Indexing Biomedical Documents With a Possibilistic Network

Wiem Chebil
Normandie Université, LITIS-TIBS EA 4108, Rouen University, Cour Leschevin, Porte 21, 1 rue de Germont, Rouen Cedex 76031, France; Research Unit MARS, Faculty of Sciences of Monastir, Monastir University, Avenue of the Environment, Monastir 5000, Tunisia. E-mail: wiem.chebil@yahoo.fr

Lina Fatima Soualmia
Normandie Université, LITIS-TIBS EA 4108, Rouen University, Cour Leschevin, Porte 21, 1 rue de Germont, Rouen Cedex 76031, France; LIMICS UMR 1142, French National Institute for Health, INSERM, 15, rue de l'école de médecine, 75006, Paris, France. E-mail: Lina.Soualmia@chu-rouen.fr

Mohamed Nazih Omri
Research Unit MARS, Faculty of Sciences of Monastir, Monastir University, Avenue of the Environment, Monastir 5000, Tunisia. E-mail: Mohamednazih.omri@fsm.rnu.tn

Stéfan Jacques Darmoni
Normandie Université, LITIS-TIBS EA 4108, Rouen University, Cour Leschevin, Porte 21, 1 rue de Germont, Rouen Cedex 76031, France; LIMICS UMR 1142, French National Institute for Health, INSERM, 15, rue de l'école de médecine, 75006, Paris, France. E-mail: Stefan.Darmoni@chu-rouen.fr

In this article, we propose a new approach for indexing biomedical documents based on a possibilistic network that carries out partial matching between documents and biomedical vocabulary. The main contribution of our approach is to deal with the imprecision and uncertainty of the indexing task using possibility theory. We enhance estimation of the similarity between a document and a given concept using the two measures of possibility and necessity. Possibility estimates the extent to which a document is not similar to the concept. The second measure can provide confirmation that the document is similar to the concept. Our contribution also reduces the limitation of partial matching. Although the latter allows extracting from the document other variants of terms than those in dictionaries, it also generates irrelevant information. Our objective is to filter the index using the knowledge provided by the Unified Medical Language System®. Experiments were carried out on different corpora, showing encouraging results (the improvement rate is +26.37% in terms of main average precision when compared with the baseline).

Received May 18, 2014; revised September 4, 2014; accepted September 22, 2014

 DOI: 10.1002/asi.23435

Introduction

To improve the performance of an information retrieval system, it is essential to develop an automatic indexing system that is able to have as output the most representative index of a document. The latter may be represented by free or controlled terms. Controlled vocabulary may be a simple set of terms or a structured set of terms (or concepts) linked by hierarchical or associative relations (synonymy, broader than, narrower than, and so on), such as the Medical Subject Headings® (MeSH) thesaurus (Nelson, Johnson, & Humphreys, 2001). Using the semantic properties of a semantic resource (thesaurus, terminology, ontology, and so on), a relevant term (or concept) may be extracted although it does not occur in the document. In the case of free term indexing, keywords are extracted without using any knowledge resource. Thus, the index is not known before and extracted terms may not conform to the topic to which the document belongs. Moreover, relevant terms that do not occur in the document cannot be extracted.

To improve estimation of the similarity between a document and a given concept and based on the fact that assigning controlled vocabulary to documents is imprecise and uncertain, we propose a new approach for indexing

documents with a thesaurus based on a possibilistic network. Our approach is labeled Possibilistic Network for Documents Indexing (PoNeDI). PoNeDI is composed of four steps: pretreatment, concept extraction, filtering, and final ranking. We applied our model to the biomedical field. Thus, we used a biomedical corpus, the MeSH thesaurus, and the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT). To the best of our knowledge, the possibility theory has not yet been applied to the task of concept extraction from textual documents. In addition, the possibilistic network (PN) is a powerful method for dealing with imprecision and uncertainty and has been efficiently applied in different fields, including information retrieval (IR) (Boughanem, Brini, & Dubois, 2009). However, the IR model based on PN does not use a semantic resource; thus, indexing is not conceptual and the semantic information is not exploited. In our model, we consider the controlled indexing process as an IR process. Consequently, terms are extracted using the PN model proposed by Boughanem et al. (2009), replacing the query by a controlled term. Thereafter, a set of concepts is assigned to the extracted terms. The similarity between a document and a given term is estimated using the two measures of possibility and necessity. The first estimates the extent to which a document is not similar to the term. The second allows confirmation that the document is similar to the term. To compute these two measures, and based on the fact that a term is a set of words, two other measures are also estimated, which are the possibility and the necessity of a word of a controlled term given a document. The first eliminates all the nonrepresentative words for the document. The second strengthens the importance of the relevant words. The maximum value of possibility and necessity of the terms of a concept is assigned to that concept. Consequently, the document is described through a conceptual representation.

In addition, extracting terms may be involve exact matching (EM) or a partial matching (PM). EM allows finding in a document only the controlled vocabulary. Partial (or approximate) matching allows: (1) finding in the document other variants of terms that are different from those existing in dictionaries by applying a stemming or lemmatization process. Stemming reduces words (in the document and in the controlled resource) to their stems (or roots) (e.g., reacts, reacting, reacted, are reduced to react). Lemmatization reduces words to their base form (e.g., operation and operated are reduced to operate). PM also allows (2) extraction of multiword terms that share a subset of their words with the document. The terms extracted in the two cases may be relevant, which leads to improvement in recall. These terms may also be irrelevant, which leads to a decrease in precision. For example, in case 2, the term "breast cancer" in a document may yield the MeSH terms "testicular cancer" and "stomach cancer" because the three terms share the word "cancer" (Trieschnig et al., 2009). The PN model that we propose for indexing is based on partial matching and extracts terms in case 1 and 2. In this article, we focus on reducing the irrelevant information generated in case 2 through the exploitation of knowledge provided by the Unified Medical Language System® (UMLS) (Bodenreider, 2004).

The article is organized as follows: the section Related Work presents the related work. *Motivations for the Proposed Work* describes the motivations of the proposed model. *Biomedical terminologies* defines the biomedical terminologies we used in this study. Possibility Theory is presented next. This is followed by details of the steps of the proposed indexing approach. In Experimental Evaluations and Results, we describe the experiments and the results generated, which is followed by the Analysis of Results and Discussion. Finally, in the Conclusion section, we conclude and present some future work.

Related Work

Several research approaches for indexing documents have been proposed. We classify them as approaches based on free indexing and approaches based on controlled indexing. Owing to the fact that PoNeDI uses biomedical terminology for indexing, we focus essentially on approaches based on biomedical controlled vocabulary.

Approaches Based on Free Term Indexing

Bracewell, Ren, and Kuroiwa (2005) used natural language processing (NLP) methods for extracting keywords from a document. The first step of their approach is morphological analysis, which consists of segmenting documents to words and tagging the segmented documents to parts of speech. The next steps are extracting noun phrases, removing stop words, and clustering together the noun phrases with common noun terms. Finally, the clusters are ranked using a score based on frequency of terms and noun phrases. Moreover, keywords extraction can be seen as supervised learning. Machine learning approaches employ the keywords extracted from training documents to learn a model and apply the model to finding keywords from new documents. This approach includes naïve Bayes (Frank, Paynter, & Witten, 1999), support vector machine (Zhang, Xu, Tang, & Li, 2006), and conditional random fields (CRFs) (Chengzhi, 2008; Fkih & Omri, 2012). Keywords may also be extracted using statistical information. Matsuo and Ishizuka (2004) considered that two terms are co-occurrent if they occur in the same phrase. The authors computed the co-occurrence frequencies of pairs of terms. A co-occurrence matrix was thus obtained. This method showed comparable performance to term frequency–inverse document frequency (TF-IDF) (Salton, Wu, & Yu, 1981). Bookstein and Swanson (1974) described probabilistic models that help explain why some words in a document are relevant for indexing the document whereas others are not. Newman, Koilada, Lau, and Baldwin (2012) exploited an unsupervised Bayesian model and applied the method Dirichlet process segmentation for extracting keyphrases from a document. Jusoh and Al Fawareh (2011) proposed to

DOI: 10.1002/asi

combine linguistic and statistic methods to extract semantic keyphrases from documents. In fact, they used possibility theory to disambiguate the words after a step of a part of speech tagging.

Approaches Based on Controlled Indexing

Happe, Pouliquen, Burgun, Cuggia, and Le Beux (2003) computed a statistical weight based on TF-IDF for each term automatically extracted from the document using a method based on NLP. These terms were then matched with the terms of the ADM (Assistance with the Medical Diagnosis) dictionary. The indexing technique of Aronson, Mork, Gay, Humphrey, and Rogers (2004) noted by Medical Text Indexer (MTI) is based on three methods. The first matches the document terms with UMLS terms using MetaMap (a software tool for English language that allows mapping of a document to the concepts included in the UMLS). The second compares the phrases of the document with the phrases of the concepts using the trigram method. The third extracts MeSH terms from the k-nearest neighbors (kNN) of the document to be indexed and ranks them using a statistical weight. The indexing method of Névéol (2004) combines a linguistic method and kNN. Couto, Silva, and Coutinho (2005) computed likelihood between gene ontology terms and a document using the evidence content (EC) of a term, which is the sum of all the EC of its words. The EC of a word corresponds to its weight in the ontology. Zhou, Zhang, and Hu (2006) annotated documents with only the most significant words in the UMLS Metathesaurus®. Ruch combined, in his approach denoted by EAGL (Ruch, 2006), two models. The first is the vector space model (VSM) and the second is a regular expression pattern matcher. De Campos, Fernández-Luna, Huete, and Romero (2007) proposed a Bayesian network (BN)-based model for indexing documents with a thesaurus. Sohn, Kim, Comeau, and Wilbur (2008) classified biomedical documents with MeSH terms using a supervised BN. Only 20 MeSH terms were used for classification of documents owing to the complexity of the training task. Leung and Kan (1997) proposed a statistical learning approach for assigning controlled index terms. Mukherjea et al. (2004) developed a new tool for indexing biomedical documents called BioAnnotator. This subsequently used a parser to identify noun phrases from a document and then matches them to UMLS concepts using a rule engine. Hliaoutakis, Zervanou, and Petrakis (2009) proposed the AMTEx (automatic MeSH term extraction) model. The first step of this model is to apply the C/NC value method, which allows extraction of composed terms from the text combining statistic and linguistic information. The second step is to rank the terms according to the value of C/NC. Only terms corresponding to MeSH terms are kept. Jonquet et al. (2011) applied the Mgrep tool for extracting concepts using 200 biomedical ontologies and computed a score for each generated annotation according to its origin (preferred term, nonpreferred term, synonym term, and so on). Dinh and Tamine (2011) combined VSM (Singhal, 2001) with a proposed similarity between the terms and the document that takes word order into account. The ConceptMapper (Tanenblatt, Coden, & Sominsky, 2010) matches the words of documents to words in dictionary entries based on configurable parameters. SpeedRead (Al Rafou & Skiena, 2013) is a pipeline for extracting named entities using NLP tools. The first step is tokenization followed by part-of-speech (POS) tagging, which is an essential feature to decide the boundaries of the named entity phrases. Thereafter, a classifier helps to choose the category to which these named entities belong Prokofyev, Demartini, and Mauroux (2014) applied POS tagging to pretreated documents and used the n-gram method to extract named entities. This approach was tested on 100 documents from the proceedings of the 2012 SIGIR conference and on 100 documents from high energy physics (hep-ph) from the arXiv.org preprint repository. The entities extracted from the two data sets are linked to DBpedia and Wikipedia entities. BioDI (Chebil, Soualmia, & Darmoni, 2013) reduces the limitation of partial matching through filtering concepts, which are extracted using VSM. Takachenko and Simanovsky (2012) exploited the CRF for a supervised named entities recognition and tested their approach on CoNLL2003, OntoNotes version 4, NLPBA 2004, and DBpedia data sets. MaxMatcher+ (Dinh & Tamine, 2012) exploits the BM25 weight for ranking the concepts extracted using MaxMatcher (Zhou et al., 2006), which annotates documents with only the most significant words in the UMLS Metathesaurus.

Motivations for the Proposed Approach

Our motivations behind (a) exploiting controlled vocabulary, (b) using an unsupervised method (especially PN), and (c) proposing a method to improve partial matching are:

- The research of Leonard (1977) and Markey (1984) showed that indexing consistency is increased by 5% when controlled vocabulary is used instead of simple extraction of keywords, owing to the reasons cited in the Introduction.
- As described in the related work, most indexing approaches that use controlled vocabulary are based on an unsupervised method. This may be explained by the fact that supervised methods are not suitable for controlled indexing owing to the complexity of training a system for huge numbers of classes (approximately 24,000 classes when MeSH thesaurus is used). In addition, a supervised method depends on using a large collection of manually annotated training data. The work of Dinh and Tamine (2011) is based on the unsupervised method, VSM, which outperformed the well-known MTI method (Aronson et al., 2004) (0.234% for MTI vs. 0.273% for the method based on VSM in terms of average precision). Moreover, De Campos et al. (2007) showed that the indexing of documents using BN outperforms VSM (the average 11-point precision of VSM is 0.17 vs. 0.34). This result highlights the interest of using the graph model for indexing. Thus, we propose exploiting the idea of a possibilistic graph.

DOI: 10.1002/asi

- Possibility theory has not yet been exploited for dealing with the controlled indexing task. It showed good results when used in IR by Boughanem et al. (2009) (0.24% using PN in IR vs. 0.22% for BM25 [Robertson & Walker, 1994] in terms of average precision).
- Several indexing approaches are based on IR methods to extract concepts from documents and have achieved good results (Dinh & Tamine, 2011; Ruch, 2006). These approaches may be improved, essentially in terms of precision, by reducing the limitations of partial matching on which they are based.
- All the cited approaches that use controlled vocabulary are based on EM or PM. These two cases of matching have limitations, as described in the Introduction. We propose to reduce these limitations in our approach during the filtering step.

Biomedical Terminologies

Several biomedical terminologies are used for indexing biomedical documents. Here, we describe the UMLS, the MeSH thesaurus, and the SNOMED CT.

The MeSH Thesaurus

The MeSH thesaurus is a controlled vocabulary created by the U.S. National Library of Medicine (NLM) and is used for indexing the documents in MEDLINE (which indexes over 20 million biomedical articles). In its 2014 version, MeSH contains 27,149 main headings (descriptors) and 83 subheadings. Descriptors are used to describe biomedical articles and indexing citations. Each descriptor consists of a set of entry terms. These latter can be nonpreferred terms or preferred terms (PTs). A nonpreferred term can be narrower (NT) or broader (BT) than the descriptor or related (RT) to it. The subheadings define the meanings of the descriptors. For example "Abortion, induced" is a descriptor and its PT, NT, BT, RT, and subheading are, respectively "Abortion, induced," "Abortion, Rivanol," "Fertility Control, Post conception," "Abortion Failure," and "Adverse effects."

The SNOMED CT

The SNOMED CT is a multilingual clinical healthcare terminology. It is created and supported by the College of American Pathologists, owned, maintained, and distributed by the International Health Terminology Standards Development Organization. The SNOMED CT allows the recording of all disease entities through a set of comprehensive clinical terms linked with associative and hierarchical relations. This terminology is used also for the interoperability of electronic health records across care settings, which are beneficial for patients. The SNOMED CT covers all fields of medicine, human dentistry, and veterinary medicine.

The UMLS

The UMLS is a program launched by the NLM that offers knowledge resources for facilitating access to biomedical information. The UMLS Semantic Network is one of the knowledge sources in UMLS and consists of a set of generic semantic types linked by directed binary semantic relationships. The UMLS concepts, which are included in the Metathesaurus, are assigned to these semantic types. The concepts may be potentially connected by the same semantic relationships that link their semantic types. For example, the MeSH concepts "imaging, Three-Dimensional" and "coronary artery disease" are linked with the semantic relation "diagnoses" because their semantic types, respectively, "Diagnostic Procedure" and "Disease or Syndrome" are linked with the same relation. The UMLS also contains a table of co-occurrences between concepts in MEDLINE and other databases. Each line of the table is a pair of concepts with its co-occurrence frequency in a database. For example, the MeSH concept "Endocarditis, Bacterial" co-occurs 100 times with the MeSH concept "Penicillins" in MEDLINE. In our model, the documents are indexed with the UMLS concepts that correspond to MeSH descriptors. We also consider that the terms of MeSH and SNOMED CT descriptors are the terms of their UMLS concepts.

Possibility Theory

Possibiliy theory was introduced by Zadeh (1978) as an extension of the fuzzy logic theory and developed further by Dubois and Prade (1988). It allows handling with incomplete information and uncertainty in the interval [0, 1]. It differs from the probability theory.

Possibility Distribution

Possibility distribution π is a function from the universe of discourse X to [0, 1]. The function $\pi(x)$ evaluates the possibility that x is the actual value of some variable to which π is attached. If $\pi(x) = 1$, then x is totally possible (or unsurprising). If $\pi(x) = 0$, then x is rejected as impossible. If an event is not possible, it does not only imply that the opposite event is possible but also that it is certain. The normalization condition takes the form $max_{x \in X}(\Pi(x)) = 1$.

The Two Measures of Possibility and Necessity

The possibility of an event A, denoted $\Pi(A)$, evaluates and reflects the situation in which A is true and relevant and it is obtained by the formula $\Pi(A) = max_{x \in A}\pi(x)$. The necessity of an event A, denoted by $N(A)$, evaluates and reflects the situation in which A is false and is defined by the formula $N(A) = min_{x \notin A}(1 - \pi(x)) = 1 - \Pi(\neg A)$.

A Possibilistic Graph

A possibilistic graph is characterized by a qualitative and a quantitative component. The first is an oriented acyclic graph consisting of a set of variables $V = \{A_1, A_2, \ldots A_n\}$ that correspond to the nodes and of a set of relations that link the nodes. The second is the conditional possibilistic

DOI: 10.1002/asi

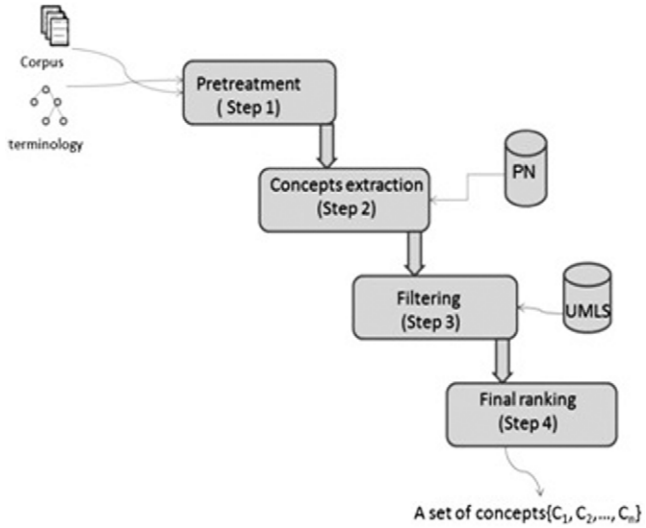

FIG. 1. The general process of our indexing approach PoNeDI.

distribution that quantified the links between a node and its parents. These possibility distributions should respect normalization.

For each variable A_i:

– If A_i is a root, which means that $Parents(A_i) = \varphi$, and the domain of A_i is dom_{A_i} then

$$\max_{\alpha} \Pi(a_i) = 1, \forall a_i \in dom_{A_i}$$

– If A_i has parents, which means that $Parents(A_i) <> \varphi$, then $max_{a_i}(\Pi(a_i/\theta_{A_i})) = 1$

$$\forall a_i \in dom_{A_i}$$

where:

dom_{A_i}: The domain of A_i.
θ_{A_i}: The set of possible configurations of the parents of A_i.

The Proposed Indexing Approach

Our approach is composed of four steps as shown in Figure 1:

1. Pretreatment
2. Concept extraction
3. Filtering
4. Final ranking

Step 1: Pretreatment

The pretreatment step consists of five tasks: (1) dividing the document indexed D_i into phrases; (2) removing punctuation; (3) pruning stop words; (4) stemming; and (5) dividing phrases to words. Tasks 2, 3, 4, and 5 are also applied to each controlled vocabulary. For example: "The binding of acetaldehyde to the active site of ribonuclease: alterations in catalytic activity and effects of phosphate" is the title of a document. After pretreatment, this title becomes "bind acetaldehydactiv site ribonucleas alter catalytactiv effect phosphat." Stemming was carried out using the Porter Algorithm (Porter, 1980). This choice is justified by the fact that Porter can be implemented in different languages,[1] including English and French, which allowed us to test our approach on English and French corpora. During the stemming process, as mentioned in the Introduction, a short stem can be confused with an acronym. Thus, as in Chebil et al. (2013), the stemming process is only applied on words that the length of their stems is equal or upper than a threshold Ts equal to 5.

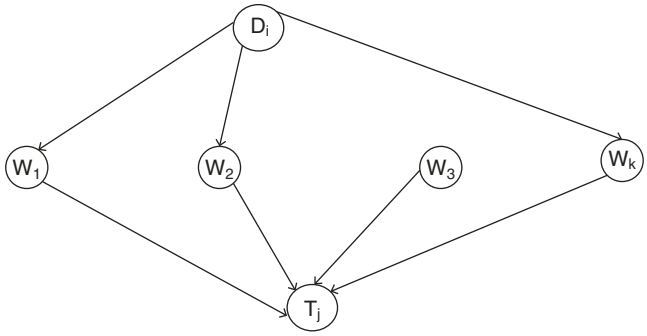

FIG. 2. The possibilistic network graph for term extraction.

Step 2: Concept Extraction

The concept extraction step begins with the extraction of terms of concepts. To extract terms, we use a possibilistic network, which allows computation of the score (Equation 1) of each term. Candidate terms are those with non-null score. The corresponding concepts are then assigned to the terms. The score of a concept is the score of its terms. If a concept corresponds to more than one term among the candidate terms, it takes the highest score. The term that gives its score to the concept is denoted the representative term (RepT).

T_j: is a term of a concept belonging to MeSH and SNOMED CT.
W_k: is a word belonging to the document to be indexed or to a term.
D_i: is the document to be indexed.
←: is a relation between two nodes.

A possibilistic network for term extraction. Terms are extracted using a possibilistic network. The architecture of the possibilistic indexing model is based essentially on the possibilistic IR model of Boughanem et al. (2009) with replacing the query Q by a term T_j. The graphic component (Figure 2) represents the following nodes: (1) a document D_i

[1]http://snowball.tartarus.org/

DOI: 10.1002/asi

(a pretreated document); (2) a word W_k belonging to the document D_i or to the term T_j; and (3) the (in) dependence relations that exist between the nodes. The domain of a document is $dom(D_i) = \{d_i, \neg d_i\}$. If the document is similar to the instantiated term, then $D_i = d_i$, if not $D_i = \neg d_i$. The domain of a word is $dom(W_k) = \{w_k, \neg w_k\}$. If the word occurs in the document or in the term, then $W_k = w_k$. If the word is absent in the document or in the term, then $W_k = \neg w_k$. The domain of a term is $dom(T_i) = \{t_i, \neg t_i\}$. If T_j is instantiated, then $T_j = t_j$, if not, then $T_j = \neg t_j$. We are interested in the instantiation of the term, so we consider only $T_j = t_j$. Thus, the term is denoted as T_j.

Evaluation of a term. Our proposed model evaluates the similarity between a document and a given term. Evaluation is carried out through propagation of the information given by the term in the network when it is instantiated. The links are activated by this instantiation from the term to the document. Two measures are computed. The first is the possibility of the document being indexed given a term (Equation 2). The second is the necessity of the document given a term (Equation 4). As Equations 1–5 show:

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

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

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

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

with:

$$\Pi(T_j) = max\left(\Pi(T_j \wedge \overline{d_i}), \Pi(T_j \wedge d_i)\right) \tag{5}$$

According to the graph topology and supposing that the words are independent, $\Pi(T_j \wedge D_i)$ is computed as follows (Equation 6):

$$\Pi(T_j \wedge D_i) = \max_{\forall \theta^l \in \theta^L} \left[\Pi(T_j|\theta^l) \times \operatorname*{prod}_{W_k \in W(T) \wedge W(D)} \left(\Pi(\theta_k^l|D_i)\right) \times \Pi(D_i) \times \operatorname*{prod}_{W_k \in W(T)/W(D)} \left(\Pi(\theta_k^l)\right) \times a \right] \tag{6}$$

- θ^L: all possible configurations of the set of parents of T_j.
- θ^l: represents a possible configuration of θ^L. θ_k^l is the instantiation of W_k in the configuration θ^l.
- The instantiation θ_1^1 of the word W_1 in the first configuration $\theta^l = \{w_1, w_2\}$ is $\theta_1^1 = w_1$.
- The possible configurations of the words of the term $T_1 = \{w_1, w_2\}$ are $\theta^L = \{\{w_1, w_2\}, \{\neg w_1, w_2\}, \{w_1, \neg w_2\}, \{\neg w_1, \neg w_2\}\}$.
- We consider that $\Pi(d_i) = \Pi(\neg d_i) = 1$ owing to the absence of any information concerning the documents.
- *prod:* means product.
- a: is a coefficient and its value belongs to]0, 1]. We consider that $a = 1$ if the words of terms are in the same phrase at least once, and $a < 1$ if not (a is experimentally tuned in *Experimental Evaluations and Results*). In fact, we hypothesize that words in the same phrase are more likely to cover the same meaning.

Aggregation of words of terms (computing $\Pi((T_j|\theta^l))$). The five canonical forms proposed by Turtle (1991) can also be adopted for the aggregation of words of terms by replacing the query by the term T_j. Thus, the words of terms can be connected with the Boolean operators (OR, AND, NOT) and probabilistic sum or one of its variations the weighted sum. We present the disjunction form (in this case, the words of terms are connected with OR) in our possibilistic model.

Disjunction. If we consider a term as a disjunctive Boolean query, then the terms having at least one word in the document are considered as candidate terms, which corresponds to a partial matching. Thus, for a term T_j having p words, we have $T_j = \{w_1 \vee w_2 \vee \ldots \vee w_p\}$. In our approach, we use PM between documents and terms, thus we consider the disjunction form for computing $\Pi((T_j|\theta^l)$ in Equation 6.

Possibility distribution. To define the representativeness of a word in a document, we consider two cases:

1. The greater the frequency of a word in the document, the higher it is possibly representative of the document.
2. The greater the frequency of a word in the document and the lower its frequency in the other documents of the collection, the higher it is necessarily representative of the document.

According to case 1, $\Pi(w_k|d_i)$ is computed as follows (Equations 7 and 8):

$$\Pi(w_k|d_i) = WW_{ki} \tag{7}$$

$$WW_{ki} = \frac{WFP_{ki}}{\max\limits_{w_r \in d_i}(WFP_{ri})}$$

$$WFP_{ki} = \sum_{y=1}^{3} FQ_{kiy} \times C_y \tag{8}$$

- WW_{ki}: The weight of a word W_k in the document D_i.
- WFP_{ki}: Word frequencies in positions.
- FQ_{kiy}: Frequency of w_k in a document D_i in the position y.
- y = 1: Title; y = 2: Abstract; y = 3: Paragraph.
- C_y: The coefficient of the position y.

For computing WW_{ki}, we use a weight based on the frequency of a word in the document that takes into account the

DOI: 10.1002/asi

TABLE 1. The conditional possibility of a word W_k given a document D_i: $\Pi(W_k \wedge D_i)$.

	d_i	$\neg d_i$
W_k	WW_{ki}	1-φ_{ki}
$\neg w_k$	1	1

position of the word in the document (in the title, in the abstract, or in the paragraphs). This weight is denoted word frequencies in positions *WFP* (Equation 8). We consider that the keywords are more dissipated in the paragraphs and mixed with nonrelevant words (compared to the abstract and title), whereas keywords are more condensed in the title (compared to abstract and paragraphs). Thus, we assigned the following coefficients to each position in the document: position coefficient (y) = 8 to the title, y = 4 to the abstract, and y = 2 to the paragraphs. For normalization, *WFP* is divided by the maximum value of *WFP* in a document.

If $WW_{ki} = 0$, then the word W_k is not representative for the document. If $WW_{ki} = 1$, then the word W_k is relevant for the document (the measure of possibility is normalized and its maximum value is 1).

According to case 2, $N(w_k|d_i)$ is computed as follows (Equations 9 and 10):

$$N(w_k|d_i) = \varphi_{ki} \tag{9}$$

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

N: the number of documents in the collection.
n_k: the number of documents where the word w_k occurs.

Having $\Pi(\neg d_i) = 1$, thus, $\Pi(w_k|\neg d_i) = \Pi(w_k \wedge \neg d_i) = 1 - N(w_k|d_i) = 1 - \varphi_{ki}$ and $\Pi(\neg w_k \wedge \neg d_i) = 1$

Table 1 summarizes the conditional possibility of a word W_k given a document D_i: $\Pi(W_k \wedge D_i)$

Estimating the Absence of a Term Word in the Document $\Pi(\theta_k^l)$

As in Boughanem et al. (2009) and with replacing the query by a term belonging to the MeSH or to the SNOMED CT, we consider that if a discriminate word of a term is absent from a document, it decreases the relevance of the term. To estimate whether or not a word is discriminate, we use the inverse frequency of the word in the collection. Thus (Equation 11):

$$\forall w_k \notin w(d_i) \quad \Pi(\theta_k^l) = 1 \quad if\ \theta_i^T = \overline{w_k} \qquad = 1 - IDF_k \quad else \tag{11}$$

With $IDF_k = \frac{log\frac{N}{n_k}}{log(N)}$

The score of a concept. The score of a concept is the maximum score of its terms (Equation 12). The term having the maximum score is the RepT.

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

$T(C_f)$: the set of terms of a concept MeSH C_f.

We consider that the possibility and necessity of a RepT are also the possibility and necessity of its concept.

Step 3: Filtering

The aim of this step is to keep only the relevant concepts among those with their RepTs having a subset of their words not occurring in the document. In fact, we classified the nonextraction of these relevant concepts as a category of indexing errors in a previous work (Chebil, Soualmia, Dahamna, & Darmoni, 2012). As in Chebil et al. (2013), this step consists in dividing the set of concepts generated in the previous step into two sets of concepts: The first set and the second set are denoted principal index *(PI)* and secondary index *(SI)*, respectively. The *PI* contains the concepts that their RepTs having all theirs words in the document. These concepts are denoted principal concepts *(PC)*. The *SI* contains the concepts that their RepTs having a subset of their words not occurring in the document. These concepts are denoted secondary concepts *(SC)*. We separate the *PC* and the *SC* because we are based on the assumption that terms having all their words in the document are more likely to be correct. Then, the relevant concepts in the *SI* are added to the PI. To perform this task, the concepts in the PI are ranked using the score (12). Thus, $PI = \{PC_1 \ldots PC_e \ldots PC_v\}$, PC_e is a PC having the rank e and v is the number of PCs in the PI. Then, we propose to compute a score s for each SC (Equation 12). S is based on the co-occurrences of MeSH concepts in MEDLINE and on the semantic relations between the concepts. In fact, our assumption is that a *SC* is more likely to be correct if it is more co-occurrent and has more semantic relations with exactly the L first *PC* in the *PI* that are considered the most relevant. L is the length of a window that contains the L first PCs. For example, according to the proposed score (13), if we fix $L = 1$, that means *S(SC)* is equal to the sum of the number of co-occurrences and relations between the *SC* and the *PC* having the rank 1*(PC1)*. If $L = 2$, that means *S(SC)* is equal to the sum of the number of co-occurrences and semantic relations between the *SC* and the two PCs having the rank 1 and 2 *(PC1* and *PC2)*. If a *SC* does not co-occur or does not have any semantic relation with one of the L PCs, or its score S is lower than

DOI: 10.1002/asi

TABLE 2. Statistics about the corpora used in tests.

	OHSUMED	CISMeF
Total number of documents	120,000	50,000
Average number of words in titles	11.2	10.5
Average number of words in abstract	132.3	100.4
Number of documents for lay people	—	22,000
Number of teaching materials	—	13,000
Number of clinical guidelines	—	15,000

a fixed threshold *Th*, it is not added to the *PI*, which is different from the filtering method in our previous work (Chebil et al., 2013). In fact, the filtering method in Chebil et al. (2013) allows the addition of a *SC* that co-occurs or has semantic relations, but not both (Equation 13).

$$S(SC) = \sum_{i=1}^{L} CF(SC, PC_i) + \sum_{i=1}^{L} NR(SC, PC_i) \quad (13)$$

CF: Co-occurrence frequency; *NR*: Number of semantic relations.

Step 4: Final Ranking

The *SC* selected in the previous step is added to the PI; the final index *(FI)* is thus constructed. The concepts of *FI* are ranked using the score (12).

Experimental Evaluations and Results

Corpora Used for the Evaluation

To test our approach, we used a subset of the OHSUMED collection[2] selected randomly and composed of 120,000 MEDLINE citations. Each selected citation is composed of title and abstract. The content of the title was merged with the content of the abstract when indexing the citations. A citation is composed of six fields: title (.T), abstract (.W), indexed concepts (.M), author (.A), source (.S), and publication (.P). We also used another corpus, which was composed of titles and abstracts of 50,000 resources selected randomly from the resources of CISMeF (Catalog and Index of Medical sites in French) (Douyère et al., 2004). Three types of documents are indexed in CISMeF: documents for lay people, clinical guidelines, and teaching material. Some statistics about the collection are given in Table 2 and Figure 3. In order to have better results and when applying PoNeDI on CISMeF corpus, we made a table of co-occurrence of concepts in the CISMeF corpus to carry out step 3 (the filtering step).

[2]http://trec.nist.gov/data/t9_filtering.html

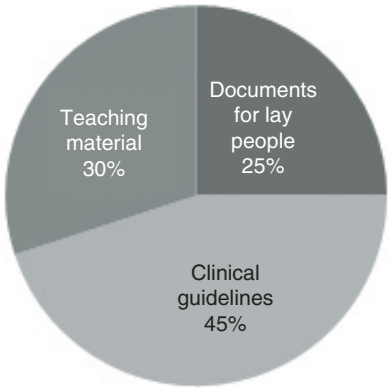

FIG. 3. Statistics about the CISMeF corpus.

In all the experiments, we kept only the first 15 concepts in the FI. In fact, the average number of concepts in the manual indexes in OSHUMED is 15 (Ruch, 2006).

Evaluation Measures

To evaluate the indexing approach, we used the mean average precision (MAP), precision, and F-score, which combines the precision and recall with an equal weight (Manning & Schütze, 1999). Moreover, we computed the ΔMAP to determinate the added value of our contributions.

In addition, to highlight the statistically significant improvements, we computed the paired-sample *t*-tests between means of each ranking obtained by each experimented method and the baseline. We consider that the difference between two given rankings is significant if $p < 0.05$ (noted*), very significant if $p < 0.01$ (noted**), and extremely significant if $p < 0.001$ (noted***).

Description of the Experiments

The experiments conducted in the implementation and evaluation stages can be classified as two categories of experiments. The first aims to tune the parameters and coefficient in order to maximize the performance of the proposed approach. The second category aims to highlight the added value of each contribution of PoNeDI. The sets of experiments belonging to the first category are (1) tuning the coefficient *a* (in Equation 6), which allows to give less. The set of experiments belonging to the first category are (1) tuning the coefficient *L* in the step of filtering. The set of experiments belonging to the second category are (1) comparing the performance of the new weight *ww/idf* with the classical weights, (2) testing the performance of PoNeDI with and without filtering, and (3) comparing the performance of PoNeDI to the performance of some existing approaches.

DOI: 10.1002/asi

TABLE 3. The final index of the citation d_1 having the PMID = 3655403 and generated using PoNeDI.

(CUI; Cf)	*((Π(d_1|Cf); N(d_1|Cf))*
(C0030842; Penicillines)	(1; 0.55)
(C0011805; Dextranase)	(1; 0; 44)
(C0061622; Glycocalyx)	(1; 0.38)
(C0014121; Endocarditis, bacterial) (the SC)	(0.33; 0)
(C0014118; Endocarditis)	(0.33; 0)
(C0003062; Animals)	(0.30; 0)
(C0053749; Viridans, Streptococci)	(0.29; 0)
(C0067086; Polysaccharides)	(0.29; 0)
(C0318157; organisation and administration)	(0.21; 0)
(C0032594; Myxococcus Xanthus Antibiotic)	(0.15; 0)
(C0965970; Et protocol)	(0.15; 0)
(C0053749; bisphenol a-glycidyl methacrylate)	(0.15; 0)
(C0035820; Role)	(0.15; 0)
(C0017963; Glycoprotein hormones, alpha subunit)	(0.15; 0)
(C0018801; Heart Failure)	(0.08; 0)

The gold standard is the manual indexing of the full document (not only the title and abstract). For the two corpora of test, the matching between the generated indexes and the gold standard is exact. For example, if the concept «Viridans, Streptococci» exists in the manual index and the concept « Viridans » exists in the automatic index and does not occur in the manual one, « Viridans » is not considered for indexing. Moreover, two indexing rules are applied by the indexers of CISMeF: (1) If a concept is the ancestor of another concept (e.g., «Endocarditis» and «Endocarditis, bacterial») and these latters occur in the same document, the first concept is not considered for indexing; and (2) if a concept is a subset of another concept and these latters occur in the same document, only the longest one is considered for indexing. Rules 1 and 2 are not the cases of the manual indexing of OSHUMED. In fact, if a concept is the ancestor or a subset of another, the two concepts will be considered for indexing. In order to be sure that the automatic indexing is performed in the same way as the manual one, the algorithm of PoNeDI follows the same indexing rules of each corpus.

Example of Indexing Citation

The citation below having the PMID (PubMed[3] identifier) 3655403[4] and belonging to OSHUMED (denoted d_1) was indexed using PoNeDI at $L = 3$ (Table 3).

Title: Enzymatic modification of glycocalyx in the treatment of experimental endocarditis due to viridans streptococci.
Abstract: The presence of abundant surface polysaccharide, or glycocalyx, on viridans streptococci has been associated with failure to eradicate the organism from experimental cardiac vegetations during penicillin treatment. The role of glycocalyx in retarding sterilization was tested by in vivo administration of dextranase, an endohydrolase that attacks internally situated alpha (1-6) linkages. Dextranase and penicillin, either singly or in combination, were used to treat experimental endocarditis. After two days of therapy, 100% of animals treated with penicillin or dextranase alone had infected vegetations, whereas only 25% treated with penicillin and dextranase had infected vegetations (P less than .01). After five days of therapy, 100% of the animals treated with penicillin had infected vegetations, versus none that were treated with penicillin and dextranase (P less than .01). We conclude that glycocalyx acts to retard antibiotic activity in vegetations and that partial enzymatic digestion of the glycocalyx facilitates penicillin sterilization of the infected valve.

The index of the citation is composed of a set of concepts (CUI[5]; *Cf*). The manual index of the same citation is as follows:

Manual Index: (C0003062; Animals); (C0011805; Dextranase); (C0014121; Encaditis, Bacterial); (C0017968; Glycoproteins); (C0026020; Microscopy, Electron, Scanning); (C0030827; penicillin G); (C0033218; Procaine); (C0032594; Polysaccharides); (C0032595; Polysaccharides, Bacterial); (C0034493; Rabbits); (C0038395; Streptococcal Infection); (C0038402; Streptococcus); (C0038412; Streptococcus Sanguis).

Results

Tuning the coefficient a. During this experiment, we tuned the value of *a* when the words of a term were not in the same phrase (Table 4). The different values of *a* tested were 1; 0.9; 0.8; 0.7; 0.6; 0.5; 0.4; 0.3; 0.2; and 0.1. For each of these values, we computed the MAP and the F-score using the two corpora OSHUMED and CISMeF. To carry out this experiment, we fixed an approximate value of *L*, which was 3. This value (the value of *L*) was tuned in the next experiment. In this experiment, the value of the threshold *Th* was fixed at its minimum value, which was 2.

Evaluation of the filtering step. To evaluate the filtering step, we generated the results (MAP and F-score) of PoNeDI for different values of *L* (Table 5). For each value of *L*, we computed the MAP and the F-score using the two corpora OSHUMED and CISMeF. The baseline considered in this

[3]The main search engine allowing the access to MEDLINE.
[4]http://www.ncbi.nlm.nih.gov/pubmed/3655403
[5]CUI = Concept Unique Identifier.

DOI: 10.1002/asi

TABLE 4. Tuning the coefficient *a*.

	1 MAP F-score	0.9 MAP F-score	0.8 MAP F-score	0.7 MAP F-score	0.6 MAP F-score	0.5 MAP F-score	0.4 MAP F-score	0.3 MAP F-score	0.2 MAP F-score	0.1 MAP F-score
OSHUMED	0.547	0.552	0.563	0.571	0.575	0.567	0.559	0.552	0.541	0.540
	0.660	0.663	0.671	0.689	0.693	0.685	0.674	0.666	0.657	0.648
CISMeF	0.544	0.551	0.561	0.569	0.573	0.561	0.551	0.556	0.547	0.537
	0.657	0.664	0.672	0.681	0.689	0.677	0.664	0.652	0.638	0.625

TABLE 5. Evaluation of the filtering step.

		OSHUMED corpus	CISMeF corpus
PoNeDI without filtering (baseline)	MAP	0.462	0.452
	F-score	0.596	0.593
	P@3	0.791	0.794
	P@10	0.652	0.652
	P@15	0.390	0.401
PoNeDI at L = 1	MAP (Δ%)	0.489 (+1.14)	0.458 (+1.32)
	F-score (Δ%)	0.597 (+0.16)	0.598 (+0.84)
	P@3 (Δ%)	0.798 (+0.88)	0.797 (+0.37)
	P@10 (Δ%)	0.673 (+0.14)	0.676 (+0.14)
	P@15 (Δ%)	0.497 (+1.22)	0.498 (+1.21)
PoNeDI at L = 2	MAP (Δ%)	0.493 (+2.28)	0.492 (+8.84)
	F-score (Δ%)	0.616 (+3.35)	0.612 (+0.32)
	P@3 (Δ%)	0.841 (+6.32)	0.849 (+6.92)
	P@10 (Δ%)	0.752 (+11.90)	0.753 (+11.85)
	P@15 (Δ%)	0.505 (+2.85)	0.511 (+3.86)
PoNeDI at L = 3	MAP (Δ%)	0.575 (+24.45)*	0.573 (+25.10)*
	F-score (Δ%)	0.693 (+16.26)	0.689 (+16.18)
	P@3 (Δ%)	0.936 (+18.33)	0.9310 (+17.25)
	P@10 (Δ%)	0.815 (+12.27)*	0.808 (+19.70)*
	P@15 (Δ%)	0.579 (+17.92)*	0.570 (+15.85)*
PoNeDI at L = 4	MAP (Δ%)	0.527 (+9.33)	0.523 (+8.05)
	F-score (Δ%)	0.648 (+8.72)	0.643 (+8.43)
	P@3 (Δ%)	0.865 (+9.35)	0.862 (+8.56)
	P@10 (Δ%)	0.741 (+10.26)	0.740 (+9.62)
	P@15 (Δ%)	0.533 (+8.55)	0.531 (+8.14)
PoNeDI at L = 5	MAP (Δ%)	0.518 (+7.46)	0.509 (+5.16)
	F-score (Δ%)	0.619 (+3.85)	0.614 (+3.54)
	P@3 (Δ%)	0.833 (+8.22)	0.831 (+7.89)
	P@10 (Δ%)	0.702 (+7.25)	0.704 (+7.32)
	P@15 (Δ%)	0.509 (+8.12)	0.499 (+7.11)
PoNeDI at L = 6	MAP (Δ%)	0.497 (+1.54)	0.492 (+1.95)
	F-score (Δ%)	0.599 (+3.85)	0.595 (+0.03)
	P@3 (Δ%)	0.801 (+1.26)	0.804 (+12.59)
	P@10 (Δ%)	0.682 (+1.48)	0.680 (+0.07)
	P@15 (Δ%)	0.498 (+1.42)	0.493 (+0.02)

Note. *A significant change at $p < 0.05$.

experiment was our PoNeDI approach without the filtering step. We carried out this experiment at $a = 0.6$ (PoNeDI generated the best results at $a = 0.6$ in the previous experiment). *Th* was fixed at 2.

Comparison Between Using ww/idf *and Some Existing Weights for Computing the Two Measures of Possibility and Necessity*

To focus on the added value of using the proposed weight *ww/idf* for computing the two measures of possibility and necessity, we conducted three experiments. These experiments are testing PoNeDI using *ww/idf*, *tf/idf*, and then *tf/BM*$_{25}$ on both test collections. The results are detailed in Table 6.

Comparison against other approaches. To highlight the effectiveness of our indexing approach, we compared the performance of PoNeDI (final results) with the performance of other approaches (Table 7). We computed, for each approach, the MAP, F-score, and precision at ranks 5, 10, and 15. We considered that MaxMatcher+ was the baseline against which the other approaches were compared. The choice of MaxMatcher+ as a baseline is based on the fact that it is among the most recent tools developed for extracting concepts from biomedical documents. Table 8 and Figure 4 put the stress on the total number of extracted concepts, the number of the extracted relevant concepts, and the number of extracted concepts having at least one word in the document for each approach compared to the gold standard.

Analysis of Results and Discussion

The results of the IR model based on a possibilistic network (Boughanem et al., 2009) showed the effectiveness of this model, compared to the baseline (the BM25 model). In fact, the use of the two measures of possibility and necessity to estimate the relevance between a query and a document allows ranking relevant documents at the top of the retrieved documents. Similar to this model, we expected, when applying the possibilistic network to indexing documents, that relevant concepts will be at the first ranks. Moreover, the IR model, based on the PN, uses the classical measure of *tf/idf* for computing the possibility and necessity. In our model, the *tf* measure is replaced by the weight *ww/idf*, which will improve the estimation of the relevance of concepts.

When dealing with the limitation of the partial match, we proposed a filtering step. The stemming process is also added to our model, which is applied only on words with the length of their stem being greater than 5. Thus, we expected that more relevant concepts would be extracted by avoiding, at the same time, the presence of the irrelevant concepts in the index.

DOI: 10.1002/asi

TABLE 6. Comparison between the use of *ww/idf* and some existing weights for computing the two measures of possibility and necessity.

	OHSUMED				CISMeF			
	P@5	P@10	P@15	MAP	P@5	P@10	P@15	MAP
PoNeDI using *ww/idf*	0.935	0.815	0.579	0.575	0.931	0.808	0.570	0.579
PoNeDI using *tf/idf*	0.857	0.755	0.535	0.513	0.853	0.752	0.515	0.517
PoNeDI using *tf/BM*$_{25}$	0.873	0.773	0.536	0.525	0.874	0.778	0.531	0.519

TABLE 7. Comparison of the performance of PoNeDI with other approaches.

		OSHUMED corpus	CISMeF corpus
MaxMatcher+ (baseline)	MAP	0.455	0.459
	F-score	0.585	0.588
	P@5	0.771	0.775
	P@10	0.651	0.652
	P@15	0.459	0.457
ConceptMapper	MAP (Δ%)	0.454 (–0.02)	0.414 (–9.80)
	F-score (Δ%)	0,559 (–4.44)	0.583 (–1.04)
	P@5 (Δ%)	0.757 (–1.81)	0.606 (–0.85)
	P@10 (Δ%)	0.648 (–.0.04)	0.641 (–1.68)
	P@15 (Δ%)	0.444 (–3.26)	0.440 (–4.13)
AMTEx	MAP (Δ%)	0.393 (–13.62)	0.398 (–13.28)
	F-score (Δ%)	0.502 (–14.18)	0.505 (–14.28)
	P@5 (Δ%)	0.691 (–10.37)	0.693 (–10.58)
	P@10 (Δ%)	0.551 (–15.36)	0.559 (–12.79)
	P@15 (Δ%)	0.407 (–11.32)	0.403 (–11.81)
PoNeDI	MAP (Δ%)	0.575 (+26.37)*	0.573 (+24.83)*
	F-score (Δ%)	0.693 (+18.46)	0.689 (+17.17)
	P@5 (Δ%)	0.965 (+25.16)	0.962 (+24.12)
	P@10 (Δ%)	0.809 (+24.27)*	0.807 (+23.77)*
	P@15 (Δ%)	0.565 (+23.09)*	0.563 (+23.19)*
BioDI	MAP (Δ%)	0.502 (+10.32)	0.503 (+9.58)
	F-score (Δ%)	0.612 (+4.61)	0.613 (+4.25)
	P@5 (Δ%)	0.851 (+10.37)	0.852 (+9.93)
	P@10 (Δ%)	0.747 (+15.27)	0.749 (+14.87)
	P@15 (Δ%)	0.495 (+7.84)	0.499 (+9.19)

Note. *A significant change at $p < 0.05$.

We present, in the following subsections, the results of each contribution and we discuss also the comparative experiments between PoNeDI and some existing indexing tools.

The Interest of Using Coefficient a *(Equation 6)*

According to Table 3, the best values of MAP and F-score are seen when $a = 0.6$ on both test collections. We can deduce the effectiveness of giving more importance to terms with all their words in the same phrase. It is also clear that some terms without their words in the same phrase may also be relevant; indeed, extraction of these terms allows extraction of more multiword terms, which characterize the biomedical terminologies. Tuning *a* allows for the maintaining of a maximum number of these terms (terms without their words in the same phrase, but relevant).

The Value of Exploiting the Weight ww/idf *for Computing Possibility and Necessity*

As shown in Table 5 and as expected, it is clear that the maximum performance of PoNeDI is seen when the weight of a word *ww/idf* is used on both test collections. In fact, the MAP and the precision at different ranks of our approach are higher when the possibility and necessity are computed with *ww/idf* than with the classical weights *tf/idf* and *tf/BM*$_{25}$ (MAP = 0.575 for *ww/idf* vs. MAP = 0.5132 for *tf/idf* vs. MAP = 0.525 for *tf/BM*$_{25}$ when the OHSUMED corpus is used). These results highlight the value of assigning coefficient to document positions, which contributes to improve the estimation of the relevance between the concepts and the documents and to better rank the extracted concepts. Moreover, these results prove that making the difference between words in the title and words in the abstract is necessary for evaluating the representativeness of the words in the document.

The Added Value of the Filtering Step

Table 6 shows, when *L* is equal to 1 and 2, that there is no significant reduction of irrelevant information in the FI. In fact, there is no notable increase of precision and F-score, compared to the precision, before expansion of the PI (the improvement rate). This result is explained by the fact that the filtering at $L = 1$ and $L = 2$ allows the PI's expansion with relevant concepts as well as with irrelevant concepts. However, it is clear that filtering at $L = 3$ gives a significant increase to the performance of PoNeDI (the improvements rates of MAP, F-score, P@5, P@10, and P@15 are, respectively: +24.45%, +16.26%, +18.33%, +12.27% and +17.92%. Moreover, only at $L = 3$ is PoNeDI statistically significant, compared to the baseline ($p = 0.0231$, $df = 43$, $t = 2.241$, M = 0.782).

We can observe, also according to Table 6, that the filtering at a different value of *L* resulted in a small improvement after PI's expansion when the three first concepts are retrieved. Nonetheless, at ranks 10 and 15, a consistent improvement of precision can be seen. Obviously, a concept that no subset of its RT occurs in the document does not have the best weight. In addition, when $L = 6$, there is a significant decrease in results, compared to $L = 3$ (improvement rates are 0.49% for MAP and 16.38% for P@10% and P@15). Indeed, at $L > 5$, there is less likelihood of finding a secondary concept that co-occurs and has semantic relations with exactly the first *L* principal concepts.

DOI: 10.1002/asi

TABLE 8. Details about the extracted concepts for each indexing system compared to the gold standard.

		Total no. of EC	No. of RC	No. of CSTD	No. of relevant CSTD @15	No. of irrelevant CSTD @15
PoNeDI	OSHUMED	2,269,684	1,478,647	482,000	481,500	NEC
	CISMeF	780,569	366,875	108,120	108,070	
MaxMatcher+	OSHUMED	2688450	1,209,802	513,145	14,5879	360,000
	CISMeF	781,122	356,216	120,256	37,023	80,020
ConceptMapper	OSHUMED	1,928,608	1,088,925	NEC	NEC	NEC
	CISMeF	795,486	352,291			
AMTEx	OSHUMED	2,169,795	1,016,325	NEC	NEC	NEC
	CISMeF	782,005	312,662			
BioDI	OSHUMED	2,114,281	1,272,879	238,125	376,762	NEC
	CISMeF	781,875	361,223	106,012	99,963	
Gold standard	OSHUMED	2,215,258		837,250		
	CISMeF	70,700		25,150		

Note. EC = extracted concepts; RC = relevant concepts; CSTD = concepts with a subset of their RT terms in the document; CTD = concepts with their RT terms in the document; NEC = no extracted concepts.

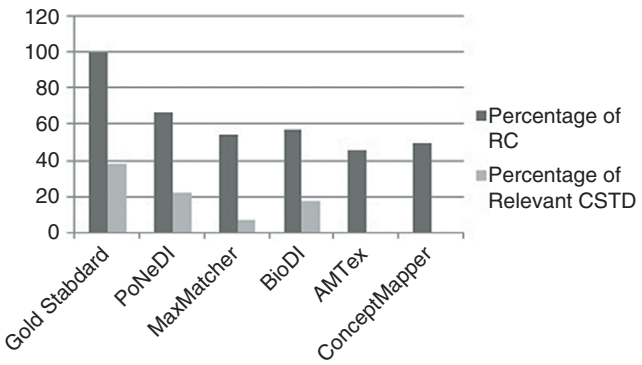

FIG. 4. Percentage of RC and percentage of relevant CSTD compared to the gold standard when using the OSHUMED corpus.

The Effectiveness of Using Possibility Networks for Concepts Extraction and Results Analysis of the Tested Approaches

When analyzing Table 7, it is clear that only the possibilistic network-based approach outperforms the baseline and all the other approaches in terms of MAP, F-score, and precision in different ranks on both test collections. Moreover, we observe that only in the case of indexing using PoNeDI (compared to all other approaches) the improvement rate at p@5 is higher than the improvement rate at p@10, and this latter is also higher than the improvement rate at p@15 (+25.16% for P@5 and +24.27% and +23.09%, respectively, for p@10 and p@15). In addition, only our approach is statistically significant, compared to the baseline, among the tested approaches ($p = 0.0257$, $df = 23.89$, $t = 2.377$, $M = 0.775$). We observe also that PoNeDI outperformed BioDI (the improvement rates of BioDI are +10.32%, +4.61%, +10.37, +15.27%, and +7.84% in terms of MAP, F-score, P@5, P@10, and P@15 when the OSHUMED corpus is used) and that BioDI outperformed the remaining approaches. Consequently, owing to the fact that BioDI and PoNeDI share the same step of filtering, this observation proves the effectiveness of the concept extraction step, which was carried out using the possibilistic network. These results confirm what was expected. In fact, using the possibility modeled by the necessity degree contributes to improve the extraction and ranking of relevant concepts. The recall of PoNeDI is also higher than the recall of MaxMatcher+ (the F-score and precision of PoNeDI are higher than the recall and F-score of the baseline), although that the baseline keeps all concepts partially matched to the document. This result is owing to the fact that PoNeDI applies the stemming process, which is not the case for MaxMatcher+.

The significant results achieved by PoNeDI are explained by two reasons: The first is the effectiveness of the contributions proposed in PoNeDI; the second is the limitations of the tested approaches. In fact, MaxMatcher+ is a PM-based approach; thus, concepts having a subset of their words in the document may be extracted, which decrease the precision. Moreover, MaxMatcher+ uses BM25 for ranking concepts, which is less efficient than *ww/idf* according to the results detailed in Table 4. The ConceptMapper (having an improvement rate equal to –0.02%, compared to the baseline, at P@10 when the OSHUMED corpus is used) is an EM-based approach, which allows extracting only concepts in the dictionary entry. In addition, the extracted concepts are not ranked using a weight that is a major limitation of the tool. The AMTEx (having an improvement rate equal to –13.62%, compared to the baseline, in terms of MAP and using the OSHUMED corpus) extracts also only terms in the thesaurus and exploits the C-value weight, which applies linguistic rules that depend on the corpus.

Table 8 showed that the total number of extracted concepts (NEC) using PoNeDI is lower than the NEC using MaxMatcher+ and higher than the NEC when using ConceptMapper. This result is attributable to filtering. In fact, PoNeDI, compared to MaxMatcher+, extracted only relevant concepts among those extracted owing to the PM. Moreover, AMTEx and ConceptMapper extract only concepts having all their words in the document. One can see also that the number of relevant concepts extracted using

DOI: 10.1002/asi

PoNeDI is the highest (Table 8 and Figure 4). In addition, as shown in Table 8 and Figure 4, the number of relevant concepts extracted by ConceptMapper is higher than MaxMatcher+, although the precision at rank 5 of MaxMatcher+ is higher. This could be explained by the fact that ConceptMapper does not rank the extracted concept, which leads to lower precision at the first ranks. Table 8 puts the stress also on the added value of the partial match and the step of filtering, which characterize PoNeDI.

In fact, 36% and 37% of the gold standard are concepts having a subset of their words in the document for, respectively, the OHSUMED and CISMeF corpora: approaches based on PM (PoNeDI and MaxMatcher+) generate these concepts. Moreover, it can be seen that 54% of relevant concepts with a subset of their RT terms in the document (CSTD) are extracted by PoNeDI versus 20% extracted by MaxMatcher+ and 45% extracted by BioDI in the same rank 15 on both test collecions. MaxMatcher+ retrieves also irrelevant CSTD, which is not the case for PoNeDI.

We can deduce, through this analysis, that filtering and using a possibilistic network in concept extraction contribute all the more to improve the effectiveness of PoNeDI. In fact, PoNeDI is statistically significant in the two cases (1) when filtering at $L = 3$ and (2) compared to the tested approaches, especially BioDI. We observe also that the performance of our approach depends on the parameters L, a, which must be well tuned to allow PoNeDI to significantly outperform the other approaches.

Conclusion

In this article, we proposed a new approach for indexing biomedical documents based on a possibilistic network. This approach is composed essentially of four steps: pretreatment, concept extraction, filtering, and final ranking. Our main contribution is to use a possibilistic network to extract concepts, which allowed estimation of the similarity between a document and a given concept, using two measures. We also propose giving more importance to extracted terms with all their words in the same phrase. In addition, our contribution in step 3 is to keep only relevant concepts among those with a subset of the words of their RTs not occurring in the document by using the UMLS. The experiments clearly showed the value of our indexing approach, which can be improved by adding other steps, such as detecting acronyms. In a future work, we aim to test the proposed approach by computing the score (Equation 13) between a *SC* and all the possible combinations of the first principal concepts with different values of *Th*. In addition, we are working on applying our approach to the CISMeF and OSHUMED corpora using biomedical terminologies other than MeSH and SNOMED CT (Soualmia et al., 2013).

Acknowledgments

The authors are grateful to Nikki Sabourin-Gibbs, Rouen University Hospital, for editing the manuscript. The authors are grateful also to the referees for their helpful comments.

References

Al Rafou, R., & Skiena, S. (2013). SpeedRead: A fast named entity recognition Pipeline. In M. Kay & C. Boitet (Eds.), Proceedings of CoRR (pp. 51–66). Mumbai, India: The COLING 2012 Organizing Committee. 2013.

Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., & Rogers, W.J. (2004). The NLM indexing initiative's medical text indexer. Medical Health Informatics, 11(1), 268–272.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. Nucleic Acids Research, 32(4), 267–270.

Bookstein, A., & Swanson, D. (1974). Probabilistic models for automatic indexing. Journal of the American Society for Information Science, 25(5), 312–318. 1974.

Boughanem, M., Brini, A., & Dubois, D. (2009). Possibilistic networks for information retrieval. International Journal of Approximate Reasoning, 50(7), 957–968.

Bracewell, D.B., Ren, F., & Kuroiwa, S. (2005). Multilingual single document keyword extraction for information retrieval. In F. Ren & Y. Zhong (Eds.), Proceedings of Natural Language Processing and Knowledge Engineering (NLP-KE) (pp. 517–522). Wuhan, China: IEEE.

Chebil, W., Soualmia, L.F., Dahamna, B., & Darmoni, S.J. (2012). Indexation automatique de documents en santé: Évaluation et analyse de sources d'erreurs. BioMedical Engineering and Research, 33(5–6), 129–136.

Chebil, W., Soualmia, L.F., & Darmoni, S.J. (2013). BioDI: A new approach to improve biomedical documents indexing. In H.Decker, L. Lhotska, S. Link, J. Basl & A.M. Tjoa (Eds.), Proceedings of the 24th International Conference on Database and Expert Systems Applications (DEXA) (pp. 78–87), Lecture Notes in Computer Science. Prague: Springer.

Chengzhi, Z. (2008). Automatic keyword extraction from documents using conditional random fields. Journal of Computational and Information Systems, 4(3), 1169–1180.

Couto, F.M., Silva, M.J., & Coutinho, P.M. (2005). Finding genomic ontology terms in text using evidence content. BMC Bioinformatics, 6(1), 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. Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU) (pp. 865–877). Hammamet, Tunisia: Springer.

Dinh, D., & Tamine, L. (2011). Combining global and local semantic contexts for improving biomedical information retrieval. In P. Clough, C. Foley, C. Gurrin, H. Lee, G.J.F. Jones, W. Kraaij, H. Lee & V. Mudoh (Eds.), Proceedings of the 33rd European Conference on Information Retrieval (ECIR) (pp. 375–386). Dublin, Ireland: Springer.

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, 12–13, 41–52.

Douyère, M., Soualmia, L.F., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.P., . . . Darmoni, S.J. (2004). Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Information Library Journal, 21(4), 253–261.

Dubois, D., & Prade, H. (1988). Possibility theory. New York: Plenum.

Fkih, F., & Omri, M.N. (2012). Complex terminology extraction model from unstructured web text based linguistic and statistical knowledge. International Journal of Information Retrieval Research, 2(3), 1–18.

Frank, E., Paynter, G.W., & Witten, I.H. (1999). Domain-specific keyphrase extraction. In T. Dean (Ed.), Proceedings of International Joint Conference on Artificial Intelligence (pp. 673–688). Stockholm, Sweden: Morgan Kaufmann Publishers.

DOI: 10.1002/asi

Happe, A., Pouliquen, B., Burgun, A., Cuggia, M., & Le Beux, P. (2003). Automatic concept extraction from spoken medical reports. International Journal of Medical Informatics, 70(2–3), 255–263.

Hliaoutakis, A., Zervanou, K., & Petrakis, E.G.M. (2009). The AMTEx approach in the medical document indexing and retrieval application. Data Knowledge Engigneering, 68(3), 380–392.

Jonquet, C., Lependu, P., Falconer, S., Coulet, A., Noy, N.F., Musen, M.F., & Shah, N.H. (2011). NCBO resource index: Ontology-based search and mining of biomedical resources. Journal of Web Semantics, 9(3), 316–324.

Jusoh, S., & Al Fawareh, H.M. (2011). Semantic extraction from texts. In M. Othman & Y. Xie (Eds.), International Conference on Computer Engineering and Applications IPCSIT (pp. 595–601). Manila, Phillippines: IACSIT.

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 (p. 131).

Leung, C.H., & Kan, W.K. (1997). A statistical learning approach to automatic indexing of controlled index terms. Journal of the American Society for Information Science, 48(1), 55–66.

Manning, C.D., & Schütze, H. (1999). Fondations of statistical natural language processing (pp. 534–536). Cambridge, MA: MIT Press.

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, 2(6), 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, 13(1), 157–169.

Mukherjea, S., et al. (2004). Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. IBM Journal of Research and Development, 48(5–6), 693–701.

Névéol, A. (2004). Automatisation des tâches documentaires dans un catalogue de santé en ligne. Ph.D. thesis, Institut National des Sciences Appliquées de Rouen.

Nelson, S.J., Johnson, W.D., & Humphreys, B.L. (2001). Relationships in medical subject heading. In C.A. Bean (Ed.), Relationships in the organization of knowledge (pp. 171–184). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Newman, D., Koilada, N., Lau, J.H., & Baldwin, T. (2012). Bayesian text segmentation for index term identification and keyphrase extraction. The 23rd International Conference on Computational Linguistics (COLING) (pp. 2077-2092). Mumbai, India: The COLING 2012 Organizing Committee.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14(3), 130–137.

Prokofyev, R., Demartini, G., & Mauroux, P.C. (2014). Effective named entity recognition for idiosyncratic web collections. In C.W. Chung (Ed.), Proceedings of the 23rd International Conference on World Wide Web (pp. 397–408). Seoul: IW3C.

Robertson, S., & Walker, S. (1994). Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In W.B. Croft & C.J.V. Rijsbergen (Eds.), Proceedings of the International ACM-Special interest Group on Information Retrieval Conference (pp. 232–241). Dublin: ACM.

Ruch, P. (2006). Automatic assignment of biomedical categories: Toward a generic approach. Bioinformatic Journal, 22(6), 658–664.

Salton, G., Wu, H., & Yu, C.T. (1981). The measurement of term importance in automatic indexing. Journal of the American Society for Information Science, 32(3), 175–186.

Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Engineering Bulletin, 24(4), 35–43.

Sohn, S., Kim, W., Comeau, D.C., & Wilbur, W.J. (2008). Optimal training sets for Bayesian prediction of MeSH assignment. Journal of American Medical Informatic Association, 15(4), 546–553.

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, 1(1), 1–8.

Takachenko, M., & Simanovsky, A. (2012). Named entity recognition: Exploring feature. In J. Jancsary (Ed.), Proceedings of KONVENS 2012 (pp. 118–127). Vienna: ÖGAI.

Tanenblatt, M.A., Coden, A., & Sominsky, I.L. (2010). The ConceptMapper approach to named entity recognition. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), The International Conference on Language Resources and Evaluation (LREC) (pp. 85–96). Valletta, Malta: Springer.

Trieschnig, D., Pezik, P., Lee, V., de Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: Effective MeSH text classification for improved document retrieval. BMC Bioinformatics, 25(11), 1412–1418.

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(12), 3–28.

Zhang, K., Xu, H., Tang, J., & Li, J. (2006). Keyword extraction using support vector machine. In J.X. Yu, M. Kitsuregawa, & H.V. Leong (Eds.), Proceedings of the Seventh International Conference on Web-Age Information Management (WAIM) (pp. 85–96). Hong Kong: European Languages Resources Association (ELRA).

Zhou, X., Zhang, X., & Hu, X. (2006). MaxMatcher: Biological concept extraction using approximate dictionary lookup. In Q. Yang & G. Webb (Eds.), Pacific Rim International Conferences on Artificial Intelligence (PRICAI) (pp. 145–149). Guilin, China: Springer.

DOI: 10.1002/asi