International Journal on
"Technical and Physical Problems of Engineering"
(IJTPE)
Published by International Organization of IOTPE

ISSN 2077-3528
IJTPE Journal
www.iotpe.com
ijtpe@iotpe.com

June 2022 Issue 51 Volume 14 Number 2 Pages 375-379

INFORMATION RETRIEVAL SCHEME VIA SIMILARITY TECHNIQUE

H.A. Taher [1] M.H. Abdulameer [1] B. Mahdi [2]

1. Department of Computer Science, Faculty of Education for Girls, University of Kufa, Najaf, Iraq
hawraaa.alshimirty@uokufa.edu.iq, mohammed.almayali@uokufa.edu.iq
2. Department of Computer Science, Faculty of Education, University of Kufa, Najaf, Iraq
bushram.alhashimi@uokufa.edu.iq

Abstract- In computer science, retrieving Arabic information has become a major research topic. Searching and retrieving knowledge-based information from databases is known as information retrieval (IR). However, the influence of the stop word removal part has a high impact on the retrieved Arabic text. In order to demonstrate the effect of stop words in Arabic text, we used one of the most commonly used techniques, the Jaccard technique, in this paper. We used the Jaccard technique in two ways, the first with Arabic stop words and the second way without stop words. In the experiments, we collected property Arabic datasets manually. The Arabic information retrieval system showed promising results for Arabic texts in similarity between documents and texts. It gave higher accuracy while removing stop words compared to texts containing stop words.

Keywords: Arabic Language, Jaccard Similarity, Information Retrieval, Natural Language Processing.

1. INTRODUCTION

It is the process of representing, storing and searching for a set of data or information for the purpose of knowledge discovery and retrieval for user, in response to user request [1]. Also, information retrieval (IR) is retrieving the pertinent documents from a set of documents. This task has two errors types, either retrieving non-relevant documents, or missing the relevant documents. These errors result from many reasons such as word sense, synonyms and many others semantic problems [2, 3]. Main research interests have focused on official language retrieval, generally in the news domain, document retrieval for OCR and language retrieval. Efforts were made in aspects of the Arabic language retrieval that had interests including: (image retrieval, social media, speech search, internet search, and filtering) [4].

The efforts made on various aspects to recover the Arabic language are still insufficient and the efforts made in other languages are severely lacking. Arabic is the seventh largest language on the Internet. However, the Arabic language was the fastest growing in the past period in terms of the number of users, despite the difficulties and challenges we face in the language [5].

Given the current growth rate of internet penetration among the Arabic speaking population, should have the fourth largest of Arabic language users number of users on the Internet by 2020. This gives special importance to the language and emphasizes the need for effective infrared approaches to enable effective search for Arabic documents [6]. Information retrieval (IR) systems were provided to assist in the management of large amounts of data. Numerous universities, public libraries and modern companies have used IR systems to facilitate access to books, magazines and required documents. The information retrieval system has been used nowadays in various important applications [1], [7]. Among the common and important applications of the information retrieval system are search engines, digital library, and research about media. In the information retrieval system, the semantic similarity measurement between words is used as it is an important measure, and it is also important in web or internet exploration and in natural language processing (NLP) [8], [9].

Many researchers have developed different types of Arabic information retrieval methods for example Kanaan et al. The [10] proposed a way to improve the Arabic information retrieval system by using a part of speech markers (POS), thus reducing the burden of indexing storage accordingly. To speed up the retrieval process. In addition, Larky and colleagues [11] developed several light derivatives of the Arabic language, compared the light stemming by many derivatives based on morphological analysis and measured the information retrieval efficacy using standard TREC data. The light 10 stemmer was superiors in comparison to other styles. It's part of the Lemur toolset and is commonly used for retrieving Arabic information. Moreover, Tan, et al. [12] developed an innovative approach for retrieving information from the Internet using handwritten texts, belonging to three text families; Arabic, Roman and Tamil texts. Results with an accuracy of 93.3%. Also, Sembok, et al. [13]. He focused in his research on finding the derivation mechanism that improved and increased the effectiveness of the IR system, which he applied to Arabic and Malaysian documents. And in all common languages, words generally include suffixes, prefixes, and suffixes. Examples: Use, Useful, Useless, User, and other examples. The important thing here is converting both the user query and stored database words into one standard form, known as Conflation.

On the other hand, Abu Salih [14] used a Vector Space Model (VSM) for the basic IR information retrieval system. He chose VSM for his project because where the weighting system is a term, and retrieval documents can be sorted according to their suitability. Also, another important feature for this technology is the ability to obtain relevant feedback from system users. Where users can judge and respond to whether or not the recovered document relates to their need.

2. MATERIALS AND METHOD (THE PROPOSED ARABIC IR SYSTEM)

The proposed system consists of two phases as in Figure 1.

2.1. Pre-processing Phase

This phase, consist of the following steps:

A. Stop words removal: It is the process of deleting unwanted words, which represents repeated words such as prepositions, which do not affect the result. In our work, the stop words were removed from all sites, and the result was compared with the text containing stop words.

B. Stemming: It is regarded as a key tool that, in addition to normalizing, is used in information retrieval to address the problem of vocabulary mismatch [11]. in addition, the process of reducing inflected words to their origin, base, or root form. Stemming enhances retrieval performance by decreasing word variations, which is especially important for high-impact languages like Arabic. The pre-processing of texts is depicted in Table 1.

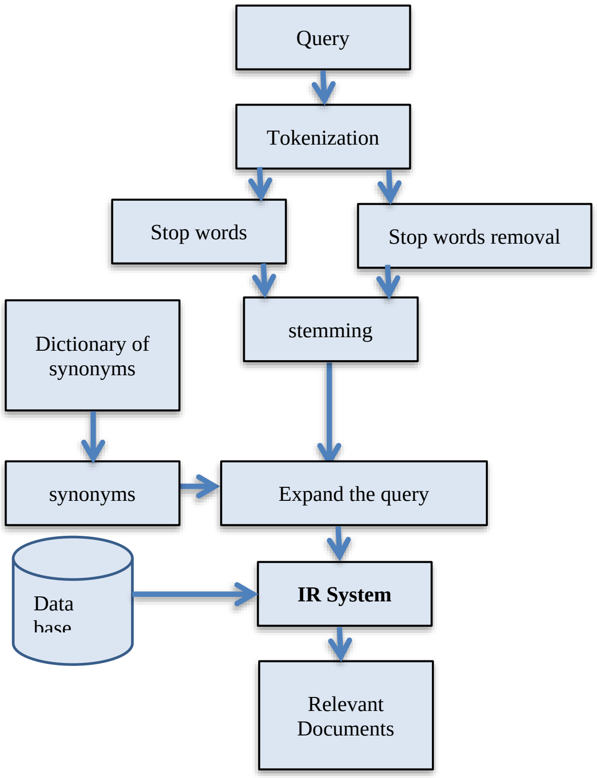


Figure 1. Diagram of the proposed Arabic IR System

Table 1. Preprocessing of Text

Normalize Arabic	Stemming	Stop word removal	Query	
['رايت', 'قرش', 'انهار', 'جارية']	['رأيت', 'قرش', 'انهار', 'جارية']	['رأيت', 'قرش', 'أنهار', 'جارية']	رأيت قرش في أنهار جارية	Q_1
['تشاهد', 'شيء', 'واقع']	['تشاهد', 'شيء', 'واقع']	['تُشاهد', 'شيء', 'بالواقع']	تُشاهد كل شيء بالواقع	Q_2
['سيف', 'طالب', 'ذكي']	['سيف', 'طالب', 'ذكي']	['سيف', 'طالبٌ', 'ذكي']	سيف طالبٌ ذكي	Q_3
['شرط', 'علي', 'بعض', 'شروط']	['شرط', 'علي', 'بعض', 'شروط']	['شرطً', 'عليهم', 'بعض', 'الشروط']	هو شرطً عليهم بعض الشروط	Q_4
['خسر', 'مال', 'كله']	['خسر', 'مال', 'كله']	['خَسر', 'المال', 'كله']	خَسر المال كله	Q_5
['سافر', 'باخرة']	['سافر', 'باخرة']	['سافر', 'بالباخرة']	سافر بالباخرة	Q_6

2.2. Similarity Phase

This phase, consist of the following steps:

A. Expand the query: After acquiring the meanings of all words, synonyms of all words are added based on their meanings from a small dictionary created for this purpose, and those with only one meaning are also included as synonyms. The query now has more words than it had in origin and it may reach multiples. The extended query is sent to the information retrieval system to be compared to documents that match the query.

B. Jaccard similarity Technique: The Jaccard Similarity is a measure of how similar two data sets are, and it's obtained by dividing the number of shared features by the total number of characteristics. Therefore, it measures the similarity between two texts since the intersection is separated by the object's union, that means the division of similar words on the union of words. The standard definition of Jaccrard similarity is shown below [15], [16], [17].

$$S_j(t_a.t_b) = \frac{\vec{t}_a.\vec{t}_{b|}}{|\vec{t}_a| + |\vec{t}_b| - \vec{t}_a.\vec{t}_{b|}} \tag{1}$$

$$S_j(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

$$S_j(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{3}$$

Figure 2, shows the algorithm of the Jaccard similarity for Arabic information retrieval system.

Algorithm A:
Input :Q//Query
Output: Relevant Documents

Step 1:preprocessing**:**

- Tokenization
- Remove stops word or with stop words
- Stemming (Get all word roots)
- Normalize

Step 2: expanding the query according to synonyms

Step 3: input the new query to Arabic IR System(Jaccard Similarity Technique)

Figure 2. Algorithm of Jaccard Similarity Technique for Arabic IR system

2.3. Data Set

By the database, we mean storing a set of documents for their data contents, and a presentation is created for each document by extracting the contents of the document [18]. There is a small database and there is a big database which is a large set of data that is difficult to manage and analyze using traditional tools [19]. in this paper, a set of special data was used in the Arabic information retrieval system, which was collected manually. A database was formed consisting of a set of documents that were taken from Al-Sabah newspaper, which is an official newspaper in Iraq.

3. RESULTS AND DISCUSSION

We use six queries for testing, and Python language is used to retrieve all required files. The Jaccard similarity algorithm was used for information retrieval system and applied to queries after preprocessing and then extending the query. The results of the Jaccard similarity algorithm for the Arabic IR system with and without stop words as shown in table 2 that show number of the documents that returned for both with/ without stop word.

Table 2. Jaccard similarity results for the Arabic IR System

Number of Relevant Document with remove stop words	Number of Relevant Document with stop words	Preprocessing & expanding	Query
2	5	['رايت', 'قرش', 'سمك', 'انهار', 'جارية', 'متدفقة']	Q_1
5	5	['ترى', 'تشاهد', 'تنظر', 'شيء', 'واقع', 'حقيقة']	Q_2
1	1	['أسم', 'سيف', 'حسام', 'بتار', 'صارم', 'قاطع', 'فاتك', 'حاد', 'سيف']	Q_3
2	4	['شرط', 'قيد', 'فرض', 'شروط', 'ضوابط', 'قيود']	Q_4
3	3	['خسر', 'فقد', 'نقود', 'مال']	Q_5
2	2	['سافر', 'ذهب', 'هاجر', 'غادر', 'رحل', 'سفينة', 'باخرة']	Q_6

Where we conclude Advantages of applied Jaccard Similarity with remove stop words are more accurate as only relevant documents are returned compared with the results of texts containing stop words, as disadvantage less accurate and returned other documents not relevant. for comparison of different algorithms and techniques on Information Retrieval in related work (introduction part) with we work.

Table 3. Comparison of different algorithms and techniques on Information Retrieval

Author	Algorithm	Language	Different with Similarity Technique	advantage
Kanaan,, Al-Shalabi and Sawalha 2005[10]	Part of Speech Tagging	Arabic	Not remove stop words, and not expand queries	reducing the burden of indexing storage accordingly
Larkey, Ballesteros& Connell 2007[11]	Light Stemming	Arabic	Remove stop words,and expand queries with light stemmer	Improve information retrieval
Tan, Gaudin and Kot 2009 [12]	Tf-Idf	Arabic, Roman and Tamil	Not remove stop words, and not expand queries	Improve information retrieval on handwritten texts
Sembok & Ata 2013 [13]	Stemming Algorithms	Arabic, Malaysian	Remove stop words, and not expand queries	improved and increased the effectiveness of the IR system
Abu-Salih 2018 [14]	Vector Space Model	Arabic	Remove stop words, and not expand queries	ability to obtain relevant feedback from system users.

After analyzing the results, we need to evaluate the result. The evaluation process for the proposed method and measuring the quality of information retrieval in the retrieval of relevant documents requested by the user are composed of three measures:

1) Precision: is the percentage of documents recovered, and is really relevant to the query. Also, the precision is defined the sum of number relevant documents recovered over the sum of number documents recovered. Precision is part of the true positive examples which means the number of "true positives" divided by number of "false positives" plus "true positives", as follows [20] [21].

$$P = \frac{TP}{TP + FP} \tag{4}$$

2) Recall: It is to determine the percentage of documents related to the query that were truly retrieved. Also, it is defined as sum of number related document retrieved over sum of number related documents in the database. and define the recall is calculated as follows: the number of "true positives" divided by the total number of "true positives" plus "false negatives" as in equation below [20] [21].

$$R = \frac{TP}{TP + FN} \tag{5}$$

where, *TP* (True Positives) is relevant document in the ranking, *FP* (False Positives) is non-relevant document in the ranking, *FN* (False Negatives) is relevant document but not retrieved, and *TN* (True Negatives) is non relevant document and not retrieved.

3) *F*-Measure: This is a metric for categorizing examples as "positive" or "negative" in order to determine model accuracy in a dataset. The *F*-score is a method for integrating precision and recall that is defined as a harmonic means of accuracy and recall [20], [21]:

$$F = 2\left(\frac{P.R}{P + R}\right) \tag{6}$$

These concepts can be illustrated through Table 4 and Figure 3 that show *F*-measure for each query for both (stop words, remove stop words).

Table 4. *F*-measure for each query

Query	*F*-measure with remove stop words	*F*-measure with stop words
Q_1	0.67	0.33
Q_2	0.33	0.33
Q_3	1.0	1.0
Q_4	0.67	0.4
Q_5	0.8	0.8
Q_6	1.0	0.8

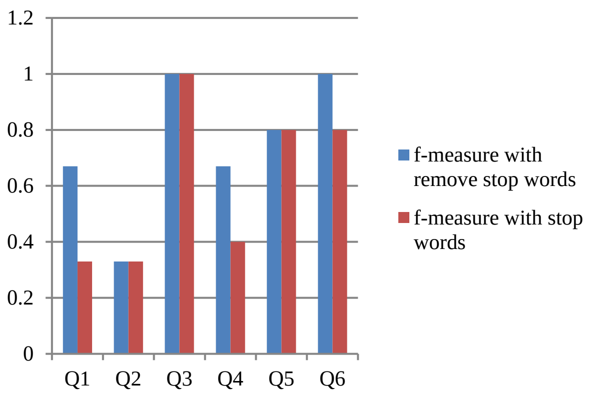


Figure 3. Comparative for each query

Also, we note the accuracy of the results in the texts after remove the stop words and expanding the query, table 5 and figure 4, show comparative for all queries according precision, recall and f-measure:

Table 5. Precision, recall and *F*-measure for all queries

All queries	Remove Stop Words	With Stop Words
Precision	0.53	0.4
recall	1.0	1.0
f-measure	70%	57%

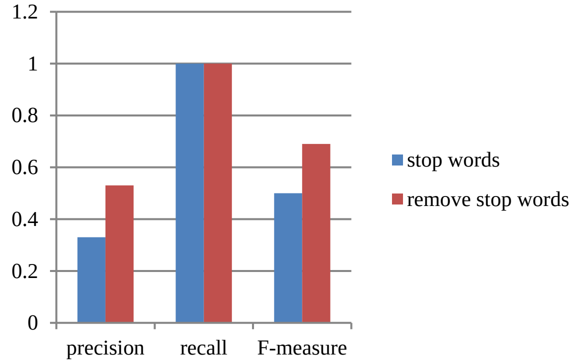


Figure 4. Comparative for All Queries

ACKNOWLEDGEMENT

The authors would like to thank Department of Computer Science, Faculty of Education for Girls, University of Kufa, Najaf, Iraq for providing them with the required documents and lab.

4. CONCLUSION

In this paper, we suggest information retrieval system based on similarity measure (Jaccard Similarity), for Arabic language, Despite the difficulties and challenges that we face in the Arabic language such as (morphology, vocabulary, word order, short and long vowels, prefixes and suffixes Diacritics, and et). The effects of stop words on Arabic retrieval were examined and compared with stop words removal, Where, we find stop words obviously influences information retrieval performance. And we got best performing result for retrieval in the Arabic language with stop words removal that gave *F*-measure is 70%, and gave f-measure is 57% when not remove stop words. We suggestion, solving the ambiguity in word sense is very useful for information retrieval that determine exact meaning of word and increase accuracy for document relevant.

REFERENCE

[1] M. Sharma, R. Patel, "A Survey on Information Retrieval Models, Techniques and Applications", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, pp. 542-545, Nov. 2013.

[2] G.Jena, S. Rautaray, "A Comprehensive Survey on Cross-Language Information Retrieval System", Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), Vol. 14, pp. 127-134, April 2019.

[3] B. Dhivakar, S.V. Saravanan, R.A. Krishnan, "Statistical Score Calculation of Information Retrieval Systems Using Data Fusion Technique", Computer Science and Engineering, Vol. 2, pp. 43-45, 2012.

[4] N. Atashafrazeh, A. Farzan, "A Review of Using Machine Learning Algorithms for Image Retrieval Words", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 20, Vol. 6, No. 3, pp. 139-144, September 2014.

[5] A. Jihad, A. Abdalkafor, "A Framework for Sentiment Analysis in Arabic Text", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 16, pp. 1482-1489, Dec. 2019.

[6] K. Darwish, W. Magdy, "Arabic Information Retrieval", Foundations and Trends in Information Retrieval, Vol. 7, No. 4, pp. 239-342, 2014.

[7] A. Roshdi, A. Roohparvar, "Information Retrieval Techniques and Applications", International Journal of Computer Networks and Communications Security, Vol. 3, pp. 373-377, Sep 2015.

[8] R. Karthikeyan, V. Udhayakumar, "A Web Search Engine-Based Approach to Measure Semantic Similarity between Words", International Journal of Emerging Research in Management and Technology, Vol. 4, pp. 102-105, April 2015.

[9] S. Tongphu, B. Suntisrivaraporn, P. Aimmanee, "Toward Semantic Similarity Measure Between Concepts in An Ontology", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 14, pp. 1356-1372, June 2019.

[10] G. Kanaan, R. Al Shalabi, M. Sawalha "Improving Arabic Information Retrieval Systems Using Part of Speech Tagging", Information Technology Journal, vol. 4, pp. 32-37, 2005.

[11] L.S. Larkey, L. Ballesteros, M.E. Connell, "Light Stemming for Arabic Information Retrieval", Arabic Computational Morphology, pp. 221-243, Springer, Dordrecht, 2007.

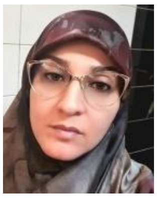
[12] G.X. Tan, C.V. Gaudin, A.C. Kot, "Information Retrieval Model for Online Handwritten Script Identification", The 10th IEEE International Conference on Document Analysis and Recognition, pp. 336-340, July 2009.

[13] T. Sembok, B. Ata, "Arabic Word Stemming Algorithms and Retrieval Effectiveness", Proceedings of the World Congress on Engineering, Vol. 3, pp. 3-5, 2013.
[14] B. Abu Salih, "Applying Vector Space Model (VSM) Techniques in Information Retrieval for Arabic Language", Preprint arXiv:1801.03627, 2018.
[15] S. Chauhan, P. Arora, P. Bhadana, "Algorithm for Semantic Based Similarity Measure", Int. J. Eng. Sci. Invent, Vol. 2, pp.75-78, 2013.
[16] S. Niwattanakul, J. Singthongchai, E. Naenudorn, S. Wanapu, "Using of Jaccard Coefficient for Keywords Similarity", Proceedings of International Multiconference of Engineers and Computer Scientists, Vol. 1, No. 6, pp. 380-384, March 2013.
[17] M.S.C. Sapul, R. Setthawong, P. Setthawong, "New Hybrid Flower Pollination Algorithm with Dragonfly Algorithm and Jaccard Index to Enhance Solving University Course Timetable Problem", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 20, pp. 1556-1568, Dec. 2020.
[18] M. Chahal, "Information Retrieval Using Jaccard Similarity Coefficient", International Journal of Computer Trends and Technology, Vol. 36, pp. 140-143, 2016.
[19] B.J.N Falih, B. Jabir, "Big Data Analytics Opportunities and Challenges for the Smart Enterprise", International Journal on Technical and Physical Problems of Engineering (IJTPE), Issue 47, Vol. 13, No. 2, pp. 20-26, June 2021.
[20] M. Arora, U. Kanjilal, D. Varshney, "Evaluation of Information Retrieval: Precision and Recall", International Journal of Indian Culture and Business Management, Vol. 12, pp. 224-236, 2016.
[21] F. Harrag, A.H. Cherif, A.M.S. Al Salman, E. El Qawasmeh, "Experiments in Improvement of Arabic Information Retrieval", The 3rd International Conference on Arabic Language Processing (CITALA), Rabat, Morocco, pp. 71-81, May 2009.

BIOGRAPHIES

Hawraa Ali Taher was born in Najaf, Iraq, on May 25 1987. She graduated from Computer Science and Mathematics Department, Kufa University, Iraq with the B.Sc. degree in computer science in 2009, and the M.Sc. degree in computer science from the in 2020. She is currently employed as an Assistant Teacher at Kufa University. Her research interests are artificial intelligence, and natural language processing.

Mohammed Hasan Abdulameer was born in Iraq, on January 1, 1978. He graduated from Al Maamoun University, Iraq with the B.Sc. degree in computer science in 2002, the M.Sc. degree in computer science from the Iraqi Commission for Computer and Informatics, Iraq in 2006, and the Ph.D. degree in computer science from National University of Malaysia, Malaysia in 2015. He is currently employed as an Assistant Professor at Kufa University, Najaf, Iraq. His research interests are artificial intelligence, natural language processing, and face recognition.



Bushraa Mahdi was born in Bagdad, Iraq, on August 15, 1977. She graduated from Computer Science and Mathematics Department Kufa University, Najaf, Iraq with the B.Sc. degree in computer science in 2009, and the M.Sc. degree in computer science in 2020. She is currently employed as an Assistant Teacher at Kufa University Najaf, Iraq. Her research interests are artificial intelligence, and image processing