Arabic Information Retrieval Using Semantic Analysis of Documents

Mohammad Khaled A. Al-Maghasbeh-[1], Mohd Pouzi Bin Hamzah-[2]

Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia [1,2]

Abstract

Arabic language is one of most enrich Semitic languages due to it has many concepts which related together in different semantic relationships. That makes Arabic information retrieval face more challenge to access the information needs. In addition, Arabic language more efficient of retrieval systems to be able of understanding, analysing texts, and extracting semantic relationships between concepts. This paper aims to introduce a method of improving of the information retrieval in Arabic documents. The proposed method is a semantic analysis of Arabic texts which helps to access the target documents that spread over the web. This approach used the Arabic WordNet to analyse both query, and documents in standard Arabic test collection to facilitate retrieval the related documents. The core goal of this approach is an enhancement of the evaluation measurements for Arabic information retrieval systems. Furthermore, this study used an other traditional method called words-matching to compare the results of the proposed method in selected queries from standard test collection. The experimental results of 5-queries show the degree of the performance of the proposed method. The proposed approach showed an efficient results of information retrieval measurement through compare recall, and precision with other traditional method that has been applied in this same test collection. The mean average was about our algorithms, keyword searching, and query-expansion were 0.826364, and 0.576666457, respectively. That indicates a slightly improvement of the performance of proposed semantic analysis approach in the information retrieval in Arabic documents.

Keywords

Information retrieval, Arabic information retrieval, Arabic WordNet, semantic analysis, Arabic semantic information retrieval.

1. Introduction

Information is concerned with how to access the relevance document or information which spread over the web. Information retrieval has a problem is how to retrieve documents in several languages due to the current used methods are doing by matching between the query with relevant document. This operation represents a more challenge for researcher and the search engines because that there are more documents which haven't the same words in the specific query not retrieved [1]. Other challenge is faced the retrieval system is a multilingual retrieval. In that case, the both query, and the document in test collection have been translated into common language to match between them. As result to that, to reduce and fix this problem, it has been designed a method to retrieve the document in Multilanguage via use multilingual ontology technique which it's described as there's one an ontology related with several dictionaries; each dictionary belongs to certain language and contains a special concept about the ontology. The ontologies related with each other through relations of shared conceptualization. The results were conducted using the multilingual ontology on some documents, whereas it has been retrieved documents in different languages [2].

This paper is organized as follows. Section 2 briefly describes the related works in the area of Arabic information retrieval. Section 3 provides the Arabic test collection used in this paper. Section 4 shows the Arabic information retrieval limitations. Section 5 is about proposed model. Section 6 provides briefly explanation, and discussion of the results. In section 7 summarizes of the work.

2. Related Works

Abderrahim, developed an approach to enhance information retrieval in Arabic documents. His study was depends on semantic indexing using Arabic WordNet to enhance of extracting the best concept those having a single senses in both query, and documents. He used a word sense disambiguation (WSD) technique to assign the appropriate sense of the text [3]. Menai, applied his study to assign word sense disambiguation using Evolutionary algorithms (EAs) such as Genetic (GA), and Metric algorithms (MA) to enhance the performance of NLP applications [4].

Uddin.et al., in their study proposed a model to support the information retrieval performance through using enhancing the semantic similarity among tags. They used a WordNet to extract the semantic relation between sysnset using an enriched VSM [5]. Al-Kabi. et al., proposed a novel a light Arabic stemmer to generate the root forms of Arabic words. Their proposed approach was based on two well-known Arabic stemmer know as heavy, and light root stemmer that created by Khoja, and Ghwanmeh respectively [6].

Manuscript received May 5, 2018
Manuscript revised May 20, 2018

Harrag, et al, in their study showed the Arabic text and the traditional methods which used to classify and propose the new model that depends on neural networks (NN) that has little studies that is applied in this field. In this study applied the three hidden layer of neural networks (NN) combination with singular value decomposition (SVD) to extract the feature and then minimize the size of data to facilitate a classification it into some groups based on some features, whereas in the previous studies depended on the feature vector property by took some document criteria's such as title, special word, etc, and then represented as vector. After that, it is dealt with vectors to compute weigh, distance, and relative rate [7].

3. Arabic Test Collection

The Arabic corpus is a sample of Arabic textual document that taken from Arabic test collection named as EveTAR. It was built in 2016. The original test collection includes a documents collection, as topics to describe information needs. So more, it also has some relevance judgments belongs the topics relevance's, and 66-standard queries [8].

4. Arabic Information Retrieval Limitations

Arabic language is one of the most common languages, whereas it has more than 420 million speakers over the world. It written from right to left with no upper case like English language. Arabic language is processed through different methods, and using several textual features due it represents a rich environment of morphology, concepts, and an ontology. There are some challenges and limitations in current web such as: The understandability that represent a problem with machine in searching, (eg. when the user search about something, the result may be different due to it isn't have semantic relations). So, it led to appear the semantic web idea to solve the ambiguity of results retrieval [9].

The search process around some documents in web faces challenges, whereas the searched document may be written in different language of query, so, the semantic web search engine becomes using the cross- multiannual techniques to search in different languages. There several studies showed cross- multilingual method for search correspond document, and results for queries through linguistic analysis of language. This process involves different phases part-of-speech through morphological analysis, and normalization of to facilitate retrieve the documents in more than one language [10]. The semantic web search engine is the common information retrieval systems. It depends on the different natural languages (NL) with a lot of concepts, and ontology [11]. The texts contain a lot of features such as ambiguity that represents a problem for the readers' due to the hidden text contents such as events, things specification. So the knowledge in these texts must be represented to facilitate the understanding by the users. Previous study showed the Naïve Semantic as one of the methods that used to represent the knowledge [12]. The more importance of these limitations is the capacity and effectiveness of Arabic information retrieval and their ability to face the user needs. So, in this search a new approach has been processed to decrease some of the challenge and limitations.

5. Proposed System

The proposed system is composed of several phases, each one made up of one or several resources. The proposed system architecture is shown below:

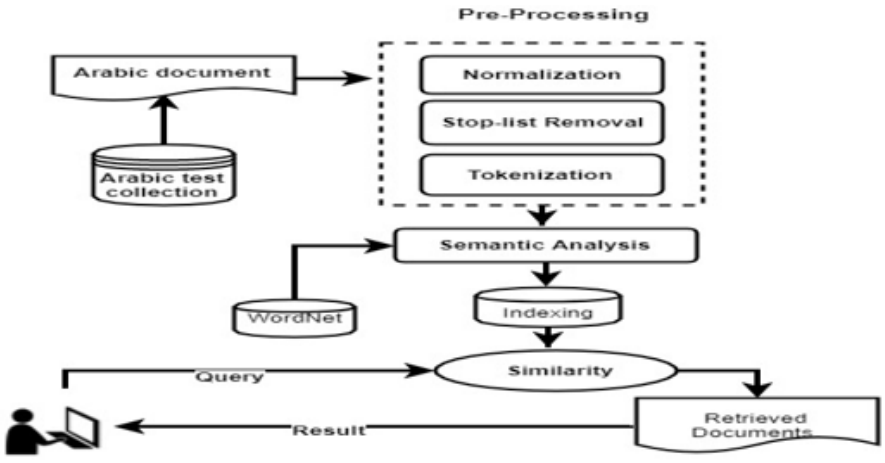

Fig. 1 proposed system architecture

The proposed approach that mentioned above in Figure (1) is contained with three main steps. It stats from pre-processing, extract the concepts, semantic similarity. These main steps are applied in both the documents, and query. The second step is term extraction from both query, and documents. The final step is semantic similarity among of them.

5.1. Preprocessing Phase

Pr-processing is the main phase in common natural languages processing (NLP) applications like information retrieval. The core goal of pr-processing phase is to make the text clearer through remove the insignificant words, stop words, normalize Arabic words, or characters.
The first sub-step is text normalization. Normalization task is applied to transfer the inconsistence text to be more consistency. In the Arabic language was used normalization to remove the diacritics marks, and normalize the other specific characters [13]. The second one is text tokenization. It is a process to split the plain text into tokens to remove the noise from the text [14].

5.2 Arabic WordNet

Arabic WordNet is a lexical database of Arabic words, or concepts with semantic relations between them. This WordNet comprises many relationship types between different concepts such as hyponymy, synonymy, domain, and other. These relations express the degree of how similar of meaning among concepts [15].

5.3 Similarity Calculation

Semantic similarity is a method to measure the similarity degree between sentences. In other words; Semantic similarity (SS) is a process to measure the distance between two documents, using extract the semantic distance among terms or concepts nodes in these documents. SS is "is-a" relationship that helps to extract the relations between sentence, and documents. In addition, it just isn't computed in documents level, whereas can be applied in different levels of sentences, and words too. In semantic similarity, the WordNet, ontologies, and other thesauruses have been used to obtain the semantic similarity among documents [15].

6. Results discussion

The proposed approach is applied on a sample of 30-documents, and 5- queries from EveTAR test collection. In addition, in the same time, we are applied the traditional words matching in the same sample of corpus. The table shown below contains the queries with their related top documents that retrieved.

Table 1: Retrieved Dcoument Details by Two Different Methods

Query	***Number of retrieved documented by Semantic analysis method.***	***Number of retrieved documented by words matching method.***
Query #1	20- documents	17- documents
Query #2	23- documents	20- documents
Query #3	27-documents	21-documents
Query #4	19-documents	14-documents
Query #5	21-documents	13-documents

The figures shown below are the obtained results of information retrieval evaluation measurements of the proposed system.

Recall	precision of words matching
0	1
0.1	0.94332217
0.2	0.832810075
0.3	0.75329798
0.4	0.673785885
0.5	0.591427379
0.6	0.471761696
0.7	0.398249601
0.8	0.305737507
0.9	0.216225412
1	0.156713317

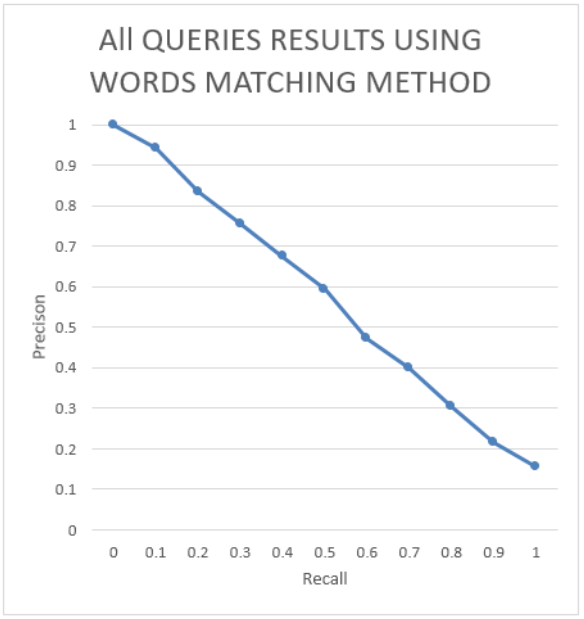

Fig. 2 The Results using Words-Matching Method

Figure (1) is shown the obtained results from applying the words- matching method. The experimental results were about 5- queries The mean average was about 0.576666457.

In contrast, the same sample is conducted by the proposed method, where is the results were as shown in the following figure.

Recall	Average precision
0.00	1.00
0.10	1.00
0.20	1.00
0.30	0.94
0.40	0.90
0.50	0.88
0.60	0.86
0.70	0.74
0.80	0.65
0.90	0.58
1.00	0.54

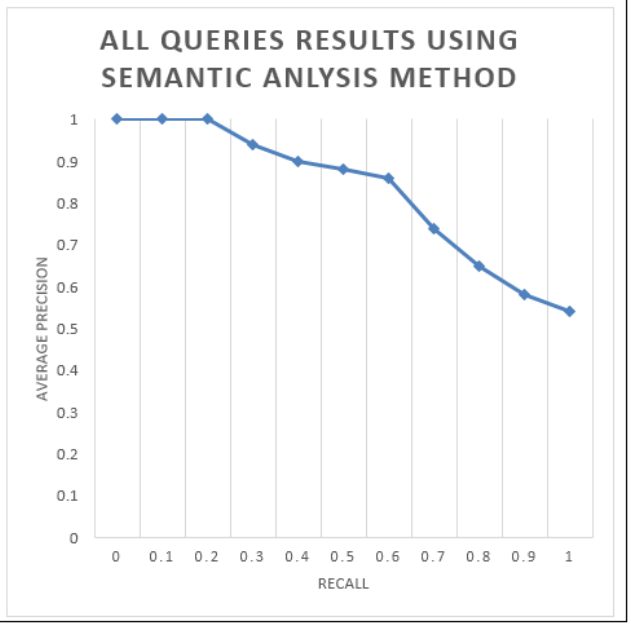

Fig. 3 The Results using Semantic Analysis Method

The experimental results of 5- queries show the degree of the performance of the proposed method. The proposed approach showed a good results of information retrieval measurement through compare recall, and precision with other studies that have been applied in this area. The mean average was about 0.826364. This indicates a good

improvement of the performance of the information retrieval in Arabic documents. In addition to that, the semantic analysis of document shows a significant enhancement than other rational models that conducted in Arabic information retrieval.

Comparison of The Results

The two figures shown above of the obtained results of information retrieval evaluation measurements of the proposed system compared with words-matching method:

Recall	precision of words matching
0	1
0.1	0.94332217
0.2	0.832810075
0.3	0.75329798
0.4	0.673785885
0.5	0.591427379
0.6	0.471761696
0.7	0.398249601
0.8	0.305737507
0.9	0.216225412
1	0.156713317

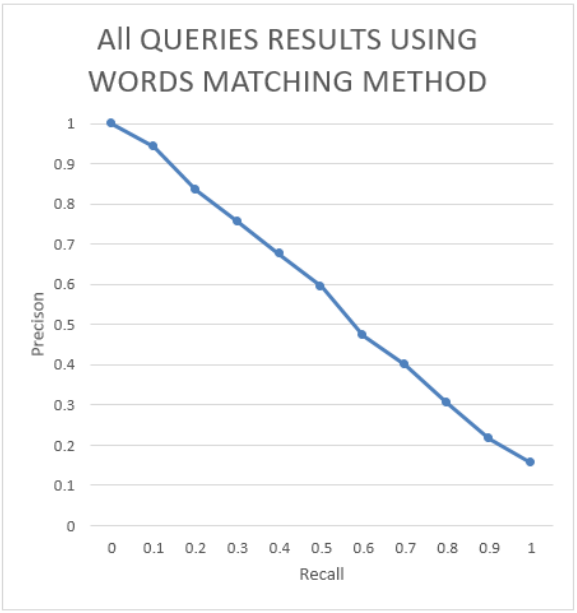

Fig. 4 Comparison between semantic analysis, and words- matching methods

7. Conclusion

Information retrieval still faces many challenges due to the nature of language. So, to solve these challenges, should be enhance the information retrieval systems accuracy, and recall measurement. One of these suggested solution to improve the efficiency of Arabic information retrieval system is a semantic analysis. A semantic analysis helps information retrieval systems to access the target information domain. This paper showed the semantic analysis to improve the information retrieval from different resources. The experimental results of 5- queries showed the degree of the performance of the proposed method. The proposed approach showed an efficient results of information retrieval measurement through compare recall, and precision with words-matching method that has been applied in this same test collection.

Acknowledgment

First and foremost, thanks to God for the completion of this study. We would like to thank Universiti Malaysia Terengganu (UMT) for providing us with good environment and facilities to complete this work.

References

[1] Hasan, K. S., & Ng, V. (2014). Automatic keyphrase extraction: A survey of the state of the art. Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

[2] Falquet, G., Guyot, J., & Radhouani, S. (2006). Ontology-based multilingual information retrieval.

[3] Abderrahim, M. A., Abderrahim, M. E. A., & Chikh, M. A. (2013). Using Arabic wordnet for semantic indexation in information retrieval system. arXiv preprint arXiv:1306.2499.

[4] Menai, M. E. B. (2014). Word sense disambiguation using evolutionary algorithms–Application to Arabic language. Computers in Human Behavior, 41, 92-103.

[5] Uddin, M. N., Duong, T. H., Nguyen, N. T., Qi, X.-M., & Jo, G. S. (2013). Semantic similarity measures for enhancing information retrieval in folksonomies. Expert Systems with Applications, 40(5), 1645-1653.

[6] Al-Kabi, M. N., Kazakzeh, S. A., Ata, B. M. A., Al-Rababah, S. A., & Alsmadi, I. M. (2015). A novel root based Arabic stemmer. Journal of King Saud University-Computer and Information Sciences, 27(2), 94-103.

[7] Harrag, F., & Al-Qawasmah, E. (2010). Improving Arabic Text Categorization Using Neural Network with SVD. JDIM, 8(4), 233-239.
[8] Almerekhi, H., Hasanain, M., & Elsayed, T. (2016). Evetar: A new test collection for event detection in arabic tweets. Paper presented at the Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval.
[9] Al-Khalifa, H. S. (2014). Introduction to the special issue on Arabic NLP: Current state and future challenges: Elsevier.
[10] Semmar, N., Laib, M., & Fluhr, C. (2006). A Deep Linguistic Analysis for Cross-language Information Retrieval. Paper presented at the Proceedings of the International Conference on Language Resources and Evaluation, LREC.
[11] Pulido, J., Ruiz, M., Herrera, R., Cabello, E., Legrand, S., & Elliman, D. (2006). Ontology languages for the semantic web: A never completely updated review. Knowledge-Based Systems, 19(7), 489-497.
[12] Dahlgren, K., McDowell, J., & Stabler, E. P. (1989). Knowledge representation for commonsense reasoning with text. Computational linguistics, 15(3), 149-170.
[13] Habash, N. Y. (2010). Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1), 1-187.
[14] Attia, M. A. (2007). Arabic tokenization system. Paper presented at the Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources.
[15] SRAVANTHI, P., & Srinivasu, B. (2017). Semantic Similarity between Sentences. International Research Journal of Engineering and Technology (IRJET), 4(1), 156-161.