en effet très proche de celle obtenue avec M3 qui est de 27,2%. La précision avec M1 est de 24,2%. Par contre, le rappel du système de filtrage est de 16,3% avec M2, 14% avec M3 et 11,4% avec M1. Ces résultats s'expliquent clairement par la gestion des seuils qui contraint le système à ne pas descendre en dessous d'une certaine précision. Comme nous l'avons vu, les termes extraits par M2 sont plus précis. Ils permettent, de ce fait, d'éviter la sélection de

documents non pertinents qui aurait pour effet une augmentation du seuil et une baisse du rappel.

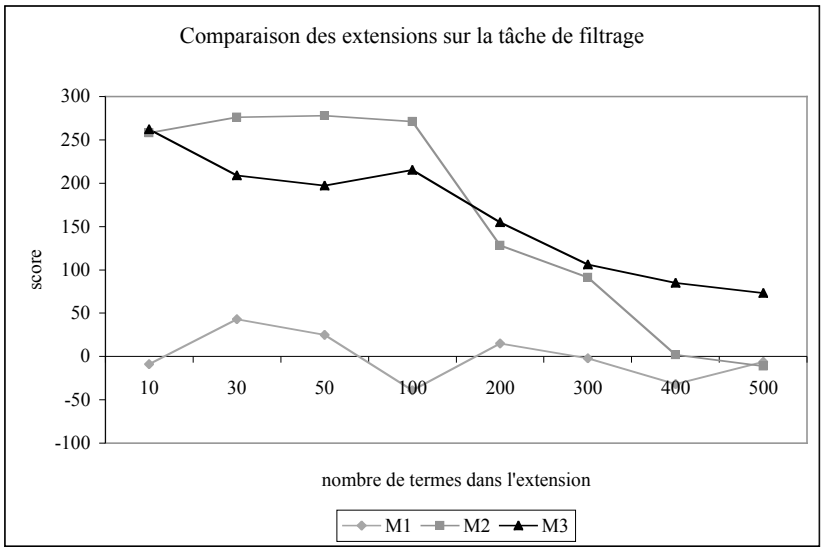

Figure 5. *Résultats de la comparaison des extensions correspondant aux différentes mesures d'associations sur la tâche de filtrage. Le score correspond au nombre de documents pertinents sélectionnés par le système multiplié par 2 auquel on retranche le nombre de documents non pertinents sélectionnés.*

Enfin, si l'on veut rapprocher ces résultats de ceux de l'expérience préliminaire, on peut calculer le nombre de termes ajoutés en moyenne à chaque sélection de document pertinent. Ce nombre moyen est largement lié à la fréquence marginale des termes extraits par les différentes mesures. On sélectionne ainsi plus de termes avec M1 qu'avec M3 et plus de termes avec M3 qu'avec M2. Par exemple, pour une extension de 100 termes avec M2, 1,65 nouveaux termes en moyenne sont ajoutés à chaque sélection de document pertinent contre 3,73 pour M3 et 6,65 pour M1. Pour une extension de 50 termes avec M2, 0,4 nouveaux termes en moyenne sont ajoutés à chaque sélection de document pertinent contre 1,10 pour M3 et 1,93 pour M1. Les résultats pour M2 sont ainsi contrairement à ceux de M1 et M3, pour un nombre de termes ajoutés compris entre 1 et 2, équivalents à ceux obtenus lorsque le premier ou les deux premiers termes du document étaient ajoutés dans l'expérience préliminaire.

5. Conclusion & Perspectives

Notre étude permet donc de mettre en évidence l'importance de l'orientation des mesures d'association considérées. Elle présente une analyse de la mesure du cosinus comme le produit de deux associations orientées dont l'une nous semble vouée à l'échec du fait de sa sensibilité aux fréquences marginales des termes et dont

l'autre nous semble offrir plus de garanties sur sa capacité à extraire des termes véritablement liés sémantiquement. En imposant des valeurs minimales sur la fréquence des termes dans le corpus nous avons pu obtenir de meilleurs résultats que le cosinus en considérant une orientation particulière.

Un des intérêts du corpus que nous utilisons est qu'il s'accompagne du thésaurus MESH[3]. Ce thésaurus est construit manuellement et se présente comme une arborescence dans laquelle les termes deviennent de plus en plus spécifiques à mesure que l'on descend dans l'arborescence. Cette spécialisation correspond dans de nombreux cas à une relation d'hyponymie (sorte de) mais elle peut correspondre aussi à une relation de méronymie (partie de) ou encore à d'autres types de relations. L'étude d'un échantillon de termes extraits par M2 nous a semblé bien correspondre à une relation de spécificité. Par exemple, le terme "lipide" est étendu au terme "triglycéride" par M2. Or les triglycérides sont décrits comme un type de lipide dans MESH. Aussi, nous souhaitons par la suite mettre en évidence le lien privilégié qui existe entre la relation de spécificité et la mesure d'association M2 en comparant l'intersection des spécialisations d'un terme dans MESH (fils dans l'arborescence) avec les extensions pour les différentes mesures d'associations. Cette perspective de recherche se situe dans une problématique un peu différente (bien que liée) de celle abordée dans cette étude. En effet, elle focalise comme l'étude de Cherfi et Toussaint (2002) sur une interprétation sémantique d'une mesure statistique. L'étude que nous avons présentée ici focalise plus particulièrement sur l'extension de requêtes et l'intégration de la mesure d'association dans un processus de recherche d'information.

Par ailleurs, la prise en compte de l'association M2 dans la sélection des termes à ajouter a donné des résultats équivalents au choix des premiers termes du document. Une façon de prendre en compte l'avantage de pouvoir compter sur la présence du thésaurus avant même le traitement du premier document reste à élaborer. L'objectif principal n'était cependant pas ici de proposer une amélioration du fonctionnement des systèmes de filtrage adaptatif mais de les utiliser en vue de comparer des mesures d'association.

Enfin, une partie de cette étude consistait à élaborer un test pour les mesures d'association. Nous avons proposé une extension contrôlée dans le cadre d'un système de filtrage avec bouclage de pertinence et n'intégrant pas dans le poids des termes des connaissances sur leurs distributions dans le corpus.

6. Références

Arampatzis A., Beney J., Koster C.H.A., van der Weide T.P., «Incrementality, Half-life, and Threshold Optimization for Adaptive Document Filtering», *Proceedings of the Text Retrieval Conference (TREC9)*, 2001, p. 589-600.

[3] Le thésaurus MESH est accessible en ligne à l'adresse : www.ncbi.nlm.nih.gov/sites/entrez?db=mesh

Brouard C., RELIEFS : «un système d'inspiration cognitive pour le filtrage adaptatif de documents textuels», *Revue des Sciences et Technologies de l'Information*, vol7, no1/2, 2002, p. 157-182.

Bruandet M.F., Chevallet J.P., «Utilisation et construction de bases de connaissances pour la Recherche d'Informations», in *Assistance Intelligente à la Recherche d'Information* , M.-H. Stefanini, E. Gaussier, Hermes, 2003, chapter 3, p. 85-118.

Cherfi H., Toussaint, Y., «Interprétation des règles d'association extraites par un processus de fouille de textes», *actes du congrès francophone de Reconnaissance des Formes et d'Intelligence Artificielle (RFIA'02)*, 2002, vol. 3, p. 975-983.

Chung Y. M., Lee, J. Y., «A corpus-based approach to comparative evaluation of statistical term association measures», *Journal of the American Society for Information Science and Technology (JASIST)*, 52, 4, 2001, p. 283-296.

Frakes W. B., Yates, R. B., *Information retrieval: data structures and algorithms* (p. 168-173). Prentice-Hall, 1992.

Grefenstette G., *Explorations in automatic thesaurus discovery*. Kluwer Academic Publishers, Boston, 1994.

Haddad H., Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information, Ph.D. thesis, Université Joseph Fourier, 2002.

Kim M., Choi K., «A comparison of collocation-based similarity measures in query expansion». *Information Processing and Management*, vol 35, 1, 1999, p. 19-30.

Miller G.A., «WordNet: A Lexical Database for English». *Commununications of the ACM*, 38(11), 1995, p. 39-41

Miyamoto S., «Information retrieval based on fuzzy association». *Fuzzy Sets and Systems*, vol 38, 1990, p. 191-205.

Peat H.J., Willet P., «The limitations of term co-occurence data for query expansion in documents retrieval systems». *Journal of the American Society for Information Science*, 42(5), 1991, p. 378-383.

Qiu Y., Frei H.P., «Concept-based query expansion». *In proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval*, Pittsburg, US, 1993, p.160-169.

Robertson S.E., «On term selection for query expansion». *Journal of Documentation*, vol 46, 1990, p. 359-364.

Robertson S., Hull. D., «The TREC-9 filtering track final report». *Proceedings of the Text Retrieval Conference (TREC9)*,. NIST Special Publication, 2001, p.25-33.

Robertson S. E., Walker S., «Microsoft Cambridge at TREC-9: Filtering Track», *Proceedings of the Text Retrieval Conference (TREC9)*, 2001, p. 361-368.

Rocchio J.J., «Relevance Feedback in Information Retrieval». In *The SMART Retrieval. System*, G. Salton (Ed), Prentice Hall, Ine, p.313-323, 1971.

Wu L., Luang X., Guo Y., Zhang Y., «FDU at TREC-9: CLIR, Filtering and QA Tasks», *Proceedings of the Text Retrieval Conference (TREC9)*, 2001, p. 189-202.