Fuzzy Ontology-Based Possibilistic Approach for Document Indexing Using Semantic Concept Relations

Kabil Boukhari(✉) and Mohamed Nazih Omri(✉)

MARS Research Laboratory, LR17ES05, University of Sousse, Sousse, Tunisia
Mohamednazih.omri@eniso.u-sousse.tn

Abstract. To overcome the weaknesses of current information retrieval system and to utilize the strengths knowledge extraction a novel approach based on fuzzy ontology and possibility theory is proposed for indexing documents. Possibility theory allows to model and quantify the relevance of a document given a controlled vocabulary through two measures: necessity and possibility. Fuzzy ontology is used to improve the indexing process on information retrieval by means of external resource. Besides, the fuzzy approach has been proposed in order to make the availability of terms representations in a document more flexible. It allows a formal representation of a knowledge domain in the form of a hierarchical terminology provided with semantic relationships. As a result, the proposed approach has been made on different corpora, had better performance than other indexing approaches and it prove important results.

Keywords: Knowledge extraction · Possibility theory · Data analysis · Fuzzy ontology · Information retrieval · Document indexing

1 Introduction

The amount of data and textual information on the web is constantly increasing and the manual processing of information is too expensive and time consuming, especially when it comes to a specific domain such as medical domain. Thus, any automatic indexing process is at the heart of documentary research. Classic information search models are based only on words found in the document. However, these which are not systematically relevant share a set of words with the query, in the form of a set of words, which can thus be returned to browsers. The success of a documentary retrieval system depends mainly on three tasks: (i) document representation (indexing), (ii) query representation and (iii) query/document correspondence. Document indexing tends to select and extract words that match the content of the document, making it easier to find information. Much research has been proposed in this context to improve information retrieval models.

Several reasons that motivated us to opt for the integration of the theory of possibilities and fuzzy ontology. Possibility Theory is an efficient and robust

C. Strauss et al. (Eds.): DEXA 2021, LNCS 12924, pp. 264–269, 2021.
https://doi.org/10.1007/978-3-030-86475-0_26

method for uncertain processing tasks such as matching documents and external resources in a reliable and efficient manner. This theory is known as a robust method for dealing with uncertain tasks. As for the use of fuzzy ontology, this makes it possible to increase the performance of information search since they offer semantic links between the different concepts handled.

The rest of this paper is organized as follows: Sect. 2 presents the related work. Section 3, details the description of the proposed approach. In the next Sect. 4 we describe the experimental evaluations and we analyze the obtained results. Section 5 summarizes different steps of the proposed approach and gives the main prospects for this work.

2 Related Work

Several approaches have been proposed for knowledge extraction and documents indexing [4,5]. We can classify these approaches into two families: (i) Approaches based on free terms extraction and (ii) Approaches based on the controlled vocabulary.

Approaches based on free terms extraction use only terms existing in the document [12]. In [8] an approach has been proposed to extract complex terms from texts, they exploited statistical and linguistic methods. In [15], authors presented an approach for extracting keywords based on CRF (Conditional Random Fields). The work in [6] used Natural Language Processing (NLP) to extract keywords from a document. In [14] the authors proposed an approach to extract automatically keywords from scientific documents. It generates the candidate expressions based on a word expansion algorithm and introduces document frequency feature. The work in [10] has proposed a keyword extraction approach which only uses the document to index without using the whole collection.

Approaches based on controlled vocabulary use external resources and exploits specific terminology for indexing document. The author of [13] presented the approach "QuickUMLS" which uses a dictionary to extract medical concept based on approximate matching. In [9], authors used more than 200 ontologies to facilitate the space of medical discoveries by providing to scientists unified view of this diverse information. In [7], the authors proposed an indexing approach for biomedical documents by using the MeSH thesaurus, the basic idea is to use the VSM method for extracting concepts, and combines a static and a semantic method to estimate the concept's relevance for a given document.

3 Fuzzy Ontology-Based Possibilistic Proposed Approach

The proposed approach consists of 3 stages. (i) Pre-processing, (ii) concept extraction and (iii) filtering task.

Pre-processing. The pre-processing step consists of 5 tasks: (i) divide the document into sentences, (ii) remove punctuation, (iii) remove stop words, (iv) desuffix words and (v) divide each sentence into words. The four tasks (ii), (iii),

(iv) and (v) are also applied to each term of the MeSH thesaurus. We used the RAID algorithm [3] an improved version of SAID [1], for stemming.

Concepts Extraction. In this part, we present the concepts extraction phase which is based on the two methods: Possibility theory and Fuzzy ontology.

Possibilistic Concepts Extraction. In this part, the concepts extraction is done first to extract the terms that compose a concept. For this, we use possibility theory that allows us to calculate a score for each term. The extracted candidate terms are those having a non-zero score. Besides, the corresponding concepts are assigned to the terms and the concept score corresponds to the score of its term. If a concept matches more than one term among the candidate terms, we associate it the maximum score. The term having the same score as its concept is denoted representative term.

The conditional possibility allows to calculate two measures: The document possibility giving a term (see Eq. 1) and The document necessity giving a term (see Eq. 2).

Equations 1 and 2 are calculated between the document to be indexed D_i and each term T_k. Terms with non-zero values are ranked according to the score Sc (Eq. 1). The score Sc in term k and in a document$i$ is the sum of the possibility and necessity values. The addition of the necessity and possibility measures has already been adapted in other approaches for the relevance calculation. Indeed, we have exploited this method to disambiguate words and the measures combination gave an interesting result.

$$\begin{aligned} Sc(T_k, D_i) &= \prod(D_i|T_k) + N(D_i|T_k) \\ \prod(D_i|T_k) &= \frac{\prod(D_i \wedge T_k)}{\prod(T_k)} \end{aligned} \tag{1}$$

$$N(D_i|T_k) = 1 - \prod(\overline{d_i}|T_k) \tag{2}$$

We represent word in the document in two cases: (i) More a word is frequent in the document, more it is likely to be a representative for this document. (ii) More a word is frequent in the document and it is less frequent in other documents of the collection, more it is necessarily to be a representative for this document.

In this step, we calculate the concept score from the terms that it compose. This is the higher value of its terms scores, the term having the maximum score is considered as representative term of document.

$$Score(C) = \max_{T_k \in (C)} (Sc(T_k)) \tag{3}$$

$T(C)$: Set of Concept Terms

A representative term take the same values of the concept possibility and necessity.

Fuzzy Concepts Extraction. The creation of the fuzzy ontology goes through four stages: (i) Corpus/External vocabulary pre-treatment, (ii) Concept identification, (iii) Fuzzy membership assignment and fuzzy ontology creation and (iv) Concept extraction based on fuzzy ontology.

This phase of "Corpus/External vocabulary pre-treatment" is described in the section "Pre-processing".

To identify the relationships among the concepts we have used the MeSH thesaurus. In this work we have fuzzified the semantic relationships (Synonyms, described by, preferred term, non-preferred term, preferred concept, non-preferred).

The relationship between two concepts is assigned to the given relationship in the range of [0–1]. In fuzzy ontology, membership scores are assigned to the relationships to show its robustness. Here we defined the weights to the type of semantic relationship between the concepts.

All relationships and scores are assigned based on the MeSH thesaurus. For each concept C_1 having relationship R with concept C_2 and Fuzzy Membership, we create two nodes C_1 and C_2 by adding the semantic relation on the link between the nodes and the weight. Finally, now our fuzzy ontology has been created.

In the last part we use the fuzzy ontology to find the most associated concept, those having the highest membership. We give the document word to find the most related concept for the proposed word using the fuzzy ontology. In this work, we select the top semantically related concepts from the ontology, given a document words, those having the highest membership in between [0,1].

Filtering and Final Ranking. To refine this approach, we divided this bag of the candidate concepts into two others: (i) Principal concepts bag: contains all concepts in common and those which all their words are present in the document. (ii) Secondary concepts bag: here we find concepts that some words of their term are present in the document. To build the final list of concepts, we start by filtering the two lists by adding concepts from the secondary list to the main list by exploiting the Unified Medical Language System (UMLS) medical terminology and the MeSH thesaurus architecture.

4 Analysis of Experimental Results and Discussion

To evaluate the proposed approach, we used a subset of 150,000 documents of the OHSUMED 88 collection relating to scientific articles from Pubmed [1]. Moreover, we used three evaluation measures (i) the Precision that represents the ratio between the Number of Correct Concepts and the total Number of Extracted Concepts, (ii) the Recall defined as the ratio between the Number of Correct Concepts and the Number of concepts that correspond to Manual Indexing and (iii) the F-score that combines precision and recall with an equal weight concepts.

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

For the evaluation, the approaches are classified according to their families: those based on controlled language using controlled vocabulary [2], those based on partial correspondence (MaxMatcher [16]/QuickUMLS [13]) and others based on semantic algorithm with exact matching (BioAnnotator [11]). By analyzing

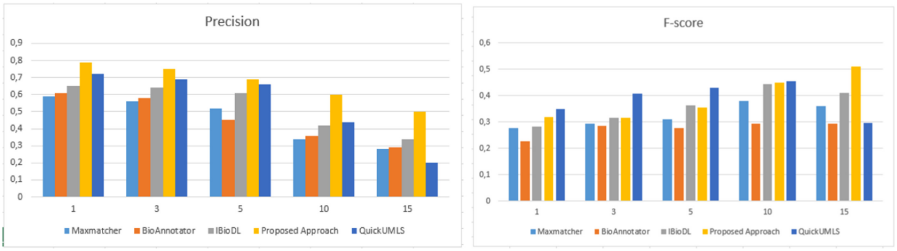

Fig. 1. Precision and F-score value for different approaches

Fig. 1, we notice that the proposed approach gives high performance and good results comparing to other approaches. The precision value shows the robustness of our approach in different ranks. These results highlight the interest of using the fuzzy logic and the possibility theory. A fuzzy membership is assigned for all type of semantic relationships to deal with the vocabulary mismatch problem among the concepts. In addition, the possibility theory gives more relevance in the concept extraction phase. Indeed, it is modeled by the necessity degree which contributes to improve the both extraction and classification of the relevant concepts. The recall value of our approach is higher than the recall of MaxMatcher approach, although the base keeps all the concepts partially matched to the document. This result is due to the filtering step applied to the proposed approach which is not the case for MaxMatcher approach.

The filtering step represents a robust solution to minimize incorrect concepts generated by partial matching. Also, the stemming process is applied to the proposed approach only on words with the stem length greater than 4.

5 Conclusion and Prospects

In this paper, we proposed a new document indexing approach by exploiting fuzzy ontology and possibility theory. The obtained results and the evaluation test show clearly the importance of our approach, which give more representative concepts compared to other approaches especially with the use of the fuzzy logic and the possibilistic theory.

The future works can be articulated around two new directions. The first is to conduct a more in-depth comparative study. In a second direction, we plan to exploit new source of controlled vocabulary as external terminologies. Moreover, we are working to evaluate the proposed approach on Big Data corpora.

References

1. Boukhari, K., Omri, M.N.: SAID: a new stemmer algorithm to indexing unstructured document. In: The International Conference on Intelligent Systems Design and Applications, pp. 59–63 (2015)
2. Boukhari, K., Omri, M.N.: Information retrieval based on description logic: application to biomedical documents. In: Conference: International Conference on High Performance Computing and Simulation (HPCS 2017), vol. 15, pp. 1–8 (2017)
3. Boukhari, K., Omri, M.N.: RAID: robust algorithm for stemming text document. Int. J. Comput. Inf. Syst. Ind. Manage. Appl. **8**, 235–246 (2016)
4. Boukhari, K., Omri, M.N.: Approximate matching-based unsupervised document indexing approach: application to biomedical domain. Scientometrics **124**(2), 903–924 (2020). https://doi.org/10.1007/s11192-020-03474-w
5. Boukhari, K., Omri, M.N.: DL-VSM based document indexing approach for information retrieval. J. Ambient Intell. Human. Comput. 1–25 (2020)
6. Bracewell, D., Ren, F., Kuroiwa, S.: Multilingual single document keyword extraction for information retrieval. In: Proceedings of Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 517–522 (2005)
7. Chebil, W., Soualmia, L.F., Omri, M.N., Darmoni, S.J.: Biomedical concepts extraction based on possibilistic network and vector space model, pp. 227–231 (2015)
8. Fkih, F., Omri, M.N.: Complex terminology extraction model from unstructured web text based linguistic and statistical knowledge. Int. J. Inf. Retrieval Res. **2**(3), 1–18 (2012)
9. Jonquet, C., et al.: NCBO resource index: ontology-based search and mining of biomedical resources. Web Semant. **9**(3), 316–324 (2011)
10. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools **13**, 1–13 (2004)
11. Mukherjea, S., et al.: Enhancing a biomedical information extraction system with dictionary mining and context disambiguation. IBM J. Res. Dev. **48**(5–6), 693–702 (2004)
12. Omri, M.N., Chenaina, T.: Uncertain and approximative knowledge representation to reasoning on classification with a fuzzy networks based system. In: IEEE International Fuzzy Systems Conference, pp. 1632–1637 (1999)
13. Soldaini, L., Goharian, N.: QuickUMLS: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, SIGIR, pp. 1–4 (2016)
14. You, W., Fontaine, D., Barthès, J.P.: An automatic keyphrase extraction system for scientific documents. Knowl. Inf. Syst. **34**(3), 691–724 (2013)
15. Zhang, C., Wang, H., Liu, Y., Wu, D., Liao, Y., Wang, B.: Automatic keyword extraction from documents using conditional random fields. J. Comput. Inf. Syst. **4**(3), 1169–1180 (2008)
16. Zhou, X., Zhang, X., Hu, X.: MaxMatcher: biological concept extraction using approximate dictionary lookup*. In: Pacific Rim International Conference on Artificial Intelligence, pp. 1145–1149 (2006)