Extraction possibiliste de concepts MeSH à partir de documents biomédicaux

Wiem Chebil[1,2], Lina F. Soualmia[1,3], Mohamed N. Omri[2], Stéfan J. Darmoni[1,3]

1. Normandie Université, LITIS-TIBS EA 4108, Norma STIC CNRS FRE 3638, Université de Rouen, France

2. Unité de recherche MARS, Université de Monastir, Tunisie

3. INSERM, LIMICS UMR 1142, Paris, France

wiem.chebil@yahoo.fr
Lina.Soualmia@chu-rouen.fr;Stefan.Darmoni@chu-rouen.fr
Mohamednazih.omri@fsm.rnu.tn

RESUME. *Nous proposons dans cet article une nouvelle approche d'indexation de documents biomédicaux basée sur les réseaux possibilistes permettant de les apparier partiellement aux termes du thésaurus MeSH* (Medical Subject Headings). *La principale contribution de notre approche est le traitement de l'imprécision et de l'incertitude liée à la tâche d'indexation à travers l'utilisation de la théorie des possibilités. En effet, nous proposons d'améliorer l'estimation de la pertinence d'un document étant donné un concept en utilisant deux mesures : la possibilité et la nécessité. La possibilité estime le degré de rejet d'un document non pertinent étant donné un concept. La nécessité de la pertinence d'un document évalue dans quelle mesure un document est pertinent pour le concept. De plus, notre approche permet de réduire les limites de l'appariement partiel qui génère de l'information inutile, bien que ce type d'appariement permette de trouver dans le document d'autres variantes du vocabulaire contrôlé. Pour ce faire, nous proposons de filtrer l'index en utilisant les connaissances fournies par l'UMLS* (Unified Medical Langage System). *Filtrer l'index permet de ne garder que les concepts pertinents parmi ceux ayant un sous-ensemble de leurs mots dans le document. Les expérimentations réalisées sur les deux corpus OSHUMED et CISMeF ont montré des résultats encourageants.*

ABSTRACT. *We propose in this paper a new approach for indexing biomedical documents based on the possibilistic network, which carries out a partial matching between documents and the MeSH thesaurus (Medical Subject Headings) terms. The main contribution of our approach is to deal with the imprecision and the uncertainty of the indexing task by using the possibility theory. In fact, we propose to enhance the estimation of a document relevance given a concept by using the two measures of possibility and necessity instead of only one measure used by common approaches. The possibility measure estimates the degree of rejection of an irrelevant document given a concept. The necessity of the relevance of a document estimates what extent a document is relevant for a given concept. Our contribution*

also consists in reducing the limitation of the partial matching that generates irrelevant information although it allows finding in the document other variants of terms than those in the dictionaries. In fact, we propose to filter the index using the knowledge provided by the Unified Medical Language System (UMLS). The filtering allows keeping relevant concepts among those having a subset of their words terms in the document. The experiments carried out at the different steps of our approach and on different corpora showed very encouraging results.

MOTS-CLES : indexation de documents biomédicaux, réseaux possibilistes, appariement partiel, vocabulaires contrôlés.

KEYWORDS: biomedical documents indexing, possibilistic networks, partial matching, controlled vocabularies.

DOI:10.3166/RIA.28.727-750 © 2014 Lavoisier

1. Introduction

Le but d'un système de recherche d'information (SRI) est de trouver les documents les plus pertinents qui correspondent aux requêtes des utilisateurs. Cette tâche est devenue de plus en plus difficile vu le grand nombre de documents sur Internet. Pour améliorer la performance d'un SRI, il est essentiel de développer un système d'indexation capable de générer l'index le plus représentatif du document. Plusieurs approches pour l'indexation des documents ont été proposées. On peut classer ces approches en deux grandes familles :

1) les approches basées sur l'extraction libre des termes (ou mots clés) (Frank et Paynter, 1999 ; Matsuo et Ishizuka, 2003 ; Bracewell *et al.*, 2005 ; Zhang *et al.*, 2006 ; Chengzhi, 2008) ;

2) les approches basées sur l'extraction contrôlée des termes (indexation contrôlée) (Happe *et al.*, 2003 ; Aronson *et al.*, 2004 ; Névéol, 2005 ; Couto *et al.*, 2005 ; Ruch, 2006 ; Zhou *et al.*, 2006 ; De Campos, 2007 ; Sohn *et al.*, 2008 ; Harrathi, 2009 ; Mukherjea *et al*, 2009 ; Hliaoutakis *et al*., 2009 ; Dinh & Tamine, 2011 ; Jonquet *et al.*, 2011).

Dans le cas de l'indexation contrôlée par des termes ou des concepts, le vocabulaire est fixé au préalable. Le vocabulaire contrôlé peut être un ensemble de termes ou un ensemble structuré de termes liés par des relations hiérarchiques ou associatives (synonymie, plus large que, plus étroit que, etc.) comme le thésaurus MeSH (*Medical Subject Headings*) (Nelson *et al.*, 2009). En utilisant une ressource contrôlée, un terme pertinent peut être extrait même s'il n'apparaît pas dans le document. Dans le cas de l'extraction libre des mots-clés, ces derniers sont extraits sans utiliser de connaissances du domaine. Ainsi, l'index n'est pas connu à l'avance et les termes extraits peuvent ne pas être conformes au domaine auquel appartient le document. De plus, seuls les termes présents dans le document peuvent représenter le document. Les termes pertinents qui ne sont pas présents dans le document et qui ont des synonymes dans le document ne peuvent pas être extraits.

Dans le but d'améliorer l'estimation de la pertinence d'un concept étant donné un document, nous proposons une nouvelle approche pour l'indexation des

documents avec un vocabulaire contrôlé basée sur un réseau possibiliste (RP). Bien que la tâche d'indexation soit imprécise et incertaine, à notre connaissance, la théorie des possibilités n'a jamais été appliquée à la tâche d'extraction de concepts issus d'un vocabulaire contrôlé à partir de documents. Les RP sont des méthodes performantes pour le traitement de l'imprécision et l'incertitude et ont été appliquées efficacement dans différents domaines dont notamment la recherche d'information (RI) (Boughanem *et al.*, 2009). Boughanem *et al.*, (2009) estiment qu'une unique mesure n'est pas suffisante pour exprimer la sémantique liée à la notion de pertinence d'un document étant donnée une requête. Leur modèle est basé sur deux mesures proposées par la théorie des possibilités qui sont la possibilité et la nécessité pour estimer la pertinence d'un document étant donnée une requête. Cependant, ce modèle n'exploite pas de système d'organisation de connaissances tel qu'une terminologie, une ontologie ou un thesaurus. Ainsi, l'information sémantique fournie par la ressource contrôlée n'est pas exploitée. Notre approche d'indexation est basée essentiellement sur le modèle de Boughanem *et al.*, (2009) en remplaçant la requête par un terme d'une ressource contrôlée. La pertinence d'un document étant donné un terme est estimée en utilisant les deux mesures de possibilité et de nécessité. La pertinence possible d'un document étant donné un terme estime dans quelle mesure le document est pertinent pour ce terme. La pertinence nécessaire d'un document étant donné un terme permet de trouver la certitude liée à la non pertinence de document. Pour estimer la valeur de ces deux mesures deux autres mesures sont également utilisées. La première est la mesure de possibilité d'un mot appartenant à un terme étant donné un document. Celle-ci élimine tous les mots non représentatifs du document. La seconde est la mesure de nécessité qui met en évidence l'importance des mots pertinents. Le score maximal des termes d'un concept, qui est composé des deux mesures de possibilité et de nécessité, sera attribué à ce concept. Ainsi, le document est décrit à travers une représentation conceptuelle.

De plus, l'extraction des concepts peut être un appariement partiel ou exact. L'appariement exact (AE) (Happe *et al.*, 2003 ; Névéol, 2005 ; Harrathi, 2009 ; Mukherjea *et al.*, 2009 ; Hliaoutakis *et al.*, 2009 ; Jonquet *et al.*, 2011) entre le vocabulaire contrôlé et un document permet de ne trouver dans le document que les concepts qui existent dans le vocabulaire. L'appariement partiel (AP) (ou approximatif) (Aronson *et al.*, 2004 ; Zhou *et al.*, 2006 ; Harrathi, 2009 ; Dinh & Tamine, 2012) permet de trouver dans le document d'autres variantes du vocabulaire contrôlé en appliquant la dé-suffixation ou la lemmatisation. La dé-suffixation réduit les mots (dans le document ainsi que dans le vocabulaire contrôlé) à leurs racines (par exemple : *continuait*, *continuant*, et *continuation* sont réduits à *continu*). La lemmatisation réduit les mots à leur forme de base (par exemple : *vacciner* et *vaccination* sont réduits à *vaccin*). L'appariement partiel permet également d'extraire les termes composés qui partagent un sous-ensemble de leurs mots avec le document. Les termes extraits dans les deux cas (identification de variantes et extraction de termes composés) peuvent être pertinents, ce qui permet d'améliorer le rappel. Mais ils peuvent être également non pertinents entraînant une baisse de la précision. Par exemple, le terme « cancer du sein » dans le document peut entraîner l'extraction des deux termes MeSH « cancer du testicule » et « cancer de l'estomac » car les trois

termes partagent le mot « cancer » (Trieschnigg *et al.*, 2009). Le modèle de réseaux possibilistes que nous proposons est basé sur l'appariement partiel et extrait les termes variantes et termes composés. Dans cet article, nous proposons de réduire l'information inutile générée dans le cas de termes composés. Pour ce faire, nous exploitons les connaissances fournies par l'*Unified Modelling Language System* (UMLS) (Bodenreider, 2004) pour filtrer les concepts extraits. Le filtrage permet ainsi de ne retenir que les concepts pertinents parmi ceux extraits.

La suite de cet article est organisée comme suit. La section 2 décrit un état de l'art sur l'extraction des termes à partir des documents. La section 3 présente les motivations de l'approche proposée, que nous nommons « réseau possibiliste pour l'indexation des documents » (RéPIDo). Dans la section 4 nous rappelons les définitions de la théorie des possibilités sur lesquelles nous nous appuyons. La section 5 détaille l'approche d'indexation proposée en 4 étapes : (1) le prétraitement du corpus et du vocabulaire contrôlé, (2) l'étape d'extraction de concepts qui exploite un réseau possibiliste, (3) l'étape de filtrage et (4) le classement final des concepts extraits. La section 6 est dédiée aux expérimentations menées sur un corpus de documents biomédicaux. Les résultats obtenus sont discutés dans la section 7. La section 8 conclut cette étude et présente quelques pistes pour de futurs travaux.

2. État de l'art

Plusieurs approches pour l'indexation des documents ont été proposées. Nous classons ces approches en approches basées sur l'extraction libre des termes et en approches basées sur l'extraction contrôlée des termes. Étant donné que notre approche utilise une terminologie biomédicale pour l'indexation, nous nous sommes concentrés sur les méthodes basées sur l'indexation contrôlée dans le domaine biomédical.

2.1. Approches basées sur l'extraction libre des termes

Bracewell *et al.*, (2005) se sont basés sur les méthodes du TALN (traitement automatique de la langue naturelle) pour l'extraction des mots clés à partir de documents. La première étape de leur approche est l'analyse morphologique qui consiste à segmenter le document en mots et à identifier leurs catégories grammaticales. Les étapes suivantes sont respectivement l'extraction des phrases nominales, la suppression des mots vides. Ensuite, les phrases nominales ayant des termes nominaux en commun sont regroupées. Enfin, les groupements sont classés selon un score basé sur la fréquence des termes et les phrases nominaux. De plus, l'extraction des mots clés peut être considérée comme un apprentissage supervisé. Les méthodes d'apprentissage utilisent les mots clés extraits à partir du corpus de l'apprentissage pour déduire un modèle de système d'indexation qui est ensuite appliqué pour indexer de nouveaux documents. Cette approche peut exploiter plusieurs méthodes d'apprentissage. On peut citer le réseau bayésien naïf (Frank et Paynter, 1999), les machines à vecteur de support (SVM) (Zhang *et al.*, 2006), le

CRF, *The Conditional Random Fields* (Chengzhi, 2008), *etc*. Les mots clés peuvent être extraits aussi en utilisant des informations statistiques. En effet, Matsuo et Ishizuka (2003) considèrent que deux termes sont cooccurrents s'ils sont dans la même phrase. Ils calculent les fréquences de cooccurrence de couples de termes. Une matrice de cooccurrence est ainsi obtenue. Cette méthode montre une performance comparable à celle du TF-IDF (*Term Frequency Inverse Document Frequency*) (Singhal, 2001).

2.2. Approches basées sur l'indexation contrôlée

Dans (Happe *et al.*, 2003), les auteurs calculent un poids statistique basé sur la mesure TF-IDF pour chaque terme automatiquement extrait du document avec des techniques de TALN. Ces termes sont ensuite filtrés pour ne retenir que les termes du dictionnaire ADM (aide au diagnostic médical). La technique d'indexation de (Aronson *et al.*, 2004) est basée sur trois méthodes. La première apparie le document avec les termes d'UMLS en utilisant l'outil MetaMap. La deuxième méthode compare les phrases de document et les concepts en utilisant la méthode de tri-gram. La troisième méthode se base sur l'apprentissage pour extraire les descripteurs MeSH des *k* documents les plus proches du document à indexer. Couto *et al.*, (2005) calculent une probabilité entre la *Gene Ontology* et un document en utilisant l'évidence de contenu d'un terme, qui est la somme de toutes les évidences du contenu des mots qui le composent. L'évidence de contenu d'un mot correspond à son poids dans l'ontologie.

L'approche d'indexation de Névéol (2005) combine deux méthodes. La première est une méthode linguistique qui utilise un analyseur syntaxique pour extraire les termes simples et les termes composés. La deuxième est basée sur l'apprentissage et exploite les kPP (les k plus proches voisins). Zhou *et al.*, (2006) proposent d'annoter les documents avec les termes ayant les mots les plus significatifs de l'UMLS en calculant un score statistique.

Ruch (2006) décrit une approche d'indexation nommée Eagl qui combine deux modèles : le *Vector Space Model* (VSM) et une méthode basée sur les expressions régulières. De Campos, (2007) propose un modèle d'indexation contrôlée des documents qui se base sur les réseaux bayésiens non supervisés et exploite la structure d'un thésaurus. Harrathi (2009) a proposé une approche statistique indépendante de la langue qui extrait des termes simples et composés en utilisant des informations mutuelles et leur fait correspondre ensuite les concepts de l'UMLS. Mukherjea *et al.*, (2009) ont développé l'outil BioAnnotator pour l'indexation des documents biomédicaux. Ils utilisent un analyseur syntaxique pour identifier des syntagmes nominaux d'un document. Ils attribuent ensuite les concepts UMLS à ces termes en utilisant un moteur de règles. Hliaoutakis *et al.*, (2009) proposent le modèle AMTX (*Automatic MeSH Terms Extraction*). La première étape de ce modèle est l'application de la méthode C/NC-value qui combine l'information statistique et linguistique pour l'extraction de termes composés du texte. La deuxième étape est le classement des termes selon la valeur de C/NC-value. Seuls les termes qui correspondent aux descripteurs MeSH sont retenus. Dans (Jonquet *et*

al., 2011) les auteurs appliquent l'outil Mgrep pour extraire des concepts à partir de 200 ontologies biomédicales et calculent un score pour chaque annotation produite selon son origine (le terme préféré, le terme non préféré ou le terme synonyme). (Dinh et Tamine, 2011) combinent le VSM avec une similarité entre des descripteurs MeSH et le document qui prend en compte l'ordre des mots dans le document.

Nos motivations pour l'approche proposée se basent essentiellement sur les points faibles des approches existantes qui vont être discutés dans la section suivante.

3. Motivations pour l'approche RéPIDO

Les différentes raisons qui nous ont motivés à (1) utiliser un vocabulaire contrôlé**,** (2) exploiter une méthode non supervisée pour l'extraction des termes, particulièrement le réseau possibiliste et (3) proposer une méthode pour améliorer l'appariement partiel, sont les suivantes :

– Les recherches de (Leonard, 1977) et (Markey, 1984) ont montré que la consistance d'indexation augmentait de 5 % quand un vocabulaire contrôlé était utilisé au lieu d'une extraction simple de mots clés. Ce résultat est justifié par les différentes raisons que nous citons dans l'introduction.

– La plupart des approches d'indexation contrôlées décrites dans l'état de l'art se basent sur une méthode non supervisée. Cela est expliqué par le fait que les méthodes supervisées comme le SVM et le réseau bayésien naïf ne sont pas appropriées pour l'indexation contrôlée vu qu'il y a un grand nombre de classes (24 000 classes par exemple quand le thesaurus MeSH est utilisé).

– Le travail de Dinh et Tamine (2011) est basé sur la méthode non supervisée VSM qui a généré des résultats meilleurs que la méthode MTI (Aronson *et al.*, 2004), utilisée pour l'indexation de la grande base de données bibliographiques MEDLINE (0,234 % pour MTI contre 0,273 % pour la méthode basée sur VSM (Dinh et Tamine, 2011) en termes de précision moyenne). De plus, les auteurs de (De Campos *et al*., 2007) ont montré que la performance de l'indexation des documents qui utilise le réseau bayésien est meilleure que le VSM (la précision moyenne à 11 points de VSM est 0,17 versus celle (De Campos *et al.*, 2007) qui est de 0,34). Ce résultat met en évidence l'intérêt de la modélisation graphique. Ainsi nous avons proposé d'utiliser une modélisation par graphe possibiliste.

– À notre connaissance, la théorie de possibilité a été peu utilisée pour les tâches d'indexation contrôlées et son application dans la RI par (Boughanem *et al.*, 2009) était efficace (0,29 % en utilisant le réseau possibiliste dans la RI contre 0,25 % pour BM25 (Robertson et Walker, 1994)).

– Plusieurs approches d'indexation se sont basées sur des méthodes appliquées dans la RI et ont montré de bons résultats comme (Ruch, 2006) et (Dinh et Tamine, 2011). En effet, ces auteurs ont exploité le VSM qui a été initialement utilisé pour trouver les documents pertinents étant donnée une requête (Singhal, 2001). Les

résultats de ces approches peuvent être améliorés en réduisant les limites de l'appariement partiel.

– Les méthodes précédentes ont utilisé une seule mesure pour évaluer la pertinence d'un document étant donné un concept comme la similarité de Cosinus (Ruch, 2006 ; Dinh et Tamine, 2011) ou BM25 (Dinh et Tamine, 2012), ce qui n'est pas suffisant pour décrire la sémantique liée à la pertinence d'un document étant donné un concept (Boughanem *et al.*, 2009). Nous proposons d'exploiter le réseau possibiliste pour traiter cet inconvénient à l'aide des deux mesures que sont la possibilité et la nécessité.

Toutes les approches citées dans l'état de l'art qui utilisent un vocabulaire contrôlé se sont basées sur l'appariement exact ou sur l'appariement partiel. Ces deux cas présentent des limites (décrites dans l'introduction). C'est pourquoi nous proposons une méthode pour réduire ces limites dans l'étape de filtrage.

Avant de détailler l'approche proposée d'indexation RéPIDo, nous rappelons brièvement les concepts de la théorie des possibilités sur lesquels se base notre approche.

4. Théorie des possibilités et graphes possibilistes

La théorie des possibilités a été proposée par Zadeh (1981) comme une extension de la théorie de la logique floue qui a été développée ensuite par Dubois et Prade, (1988). Cette théorie permet de traiter des informations incomplètes et incertaines dans l'intervalle [0,1]. Elle diffère de la théorie des probabilités.

4.1. Distribution de possibilité et les mesures de possibilité et de nécessité

Une distribution de possibilité Π est une fonction de l'univers de discours X vers [0,1]. La fonction Π(x) évalue la possibilité que x soit la valeur actuelle d'une certaine variable à laquelle ∏ est attachée. Si, Π(x) = 1alors x est totalement possible. Si Π(x) = 0 alors x est rejeté et considéré comme impossible. Si un évènement est impossible cela n'implique pas seulement que l'évènement contraire est possible mais aussi qu'il est certain. La condition de normalisation est : $\max_{x \in X}(\Pi(x))=1$.

La possibilité d'un événement *A*, appelée Π(*A*), évalue et reflète la situation dans laquelle A est vrai et pertinent. La valeur de Π est obtenue via la formule $\Pi(A) = \max_{x \in A}(\Pi(x))$. La nécessité d'un événement *A*, notée par *N*(*A*), évalue et reflète la situation dans laquelle *A* est faux et elle est définie par la formule suivante

$$N(A) = \min_{x \notin A}(1 - \pi(x)) = 1 - \Pi(\neg A)$$

4.2. Graphes possibilistes

Un graphe possibiliste est caractérisé par un composant qualitatif et un composant quantitatif. Le composant qualitatif est un graphe acyclique orienté constitué d'un ensemble de variables $V = \{A_1, A_1,.., A_1\}$ qui correspondent aux nœuds et un ensemble de relations qui relient les nœuds. Le composant quantitatif est la distribution possibiliste conditionnelle qui quantifie les liens entre un nœud et ses parents. Ces distributions de possibilité devraient respecter la normalisation.

Pour chaque variable A_i on considère les conditions suivantes : si A_i est un nœud racine, alors $Parents(A_i)=\phi$, et le domaine de A_i est dom_{Ai} et $max_\alpha\ \pi(a_i) = 1$; si A_i possède des parents alors $Parents(A_i)\neq\phi$, et$max_{ai}\ \pi(a_i|\theta_{ai}) = 1$où dom_{Ai} est le domaine de A_i et θ_{A_i} représente les configurations possibles des parents de A_i.

5. RéPIDo : approche proposée pour l'indexation contrôlée des documents

Notre approche est composée de 4 étapes (figure 1) :

1) Le prétraitement des documents et des termes MeSH.

2) L'extraction de concepts du document prétraité en utilisant un réseau possibiliste (RP).

3) Le filtrage des concepts extraits dans l'étape précédente.

4) Le classement final.

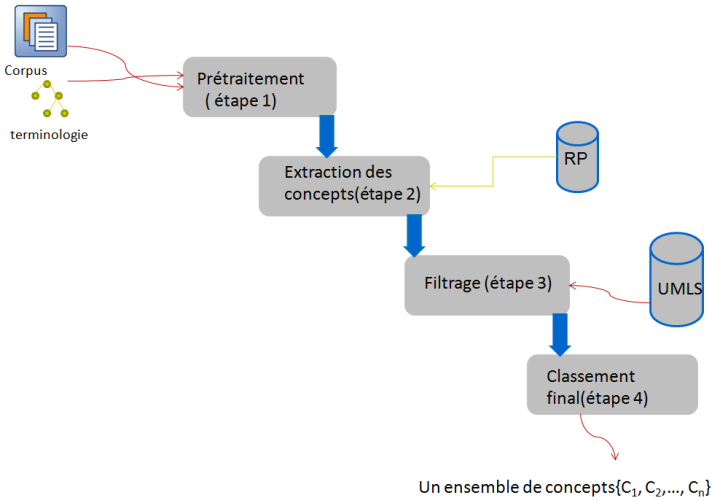

Figure 1. Processus général de notre approche d'indexation RéPIDo permettant l'extraction de concepts C_i en utilisant les réseaux possibilistes et l'UMLS

5.1. Étape 1. Prétraitement des documents et des termes MeSH

L'étape de prétraitement consiste en 5 tâches : (1) découper le document en phrases, (2) supprimer la ponctuation, (3) supprimer les mots vides, (4) dé-suffixer les mots et (5) diviser chaque phrase en mots. Les quatre tâches (2), (3), (4) et (5) sont appliquées également sur chaque terme du thésaurus MeSH. Par exemple, soit « La natation pour le traitement de l'asthme chez les enfants et les adolescents âgés de 18 ans et moins » le titre d'un document. Après le prétraitement ce titre devient « natation trait asthm enfant adolescent age 18 an ». Nous avons utilisé l'algorithme de Porter (1978) pour la dé-suffixation car cet algorithme peut être utilisé dans des langues différentes (notamment l'anglais et le français) nous permettant de tester notre approche sur des corpus en anglais et en français.

5.2. Étape 2. Extraction des concepts

L'étape d'extraction de concepts se fait d'abord par l'extraction des termes les composant. Pour extraire les termes nous utilisons un réseau possibiliste qui permet de calculer le score (formule 1) de chaque terme. Les termes candidats sont ceux ayant un score non nul. Ensuite, les concepts correspondants sont assignés aux termes. Le score d'un concept correspond au score de son terme. Si un concept correspond à plus d'un terme parmi les termes candidats, il aura le score maximal. Le terme ayant le même score que son concept est dénoté terme représentatif (TR).

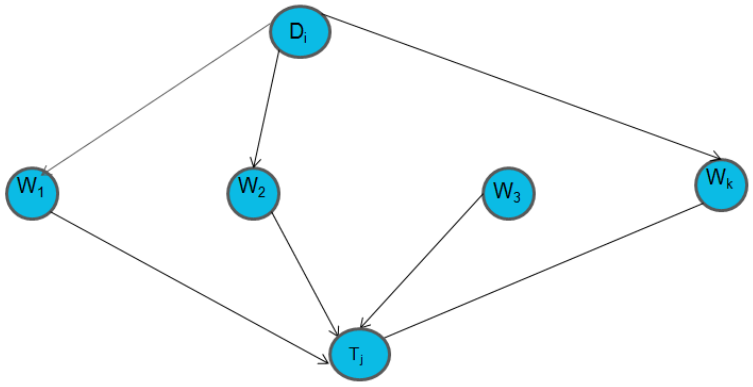

Figure 2. Le graphe représentant un réseau possibiliste pour l'extraction de termes à partir d'un document, où T_j représente un terme MeSH, W_k représente un mot appartenant au document à indexer ou à un terme et D_i est le document à indexer

5.2.1. Modèle possibiliste pour l'extraction de termes

Les termes sont extraits en utilisant un réseau possibiliste. L'architecture du modèle d'indexation possibiliste est basée essentiellement sur le modèle de

recherche d'information possibiliste (Boughanem *et al.*, 2009) en remplaçant la requête Q par le terme T_j. Les composants graphiques (figure 2) représentent les nœuds suivants : (i) le document à indexer D_i (document prétraité) (ii) un mot du document et du terme W_k (iv) le terme T_j et (v) les relations de dépendance qui existent entre les nœuds. Le domaine d'un document D_i est $Dom(D_i) = \{d_i, \neg d_i\}$. Si le document à indexer est pertinent pour un terme T_j alors $D = d$. Si le document n'est pas pertinent alors $D_i = \neg d_i$. W_k fait référence à un mot dans le document ou dans un terme. Le domaine de mots est $Dom(W_k) = \{w_k, \neg w_k\}$. Si le mot appartient au document ou au terme alors $W_k = w_k$. Si le mot est absent dans le document ou dans le terme alors $W_k = \neg w_k$. Un terme T_i prend ses valeurs dans le domaine $Dom(T_i) = \{t_i, \neg t_i\}$. Nous ne nous intéressons qu'au cas où le terme est instancié. Nous ne considérons alors que le cas $T_j = t_j$. Ainsi nous notons le terme par T_j.

5.2.1.1. Évaluation d'un terme T_j

L'évaluation est effectuée par la propagation des informations données par le terme dans le réseau quand il est instancié. Les arcs sont activés par l'instanciation du terme vers le document. Pour chaque nœud la possibilité *a posteriori* conditionnelle et marginale est calculée étant donnée la possibilité conditionnelle et marginale *a priori*. En utilisant cette possibilité conditionnelle, deux mesures sont ainsi calculées. La première est la possibilité du document étant donné un terme (formule 2). La seconde est la nécessité du document étant donné un terme (formule 3). Les formules 2 et 3 sont calculées entre le document à indexer D_i et chaque terme MeSH T_j.

$$S = \Pi(d_i | T_j) + N(d_i | T_j) \tag{1}$$

$$\Pi(d_i | T_j) = \frac{\Pi(T_j \wedge d_i)}{\Pi(T_j)} \tag{2}$$

$$\Pi(\bar{d}_i | T_j) = \frac{\Pi(T_j \wedge \bar{d}_i)}{\Pi(T_j)} \tag{3}$$

$$N(d_i | T_j) = 1 - \Pi(\bar{d}_i | T_j) \tag{4}$$

avec :

$$\Pi(T_j) = max(\Pi(T_j \wedge \bar{d}_i), \Pi(T_j \wedge d_i)) \tag{5}$$

Selon (Benferhat *et al.*, 1999) et (Dubois et Prade, 1988) nous avons :

$$\Pi(d_i | T_j) = min(1, \frac{\Pi(T_j \wedge d_i)}{\Pi(T_j \wedge \bar{d}_i)}) \tag{6}$$

Dans la suite, l'ensemble des mots du document *D* est désigné par *W(D)* et l'ensemble des mots du terme *T* est désigné par *W(T)*. Selon la topologie du graphe de la figure 2, nous avons :

$$\prod(T_j \wedge D_i) = \max_{\forall \theta^l \in \theta^L} \left(\prod \left(T_j \middle| \theta^l \right) \times \operatorname*{prod}_{W_k \in W(T) \wedge W(D)} \left(\prod \left(\theta_k^l \middle| D_i \right) \right) \times a \right) \quad (7)$$

où θ^L est l'ensemble des configurations possibles des parents de T_j. θ^l est une configuration possible de θ^L, θ_k^l est une instanciation de w_k dans la configuration de θ^l. L'instanciation θ_1^1 du mot w_k dans la première configuration $\theta^1=\{w_1,w_2\}$ est $\theta_1^l=w_1$. Les configurations possibles de $T_1=\{W_1,W_2\}$ sont $\theta^l=\{w_1, w_2\}$, $\theta^l=\{w_1, \neg w_2\}$, $\theta^l=\{\neg w_1, w_2\}$ et $\theta^l=\{\neg w_1, \neg w_2\}$.

De plus, *a* est un coefficient dont les valeurs sont comprises dans l'intervalle [0,1] : $a<1$ si les mots d'un terme ne sont pas dans la même phrase. Dans le cas où les mots d'un terme ne sont pas dans la même phrase, le coefficient *a* est fixé expérimentalement dans la section 5.

D'après Boughanem et al., (2009), nous considérons que $\prod(d_i) = \prod(\neg d_i) = 1$

Nous considérons aussi que « prod » signifie le produit car le symbole de produit est utilisé pour désigner la possibilité.

5.2.1.2. Agrégation des mots des termes

Les cinq formes canoniques proposées dans (Turtle, 1991) peuvent être adoptées également pour l'agrégation de mots de termes. Ainsi, les mots d'un terme peuvent être connectés par les opérateurs booléens (OU, ET et NON) et la somme probabiliste ou l'une de ses variantes, qui est la somme pondérée. Nous présentons la forme disjonctive (les mots d'un terme sont connectés par l'opérateur booléen OU) qui va être utilisée dans notre modèle. En particulier pour la disjonction, si nous considérons un terme *T* comme une requête booléenne disjonctive, les termes ayant au moins un mot dans le document sont considérés comme des candidats pour indexer le document D_i. Nous notons ce type d'appariement entre le terme et le document comme étant un appariement partiel. Ainsi, pour un terme T_j composé de *p* mots, $T_j = \{w_1 \vee w_2 \vee ... \vee w_p\}$. Dans notre approche nous nous sommes basés sur un appariement partiel entre des documents et des termes ainsi nous considérons la disjonction pour calculer la formule (7).

5.2.1.3. La distribution de possibilité $\prod(W_k|D_i)$

Pour définir la représentativité d'un mot dans un document nous considérons deux cas :

1) Plus un mot est fréquent dans le document plus il est probable qu'il soit un représentant du document (Formule 8), ainsi :

$$\prod(W_k|D_i)=FW_{ki} \tag{8}$$

$$FW_{ki}=\frac{freq_{ki}}{\max_{wr\in d_i}(freq_{ri})} \tag{9}$$

$Freq_{ki}$ est la fréquence d'un mot w_k dans un document d_i. Si $FW_k=0$ cela signifie que le mot ne représente pas le document. Si $FW_k=1$ cela signifie que le mot est pertinent pour le document (la mesure de possibilité est normalisée et sa valeur maximale est 1).

2) Plus un mot est fréquent dans le document et moins il est fréquent dans les autres documents de la collection plus il est nécessairement le représentatif du document (formule 10). Ainsi :

$$\varphi_{ki}=\frac{\log\frac{N}{n_k}}{\log(N)}\times FW_{ki} \tag{10}$$

où N est le nombre de documents dans la collection et n_k est le nombre de documents auxquels appartient le mot w_k.

La nécessité d'un mot (formule 11) montre la nécessité qu'un mot du terme puisse contribuer pour renvoyer un document.

$$N(w_k\rightarrow d_i)=\varphi_{ki} \tag{11}$$

avec $\prod(\neg d_i)=1$, ainsi $\prod(w_k|d_i)=\prod(w_k\wedge\neg d_i)=1-N(w_k\rightarrow d_i)=1-\varphi_{ki}$.

Le tableau 1 résume la possibilité conditionnelle d'un mot w_k étant donné un document D_i : $\prod(W_k|D_i)$.

Tableau1. La possibilité conditionnelle d'un mot W_k étant donné un document D_i : $\prod(W_k|D_i)$

	d_i	$\neg d_i$
w_k	FW_{ki}	$1-\varphi_{ki}$
$\neg w_k$	1	1

5.2.2. Le score d'un concept

Le score d'un concept C_f est le score maximal de ses termes (formule 12). Le terme d'un concept ayant le score maximal est noté un terme représentatif (TR).

$$Score(C_f) = \max_{T_j \in T(C_f)}(S(T_j)) \quad (12)$$

où $T(C_f)$ est l'ensemble des termes d'un concept C_f. On peut considérer également que la possibilité et la nécessité du terme représentatif du T_j sont aussi la possibilité et la nécessité de C_f.

5.3. Étape 3. Filtrage

Le but de cette étape de filtrage est de ne garder que les concepts pertinents parmi ceux ayant un sous-ensemble des mots de leur terme représentatif qui n'existe pas dans le document. En effet, nous avons classé la non-extraction de ces concepts pertinents comme une catégorie d'erreurs d'indexation dans une étude antérieure (Chebil *et al.*, 2012). Cette étape consiste à diviser l'ensemble de concepts extraits dans l'étape précédente en deux ensembles de concepts. Le premier ensemble est noté index principal (IP) et le deuxième est noté l'index secondaire (IS). L'index principal IP contient les concepts ayant tous les mots de leur terme représentatif présents dans le document. Ces concepts sont notés concepts principaux (CP). L'index secondaire contient les concepts ayant au moins un mot de leur terme représentatif non présent dans le document. Ces concepts sont notés concepts secondaires (CS). Nous avons distingué les concepts principaux des concepts secondaires car nous considérons que les termes qui ont tous leurs mots dans le document seront plus probablement corrects. Ensuite les concepts pertinents de l'index secondaire sont ajoutés à l'index principal. Pour ce faire, l'ensemble des concepts principaux dans l'index principal est trié (du concept le plus pertinent au moins pertinent) en utilisant le score calculé à l'équation (12). Ainsi, nous avons $IP=\{CP_1,...,CP_i...,CP_v\}$, avec CP_i un concept principal ayant le rang i et v le nombre de concepts principaux dans l'index principal. Pour chaque concept secondaire un score est calculé en utilisant l'équation (12). Ce score est basé sur les cooccurrences de concepts MeSH dans MEDLINE et les relations sémantiques entre ces concepts fournies par le réseau sémantique de l'UMLS. En effet, nous nous basons sur l'hypothèse suivante : un concept secondaire est plus probablement correct s'il est cooccurrent et s'il possède des relations sémantiques avec les L premiers concepts principaux de l'IP (les L premiers considérés les plus pertinents). Par exemple, selon la formule proposée pour SR (équation 13), si nous fixons L à 1, cela signifie que SR est égal à la somme de la fréquence de cooccurrence et du nombre de relations sémantiques entre le concept secondaire et le concept principal de rang 1 (CP_1). Si $L = 2$, cela signifie que SR est égal à la somme de la fréquence de cooccurrences et du nombre de relations sémantiques entre le concept secondaire et les deux concepts principaux ayant les rangs 1 et 2 (CP_1 et CP_2). Si le concept secondaire n'est pas co-occurrent ou n'a pas de relation sémantique avec l'un des L concepts principaux, il ne sera pas ajouté à l'index principal. Cela est différent de la méthode de filtrage proposée dans une étude antérieure (Chebil *et al.*, 2013) qui permet de ne garder que les concepts secondaires cooccurrents ou qui possèdent des relations sémantiques avec les concepts principaux et pas les deux à la fois.

$$SR(CS) = \sum_{i=1}^{L} CF(CS, CP_i) + \sum_{i=1}^{L} NR(CS, CP_i) \quad (13)$$

où *CF* est la fréquence de cooccurrence et *NR* est le nombre de relations sémantiques.

5.4. Étape 4. Classement final

Le concept secondaire choisi dans l'étape précédente est ajouté à l'index principal permettant de construire ainsi l'index final. Les concepts de l'index final sont ensuite reclassés en utilisant le score (équation13).

6. Évaluations expérimentales et résultats

Nous présentons dans cette section les objectifs des expériences, le corpus utilisé, les mesures d'évaluation. Nous détaillons ensuite les expérimentations réalisées afin d'évaluer les performances de RéPIDo.

6.1. Objectifs des expériences

L'objectif principal des expérimentations est double : régler les paramètres de notre approche et montrer la valeur ajoutée de chacune de nos contributions. Tout d'abord, nous cherchons à déterminer la meilleure valeur du coefficient *a* qui permet de donner moins d'importance aux termes dont les mots ne figurent pas dans la même phrase. Ensuite, la valeur de *L* est réglée dans l'étape de filtrage. Enfin, pour mettre en évidence l'intérêt de l'utilisation d'un réseau possibiliste dans le processus d'indexation, nous avons comparé notre approche à d'autres systèmes d'indexation.

6.2. Corpus et terminologies utilisées pour l'évaluation

Pour tester notre approche nous avons utilisé un sous-ensemble de la collection OHSUMED 88[1] composé de 49 538 citations. Chaque citation choisie est composée d'un titre et d'un résumé en anglais. Le contenu du titre est fusionné avec le contenu du résumé. Une citation est composée de six champs : titre (.T), résumé (.W), concepts indexés (.M), auteur (.A), source (.S), et publication (.P). Nous avons également utilisé un autre corpus composé de titres et résumés de 1 000 ressources de CISMeF (catalogue et index des sites médicaux en français) (Douyère *et al.*, 2004) sélectionnées au hasard. Trois types de documents sont indexés dans CISMeF : les documents pour les patients, les recommandations et les documents destinés à l'enseignement. Des statistiques de la collection sont illustrées dans le

1. http://trec.nist.gov/data/t9_filtering.html

tableau 2. Pour obtenir de meilleurs résultats pour RéPIDo en appliquant l'étape de filtrage sur le corpus CISMeF, nous avons construit une table de cooccurrences entre concepts à partir du corpus CISMeF. Nous avons utilisé le thesaurus MeSH pour l'indexation des corpus d'évaluation. Dans toutes les expériences nous n'avons pris en considération que les 15 premiers concepts dans le dernier index car le nombre moyen de concepts dans les index manuels dans OSHUMED est 15 (Ruch 2006).

Tableau 2. Statistiques des corpus de test

	OHSUMED	CISMeF
Nombre total de documents	49 538	1000
Nombre moyen de mots dans les titres	10,4	9,2
Nombre moyen de mots dans les résumés	124,7	101,1
Nombre de documents pour les patients	-	300
Nombre de documents destinés à l'enseignement	-	350
Nombre de documents de type recommandation	-	350

6.3. Mesures d'évaluation

Pour évaluer l'approche d'indexation proposée, nous utilisons la précision moyenne (la MAP) et le F-score qui combine la précision et le rappel avec un poids égal (Manning et Schütze, 1999). Nous calculons également la précision aux rangs 5, 10 et 15 ainsi que le taux d'amélioration (TA) de la précision moyenne (ΔMAP) par rapport à une base de référence qui va être précisée à chaque description d'une évaluation (formule 14) :

$$\Delta MAP = \frac{MAP_{méthode} - MAP_{baseline}}{MAP_{baseline}} \times 100 \quad (14)$$

Il faut noter que nous n'avons pris en considération que les 15 premiers concepts générés par RéPIDo car le nombre moyen des concepts de l'indexation manuelle des documents OSHUMED est 15 (Ruch, 2006). De plus, nous avons calculé les tests de Student (t-tests) par séries appariées entre les rangs entre les rangs (P@10, P@20 et MAP) obtenus par chaque méthode testées et la base de référence.

6.4. Description des expérimentations

Le « gold standard » est ici l'indexation manuelle de l'ensemble du document. Pour les deux corpus de test, l'appariement entre les index générés automatiquement et le « gold standard » est exact. Par exemple, si le concept *Viridans, les Streptocoques* existe dans l'index manuel et que le concept *Viridans* existe dans l'index automatique et n'existe pas dans le manuel, alors *Viridans* n'est pas considéré pour l'indexation. De plus, deux règles d'indexation s'appliquent par les

experts de CISMeF : (1) si un concept est l'ancêtre d'un autre concept (par exemple *Endocarditis* et *Endocarditis, bactérien*) et que ces derniers sont extraits du même document, le premier concept n'est pas considéré pour l'indexation, (2) si un concept est un sous-ensemble d'un autre concept et que ces derniers sont extraits du même document alors uniquement le plus long est considéré pour indexer. Les deux règles (1) et (2) ne sont pas appliquées pour l'indexation manuelle d'OSHUMED. En effet, si un concept est l'ancêtre ou un sous-ensemble d'un autre concept, les deux concepts sont considérés pour l'indexation. Pour être sûr que l'indexation automatique est exécutée de la même manière que l'indexation manuelle, l'algorithme de RéPIDo suit les mêmes règles d'indexation de chaque corpus.

6.5. Exemple d'indexation avec RéPIDo

Le document (noté d_1) ayant l'identifiant PMID = 3655403 (*http://www.ncbi.nlm.nih.gov/pubmed/3655403*) dans MEDLINE est un document extrait à partir de la collection OSHUMED. Ce document est indexé automatiquement par notre approche (tableau 3) en fixant $L = 3$ et manuellement par des experts.

Tableau 3. L'index final du document ayant le PMID = 3655403 généré par RéPIDo

Concept C_i	$((\prod(d_1 \mid c_f); N(d_1 \mid c_f))$
Penicillin	(1 ;0,55)
Dextranase	(1 ; 0,43)
Glycocalyx	(1 ; 0,37)
Bacterial, Endocarditis	(0,31 ; 0)
Endocarditis	(0,31 ; 0)
Animals	(0,29 ; 0)
Bisphenol a-GlycidylMethacrylate	(0,27;0)
MyxococcusXanthusAntibiotic	(0,27; 0)
Viridans, Streptococci	(0,2 ; 0)
Et protocol	(0,14 ; 0)
Therapy	(0,14 ; 0)
Role	(0,14 ; 0)
Alpha	(0,14 ; 0)
Organization and Administration	(0,12 ; 0)
HeartFailure	(0,07 ; 0)

***Index Manuel* :** Animal ; Dextranase ; Encaditis, Bacterial ; Glycoproteins ; Microscopy ; Electron ; Scanning ; penicillin G ; Procaine ; Polysaccharides ;

Polysaccharides, Bacterial ; Rabbits ; Streptococcal Infection ; Streptococcus ; Streptococcus Sanguis.

Le concept secondaire ajouté de l'index secondaire vers l'index principal est ***Endocarditis, Bacterial***. En effet, il existe trois relations issues du réseau sémantique de l'UMLS entre *Penicillins* et *Endocarditis, Bacterial*. Ces relations sont *affects*, *treats* et *diagnoses*. De plus, *Endocarditis, Bacterial* et *Dextranase* sont associés par la relation *produces*. *Glycocalyx* et *Endocarditis, Bacterial* sont associés également par la relation *location_of*. Il existe une relation entre *Endocarditis, Bacterial* et *Viridans, Streptococci* qui est *causes*. De plus, *CF(Penicillins; Endocarditis, Bacterial)* = 100, *CF(Endocarditis, Bacterial; Viridians,Streptococci)* = 43 et *CF(Endocarditis, Bacterial ; dextranase)* = 10. Ainsi, *SR(Endocarditis, Bacterial)* = 148.

6.6. Résultats

Avant de présenter les résultats obtenus pour l'étape de filtrage et la comparaison avec d'autres approches, nous précisons la façon dont nous avons réglé le paramètre *a* lorsque les mots d'un terme ne figurent pas dans la même phrase.

Nous avons expérimenté avec différentes valeurs de *a* : 1 ; 0,9 ; 0,8 ; 0,7 ; 0,6 ; 0,5 ; 0,4 ; 0,3 ; 0,2 et 0,1. Pour chacune de ces valeurs nous avons calculé la MAP et le F-score en utilisant les deux corpus OSHUMED et CISMeF. Il faut signaler que pour réaliser cette expérimentation nous avons fixé approximativement la valeur de *L* à 3. La valeur de *L* est réglée expérimentalement dans la prochaine étape des expérimentations.

Tableau 4. Réglage expérimental du coefficient a

	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
	MAP Fs	MAP Fs	MAP Fs	MAP Fs	MAP Fs	MAP Fs	MAP Fs	MAP Fs	MAP Fs	MAP Fs
OSHUMED	0,424 0,433	0,426 0,436	0,435 0,441	0,443 0,446	**0,447** **0,448**	0,443 0,440	0,438 0,429	0,430 0,417	0,427 0,408	0,421 ,400
CISMeF	0,400 0,419	0,401 0,422	0,410 0,427	0,429 0,425	**0,438** **0,440**	0,430 0,431	0,420 0,413	0,425 0,402	0,418 0,394	0,415 0.391

6.6.1. Évaluation de l'étape de filtrage

Pour évaluer l'étape de filtrage, nous avons testé notre approche pour différentes valeurs de *L* en fixant la valeur de *a* à 0.6 (la valeur qui a généré les meilleurs résultats dans l'expérimentation précédente). Pour chacune des valeurs de *L* nous avons calculé la MAP, le Fscore et la précision aux rangs 5, 10 et 15 toujours en utilisant les deux corpus OSHUMED et CISMeF.

Tableau 5. Évaluation de l'étape de filtrage comparée à la base de référence

		OSHUMED	CISMeF
RéPIDo sans filtrage (base de référence)	MAP (Δ%)	0,358	0,331
	Fscore (Δ%)	0,364	0,344
	p@5 (Δ%)	0,512	0,511
	p@10(Δ%)	0,348	0,345
	p@15(Δ%)	0,273	0,270
RéPIDo à L = 1	MAP (Δ%)	0,362 (+1,11%)	0,335 (+1,2%)
	Fscore (Δ%)	0,369(+1,37%)	0,347 (+0,87%)
	p@5 (Δ%)	0,524(+2,34%)	0,523(+2,34%)
	p@10(Δ%)	0,351(+0,86%)	0,480(+2,12%)
	p@15(Δ%)	0,279(+2,19%)	0,275(+1,85%)
RéPIDo à L = 2	MAP (Δ%)	0,373 (+4,18%)	0,352 (+6,34%)
	Fscore (Δ%)	0,381 (+4,67%)	0,360 (+4,65%)
	p@5 (Δ%)	0,545(+4,00%)	0,636(+24,46%)
	p@10(Δ%)	0,355(+0,20%)	0,354(+2,60%)
	p@15(Δ%)	0,282(+3,29%)	0,281(+4,07%)
RéPIDo à L = 3	MAP (Δ%)	0,447(+24,86%)*	0,438 (+25,86%)*
	Fscore (Δ%)	0,442(+20,87%)	0,441(+28,19%)
	p@5 (Δ%)	0,564 (+19,23%)	0,553 (+28,00%)
	p@10(Δ%)	0,364(+15,92%)*	0,360 (+19,81%)*
	p@15(Δ%)	0,301(+14,53%)*	0,290 (+11,87%)*
RéPIDo à L = 4	MAP (Δ%)	0,425(+18,71%)	0,405 (+22,35%)
	Fscore (Δ%)	0,439 (+20,60%)	0,423 (+22,96%)
	p@5 (Δ%)	0,552(+8,02%)	0,551(+7,82%)
	p@10(Δ%)	0,358(+2.87%)	0,357(+3,47%)
	p@15(Δ%)	0,284(+1,79%)	0,282(+4,44%)
RéPIDo à L = 5	MAP (Δ%)	0,412 (+15,08%)	0,394 (+19,03%)
	Fscore (Δ%)	0,424 (+16,48%)	0,412 (+19,76%)
	p@5 (Δ%)	0,541(+5,66%)	0,541(+5,87%)
	p@10(Δ%)	0,352(+1,14%)	0,351(+1,73%)
	p@15(Δ%)	0,275(+0,07%)	0,274(+1,48%)
RéPIDo à L = 6	MAP (Δ%)	0,403 (+12,56%)	0,381 (+15,10%)
	Fscore (Δ%)	0,412 (+13,18%)	0,393 (+14,24%)
	p@5 (Δ%)	0,521(+1,75%)	0,520(+1,76%)
	p@10(Δ%)	0,349(+0,02%)	0,346(+0,28%)
	p@15(Δ%)	0,274(+0,03%)	0,273(+1,11%)

***:** *un changement significatif à $p < 0.05$*

Tableau 6. Comparaison de la performance de RéPIDo avec d'autres approches

		Corpus OSHUMED	Corpus CISMeF
MaxMatcher	MAP (Δ%)	0,363(+27,81%)	0,338(+18,59%)
	Fscore (Δ%)	0,408(+34,65%)	0,407(+34,76%)
	p@5 (Δ%)	0,471(+36,12%)	0,441 (+19,18%)
	p@10(Δ%)	0,321(+27,38%)	0,309(-+22,61%)
	p@15(Δ%)	0,258(+16,74%)	0,247(+8,81%)
Eagl	MAP (Δ%)	0,357 (+25,70%)	0,324 (+13,68%)
	Fscore (Δ%)	0,395(+30,36%)	0,392(+29,80)
	p@5 (Δ%)	0,458(-+32,36%)	0,417 (+21,22%)
	p@10(Δ%)	0,299(+18,25%)	0,281(+11,50%)
	p@15(Δ%)	0,222(+0,04%)	0,245(+7,92%)
BioAnnotator	MAP (Δ%)	0,334 (+17,60%)	0,314 (+10,35%)
	Fscore (Δ%)	0,385(+27,06%)	0,383(+26,82%)
	p@5 (Δ%)	0,427 (+23,41%)	0,406 (+18,02%)
	p@10(Δ%)	0,268(+6,34%)	0,261(+3,57%)
	p@15(Δ%)	0,268(+22,27%)	0,250(+10,13%)
VSM (base de référence)	MAP (Δ%)	0,284	0,285
	Fscore (Δ%)	0,303	0,302
	p@5 (Δ%)	0,346	0,344
	p@10(Δ%)	0,252	0,252
	p@15(Δ%)	0,221	0,227
AMTex	MAP (Δ%)	0,374 (+31,69%)	0,362(+27,01%)
	Fscore (Δ%)	0,403(+33,00%)	0,401(+32,78%)
	p@5 (Δ%)	0,466 (+34,68%)	0,431 (+25,29%)
	p@10(Δ%)	0,334(+32,53%)	0,324(+28,57%)
	p@15(Δ%)	0,267(-+20,81%)	0,251(+10,57%)
RéPIDo	MAP (Δ%)	**0,447(+57,39%)***	**0,438 (+53,68%)***
	Fscore (Δ%)	**0,442(+45,87%)**	**0,441(+46,02%)**
	p@5 (Δ%)	**0,564 (+63,00%)**	**0,553 (+60,75%)**
	p@10(Δ%)	**0,364(+44,44%)***	**0,360 (+42,85%)***
	p@15(Δ%)	**0,301(+36,19%)***	**0,290 (+27,75%)***

***:** *un changement significatif à $p < 0.05$*

6.6.3. Comparaison avec d'autres approches

Pour mettre en évidence l'efficacité de notre approche d'indexation, nous avons comparé la performance de RéPIDo à celle d'autres travaux. Pour ce faire, nous avons calculé le Fscore, la MAP et la précision aux rangs 5, 10 et 15. Nous considérons que l'appariement entre le corpus de test les concepts en utilisant le

VSM est la base de référence à laquelle se comparent les autres méthodes (Trieschnigg *et al.*, 2009).

7. Analyse des résultats et discussion

7.1. Utilité de donner plus d'importance aux termes ayant tous leurs mots dans la même phrase

Le tableau 4 montre que les meilleures valeurs de précision et des F-scores sont observées quand $a = 0,6$. Nous pouvons en déduire que les termes n'ayant pas leurs mots dans la même phrase peuvent être pertinents. Ainsi le réglage du coefficient a a permis de retenir le nombre maximal de ces termes (n'ayant pas leurs mots dans la même phrase et pertinents).

7.2. Valeur ajoutée du filtrage

Selon le tableau 5, nous pouvons observer que quand L est égal à 1 et 2, il n'y a pas de réduction significative de l'information inutile dans l'index final puisqu'il n'y a pas d'augmentation notable de MAP comparée à la MAP de notre approche sans filtrage. Ce résultat est expliqué par le fait que le filtrage à $L = 1$ et $L = 2$ permet l'expansion de PI avec les concepts pertinents aussi bien qu'avec les concepts non pertinents. Cependant, le filtrage à $L = 3$ permet une augmentation significative de la performance de RéPIDo (le TA de MAP, F-Score, P@5, P@10, P@15 sont respectivement (+ 24,86 %, + 20,87 %, + 19,23 %, + 15,92 % et + 14,53 %) lorsque le corpus OSHUMED est utilisé. De plus RéPIDo est statistiquement significatif comparé à la base de référence uniquement à $L = 3$ ($p = 0,0231$; $df = 43$; $t = 2,241$; $M = 0,782$). En outre, lorsque $L = 6$ nous avons une diminution notable des résultats comparés aux résultats quand $L = 3$. En effet, quand L atteint la valeur 6 il est moins possible de trouver un concept secondaire cooccurrent et possédant des relations sémantiques avec exactement les L premiers concepts principaux. Ainsi, on peut déduire que la performance de notre approche est très dépendante de la valeur de L.

7.3. Intérêt des réseaux possibiliste pour l'extraction des termes

En analysant le tableau 6, nous pouvons observer que seule la performance de l'approche basée sur les réseaux possibilistes est meilleure que celle de la base de référence (le TA est de +44,44 % pour P@10). De plus, si on compare RéPIDo avec les autres approches on peut observer que le TA de p@5 est plus élevé que les TA de p@10 et p@15 et que le TA de p@10 est plus élevé que ce lui de p@15.Par exemple, dans le cas de RéPIDo le TA est + 63,00 % pour la P@5,+ 44,44 % pour la p@10 et + 36,19 % pour la p@15. De plus, uniquement RéPIDo est statistiquement significative ($p = 0,0251$; $df = 42$; $t = 2,221$; $M = 0,771$) par rapport à la base de référence. Ces résultats montrent bien l'intérêt de l'utilisation des deux mesures de possibilité et de nécessité pour évaluer la pertinence d'un document étant donné un

concept. En effet, ces deux mesures permettent de mieux classer les concepts pertinents aux premiers rangs que l'utilisation d'une seule mesure. Les résultats prometteurs de RéPIDo sont expliqués par deux raisons : la première est l'efficacité des contributions proposée dans RéPIDo, la seconde concerne les limitations des approches testées. En fait, la base de référence, MaxMatcher et Eagl sont des approches à base d'un appariement partiel, donc les concepts qui ont un sous-ensemble de leurs mots dans le document peuvent être extraits, ce qui diminue la précision. On peut observer que la performance de BioAnnotator (Mukherjea *et al.*, 2004) est plus élevée que la base de référence car cette méthode combine deux méthodes qui sont le moteur de règles et l'apprentissage automatique. Cependant, l'inconvénient principal de BioAnnotator est qu'il n'extrait que les termes du dictionnaire. AMTex est une approche à base d'un appariement exact qui permet uniquement d'extraire les concepts appartenant au vocabulaire contrôlé. De plus, elle exploite le poids du C-valeur qui applique l'apprentissage des règles linguistiques qui dépendent du corpus, ce qui diminue sa performance.

8. Conclusion

Dans cet article, nous avons proposé une nouvelle approche pour l'indexation de documents biomédicaux basée sur les réseaux possibilistes et sur l'appariement partiel entre les documents et le vocabulaire biomédical. Cette approche est composée essentiellement de 4 étapes : le prétraitement, l'extraction des concepts, le filtrage et le classement final. Dans la deuxième étape nous avons utilisé les réseaux possibilistes pour extraire les concepts. La pertinence des concepts est ainsi estimée par les deux mesures de possibilité et de nécessité. Nous avons également proposé de donner plus d'importance aux termes ayant tous leurs mots apparaissant dans la même phrase du document. Notre contribution dans la troisième étape est de proposer de garder les concepts pertinents parmi ceux ayant un sous-ensemble de leurs termes représentatifs dans le document en utilisant les propriétés sémantiques et statistiques de l'UMLS. L'expérimentation de chaque étape montre clairement l'intérêt de notre approche d'indexation RéPIDo. Nous signalons que nous n'avons pas considéré l'ordre des mots car un grand nombre de termes MeSH ayant l'ordre de leurs mots qui ne correspond pas à leur ordre dans le document sont pertinents. Par exemple, le terme MeSH *Peritoneal dialysis, continuous ambulatory* correspond au terme *Continuous ambulatory peritoneal dialysis* dans un document. Dans nos travaux futurs nous prévoyons de calculer le score (équation 12) entre les concepts secondaires et toutes les combinaisons possibles des *n* premiers concepts principaux. De plus, nous appliquerons notre méthode en utilisant plusieurs terminologies biomédicales. En effet, l'indexation multi-terminologique est implémentée par CISMeF depuis 2009 et montre déjà de bons résultats en recherche d'information (Soualmia *et al.*, 2013).

Bibliographie

Aronson A.R., Mork J.G., Gay C.W., Humphrey S.M., Rogers W.J. (2004).The NLM indexing initiative's medical text indexer. *Med Health Info*, vol 11, n°1, p. 268–72.

Benferhat S., Dubois D., Garcia L., Prade H. (1999). Possibilistic logic bases and possibilistic graphs. *In Proc. of the Conference on Uncertainty in Artificial Intelligence*, p. 57–64.

Bodenreider O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *NucleicAcids Research*, vol 32, n°4, p. 267–270.

Bracewell D.B., Ren F., Kuroiwa S. (2005).Multilingual Single Document Keyword Extraction For Information Retrieval. *Proceedings of NLP-KE*, p. 517-522.

Boughanem M., Brini A., Dubois D. (2009). Possibilistic networks for information retrieval. *International Journal of Approximate Reasoning*, p. 957–968.

Chebil W., Soualmia L.F., Darmoni S.J. (2013). BioDI: A new approach to improve Biomedical Documents Indexing. *Proceedings of the 24th International Conference on Database and Expert Systems Applications*, August, Lecture Notes in Computer Science, Springer, p. 78-87.

Chebil W., Soualmia L.F., Dahamna B., Darmoni S.J. (2012). Indexation automatique de documents en santé : évaluation et analyse de sources d'erreurs. *IRBM BioMedical Engineering and Research*, vol33, n°5-6, p 129-136.

Chengzhi Z. (2008). Automatic Keyword Extraction From Documents Using Conditional Random Fields. *Journal of Computational and Information Systems*.vol 4, n°3, p. 1169-1180.

Couto F., Silva M., Coutinho M.J. (2005). Finding genomic ontology terms in text using evidence content. *BMC Bioinformatic*, vol 6, n°1, p. 1-6.

De Campos L., Fernández-Luna J., Huete J., Romero A.E.(2007). Automatic indexing from a thesaurus using Bayesian networks: application to the classification of parliamentary initiatives, Lecture Notes in *Artificial Intelligence*, p. 865–877.

De Campos L., Fernández-Luna J., Huete J. (2002). A layered bayesian network model for document retrieval. *Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval*, p. 169–182.

Dinh D.,Tamine L. (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*, vol 12-13, p. 41-52.

Dinh D.,Tamine L. (2011). Combining Global and Local Semantic Contexts for Improving Biomedical Information Retrieval. *Proceedings of the 33th European Conference on Information Retrieval*, p. 375–386.

Douyère M., Soualmia L.F., Névéol A., Rogozan A., Dahamna B., Leroy J.P., Thirion B., Darmoni S.J. (2004). Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info LibrJ*, Dec, vol. 21, n°4, p. 253-261.

Dubois D., Prade H. (1988). *Possibility Theory*. Plenum.

Frank E. *et al.* (1999). Domain-specific keyphrase extraction. *Proceedings of IJCAI*, p. 688–673.

Happe A., Pouliquen B., Burgun A., Cuggia M., Le Beux P. (2003). Automatic concept extraction from spoken medical reports. *International Journal of Medical Informatics*, vol 70, n°2-3, p. 255-63.

Harrathi F. (2009). *Extraction de concepts et de relations entre concepts à partir de documents multilingues : Approche statistique et ontologique*. PhD thesis. Institut National des Sciences Appliquées de Lyon.

Hersh W.R., Bhuptiraju R.T., Ross L., Johnson P., Cohen A.M., Kraemer D.F. (2004). TREC 2004 Genomics Track Overview. *Proceedings of Text Retrieval Conference*.

Hliaoutakis A., Zervanou K., Petrakis (2009). EGM. The AMTEx approach in the medical document indexing and retrieval application. *Data Knowledge Eng*, vol. 68, n° 3, p. 380–92.

Jonquet C., Le Pendu P., Falconer S.M., Coulet A., Noy N.F., Musen M.A., Shah N.H. (2011). NCBO Resource Index: Ontology-based search and mining of biomedical resources. *Journal of Web Semantics*, vol. 9, n° 3, p 316-324.

Leonard L.E. (1977). *Inter-indexer consistency studies, 1954-1975: a review of the literature and summary of study results*. University of Illinois Graduate School of Library Science Occasional Papers (131).

Manning C.D, Schütze H. (1999). Foundations of statistical natural language processing.Cambridge, *MA: MIT Press*, p. 534–36.

Markey K. (1984). Inter indexer consistency tests : a literature review and report of a test of consistency in indexing visual materials. *Library and Information Science Research*, vol. 2, n° 6, p. 155-177.

Matsuo Y., Ishizuka M. (2004). Keyword Extraction from a Single Document using Word Co-occurrence Statistical Information, *International journal on Artificial Intelligence Tools*, vol. 13, n° 1, p. 157-169.

Mukherjea S., Subramaniam, L.V., Chanda G., Sankararaman S., Kothari R., Batra V., Bhardwaj D., Srivastava B.(2004). Enhancing a biomedical information extraction system with dictionary mining and context Disambiguation. *IBM Journal of Research and Development*, vol. 48, n°5-6, p. 693-701.

Nelson S.J., Johnson W.D., Humphreys B.L. (2001). Relationships in Medical Subject Heading. *Relationships in the Organization of Knowledge*, 2001, eds. Kluwer Academic Publishers, p. 171–184.

Neveol A. (2004). *Automatisation des taches documentaires dans un catalogue de santé en ligne*. PhD thesis, Institut National des Sciences Appliquées de Rouen.

Ruch P. (2006). Automatic assignment of biomedical categories: toward a generic approach. *Bioinform J*, vol. 22, n° 6, p 658–64.

Robertson S., Walker. (1994). Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. *Proceedings of the International ACM-SIGIR Conference*, p. 232–241.

Porter M. (1981). An algorithm for suffix stripping. *Program*, vol. 14, n° 3, p. 130-137.

Singhal A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull*, vol. 24, n° 4, p 35–43.

Soualmia L.F., Sakji S., Letord C., Rollin L., Massari P., Darmoni S.J. (2013). Improving information retrieval with multiple health terminologies in a quality-controlled gateway. *BMC Health Information Science and Systems*,n°1, p. 1-8.

Trieschnigg D., Pezik P., Lee V., de Jong F., Kraaij W., Rebholz-Schuhmann D. (2009). MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, vol. 25, n° 11, p. 1412–8.

Turtle H. (1991). *Inference Networks for Document Retrieval*. Ph.D. Thesis, University of Massachusetts.

Zadeh L.A. (1978). Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets and Systems* 1, p. 3–28.

Zhang K., Xu H., Tang J., Li J. (2006). Keyword Extraction Using Support Vector Machine. *Proceedings of the Seventh International Conference on Web-Age Information Management*, Hong Kong, China, p. 85-96.

Zhou X., Zhang X., Hu X. (2006). MaxMatcher: biological concept extraction using approximate dictionary lookup. *PRICAI Pacific Rim International Conferences on Artificial Intelligence*, p. 145-149.