Scientometrics (2020) 124:903–924
https://doi.org/10.1007/s11192-020-03474-w

Approximate matching-based unsupervised document indexing approach: application to biomedical domain

Kabil Boukhari[1] · **Mohamed Nazih Omri**[1]

Received: 2 April 2019 / Published online: 7 May 2020

Abstract
Document indexing is considered as a crucial phase in the information retrieval field because textual information is constantly increasing. With this accumulation of documents, the satisfaction of user needs becomes more and more complex. For these reasons, several information retrieval systems have been designed in order to respond to user requests. The main contribution of the current work resides in the suggestion of a novel hybrid approach for biomedical document indexing. We improve the estimation of the correspondence between a document and a given concept using two methods: vector space model (VSM) and description logics (DL). VSM performs partial matching between documents and external resource terms. DL allows representing knowledge in a relevant manner for better matching. The proposed contribution reduces the limitation of exact matching. It serves to index documents by exploiting medical subject headings (MeSH) thesaurus services with approximate matching. The latter partially matches document terms with biomedical vocabularies to extract other morphological variants in that resource. It also generates irrelevant concepts. The filtering step solves this problem and grants the selection of the most important concepts by exploiting the knowledge provided by MeSH. The experiments, carried out on different corpora, show encouraging results of around 25% improvement in average accuracy compared to other approaches studied in the literature.

Keywords Document indexing · Vector space model · Description logics · Partial matching · Stemming · Biomedical vocabulary

Introduction

The main purpose of an IR system is to find, from a query and a collection, relevant information that meets a user demand (Naouar et al. 2016, 2017; Fkih and Omri 2016a, b Dahak et al. 2017). A document is the most popular information on the internet, and even the most requested. A query is defined as a bag of words representing a demand. Classical IR models

✉ Kabil Boukhari
kabil.boukhari@gmail.com

Mohamed Nazih Omri
Mohamednazih.omri@fsm.rnu.tn

[1] MARS Research Laboratory, University of Sousse, Sousse, Tunisia

are based only on the words present in the document. However, irrelevant documents sharing a set of words with the query can be returned to the user. The IR system strength depends mainly on three tasks: document representation (indexing), query representation, and query/document matching. Document indexing tends to select and extract words that mostly match to the document content, to facilitate later the information retrieval (Baoli et al. 2018). The manual information processing is too expensive and requires enough time especially when it comes to a specific field (biomedical field) that accepts a high accuracy rate such as the medicine field. In the biomedical field (Nicolas et al. 2015; Lv et al. 2014; Song 2015; Liu and Wacholderc 2017; García et al. 2018) various works have been undertaken to amend the process of information retrieval. For example, the MEDLINE database contains more than 24 million scientific articles in the biological and biomedical science field and is indexed using the Medical Subject Headings (MeSH) thesaurus.

The indexing task is the heart of documentary information retrieval. Indeed, this process requires too much quality and relevance from the indexer, especially in terms of domain and language knowledge (Jutinico et al. 2019). As bad indexing will necessarily lead to irrelevant results, it is necessary to improve the indexing process to obtain an efficient document retrieval system.

Several works have been developed to refine the documentary research process (Hao et al. 2018) and improve access to information (Wongthongtham and Salih 2018; Abu-Salih et al. 2018a, b). Major tasks, such as the query/document representation and matching (Ru et al. 2018), and other sub-tasks, like stemming (Ali et al. 2019; Alotaibi and Gupta 2018; Karaa 2013) lemmatization and disambiguation, have been presented. The current work aims to minimize the error rate and to maximize the accuracy and relevance of provided results. Thereby, the main role of this research is to bring results closer to those achieved by manual work, replace experts and save a lot of time. Indeed, indexing approaches use various methods of automatic information treatment, such as statistical, semantic, probabilistic and possibilistic methods. These approaches go through the same indexing process: first, preparation (or pretreatment) consists in removing all the empty words, symbols, punctuations, etc, so that the automatic language processing can apply the indexing sub-tasks, namely lemmatization and stemming. The definitions of these two methods are very similar. Lemmatization consists in reducing words to their canonical form, whereas stemming is used to transform words to their radicals or stems. Second, the candidate terms extracted in this part may be free or controlled by an external source, such as a dictionary, taxonomy or thesaurus. A weighting score is assigned to each term with statistical measures. The final step is to classify terms by their order of relevance in order to select those representing the best document or query.

The rest of this paper is organized as follows: "Related work" section provides the related work. "Motivations for suggested approach" section presents the motivations for the proposed approach. The fourth section presents the main purpose of this research work and gives justifications for each choice. This is followed by a section presenting a description of the proposed approach. The next section describes the experiments and the obtained results, which is followed by a discussion part justified by figures. In the final section, we summarize the work and we present some future works.

Related work

Recently, many indexing approaches (Aravazhi and Chidambaram 2018; Yuan 2018) have been proposed, each has its own characteristics. For this, these approaches have been classified into two categories: free vocabulary based and controlled vocabulary based.

- *Approaches based on free language*

Approaches based on this type of indexing exploit only the words composing the document to represent it. In Zhang et al. (2008), the authors proposed a method for extracting keywords based on Conditional Random Fields (CRF). They began with a complete analysis of the document. All the units, namely the title, the summary, the text and even the references, were processed. After the segmentation phase and the marking of each word in the sentence, each word or sentence would correspond to a vector. CRF resumed the extraction task of representative terms, and uses the SegTag[1] Tool for processing automatic labels. The model took as input characteristic vectors and the marked data is used to form the CRF model. The document was pre-processed and its characteristics were extracted. Then, as the keyword type had been already specified using the CRF model, the document representatives were extracted. In Hassan Approach (Mahedi et al. 2018), the authors presented a semantic approach to extract important keywords from documents. The main idea of the suggested approach is to represent a document by semantically close words exploiting a similarity measure. The model presented by Fkih and Omri (2012) proposed to extract complex terms from texts, and they integrated linguistic and statistical knowledge. The authors in You et al. (2013) offered automatic keyword extraction approaches for scientific documents. This model would generate the candidate expressions based on a word expansion algorithm to reduce the size of candidate sentences. As a consequence, a document frequency feature was introduced, in addition, new features were added for sentence weight to maximize accuracy value. Bracewell et al. (2005) used Natural Language Processing (NLP) and suggested methods to extract keywords from a document. The first step was a morphological analysis, which consists in segmenting the documents into words and labeling the segmented documents into parts. The next steps were to extract nominal sentences to delete blank words and to group nominal sentences together with nominal terms in common. Finally, groups were ranked according to a score based on term frequency and nominal sentences. The approach of Matsuo and Ishizuka (2003) was a keyword extraction approach that was applied to a single document without using the entire corpus. A document was composed of a set of sentences, and a sentence was considered as a set of words separated by empty words. Matsuo and Ishizuka (2003) also included document titles, section titles and captions as sentences. Each sentence was considered as a basket ignoring the order of terms as well as grammatical information, except when extracting sequences of words. A list of frequent terms was obtained by counting their frequencies. Then a co-occurrence matrix was computed by counting the co-occurrence frequencies of each pair of terms. Thirty % of the most frequent terms were selected, and those whose Jensen-Shannon divergence was greater than a given threshold were grouped by a pair of terms. Probability was calculated by counting the number of terms that co-occured with the groups of terms prepared previously. The final word weight w served to calculate the X'^2 measure, which

[1] CNLP.Platform http://www.nlp.org.cn.

integrated the w co-occurrence frequency with the prepared groups plus the total number of terms in sentences, including w. The final list of keywords included words having the highest values of X'^2.

- *Approaches based on controlled language*

Approaches based on controlled are based on an external resource and use controlled vocabulary for indexing documents. The authors of Happe et al. (2003) proposed an indexing approach by exploiting a tool called NOMINDEX, which extracted MeSH concepts from medical texts in a natural language. The process begins with a word extraction and then proceeds to the extraction of concepts using the MeSH thesaurus. Finally, the authors exploited the UMLS[2] metathesaurus architecture for the final index generation. The final list of concepts was compared to that of manual indexing. The authors of Zhou et al. (2006) put forward an approach called MaxMatcher. The basic idea of this work was to select some significant words instead of all words referring to a specific concept. They used the approximate dictionary lookup,[3] which was exploited to match the controlled vocabulary to the words. In Jonquet et al. (2011), the authors utilized more than 200 ontologies to facilitate the space of medical discoveries by providing to scientists a unified view of this diverse information. They used the semantic properties of ontologies, such as class properties, class hierarchies and navigations between ontologies, to improve the search experience for the resource index user. BioAnnotator (Mukherjea et al. 2004) utilized the same resource of the MaxMatcher approach to identify and classify terms in documents. This approach used a biomedical dictionary to learn extraction instances for the rule engine to disambiguate terms that appeared to diverse semantic classes. Finally, a score was calculated for each produced annotation according to the original source. In Sohn et al. (2008), the authors opted for the supervised Bayesian network to classify biomedical documents. This approach could improve the links between indexing terms and biomedical documents. The authors selected 20 MeSH terms whose occurrences covered a specified range of frequencies. Each MeSH term was assigned to one of the sets: the first set was called "optimal" and included all documents with a given MeSH term (C_1 class). The second set included documents with one of its terms (class C_{-1}) and was close to the C_1 class. These small sets were used to manage MeSH assignments in the MEDLINE database. The Aronson et al. (2004) approach was based on three phases: after a document preparation step, the MetaMap tool took its role in matching document terms to UMLS terms. The second phase exploited the trigam method to compare between concepts and document sentences. It compared the first three characters of each sentence word with the first three characters of each UMLS term word. Finally, a learning phase extracted the MeSH descriptors for the indexing document. In the two IBioDI and IBioDL approaches (Boukhari and Omri 2017a, b) the authors used different tools to extract terms and they exploited the MeSH thesaurus for the correspondence step with partial matching. In the same context, Chebil et al. (2013) proposed an approach to index biomedical documents by using the MeSH thesaurus, to overcome the limits of partial correspondence. The basic idea exploited the Vector Space Model (VSM) to extract descriptors and combined a static and semantic method to estimate the descriptor relevance for a given document. The author utilized the TF-IDF (Sun

[2] https://www.nlm.nih.gov/research/umls/.

[3] This research work is supported in part from the NSF Career grant (NSF IIS 0448023). NSF CCF 0514679 and the research grant from PA Dept of Health.

et al. 2017; Arroyo-Fernández et al. 2019) measure to calculate the weights of document terms and external resources. The knowledge provided by Unified Medical Language System (UMLS) was exploited for the filtering phase, whose aimed to keep relevant descriptors. Soldaini and Goharian (2016) suggested the QuickUMLS approach, which used a dictionary for approximate matching in order to extract medical concepts. The proposed method was dedicated for big databases and large corpora.

The approaches cited in the previous section have advantages and disadvantages. The weight exploited by the first category of approaches to classify indexing terms is a unique measure. This measure is generally statistical and does not exploit the semantic properties of terminologies. This weight does not sufficiently reflect the relevance of the document given a term. In addition, the described approaches do not propose solutions to limit the frequency of the errors generated by the stemming. Moreover, existing approaches do not propose to reduce the frequency of ambiguity errors for terms with more than one concept.

The main and common disadvantage of the second category approaches, is the partial matching which allows the extraction of irrelevant terms having part of their words absent in the document. Indeed, these approaches do not offer a solution to keep only the relevant concepts. Other approaches are based essentially on an exact matching that only allows to extract the terms having all their words in the document which decreases the performance of the indexing. Besides, approaches that use UMLS concepts (MetaMap for example) cannot be applied in the case of use of terminologies other than those of UMLS. Furthemore, the term words belonging to the controlled vocabulary are often in the wrong order which qualifies the usefulness of the similarity measure "rank correlation coefficient". This similarity measure seems not very precise since it is based on the average position of a word in the document.

Motivations for suggested approach

The proposed approach combines two unsupervised concept extraction methods, The first one is based on VSM, and the second one is based on Description Logics (DL). In addition, it is based on partial matching by exploiting controlled vocabulary. Our motivations behind these choices are illustrated in the following points: (1) DL is not used in document indexing with a controlled vocabulary. Since it has shown good results in reported works, (Radhouani et al. 2008; Radhouani and Falquet 2008), it is integrated in our work. (2) The use of external resources increases the precision value and improves indexing. (3) The majority of indexing approaches that exploit controlled vocabulary use unsupervised methods. Therefore, these supervised methods are not suitable for controlled vocabulary because of the large number of classes. Dinh's work (2011) exploited an unsupervised method based on VSM, which gave good results in terms of precision. (4) Partial matching in indexing gives better results compared to exact matching. Moreover, the limits of the first technique are overexposed by the filtering phase.

Work objectives

The goal of this work is to develop a new indexing approach, for biomedical documents, using two unsupervised methods (VSM and DL) to minimize indexing errors. In our approach, we consider the controlled indexing process as an IR process, contrary to free

indexing which uses only the vocabulary existing in the document to represent. In the IR process, the correspondence is carried out between the document layer and the query layer while in our approach, the query layer is changed by the controlled vocabulary. The exploitation of the IR architecture is justified by the estimation efficiency of the similarity between a document and a given concept and by the elimination of irrelevant concepts (elimination of irrelevant queries in the IR process). Furthermore, in our contribution, concepts are extracted using VSM and DL. The correspondence between the two layers is done through partial matching. This choice is justified by the disadvantages induced by exact matching. The latter allows extracting only terms present in the document, while partial correspondence permits identifying other morphological variants that can index the document. VSM gives more importance to the term whose words all occur in the document and which can be a representative term. Indeed, if the weights of words of the external resource are equal to the weights of words of the document, the similarity degree is improved between the document and the controlled vocabulary. DL represents the terms of the external resource/document with descriptive expressions, which gives more relevance in the correspondence phase. The constructors and roles provided by DL give a lot of improvement to refine the matching between the document and the controlled vocabulary. Moreover, the DL is characterized by the robustness and high performance of its reasoners. These two methods improve the extraction of relevant concepts. Finally, the objective of the last part is to select the most relevant concepts using external resources.

VSM-DL based document indexing approach

The proposed approach consists in extracting the representative concepts of a document from an external resource. It starts with a pre-processing step followed by a phase of clearing concepts by exploiting VSM and DL. The proposed approach is based on the partial (or approximate) matching that permits to find, in the documents, other variants of the controlled vocabulary by applying stemming. It ends with a filtering step that enables filtering the extracted concepts from the previous phase and classify them according to a score. This part represents a robust solution to minimize incorrect concepts generated by partial matching. The proposed approach uses partial matching for many reasons. Firstly, exact matching allows finding within a document only the vocabulary present in the external resource (dictionary, taxonomy, ontology...). However, partial matching enables finding other variants of terms that are different from those existing in the external resource by applying the stemming process. Stemming reduces words (document/controlled vocabulary) to their roots (e.g. 'treats', 'treating' and 'treated' are reduced to 'treat'). Also, partial matching permits extracting terms that share a subset of their words with the document. For example, the term "adrenal cancer" in a document may yield the MeSH terms "bladder cancer" and "stomach cancer" because the three terms share the same word "cancer". In our approach, a specific step is proposed to filter the irrelevant concepts derived from the partial matching. The aim of this step is to keep only the relevant concepts among those with terms having a subset of their words not occurring in the document.

These steps are summarized in the following points:

- Pre-processing
- Concepts extraction with VSM
- Concepts extraction with DL

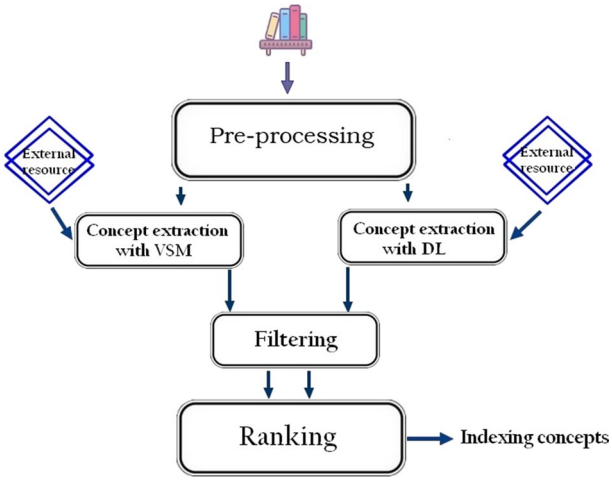

Fig. 1 General indexing process of proposed approach

- Concepts filtering
- Ranking

Pre-processing

The preparation phase, also called the pre-processing phase, consists in preparing the documents and the external resource words for indexing. It consists as well in segmenting the documents into sentences, and then breaking them into words, deleting the punctuation and empty spaces, eliminating symbols and numbers, and finally stemming words (Fig. 1).

After the pre-processing phase on a document entitled *Reductions in breath ethanol readings in normal male volunteers following mouth rinsing with water at differing temperatures*, we obtain *reduct breath ethanol read norm mal volunte follow mouth rins wat differ temper*. For that, we exploit the RAID stemmer (Boukhari and Omri 2016) services, an improved version of the SAID stemmer (Boukhari and Omri 2015), for words stemming.

Concepts extraction phase

This section presents the fundamental part of the proposed approach. It is based on two methods: the VSM method and the DL-based. Both methods give good results and increase the accuracy rate. Each method allows to extract a set of concepts for the same document. After, these two sets of concepts will be passed to the filtering stage. The objective of the use of these two methods is to give more robustness to the proposed approach and more relevance to the extracted concepts.

Concepts extraction based on VSM

This part begins with a measure of the word importance degree in the document (weight of each word in the document), followed by a measure of the word weight in the MeSH terms. Each bag of words is a vector, *VDoc* for the document words and *VTmesh* for the MeSH terms. A similarity measure will be calculated using the two previous vectors to show the

correspondence between them; this measure is added to the MeSH term weight, in the document, to calculate the final term score.

Similarity computation starts with:

- Determining the word weight in a document, noted Word Weight in the Document(WWD).
- Measuring of the word importance degree in a MeSH term, noted Word Weight in the MeSH Term (WWMeshT).

The steps of this part are represented by Algorithm 1:

Algorithm 1: Algorithm of concepts extraction based on VSM

Input: D: document; m: number of words in document; n: number of words in term; W: Word in document; WMesh: Word in MeSH term; i,j,k: counter; l: Number of terms that belong to the concept.

Output: Bag of concepts

```
while ! end_of_corpus () do
    for i ⟵ 1 to m do
        WWD_i ⟵ Word_weight_in_document(W_i) (see eq.1)
    end
end
while ! end_of_MEsH_terms () do
    for j ⟵ 1 to n do
        WWMeshT_j ⟵ Word_weight_in_MEsH_thesaurus(WMesh_j) (see eq.2)
    end
end
for i ⟵ 1 to m do
    for j ⟵ 1 to n do
        Sim ⟵ Similarity(TMesh_j, D_i) (see eq.3)
    end
end
for i ⟵ 1 to m do
    for j ⟵ 1 to n do
        Mesh_D ⟵ Mesh_term_in_document(TMesh_j, D_i) (see eq.4)
    end
end
for k ⟵ 1 to l do
    Return Max(Sim_k,Mesh_D_k) (see eq.5)
end
```

- *Computing WWD*

To measure the word weight in a document, we use measure BM25 (Jiménez et al. 2018). Since biomedical vocabulary is a set of semantically closed terms, we integrate the independence between the document words and between those in the rest of the collection. This weight is given by the following equation:

$$WWD\_i = TF_i * \left(\frac{\log\left(\frac{N - df_i + 0.5}{df_i + 0.5}\right)}{TF_i + k_1 * ((1-b) + b\frac{LD_i}{ALD_i})} \right) \quad (1)$$

with TF_i, word frequency in the document; N, number of documents in the collection. df_i, number of documents containing word W; LD_i, document length (number of words); ALD_i, the average length of the document in relation with the collection; k_1, b, free parameters.

Although a document is a citation formed by a title and a summary, we attribute the same importance to the different positions of a word in a document. Since the theme of documents is the same, we let the word deal with the whole collection. In fact, k and b are normalization factors that will be fixed during the experimental part.

- *Computing WWMeshT*

We consider the same BM25 measure of the previous part to measure the word weight in the MeSH thesaurus, which is given by the following equation:

$$WWMeshT\_i = WFMeshT_i * \left(\frac{\log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right)}{WFMeshT_i + k_1 * ((1-b) + b\frac{TL_i}{ATL_i})} \right) \quad (2)$$

with $WFMeshT_i$, word frequency in a MeSH term; N, number of terms in the MeSH thesaurus; n_i, number of terms containing word W; TL_i, term length (number of words); ATL_i, average length of a term in the MeSH thesaurus; k_1, b, free parameters.

The MeSH architecture begins with a header, called descriptor.The latter is constituted by a set of concepts $C_1, C_2, \ldots, C_n$. A concept is formed by a set of terms $T_1, T_2, \ldots, T_n$, and a term, in its turn, is composed of words $W_1, W_2, \ldots, W_n$. The thesaurus vocabulary is dependent, since it belongs to the same domain.

Equation 2 integrates all the thesaurus terms into the weight calculation to give more importance to the word to be treated. The weight is measured by the ratio between the frequency of W multiplied by its importance score throughout the thesaurus, and the frequency sum of W with its importance score in term T.

- *Compute of the similarity degree between document and MeSH term*

The similarity measure (also called distance measure) is a metric that measures the distance between two vectors and applied in the approximate search.

Document *Doc* and term MeSH *TMesh* are respectively represented by two vectors *VDoc* and *VTMesh*, and each one of them represents a bag of words:

- VDoc = WWD $_1$, WWD $_2$, ..., WWD $_n$ where WWD_i is the word weight in the document.
- VTMesh = WWMeshT $_1$, WWMeshT $_2$, ..., WWMeshT $_n$, where WWMeshT_i is the word weight in a MeSH term.

The vector representation of a document is mainly used in information retrieval models. The vector space concept is introduced in a documents space in an understandable language. It is often exploited to calculate the correspondence between two documents. In our case, the proximity between a document and a MeSH term is measured with

cosine similarity. This qualifies the similarity between *Doc* and *TMesh* by calculating the cosine between their vectors *VDoc* and *VTMesh*.

We consider X as the working universe. The cosine similarity *sim* is the function $X \times X \rightarrow R$, which helps to check the following properties:

- *Positivity* $\forall TMesh, Doc \in X, Sim(TMesh, Doc) \geq 0$ The similarity value is always positive because the two weights WWMeshT and WWD have positive values, so the similarity result with the cosine is positive.
- *Symmetry* $TMesh, Doc \in X, Sim(TMesh, Doc) = Sim(Doc, TMesh)$ The order of parameters in the cosine measure does not affect the result and returns the same value.
- *Maximality* $\forall TMesh, Doc \in X, Sim(TMesh, TMesh) \geq Sim(TMesh, Doc)$ A similarity between two vectors having the same value returns 1, which is the maximum similarity.

This degree, denoted by *Sim(TMesh, Doc)*, is given by the following equation:

$$Sim(TMesh, Doc) = \frac{\sum_{i=1}^{n} WWMeshT_i * WWD_i}{\sqrt{\sum_{i=1}^{n} (WWMeshT_i)^2 * (WWD_i)^2}} \tag{3}$$

- *Computation of MeSH term weight in document (WMeshTD)*

Based on Eq. 1, this new weight is used to measure the MeSH term weight in the document. WMeshTD is equal to the sum of words weights in the MeSH term divided by the total number of words in the term, and multiplied by a refinement coefficient, *Co*, to give more importance to the words in the same sentence (Ferjani et al. 2012).

The similarity degree, given by the cosine does not integrate the dependence notion between words in the same sentence. We give more importance to the term words that belong to the same sentence, as they admit close semantics. The ignorance of this kind of words relationship can suffer from a loss of precision. Therefore, we add this new weight to improve the relevance in terms ranking. The coefficient *Co* takes a value greater than 0 when the set of words are found together, at least once, in the document.

A new significance score assigned to a term is then calculated by the following equation:

$$WMeshTD = \frac{\sum_{i=1}^{n} WWD(W_i)}{n} * Co \tag{4}$$

- *Co > 1 if the term MeSH words are in the same sentence.*
- *Co = 1 otherwise.*

- *Concept weight calculation*

The concept weight is calculated according to the weights attributed to its terms. Once the candidate terms, representing a document, are extracted, the maximum weight of these terms is associated to the concept which they belong to. The final score of a term is the sum of its similarity to the document (Eq. 3) and its weight in the document (Eq. 4).

$$WC_k = \max_{j\in[1;n]}(WMeshTD_j + sim(TMesh_j, Doc)) \quad (5)$$

where WC is the concept weight calculation .

The number of terms that belong to the concept is indicated by n and j is a counter to denote the term being processed. After the calculation of the weights of the candidate concepts representing a document, such concepts are grouped into a single set to go through a filtering phase.

Concepts extraction based on DL

- *Description logics*

DL (Warren et al. 2019) are defined as a family of knowledge representation languages, used to represent the terminological knowledge of an application domain in a formal and structured way. Knowledge is divided by the DL into two levels: (1) The terminological Box (TBox): encodes the general knowledge of a domain and describes the semantic relationships between concepts and relations. (2) The Assertional Box (ABox): describes specific or local information about individuals belonging to certain concepts.

For DL, a common base is enriched with different extensions and each one possesses its own constructors. Kernel DL is an attributive language, which contain: concept intersection, atomic negation, limited existential quantification and universal restrictions. Its extension is the attributive language with complements which is used in our work and supports terminological knowledge representations (TBox axioms) using equivalence ($\equiv$) and subclass relationships ($\sqsubseteq$), disjunction ($\sqcup$), conjunction ($\sqcap$), negation ($\neg$), contradiction ($\bot$), tautology ($\top$), and property restrictions ($\forall$, $\exists$). In the proposed approach, the ALCQ family was used to give more expressiveness in the knowledge representation phase.

Let A and B be two atomic concepts, C and D two concept descriptions, and R an atomic role. Semantics is defined using interpretation $I(\triangle^I, .^I)$. It consists of a non-empty set $\triangle^I$, called the domain of interpretation, and function $.^I$, which assigns set $A^I \subseteq \triangle^I$ to every atomic concept A. This interpretation domain assigns a binary relation $R^I \subseteq \triangle^I \times \triangle^I$ to every atomic role R. Two concepts, C and D, are equivalent (written $C \equiv D$) if $C^I = D^I$ for all interpretations I.

Inference is generated at the terminological or factual level. Several inference engines have been proposed to reason about ABox and TBox, namely the top known ones: Pellet (Sirin et al. 2007), Racer (Haarslev and Moller 2001) et FaCT ++ (Tsarkov and Horrocks 2004).

- *Practical role of DL in indexing process*

The application of DL in the indexing process is promising because it is sufficient to consider the body of documents as a subset of the chosen discourse domain, and to represent the documents and the vocabulary controlled by concepts. Thus, each document (controlled vocabulary) will be represented in the T-Box T by its index which is a descriptive expression. *docI* is a representation of a set of documents that have the same content. The physical documents then represent the instances of *docI*. In accordance with the DL terminology, the correspondence between the controlled vocabulary and document *doc* is calculated in the subsumption hierarchy: Document *doc* is relevant to descriptor D of the controlled vocabulary if concept *docI* is subsumed by concept *cD*.

Thus, to respond to the matching task, the indexing model selects the documents whose index *docI* is subsumed by concept *cD*.

We summarize the concepts extraction phase in the Algorithm 2.

Algorithm 2: Algorithm of concepts extraction based on DL

Input: D: document; Mesh_Term: MeSH Term; m: number of words in document; n: number of words in term; i,j: counter;
Output: Bag of concepts

```
while ! end_of_corpus () do
    for i ⟵ 1 to m do
        DR_i ⟵ Document_representation(D_i)
    end
end
while ! end_of_Mesh_terms () do
    for j ⟵ 1 to n do
        MeshTR_j ⟵ Mesh_Term_representation(Mesh_Term_j)
    end
end
for i ⟵ m to l do
    for j ⟵ n to l do
        Return inference(DR_i,MeshTR_j)
    end
end
```

- *Document representation with DL*

To represent document words with DL, a statistical calculation is prepared in order to associate an importance degree to each word in the document, which will be then represented by a descriptive expression. In this part, the same weight in the previous section (WWD) is used (Eq. 1).

A document representation with a descriptive expression follows the words weighting phase: each document word (W_j) represents an indexing element (IE_j) forming a descriptive expression. The latter represents this document (D_i). Since each document is represented by an expression, the final index (FI_Doc) contains all the descriptive expressions.

The descriptive expression is defined by the following form:

$$W_j \equiv IE\_jD\_i \equiv \exists represented\_by.IE_1 \sqcap \exists represented\_by.IE_2 \sqcap \exists represented\_by.IE_3 \sqcap \ldots .. \sqcap \exists represented\_by.IE_n$$

Example

$$D_i \equiv \exists represented\_by.Skull \sqcap \exists represented\_by.Iatrogenic \sqcap \exists represented\_by.Fractures \sqcap \ldots .. \sqcap \exists represented\_by.Wounds$$

$$FI_Doc \equiv D_1 \cup D_2 \cup D_3 \cup D_4 \cup \ldots . \cup D_n$$

- *Representation of external resource terms with DL*

The MeSH thesaurus is the external resource chosen in our work. The MeSH concepts are represented by descriptive expressions in this part. MeSH is a controlled vocabulary base and a reference thesaurus in the biomedical field which includes a list of terms having hierarchical, synonymic and proximity relations. A MeSH concept is represented by a preferred term and other non-preferred, a concept forms a descriptive expression of the following form:

$$C_i \equiv PT \sqcap \exists described\_by.NT_1 \sqcap \exists described\_by.NT_2 \sqcap \exists described\_by.NT_3 \sqcap \ldots .. \sqcap \exists described\_by.NT_n$$

Example

$$\begin{aligned} &Shoulder\ Impingement\ Syndrome \\ &\equiv Shoulder\ Impingement\ Syndrome \\ &\quad \sqcap \exists described\_by.Subacromial\ Impingement\ Syndrome \\ &\quad \sqcap \exists described\_by.Shoulder\ Impingement \\ &\sqcap \exists described\_by.Syndrome,\ Shoulder\ Impingement \end{aligned}$$

The knowledge base consists of concepts *(C, PT, NT)* and roles *(described_by)*, where *C* is a MeSH concept, *PT* is a preferred term, and *NT* is a non-preferred term. This expression is a conjunction of at least one term used to identify the MeSH concept. To refine the semantic description of the concept, it is possible to contain other unprivileged terms. *Described_by* represents the role that connects these concepts and all of these expressions form the MeSH thesaurus index *(I_MeSH)*.

- *Inference*

The reasoning is conducted at the terminology level (TBox) to extract the most representative concepts from a document. In order to perform this task, we use a correspondence based on the subsumption calculation ($\sqsubseteq_S$): Concept C_1 is subsumed by concept C_2 for terminology T if and only if $C_1^I \sqsubseteq C_2^I$ for all models I of T.

The inference architecture is represented by two levels: theoretical and descriptive. The theoretical level consists of a set of documents and MeSH terms represented theoretically. Based on descriptive expressions, the descriptive level represents all the documents in a bag *(FI_Doc)* and the set of MeSH terms in another bag *(I_MeSH)*. The concepts are extracted by the correspondence between a document in *FI_Doc* and a concept in *I_MeSH*. This correspondence must check $C_i \sqsubseteq D_i$ for our knowledge base. Finally, the set of candidate concepts for a document is $\{C | C_i \sqsubseteq_S D_i\}$. For this reason, we utilize the Pellet reasoner (Sirin et al. 2007) for inference.

Concepts filtering phase

The intersection between the set of extracted candidate concepts with both VSM and DL results in a list of candidate concepts. The last step consists in regrouping all the candidate concepts into a single bag and keep only the relevant concepts that can index a document. For this purpose, the regrouped concepts are divided into two groups: the first is the

Table 1 Statistic of the test corpus

	OHSUMED	CISMeF
Total number of documents	200,000	1000
Average number of words in titles	10.4	9.1
Average number of words in summaries	125.2	103.7
Number of documents for patients	–	320
Number of documents for teaching	–	300
Number of documents of recommendation type	–	380

Main Index (MI) including concepts where all the words of its preferred term are present in the document, and the Secondary Index (SI) contains all concepts including some words of their terms (preferred and not preferred) present in the document. To add a concept C from the SI to the MI, we exploit the MeSH thesaurus architecture, so it is crucial to know whether C is related to the necessary index concepts. If concept C (from SI) belongs to the same descriptor of one of the MI concepts, it is added to the Main Index.

The first step of the filtering algorithm is a test to check if the set of the preferred term words belongs to the document. If it is the case, the concept of this term is added to the MI list. Otherwise it is added to SI list. To add a concept of the SI to the MI, we go through both groups; and if a concept of the secondary index satisfies the condition, then it will be added to the indexing list.

Ranking phase

The final phase allows classifying the concepts of the MI according to Eq. (5). Besides, the first n concepts will be selected to represent the document and then compared to manual indexing.

Experimental study and results analysis

Description of test corpora

To test our approach, a subset of the OHSUMED collection consisting of 200,000 citations for scientific articles from Pubmed is used. The OHSUMED test collection is a set of 348,566 references from MEDLINE database, which is a bibliographic database of important, peer-reviewed medical literature managed by the National Library of Medicine. The classification scheme consists of the 23 MeSH categories of cardiovascular diseases group.

A citation is composed of six fields: source (.S), title (.T), author (.A), abstract (.W), indexed concepts (.M), and publication (.P). For each selected citation, composed of a title and an abstract in English, the contents of both the title and the abstract are merged. We further utilize another corpus composed of titles and summaries of 1000 resources selected randomly from CISMeF. Three document types are indexed in CISMeF: documents for patients, recommendations, and documents intended for teaching. The statistics of the used collections are shown in Table 1. For all experiments only the first 15 concepts are kept in

the final index. In fact, the average number of concepts in manual indexes in MEDLINE is 15 (Ruch 2006). We use MeSH and SNOMED-CT as the thesaurus for evaluation .

Evaluation measures

To evaluate the proposed indexing approach, we opt for the average accuracy for each n extracted concepts. F-score combines precision and recall with an equal weight. Precision is the ratio between the number of correct concepts and the total number of extracted concepts. Recall defines the ratio between the number of correct concepts and the number of concepts that correspond to manual indexing. Concepts belonging to the manual index are considered correct.

Evaluations and results

The concepts extracted from the proposed approach are compared with those of manual indexing. The importance of combining DL and VSM are: (1) the confirmation of the relevance of the extracted concepts, (2) the use of the word weight measure by VSM in the controlled resource to improve the relevance estimation of the concepts, and (3) extracting relevant concepts that are not extracted by one of the methods.

Example We consider the following citation:

Our results suggest that ethylene oxide retention after sterilization is increased in cuprammonium cellulose plate dialyzers containing potting compound. In contrast, cuprammonium cellulose plate dialyzers without potting compound were characterized by a rapid disappearance of retained ethylene oxide after sterilization. Whether these findings explain the low incidence of SARD with cuprammonium cellulose plate dialyzers that do not contain potting material is a matter for continued study and experimentation.

In the manual indexing we find the concept "Anaphylaxis" which can represent this document. This concept is extracted by the DL method but not extracted by the VSM method. If we use the hybrid technique, this concept can represent this document. We can find other examples that a concept is extracted by the VSM method and not extracted by the DL method, and we find this representative concept in the final index using the hybrid technique.

In OHSUMED, the final index corresponds to manual indexing and is composed of a set of concepts generated by domain experts. In our experiments, a concept, found in manual and automatic indexes, is considered correct. Matching between the bags of concepts is exact; e.g., if we consider that *Esophageal Diseases* is a concept belonging to manual indexing and *Diseases* is a concept of automatic indexing, then *Diseases* is considered incorrect. The intersection between two terms of an automatic index and a manual one must be complete; i.e., the same concept exists in both indexes. In the CISMeF architecture, if two concepts C_1 and C_2, taken from the same document, are the sons of another concept C_3, the longest one between C_1 and C_2 will be considered for indexing. The obtained results for the two corpora are almost the same, and since we are mostly interested in document indexing in English, we only present the results of the OHSUMED corpus.

Table 2 F-score of terms extraction step

Stem length	>= 0	>= 3	>= 4	>= 5
F-score value	0.218	0.185	0.25	0.214

- *Term extraction*

In this part, we represent the obtained results of the first experiment, which is based on VSM. The term extraction, using the similarity measure given by the cosine function and using the BM25 measure, yields important results. The measurement parameters give more importance to the word frequency in either the document or in the term. They help to adjust the importance provided to the document size in relation to the collection/ term size and to the term set in MeSH.

Terms with a similarity value equal or greater than 0.8 are considered for indexing (Chebil et al. 2013). The number of concepts generated for each document is almost 40. The stem length fixed for a word is superior or equal to 4, giving the best result, as shown in Table 2. The values in this figure denote respectively: 0: without stemming; 3: stem >= to 3; 4: stem >= to 4; 5: stem >= to 5.

As presented in Table 2, words stemming is more important in the indexing process. Stems with a length superior or equal to 4 will be considered for next experiments. This choice is justified by the fact that: (1) the experiments give the best result, (2) the stems with a length less than 3 can be confused with acronyms or abbreviations, which allows increasing the noise and decreasing the precision value, and (3) stems with a length greater than 4 give more accuracy and more relevant terms.

- *Test and result of proposed approach*

In order to test the suggested approach performance, a comparison with other approaches in the literature was conducted. The reported approaches have been classified into controlled vocabulary [IBioDI (Boukhari and Omri 2017a), IBioDL (Boukhari and Omri 2017b) and QuickUMLS (Soldaini and Goharian 2016)], approaches based on free terms [Hasan approach (Mahedi et al. 2018)], hybrid approaches based on partial matching (MaxMatcher (Zhou et al. 2006)], and approaches based on the semantic method with exact matching [BioAnnotator (Mukherjea et al. 2004)].

The number of selected concepts representing a document is equal to 15, which is the average of keywords in manual indexing. We therefore vary the number of concepts from 1 to 15, and we note the results of precision and recall. For the f-score, we present the final results obtained for different approaches.

Figure 2 illustrates the obtained results of the precision rate as well as the difference values compared to the average accuracy. Figures 3 and 4 describe the recall rate and the f-score measure. The experiments depict the difference between the proposed approach within other approaches as well as the factors that influence the quality of results. The analysis of these results and the comparison between approaches are presented in the following section (Table 3).

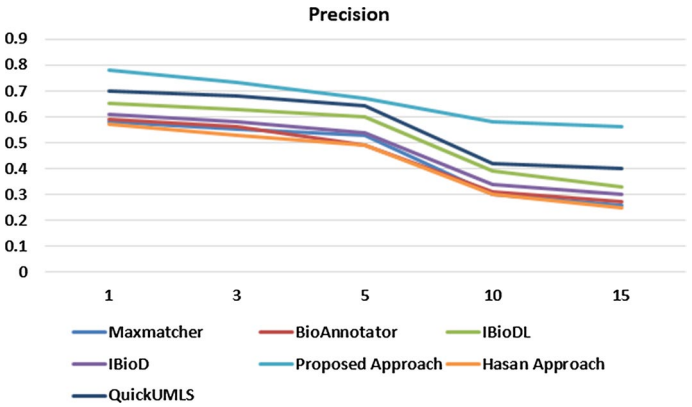

Fig. 2 Precision measurement

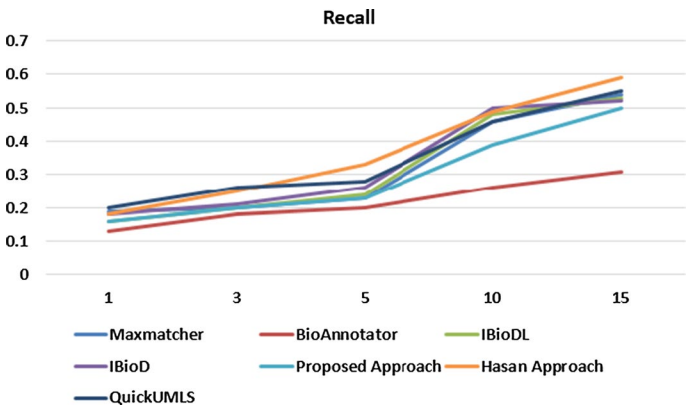

Fig. 3 Recall measurement

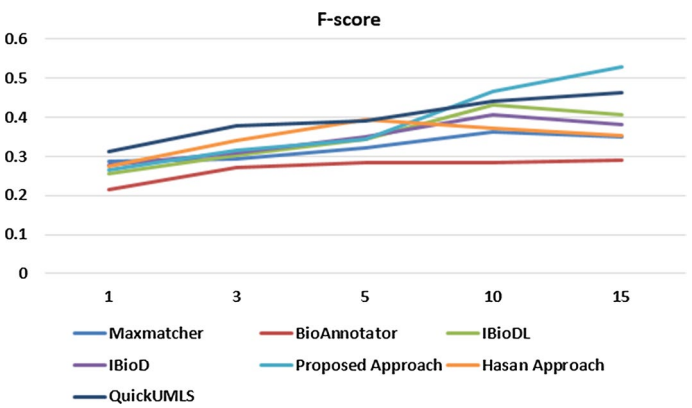

Fig. 4 F-score measurement

Table 3 Evaluation results of approaches in terms of precision, recall and F-score

Number of concepts	Approach	Precision	Recall	F-score
1	Maxmatcher	0.58	0.19	0.286
	BioAnnotator	0.59	0.13	0.213
	IBioDL	0.65	0.16	0.257
	IBioD	0.61	0.18	0.278
	Proposed approach	0.79	0.17	0.28
	Hasan approach	0.57	0.18	0.274
	QuickUMLS	0.7	0.2	0.311
3	Maxmatcher	0.55	0.2	0.293
	BioAnnotator	0.56	0.18	0.272
	IBioDL	0.63	0.2	0.304
	IBioD	0.58	0.21	0.308
	Proposed approach	0.73	0.21	0.326
	Hasan approach	0.53	0.25	0.34
	QuickUMLS	0.68	0.26	0.376
5	Maxmatcher	0.53	0.23	0.321
	BioAnnotator	0.49	0.2	0.284
	IBioDL	0.6	0.24	0.343
	IBioD	0.54	0.26	0.351
	Proposed approach	0.69	0.23	0.345
	Hasan approach	0.49	0.33	0.394
	QuickUMLS	0.64	0.28	0.39
10	Maxmatcher	0.3	0.46	0.363
	BioAnnotator	0.31	0.26	0.283
	IBioDL	0.39	0.48	0.43
	IBioD	0.34	0.5	0.405
	Proposed approach	0.61	0.38	0.468
	Hasan approach	0.3	0.49	0.372
	QuickUMLS	0.42	0.46	0.439
15	Maxmatcher	0.26	0.54	0.351
	BioAnnotator	0.27	0.31	0.289
	IBioDL	0.33	0.53	0.407
	IBioD	0.3	0.52	0.38
	Proposed approach	0.58	0.51	0.543
	Hasan approach	0.25	0.59	0.351
	QuickUMLS	0.4	0.55	0.463

- *Discussion*

By analyzing the obtained results from the experiments series developed for various approaches, we can notice the change in performances from one approach to another. The result observed from the terms extraction part demonstrates that our performances are better with a stem length equal to 4. Here, the words stemming phase plays a crucial role in the indexing process as it offers the possibility of keeping the maximum of the

relevant terms and eliminating those which are irrelevant in the filtering step. The proposed approach, compared to others, offers the best result due to DL and VSM expressiveness, which engender a better representation of documents and terms. The more meaningful and expressive representation of documents/terms, the more correct the logical deductions on the document basis and terms.

According to the results shown in the previous section, we can notice the advantages of the suggested approach in terms of important obtained value of 0.58 with 15 concepts. Moreover, partial matching between terms offers a large space to the approach to represent such a document using concepts missing in it. The document can be represented by morphological variants of words which are absent in the BioAnnotator approach. Both precision and recall measurements are not proportional. According to Fig. 3, the proposed approach gives a significant value of the recall rate, i.e., more silence (relevant concepts not extracted) in the obtained results. This allows us to deduce that there are non-extracted concepts that can represent the document. In addition, approaches based on free vocabulary do not give good results, especially for a specific field such as the biomedical domain. The syntactic and semantic analysis for the Hasan approach gives weak results because of lack of external resources.

The results illustrated in Figs. 2, 3 and 4 confirm the quality of the results found by the proposed approach. The expressiveness and the good knowledge representation in DL contribute to the relevance of the term extraction. However, the performance degradation of the MaxMatcher approach is due to the absence of a filtering step. The BioDI approach takes advantage of the *tf-idf* measure in the calculation of the similarity degree, which gives lower results compared to those of the IBioD and IBioDL approaches. The filtering step is very useful for the final classification, especially by exploiting the MeSH architecture, which makes it possible to clean extracted concepts by using its architecture. Indeed, OHSUMED generates concepts that share a few words in a document. It is the partial matching mission that we have exploited in our approaches.

The obtained results enables concluding the benefit of combining statistical and semantic methods to give more importance to all the words used in the treatment.

Conclusion

In this article, we have presented a new approach for document indexing in the biomedical field. The main contribution of this work is to combine the VSM and the description logics for concept extraction, which improves the similarity degree between a document and a given concept. The stemming process helps to group a set of words with different morphological variants into single stem, to approximately match a term as regards a document. By using the suggested vector presentation of the documents and the terms, the proposed similarity measure shows that closely associated terms have higher similarity values than others. The exploitation of the DL gives more expressivity for the knowledge representation. Besides, it facilitates combining the statistical and the semantic methods. The extracted concepts using both methods give more accuracy for the selection of concepts. The filtering step is built to overcome the non preferred concepts using the MeSH architecture. The experiments series, employing a big database with different corpora, demonstrate clearly the interest of the suggested approach compared to other approaches in the literature.

In a future work we plan to investigate new biomedical terminologies as external resources for the controlled vocabulary. Furthermore, we aim to exploit another DL family for the document representation for a better content description. Also, we are working on testing the proposed approach on other Big Data corpora.

References

Abu-Salih, B., Wongthongtham, P., & Chan, K. Y. (2018a). Twitter mining for ontology-based domain discovery incorporating machine learning. *Journal of Knowledge Management*, *22*, 949–981.

Abu-Salih, B., Wongthongtham, P., Chan, K. Y., & Zhu, D. (2018b). Credsat: Credibility ranking of users in big social data incorporating semantic analysis and temporal factor. *Journal of Information Science*, *45*, 259–280.

Ali, M., Khalid, S., & Saleemi, M. (2019). Comprehensive stemmer for morphologically rich urdu language. *The International Arab Journal of Information Technology*, *16*(1), 138–147.

Alotaibi, F. S., & Gupta, V. (2018). A cognitive inspired unsupervised language-independent text stemmer for information retrieval. *Cognitive Systems Research*, *52*, 291–300.

Aravazhi, R., & Chidambaram, M. (2018). An efficient indexing mesh term description logic using in medical subject headings. *Journal of Computer and Mathematical Sciences*, *9*(10), 1556–1567.

Aronson, A., Mork, J., Gay, C., Humphrey, S., & Rogers, W. (2004). The NLM indexing initiative's medical text indexer. *Studies in Health Technology and Informatics*, *11*(1), 268–272.

Arroyo-Fernández, I., Méndez-Cruz, C., Sierra, G., Torres-Moreno, J., & Sidorov, G. (2019). Unsupervised sentence representations as word information series: Revisiting TF-IDF. *Computer Speech and Language*, *56*, 107–129.

Baoli, H., Ling, C., & Xiaoxue, T. (2018). Knowledge based collection selection for distributed information retrieval. *Information Processing and Management*, *54*(1), 116–128.

Boukhari, K., & Omri, M. N. (2015). Said: A new stemmer algorithm to indexing unstructured document. In *The international conference on intelligent systems design and applications* (pp. 59–63).

Boukhari, K., & Omri, M. N. (2016). Raid: Robust algorithm for stemming text document. *International Journal of Computer Information Systems and Industrial Management Applications*, *8*(1), 235–246.

Boukhari, K., & Omri, M. N. (2017a). Information retrieval approach based on indexing text documents: Application to biomedical domain. In *The 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD)* (pp. 2213–2220).

Boukhari, K., & Omri, M. N. (2017b). Information retrieval based on description logic: Application to biomedical documents. In *International conference on high performance computing and simulation (HPCS)* (pp. 846–853).

Bracewell, D., Ren, F., & Kuroiwa, S. (2005). Multilingual single document keyword extraction for information retrieval. In *Proceedings of natural language processing and knowledge engineering (NLP-KE)* (pp. 517–522).

Chebil, W., Soualmia, L. F., & Darmoni, S. J. (2013). Biodi: A new approach to improve biomedical documents indexing. In *Database and expert systems applications* (pp. 78–87).

Dahak, F., Boughanem, M., & Ballaa, A. (2017). A probabilistic model to exploit user expectations in xml information retrieval. *Information Processing and Management*, *53*(1), 87–105.

Dinh, D., & Tamine, L. (2011). Combining global and local semantic contexts for improving biomedical information retrieval. In *European conference on information retrieval research* (pp. 375–386).

Ferjani, F., Elloumi, S., Jaoua, A., Sahar Ahmad Ismail, S. B. Y., & Ravan, S. (2012). Formal context coverage based on isolated labels: An efficient solution for text feature extraction. *Information Sciences-Informatics and Computer Science, Intelligent Systems, Applications: An International Journal*, *188*(1), 198–214.

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.

Fkih, F., & Omri, M. N. (2016a). Hybridization of an index based on concept lattice with a terminology extraction model for semantic information retrieval guided by wordnet. In *International conference on hybrid intelligent systems* (pp. 144–152).

Fkih, F., & Omri, M. N. (2016b). Irafca: An o(n) information retrieval algorithm based on formal concept analysis. *Knowledge and Information Systems*, *48*(2), 465–491.

García, M. A. M., Rodríguez, R. P., & Rifón, L. A. (2018). Leveraging wikipedia knowledge to classify multilingual biomedical documents. *Artificial Intelligence in Medicine*, *88*(1), 37–57.
Haarslev, V., & Moller, R. (2001). Description of the racer system and its applications. In *The international workshop on description logics* (pp. 132–141).
Hao, S., Shi, C., Niu, Z., & Cao, L. (2018). Concept coupling learning for improving concept lattice-based document retrieval. *Engineering Applications of Artificial Intelligence*, *69*(1), 56–75.
Happe, A., Pouliquen, B., Burgun, A., Cuggia, M., & Beux, P. L. (2003). Automatic concept extraction from spoken medical reports. *International Journal of Medical Informatics*, *70*(2–3), 255–263.
Jiménez, S., Cucerzan, S., González, F. A., Gelbukh, A. F., & Dueñas, G. (2018). BM25-CTF: Improving TF and IDF factors in BM25 by using collection term frequencies. *Journal of Intelligent and Fuzzy Systems*, *34*(5), 2887–2899.
Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N. F., Musen, M. A., et al. (2011). Ncbo resource index: Ontology-based search and mining of biomedical resources. *Journal of Web Semantics*, *9*(3), 316–324.
Jutinico, C. J. M., Montenegro-Marin, C. E., Burgos, D., & Crespo, R. G. (2019). Natural language interface model for the evaluation of ergonomic routines in occupational health (ilena). *Journal of Ambient Intelligence and Humanized Computing*, *10*(4), 1611–1619.
Karaa, W. B. A. (2013). A new stemmer to improve information retrieval. *International Journal of Network Security and Its Applications (IJNSA)*, *5*(4), 143–154.
Liu, Y. H., & Wacholderc, N. (2017). Evaluating the impact of mesh (medical subject headings) terms on different types of searchers. *Information Processing and Management*, *53*(4), 851–870.
Lv, X., Guan, Y., & Deng, B. (2014). Transfer learning based clinical concept extraction on data from multiple sources. *Journal of Biomedical Informatics*, *52*(3), 55–64.
Mahedi, H. H., Sanyal, F., & Chaki, D. (2018) A novel approach to extract important keywords from documents applying latent semantic analysis. In *International conference on knowledge and smart technology (KST)* (pp. 1–6).
Matsuo, Y., & Ishizuka, M. (2003). Keyword extraction from a single document using word co-occurrence statistical information. In *Proceedings of the sixteenth international Florida artificial intelligence research society conference* (pp. 392–396).
Mukherjea, S., Gaurav Chanda, L. V. S., Sankararaman, S., Kothari, R., Batra, V. S., Bhardwaj, D. N., et al. (2004). Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. *IBM Journal of Research and Development*, *48*(5–6), 693–702.
Naouar, F., Hlaoua, L., & Omri, M. N. (2016). Collaborative information retrieval model based on fuzzy confidence network. *Journal of Intelligent and Fuzzy Systems*, *30*(4), 2119–2129.
Naouar, F., Hlaoua, L., & Omri, M. N. (2017). Information retrieval model using uncertain confidence's network. *International Journal of Information Retrieval Research*, *7*(2), 34–50.
Nicolas, F., Ranwez, S., Montmain, J. M., & Ranwez, V. (2015). Usi: A fast and accurate approach for conceptual document annotation. *BMC Bioinformatics*, *16*(1), 1–10.
Radhouani, S., & Falquet, G. (2008). Description logics-based modelling for precise information retrieval. In *International workshop on description logics* (pp. 1–11).
Radhouani, S., Falquet, G., & Chevallet, J. P. (2008). Description logic to model a domain specific information retrieval system. In *International conference on database and expert systems applications* (pp. 142–149).
Ru, C., Tang, J., Li, S., Xie, S., & Wang, T. (2018). Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Information Processing and Management*, *54*(4), 593–608.
Ruch, P. (2006). Automatic assignment of biomedical categories: Toward a generic approach. *Bioinformatics Journal*, *6*(22), 58–64.
Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, *5*(2), 51–53.
Sohn, S., Kim, W., Comeau, D. C., & Wilbur, W. J. (2008). Optimal training sets for Bayesian prediction of mesh®assignment. *Journal of the American Medical Informatics Association*, *15*(4), 546–553.
Soldaini, L., & Goharian, N. (2016). Quickumls: A fast, unsupervised approach for medical concept extraction. In *Medical information retrieval (MedIR) workshop* (pp. 1–4).
Song, M. (2015). Exploring concept graphs for biomedical literature mining. In *International conference on big data and smart computing* (pp. 103–110).
Sun, P., Wang, L., & Xia, Q. (2017). The keyword extraction of Chinese medical web page based on WF-TF-IDF algorithm. In (pp. 193–198).
Tsarkov, D., & Horrocks, I. (2004). Efficient reasoning with range and domain constraints. *Description Logic Workshop DL*, *2004*, 41–50.

Warren, P., Mulholland, P., Collins, T. D., & Motta, E. (2019). Improving comprehension of knowledge representation languages: A case study with description logics. *International Journal of Human–Computer Studies*, *122*, 145–167.

Wongthongtham, P., & Salih, B. A. (2018). Ontology-based approach for identifying the credibility domain in social big data. *Journal of Organizational Computing and Electronic Commerce*, *28*, 354–377.

You, W., Fontaine, D., & Barthès, J. P. (2013). An automatic keyphrase extraction system for scientific documents. *Knowledge and Information Systems*, *34*(3), 691–724.

Yuan, L. (2018). Supporting relevance feedback with concept learning for semantic information retrieval in large owl knowledge base. In: *Knowledge management and acquisition for intelligent systems* (pp. 61–75).

Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., & Wang, B. (2008). Automatic keyword extraction from documents using conditional random fields. *Journal of Computational Information Systems*, *4*(3), 1169–1180.

Zhou, X., Zhang, X., & Hu, X. (2006). Maxmatcher: Biological concept extraction using approximate dictionary lookup. In *Pacific RIM international conference on artificial intelligence* (pp. 1145–1149).