American Journal of Applied Sciences 9 (9): 1373-1377, 2012
ISSN 1546-9239

Novel Automatic Query Building Algorithm Using Similarity Thesaurus

[1]Hayel Khafajeh, [2]Aymen Abu-Errub, [3]Ashraf Odeh and [4]Nidal Yousef
[1]Department of CIS,
Faculty of Computing and Information Technology,
Zarqa University, Zarqa, Jordan
[2]Department of CIS,
Faculty of Information Technology,
Al-Ahliyya Amman University, Amman, Jordan
[3,4]Department of CIS,
Faculty of Information Technology, AL Isra University, Amman, Jordan

Abstract: One of the most effective factors on the natural language researches is the data set which plays a significant role in designing, improving and evaluation the information retrieval systems and other applications for natural language processing. Unfortunately, building a proper data set consume time, labor and effort, in particular the query extraction from the data set documents. In this study, a novel algorithm for query extraction from any collection of documents was suggested, the algorithm elaborate the similarity thesaurus for query extraction, which leads to the ability of using the algorithm on any language, to evaluate the suggested algorithm a data set that consist of 242 Arabic documents and 60 queries was used, 48 queries was extracted 20 of them appeared in manual data set and all of them was relevant with more than one document in the used collection.

Key words: Information retrieval, natural language processing, Arabic corpora, similarity thesaurus

INTRODUCTION

The data set plays a significant role in designing and improving the data retrieval systems and other applications for natural language processing (Guo and Wang, 2011). The data set, is used in information retrieval systems evaluation processes. The values of precision and recall are calculated by using the data set that contains the classified documents as relevant and the classified ones as irrelevant. The data set is usually formed of three main parts (Abdelali, 2004), which are: the corpus, the group of queries and the relations that link the queries with the documents. When information retrieval systems are built, there must be a set of appropriate dataset through which this system is evaluated.

The Arabic data set building faces several challenges which make the task difficult because of the lack of the available Arabic corpus to build a data set and the need of this corpus to be judged, by a group of experts. This process requires a lot of time and effort from these experts that are responsible for the determination of the specific topics for each document in the data set documents. The manual judgment process from these experts becomes a difficult process for application and sometimes it is not possible due to the large quantity of the required corpus to form a corpus that forms a dataset which contains hundreds of thousands of documents. There must be automatic methods to be used for data sets building to accompany occurred changes in information science and its retrieval.

Arabic corpus: There is a slight number of the available Arabic corpuses for researchers of Arabic language processing as a natural language. These corpuses are very important to researches in the optical character recognition systems and the information retrieval systems (Abdelali, 2004) and other aspects of language processing.

The corpus must be a representative of language by its possession of used words in the recent time and the popular linguistics terms (Maamouri *et al*., 2004). These corpuses which are found in different newspapers are considered the closest to represent the language because of its possession of a great quantity of the recent used words and they are also characterized by their diversity and covering of different subjects.

What is applied on newspapers corpuses are also applied on the corpuses existed on the internet

Corresponding Author: Aymen Abu-Errub, Department of CIS, Faculty of IT, Al-Ahliyya Amman University, Amman, Jordan

comprehensive sites which cover several subjects. These sites are featured with their possession of a great and variable quantity of corpuses. One of these corpuses that a newspaper corpus needs as a resource for (Goweder and Roeck, 2001) such as the Lebanese (AL- Haya) newspaper issues for the year 1998.

Literature review: Many researchers in natural languages processing and information retrieval concern about the corpus, since the corpus represents the language which is worked on. The corpus is used instead of the language, so the corpus must be qualified to be a language representative. Despite the great importance of the corpus, the Arabic corpuses are still few and not easily available. One of the researchers who form an Arabic corpus (Hmedi *et al*., 1997) as it is considered one of the first corpuses which are presented to be used in information retrieval aspect. The researchers place a corpus which is formed of 242 research abstracts from a Saudi Arabian University. They also conduct 50 queries. The corpus and the queries are shown to the judges through whom the query relevant documents are determined. They use the results in information retrieval system evaluation. The previous corpus and the relevant data set do not represent the language well, due to the small number of the used corpus terms and the limitation of the subjects that the used abstracts cover, represent scientific papers in computer and information systems.

In their trial to overcome the problems of the corpus that does not represent the language and does not contain several subjects, the researchers in (Goweder and Roeck, 2001) prepare a corpus of 18.5 million words in seven different subjects which are corpuses of the weekly Al-Haya newspaper for the year 1998 which contains 42,591 different materials. The size of the used corpus is 268 MB.

This research shows that the Arabic language is characterized by its sparser of words compared to the English language. 70% of the corpus words are repeated five times or less and so, the researchers try to implement the conditions which make the corpus a representative of the language. This corpus is also characterized by its dependence on the corpuses of one newspaper in a relatively short period of time (one year) which decreases the possibility of its representation of the language (Yousef *et al*., 2010).

In their trail to present a comprehensive corpus for the Arabic language which is common in most of the Arabic countries, the researchers present in this research (Abdelali *et al*., 2005) initial results to their experiments on an Arabic corpus which is formed of documents from the internet. They choose a corpus from samples taken from web sites for newspapers in several Arab countries. They also suppose that there are some corpuses which are related to specific areas, so the researchers collect samples from different areas in the Arab world. They justify this by saying that the current Arabic corpuses texts are characterized by its high cost, small size and limitation for specific areas. They collect corpuses from eleven news sources and each one is the web site for a newspaper or a news agency. They also use a special program to get the corpuses text from the sources, as they collect the issues and the daily leaflets from the sites daily for three months. The researchers extract the corpuses from the internet pages and they also do not use any normalization process on the corpus, justifying this by the normalization increases the ambiguity of the corpus. They use Zipf"s law and Mandelbrot formula to evaluate the corpus. They find that after a limit size for the corpus, adding new documents does not affect the corpus nature.

The suggested corpus is characterized by its dependence on the news corpus only in a short period of time (three months). This does not guarantee the corpus biased to specific words. In addition to that, the news corpuses do not only represent the Arabic language appropriately. The researchers concentrate on using specific words in certain regions and they give examples that do not suit this, since the given examples are about using synonyms that have the same meaning or one word which has more than one meaning.

MATERIALS AND METHODS

This study presents a new method of queries construction as the query is constructed automatically without the interference of the user.

This method leads to the construction of the queries related to a great number of documents. This will be conducted with a short time and a little effort of the users. The setting up of this method requires a group of documents which are inquired manually to be checked and evaluated with the suggested method results.

The proposed methos uses the documents that are presented in (Hmedi *et al*., 1997) to check its results as the corpus consists of 242 documents, each documents is an abstract of a scientific paper talking about computer or information systems. The researchers have improved 60 queries then; they relate this query with the documents.

The corpus which is used for checking process consists of 38081 words. After deleting the repeated words 8112 different words are left. The suitable query will balance between the words repetition in a single document and its repetition in the set of documents. Because the query that depends on the documents set must not be specialized to a certain document. But, include a set of documents. The proposal method will use the weights Term Frequency-Inverse

Document Frequency (TF-IDF) (Manning *et al*., 2008). The query that leads in retrieving a great number of documents is considered an inappropriate query for this. The words that are repeated in a lot of documents are not suitable to be part of the query and so, they must be excluded as will as the words that exist in one document and their repetition are limited. The words that are used in the query are the words that come between the highly repeated words and the low repeated ones.

From Table 1 we can notice that words (1) and (7) are repeated in all documents. These words are not suitable to be in the query. While word (2) is repeated in tested document four times and it is repeated in eight documents of the corpus, so it is suitable to be in the query. Word (3) is repeated in tested document once and it is not found in others, so, it is not suitable to be part of any query. The candidate words to be part of the query must have a limited repetition in the document and a limited repetition in all documents like the words (2) and (6).

Referring to the queries constructed by (Hmedi *et al*., 1997) we will find that the query related to the tested document are the queries which have the numbers 2 and 60. These queries are (البرمجة والحاسوب) and (الحاسب والزراعة) respectively. We notice that the word (البرمجة) is repeated in the second document four times and it is repeated in eight documents of the corpus, so it is a suitable word to be in one of the queries. While, the word (الحاسوب) is repeated in the tested document thirteen times.

Figure 1 shows that the repetition of these words is similar, as most of these words are repeated once but the other words are repeated once to five times in the document. The result indicates that the whole context is closer to represent the language than the documents individually as Zipf's law assumptions (Murtra and Bernat, 2010; Ferrer and Cancho, 2005).

The candidate words for the representative query of the first document are the words that come in the areas that are not shadowy in the previous Fig. 2. That means that the candidate word for the query must be repeated more than once and does not come with the words which have a high rank repetition or a low rank one in the words of all the documents.

Suggested algorithm: The suggested query building algorithm consist of the following steps:

- Delete stop word from the documents
- Normalize rest of the words
- Apply stemming on words.
- Delete high rank repeated words in all documents and the low rank ones
- Calculate weight of each word in tested document
- Choose the two words which have the largest weight in the document to construct the query on condition that each word must be repeated at least in three documents

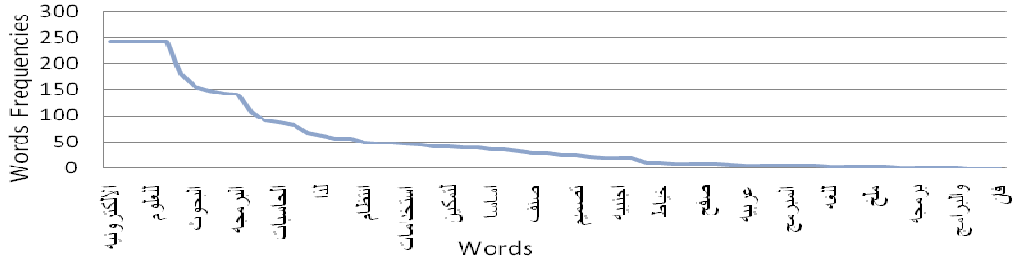

Fig. 1: Relation between words of the tested document and their repetition respectively

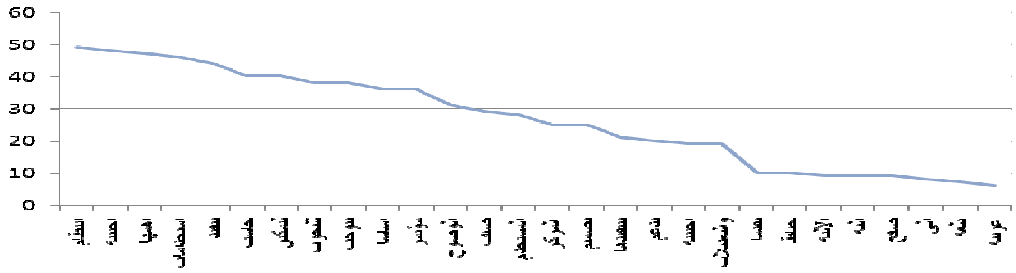

Fig. 2: Relation between candidate words of the query and their repetition in all documents

Using the thesaurus (Abuzir, 2012) to expand the query as follows:

- Search for words related to the query words- resulted of the previous steps- of the same document. If they are not available, apply the second step
- Search for a word related to the query words of the rest of the documents.
- Repeat the steps from the fifth step to each document

Table 1: Repetition of some words in tested document

No.	Word	Word freq. in tested doc.	No. of docs. containing word
1	العربيه	3	242
2	برمجه	4	8
3	سطحيا	1	1
4	استخدامات	2	7
5	حاسبات	13	8
6	تعريب	1	20
7	لغه	5	242

Table 2: Suggested Queries and their appearance in QDS

Query No.	Doc. No.	Suggested Query	Founded in queries data set
1	26	اجزاء البرامج	
2	15	اعمال الاتصالات	
3	53	الاتصالات اللمسية	
5	71	الادارة والحاسب	×
6	109	البرمجة الاجرائية	
7	206	التخطيط والادارة	×
8	169	التدريب والتعليم	×
9	156	التعريب الحاسوب	×
10	18	التعليم والتدريب	×
11	16	الحاسوب والتعليم	×
12	41	الحرف العربي	×
14	240	العلوم الشرعية	
15	10	القران الكريم	×
16	173	الكمبيوتر الهندسة	
17	155	اللغات الطبيعية	×
18	242	المعاجم الالكترونية	
19	7	المعالجات الدقيقة	
20	42	المعجم العربي	
22	12	المملكة السعودية	×
23	110	النص العربي	×
24	187	الهندسة الكمبيوتر	×
25	47	امن شبكات	
26	50	امن وسلامه	×
27	185	انظمة الامن	×
29	232	تصحيح الاخطاء	
30	238	تطوير الحاسوب	
31	88	تعريب الحاسوب	×
35	45	شبكات الاتصالات	
36	48	شركة ارامكو	
37	55	طرق التعليم	×
38	239	علوم الحاسوب	
42	85	لغات البرمجة	
43	2	لغة عربية	×
44	37	معالجة البيانات	
45	13	معلومات الجغرافية	
46	139	نظم الجغرافيا	
47	78	نظم الحاسوب	×
48	201	نظم المعرفة	

RESULTS

We will display the experiments related to the queries' building automatically for the data set documents by using the similarity thesaurus (Yousef *et al*., 2010; Khafajeh *et al*., 2010) and interactively according to the algorithm presented. We applied the algorithm on all the documents of the data set.

We could construct 48 different queries, as stated in Table 2, twenty of them appeared in queries data set (QDS) (marked with "×").

DISCUSSION

Queries that are created through the proposed algorithm showed that it can rely on this algorithm to create data sets for the purposes of the development and evaluation of information retrieval systems, where each query is linked to a number of documents in the data set, this is due to the use of similarity Thesaurus which uses the relationships between different words in the documents in the data set and in the each document, the proposed algorithm extract many queries that have been written by a group of professionals after studying the data set and this confirms the effectiveness of the proposed algorithm in the field of extraction queries automatically, which save most of the effort to create data sets for different languages in general and Arabic language in particular.

CONCLUSION

In this study, we suggest a novel algorithm for automatic query building; the suggested algorithm saves human effort and time required for query extraction in order to build a corpus. Our experiments show that we can build suitable queries from specific corpus without needing human experts' effort, by using the suggested algorithm 48 queries were built, 20 of them appeared in the set of queries that was built by a human expert and most of the rest of the 48 queries were relevant to the documents. The suggested algorithm is considered a cornerstone for a corpus automatic building, which is the further work which we think it will make use of our algorithm.

REFERENCES

Abdelali, A., 2004. Localization in modern standard Arabic. J. Am. Soc. Inform. Sci. Technol., 55: 23-28. DOI: 10.1002/asi.10340

Abdelali, A., J. Cowie and H. Soliman, 2005. Building a modern standard Arabic corpus. Proceedings of the Workshop on Computational Modeling of Lexical Acquisition, Jul. 25-28, The Split Meeting, Croatia, pp: 1-7.

Abuzir, Y., 2012. First token algorithm for searching compound terms using thesaurus database. J. Comput. Sci., 8: 61-67. DOI: 10.3844/jcssp.2012.61.67

Ferrer, I. and I. Cancho, 2005, Zipf's law from a communicative phase transition. Eur. Phys. J. B, 47: 449-457. DOI: 10.1140/epjb/e2005-00340-y

Goweder, A. and A.D. Roeck, 2001. Assessment of a significant Arabic corpus. Proceedings of the Arabic NLP Workshop at ACL/EACL, (ACL/EACL' 01), CiteSeerX.

Guo, Z. and J. Wang, 2011. Information retrieval from large data sets via multiple-winners-take-all. Proceedings of the IEEE International Symposium on Circuits and Systems, May 15-18, IEEE Xplore Press, Rio de Janeiro, pp: 2669-2672. DOI: 10.1109/ISCAS.2011.5938154

Hmedi, I., G. Kanaan and M. Evens, 1997. Design and implementation of automatic indexing for information retrieval with Arabic documents. J. Am. Soc. Inform. Sci., 48: 867-881. DOI: 10.1002/(SICI)1097-4571(199710)48:10<867::AID-ASI3>3.0.CO;2-#

Khafajeh, H., N. Yousef and G. Kanaan, 2010. Automatic query expansion for arabic text retrieval based on association and similarity thesaurus. Proceedings of the CD-ROM/Online European, Mediterranean and Middle Eastern Conference on Information Systems, (IS' 10), pp: 1-17.

Maamouri, M., A. Bies, T. Buckwalter and W. Mekki, 2004. The penn arabic treebank: building a large-scale annotated arabic corpus. University of Pennsylvania.

Manning, C.D., P. Raghavan and H. Schutze, 2008. Introduction to Information Retrieval. 1st Edn., Cambridge University Press, New York, ISBN-10: 0521865719, pp: 482.

Murtra, B.C. and R.S. Bernat, 2010. Universality of Zipf's Law. Phys. Rev. DOI: 10.1103/PhysRevE.82.011102

Yousef, N., I. Al-Bidewi and M. Fayoumi, 2010. Evaluation of different query expansion techniques and using different similarity measures in Arabic documents. Eur. J. Sci. Res., 43: 156-166.