A Method of Automatic Domain Extraction of Text to Facilitate Retrieval of Arabic Documents

Mohammad Khaled A. Al-Maghasbeh, Mohd Pouzi bin Hamzah
School of Informatics and Applied Mathematics
Universiti Malaysia Terengganu
Kuala Terengganu, Malaysia

***Abstract*—Arabic content on the internet has increased over the web because of the growth of the number of Arabic persons who use the internet in the world. Accordingly, this study introduces an automatic approach of domain extraction of information retrieval from these contents based on text classification. Text classification process makes the searching domain specific to facilitate the searching process. This paper discusses how to enhance the capacity of information retrieval in Arabic documents by classifying the unlabelled Arabic text automatically by using text classification algorithms. The classification of documents and texts is an important field in computer science and information retrieval. It aims at enhancing the retrieval process by identifying the searching-domain of retrieval systems.**

***Keywords*—Arabic information retrieval; text classification; Arabic text mining; Arabic language processing; text clustering; text classification; text categorization and classification algorithms**

I. Introduction

The Arabic language is one of the most common languages spread over the world which represents as one of the natural languages used in information retrieval field. Arabic language can be classified into three categories of dialects according to use in our life: traditional language, the formal language “Classical” and modern standard language. The first type is the language spoken or colloquial Arabic among people in their life, the second type is Holy Quran language written, and the third is a common language in the Arab world that commonly being used in literature, poetry, stories and literary writings [5].

Despite increasing the Arabic documents over the web. There are many problems still makes the Arabic information retrieval so challenge. The main problem in Arabic information retrieval is how to improve the retrieval accuracy. Hence, in this study, the new approach is proposed to merge the two applications of the NLP, namely, information retrieval and text classification.

This rest study is organized as follows. Section 2 briefly describes the related works in the area of Arabic texts classification. Section 3 describes Arabic text classification. Section 4 provides proposed approach. Section 5 shows the discussion with an example. In Section 6, it summarizes the work and future work.

II. Related Works

There are many applications of information retrieval. One of these applications is document classification that involves classifying document or text into several categories depends on some factors. Fraud et al., in their study used one of the information retrieval, namely, latent semantic analysis model with five similarity measurements to enhance the Arabic documents clustering. LSA in this study has been applied by using Singular value decomposition (SVD) to create an abstract representation of each document [9].

Mohammad Naji applied six common text classification methods such as Naïve Bayesian method (NB), support vector machines (SVM), Rocchio algorithm, k-Nearest Neighbor (KNN), neural network (NNet), Linear Least Squares Fit (LLSF) on two of set of Arabic document for training and test towards build a system to obtain a similarity degree between the training and test sets vectors based on the inner product feature. After that the proposed system computes the cosine between two vectors to find the best or appropriate class for each document has been tested. The recall and precision were the best with high similarity degree in Naïve Bayesian method [2].

In Thabtah study, the text classification has achieved based on Naïve Bayesian model with uses the mathematical statistics method that measures the correlation between two variables to check either correlated or independent. This mathematical method is known as the Chi-square method. The dataset in this study were 1562 Arabic documents from Sudia Press Agency (SPA) that is classified into six categories (Social, Cultural, Sports, Political, General, and Economic) [15].

The El Kourdi, study concerned for automatic classification of the Arabic web documents to help the search engine to deal with the continuous growth of the document via the internet. The Naïve Bayesian (NB) applied in this study to classify 300 of the Arabic web document that taken from Al-Jazeera website “the channel of Arabic News in Qatar Television” into five categories “Science, Health, Culture and Art, Business, and Sport”. The results showed accuracy in classification reaches 92.8% while the manual methods reached 62.8 [7].

Ababneh, in his study has been applied a Support Vector Machine algorithm (SV) on 5121 Arabic documents from the Saudi Newspaper (SNP) using K- Nearest Neighbor (KNN) technique to classify it’s into seven classification categories includes “Economics, Information Technology, General, Politics, Cultural, Sports, and Social”. In this work, the Arabic documents classified depend on the similarity degree, whereas it has been applied a several of experiments in (KNN) and (SVM) by using three different coefficients (Jaccard, Cosine, and Dice) to do a compression between of them. This work took the measures F1, Recall, and Precision to compute the efficiency of the two algorithms. The method has been proved that the Cosine was better than Jaccard, and Dice coefficients [1].

A study by [8] carried out some experimental for classification the Arabic texts from newswire by using mathematical or statistical classification methods. The experimental dataset was in sport, economics, politics, and culture, whereas 80% of these data used for training, and the 20% for testing

III. Text Classification

There are more than one ways that used in text mining to classify the documents into groups to represent the knowledge from them. These methods use set of factors to classify the number of documents into classes based on similarity, subject, and other characteristics [2]-[14]. Hence, there many traditional algorithms that use to classify the texts such as K-Nearest Neighbor (KNN), Naïve Bayes model, Decision tree, Support vector machine (SVM), and artificial neural networks (ANN) [10].

Text collection (TC) is a process to classify the document into groups based on the similarity, and it applies in some applications such as text filtering, web page categories, document organization, and other [11], [13]. The text classification has been become importance due to increasing the amount of data such as stories, news on the internet to facilitate the task of information retrieval when needed [15].

IV. Proposed Approach

We propose a new approach for information retrieval from Arabic documents depends on texts, or documents classification task as shown in Fig. 1. In this approach, we use the keywords extraction documents to classify them into several categories, and then in user query will extract the query terms. This model makes the system able to compute the similarity between the terms of the query with related documents through determining the domain in both, documents and user query.

The mechanism of the proposed approach has explained at the pseudo code that mentioned as follows:

```
Input: Documents collections, and General query;
Output: Retrieve the document that relevant the query;
For each query;
Begin:
Extract the query terms;
Compute the similarity between the query terms;
Determine the query domain;
Go to indexed documents, and then:
Extract all candidate keywords from each document;
Compute the similarity between document keywords;
Classify the document into categories;
Indexing each document category;
While similarity (Query domain, Doc domain) do:
Search in all the document in the same category;
Match (Query Term, Document keywords);
Retrieve all of related documents;
End
```

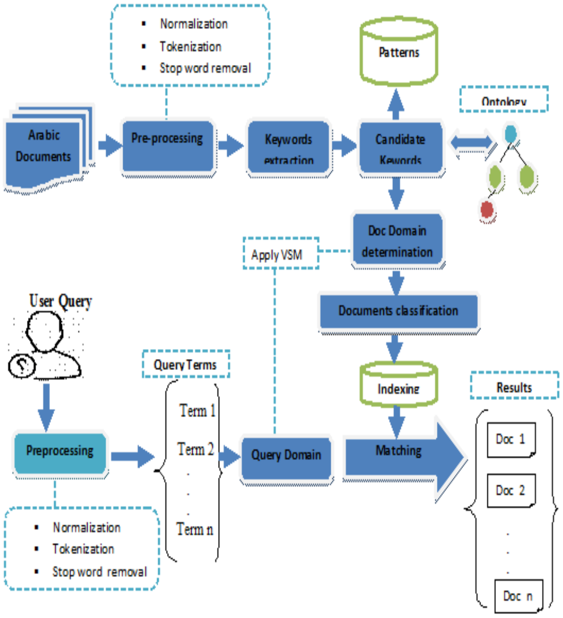

Fig. 1. Architecture of proposed approach.

The proposed model includes two main phases; each phase includes subtasks as the following:

1) *In documents processing*

- The pre-processing phase that includes the normalization, tokenization, and stops words removal.
- Extract the keywords in each document using the ontology, and the patterns that saved from other documents.
- Extract general topic or domain through computing the vectors space between the document keywords using VSM.
- Classify the documents based on Cosine similarity.

2) *In user query processing*

- Apply the preprocessing techniques like document preprocessing.
- Extract the query terms.
- Determine the query domain by computing the vector space between query terms.
- Using the VSM to compute the query vector with documents vector, and go directly into the documents that have the same domain.
- Match the documents that related to the user query using Cosine similarity.
- Rank the related documents.

A. Keyword Extraction

Keywords extraction is a method to discover the terminologies that represent the document, and text contents [8]. It is very important to aces to the main topics in the document, where it make the searching process so efficient. Keywords extraction also provides the target domain of document that helps the system to extract the relationships between concepts, and knowledge [10], [12].

B. Ontology

In this approach, we will use the Arabic ontology to access the relation between words, to candidate the keywords. Ontology includes several concepts which related with each other in class hierarchies. It concerns to determine the relevant concepts in an ontology, and semantic relations between of them**.** The ontology represents as a base infrastructure of knowledge. So the most researches of knowledge representation, semantic web concerned with ontologies. It also helps to determine the domain of the knowledge [6].

C. Preprocessing Phase

This phase is an important process in both, document, and query. The input of this phase is text of modern standard language of Arabic Newswire. It used to reduce the noise in the texts, through remove irrelevant or not important words such as stop words, prepositions, punctuation marks, digits from Arabic texts. As result, replace some Arabic letters into other letters to be more understandable and readable by computer. These subtasks can be summarized as follows:

1) Text normalization: This process is applying on several natural language texts. It represents a task 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.

2) Tokenization: Tokenization is a process to divide the plain text into tokens to remove the noise from the text. After that sent it into the morphological analyzer to continue the processing [4].

3) Stop words removal: This process is to remove the frequent Arabic words that insignificant words or don't carry important meaning.

D. Matching

The document and query represent as vectors to determine the domain of them by using SVM. This process has been done by computing the similarity between the keywords vectors, and terms in the same document, and query. As a result, to that, the document in the same query domain, and related to the query are ranked and retrieved to the user.

V. Discussion

Suppose having three document which taken from Jordan newswire named “Sarayanews”, and its URL is “sarayanews.com” as shown in Table I. Each document has different text as the following example in table:

TABLE I. SIMPLE EXAMPLE TO ELABORATE THE PROPOSED METHOD

Doc #	Document content	Translate to English	The candidate keywords	Document Domain/ Category
Doc 1	يعترض المشرفون على شؤون كرة القدم في ألمانيا على فكرة زيادة عدد الدول المشاركة في المونديال ويرون أن ذلك سيؤثر على جودة البطولة. بالمقابل تؤيد دول أخرى هذه الزيادة التي كان رئيس الفيفا قد وعد بها خلال حملته الانتخابية	**German football supervisors refuse to increase the number of countries participating in the World Cup and see that this will affect the quality of the tournament. Nevertheless, other countries support the increase that the FIFA president had promised during his campaign**	**كرة القدم, المونديال,البطولة, الفيفا** **Football, World Cup, FIFA.**	**Sport**
Doc 2	فاجأت أسرة البرنامج أعضاء لجنة التحكيم بعرض صور لهم خلال مرحلة الطفولة، غير أن اللافت أنه لم تُعرض صورة الفنانة الإماراتية أحلام على المسرح كسائر أعضاء اللجنة	**The family of the program surprised the members of the arbitration by presenting their pictures during childhood, but it is remarkable that the image of Emirati artist Ahlam was not shown on stage like the other members of the committee**	**البرنامج, لجنة التحكيم, اعضاء** **Program,Committee, Members, djudications**	**Art**
Doc 3	فقد مجلس النواب نصابه القانوني بعد أقل من نصف ساعة على بدء الجلسة الصباحية اليوم الاثنين وجاء فقدان النصاب دون رفع الجلسة المخصصة لمناقشة مشروعي قانوني الموازنة العامة وموازنات الوحدات الحكومية	**The Parliament lost its quorum since less than half an hour after the start of the morning session on Monday and the loss of the quorum without lifting the meeting to discuss the bills of the budget and budgets of government units**	**مجلس النواب, نصاب** **Load, Parliament**	**Political**

TABLE II. SAMPLE OF 3-QUERIES TO FIND RELATED DOCUMENTS

Query #	**In the Arabic language**	**Translate to English**
Q1	موعد بث برنامج ارب ايدول	**Broadcast time of program "Arab Idol"**
Q2	موازنة المملكة الاردنية الهاشمية 2017	**Budget plan of Jordan 2017**
Q3	عدد افرقة كاس العالم	**Number of World cup teams**

TABLE III. EXPECTED OUTPUT

Query #	**Terms**	**Translate to English**	**Query Domain**
Q1	بث, برنامج, موعد اربو ايدول	**Broadcast date, program, Arab Idol**	Art
Q2	موازنة, المملكة ,الاردنية, الهاشمية	**Budget, Kingdom Jordan Hashemite**	Political
Q3	عدد, افرقة, كاس, العالم	**Number, teams, cup, world**	Sport

Suppose, we have three queries as shown in Table II:

As can be seen in above Table II. When we apply the all phases, the proposed model over these queries, the analysis result will be as shown in Table III.

The proposed system will be able to search about the related document in the same domain of query directly. This work is about Arabic information retrieval analysis through text classification application to access the information, or document that need within a certain domain. Domain identification is so benefited for both search engine, and information retrieval systems to retrieve the target information.

VI. CONCLUSION

Due to the amount of the available documents and texts in websites which is a challenge for information retrieval researchers' to minimize the required time to retrieve the documents to enhance the degree of performance and accuracy in the retrieval, requires more efforts. So, to solve these challenges, the information retrieval systems, accuracy and recall measurement should be enhanced. One of these suggests a solution to improve the efficiency of the Arabic information retrieval system is using text classification. Text classification helps information retrieval systems to access the target information domain. This paper showed the importance of text classification to improve the information retrieval from different resources. It aims to enhance the performance of information retrieval systems.

ACKNOWLEDGMENT

We would like to thank University Malaysia Terengganu (UMT) for providing us with a good scientific environment to produce this simple work

REFERENCES

[1] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N. K. T., & Al-Ibrahim, A. (2014). Vector space models to classify Arabic text. International Journal of Computer Trends and Technology (IJCTT), 7(4), 219-223.

[2] Al-Kabi, M. N., & Al-Sinjilawi, S. I. (2007). A comparative study of the efficiency of different measures to classify Arabic text. University of Sharjah Journal of Pure and Applied Sciences, 4(2), 13-26.

[3] Alsaleem, S. (2011). Automated Arabic Text Categorization Using SVM and NB. Int. Arab J. e-Technol., 2(2), 124-128.

[4] 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.

[5] Belkredim, F. Z., El-Sebai, A., & Bouali, U. H. B. (2009). An ontology-based formalism for the Arabic language using verbs and their derivatives. Communications of the IBIMA, 11(5), 44-52.

[6] Brewster, C., & O'Hara, K. (2007). Knowledge representation with ontologies: Present challenges—Future possibilities. International Journal of Human-Computer Studies, 65(7), 563-568.

[7] El Kourdi, M., Bensaid, A., & Rachidi, T.-e. (2004). Automatic Arabic document categorization based on the Naïve Bayes algorithm. Paper presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.

[8] Elghannam, F., & El-Shishtawy, T. (2015). Keyphrase based Evaluation of Automatic Text Summarization. arXiv preprint arXiv:1505.06228.

[9] Froud, H., Lachkar, A., & Ouatik, S. A. (2013). Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. arXiv preprint arXiv:1302.1612.

[10] Harith, A., Kim, S., Millard, D. E., Weal, M., Hall, W., Lewis, P., & Shadbolt, N. (2003). Automatic ontology-based knowledge extraction and tailored biography generation from the web. IEEE Intelligent Systems, 18(1), 14-21.

[11] Harrag, F., & Al-Qawasmah, E. (2010). Improving Arabic Text Categorization Using Neural Network with SVD. JDIM, 8(4), 233-239.

[12] Hasan, K. S., & Ng, V. (2014). Automatic Keyphrase Extraction: A Survey of the State of the Art. Paper presented at the ACL (1).

[13] Moh'd A Mesleh, A. (2007). Chi-square feature extraction based SVMs Arabic language text categorization system. Journal of Computer Science, 3(6), 430-435.

[14] Sawaf, H., Zaplo, J., & Ney, H. (2001). Statistical classification methods for Arabic news articles. Natural Language Processing in ACL2001, Toulouse, France.

[15] Thabtah, F., Eljinini, M., Zamzeer, M., & Hadi, W. (2009). Naïve Bayesian-based on chi-square to categorize Arabic data. Paper presented at the proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt.

AUTHOR'S PROFILE

Mr. Mohammad Khaled A. Al-Maghasbeh is an PhD researcher of computer science in School of Informatics and Applied Mathematics, at Unversiti Malaysia Terengganu (UMT)- Malaysia. He obtained his M.Sc. in Computer Science from Al-Balqa'a Applied University, Jordan in 2013 with dissertation titled "Agent-Based Data mining for proteins prediction and classification". He got his B.Sc. in software Engineering from Al-Hussein Bin Talal University (AHU) in 2009. His interest and active researches are in Data mining, Big-data, Artificial Intelligence, Information Retrieval, and Natural Language Processing (NLP).