ISSN 2347 - 3983
Volume 8. No. 6, June 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter103862020.pdf
https://doi.org/10.30534/ijeter/2020/103862020

Improving The Effectiveness of Texts Retrieval using Knowledge-Based Approach

Essam S. Hanandeh[1]
[1]Department of Computer Information System, Zarqa University, Jordan,
hanandeh@zu.edu.jo

ABSTRACT

A predicate-based document query language is proposed to allow users to define the search criteria precisely and reliably, and their knowledge of the documents to be retrieved. A guided search tool is built as an intelligent user interface oriented to the natural language in order to help users formulate queries. Supported by a generator of intelligent questions, an inference engine, a query base. A problem is faced when using the modern IR systems. It's represented in the vocabulary problem. This problem is represented in the inconsistencies between the terms which are used to describe the terms and the documents that are used by the investigator for describing their need for knowledge. The researcher has an automated thesaurus. This device has been designed using the Vector Space Model (VSM). The researcher used the similarity calculation of Cosine in this method. He used the selected 242 abstract Arabic documents in this article. All these abstracts include the information and computer science process. This paper aimed at building and designing automated Arabic thesauri through the use of the term similarity which could be employed in any particular domain or field for improving the process of expansion and obtaining greater number of relevant documents for the user query. In terms of recall and precision rates, it was found that the similar thesaurus is more capable than the conventional information retrieval system to enhance the recall process and precision.

Key words: Similarity thesaurus, Query Expansion, Retrieval of information, Vector Space Model

1. INTRODUCTION

Information retrieval (IR) refers to an analysis for the way of deciding and retrieving the' portions' that is sensitive to specific information needs from a corpus of stored information [7]. It also deals with the representation of texts, data processing, text organizing and retrieving information objects that are stored. Itis connected to several disciplines. It has many things in common with many other aspects of information processing, such as: the systems of information management, systems of database management, decision support system, question-answering system, processing of natural language, and systems for the recovery of documents [8].

A thesaurus[1]is an important tool in IR process. It's used in the processes of indexing and searching. It's used as a standardized vocabulary. It's used as a way to extend or modify queries (extension of requests) [10].The domain experts or/ and document description experts build in a manual manner most of the thesauri that users use. Manual thesaurus Construction serves as a process that is expensive and time-consuming. Less subjective has been founded in the results as the person who creates the thesaurus makes choices that can affect the thesaurus structure. Techniques of automatically constructing thesauri are required. They lead to having more rational thesauri that are easier to upgrade. They lead to improvement in terms of cost and time.

2. QUERY EXPANSION

Recovery of information is connected to the storage, representation, organization, and access to items of information. This representation is important for users, because it enables them to access specific information easily [9] [15].

The area behind the user's queries is called short queries which case a lack important words or phrases. In order to solve this problem, the Query expansion

[1] The plural is thesauri

was investigated by the data recovery researcher. That was done to help the consumer formulate better queries [20].

Most users find it hard to formulate questions which are well-designed for efficient information retrieval. The modification of the query of the user could lead to enhancing the results of the recovery significantly [14] [17].

Recovery of data refers to the retrieval of user requests. Such requests are called (queries). They are used for retrieving the relevant information from a document collection [16] [19]. Question Extension is a simple approach that is adopted for solving this problem.

[18] Aimed to examine several approaches to improve the initial query formulation through the expansion of queries. They examined the term Reweighing. Such approaches are divided into 3 categories: (a) approaches based on user input data; (b) ways which are depended on information that's derived from the initially collected set of documents. They are named the regional set of documents; and (c) global information approaches. They are derived from the collection of documents. That's the goal of this paper.

3. REVIEW OF LITERATURE

Several techniques and algorithms have been developed and proposed through the literature conducted about the systems of information retrieval (IRS). They construct thesauri, and request expansion. Several approaches are suggested to build a thesaurus. Some of those approaches are based on finding terms in the documents that are similar for the query term. Other ones are based on mapping the query and mapping the documents to a thesaurus.

Different techniques are used to improve the Indonesian information retrieval efficiency. For example, stopping; tokenizing sub-words; defining right nouns and not stemming proper nouns; and adjusting existing similarity functions. Based on our experience and knowledge, the quality of the recovery can be improved by following some previous strategies, with the highest increase when gram is achieved[1]. We also present an effective method for identifying a document's language. This allows for the selective application of different information retrieval techniques depending on the language of the target documents.

Some reflections were presented in this paper on the use of thesaurus in information recovery with particular regard to information recovery structures such as databases [2]. A Thesaurus serves is an example of regulated vocabulary. It serves as a significant aid in the study of subjects. Controlled vocabulary is used for defining the subject in Knowledge Organization Systems.

We viewed several solutions to this problem and offered two examples of thesaurus use in the following databases: the Thesaurus of ERIC Descriptors and the LISTA thesaurus. These thesauri are defined in the database together with their functions.

Aimed to build an Information Retrieval System Automatic Thesaurus in Arabic language[3]. The latter study was conducted based on two hundred forty two papers obtained from the National Computer Conference of Saudi Arabia. It was conducted through using 24 questions. Their study finds that using similar thesaurus might improve using word roots by Arabic IRS.

To improve the effectiveness of the retrieval in concept-based video retrieval, an integrated semantic-based approach was adopted for conducting similarity computation[11]. The planned solution is based on a combination of knowledge-based and corpus-based semantic term similarity measures in order to retrieve the video shots of concepts not available for annotation within the system. Regarding the TRECVID 2005 dataset, it is utilized to testing related goals. The results that are obtained through the suggested method are compared with the ones obtained through individual knowledge-based and corpus-based semantic word similarity tests used in the same domain in previous studies. In terms of the average accuracy, the superiority of the integrated similarity approach is demonstrated and assessed.

This paper deals with the problem addressing automated Arabic text comprehension. Regarding our goal, it is represented in understanding a given text and answering a list of related questions[12]. In order to do that, we developed an approach that allows us to conduct an analysis for the text in a domain that is open and create logical representations from it. Regarding our goal, it is based on accepting the process of textual participation. In a system for question answering named (NArQAS), we used this approach: New Arabic Question Answering System.

4. PROCEDURES OF THE EXPERIMENTS

These steps were taken:

1. We use the vector space model to vector text query
2. Normalization: Removing stop terms that A-Shalabi et al. [] gathered and achieved 98 percent success in distinguishing a part from removing certain signs appeared (stop words are words so frequently in collection documents that are useless for retrieval purposes.
3. Stemming: the following stemming algorithms with a little modification T stands for the set characters of the Arabic surface full word Li Stand for the position of letter I in the term T. Stem stands for the term after stemming in each step Let D stands for the set definite articles (لأ) S stands for the set of suffixes
   S={"ة",Pت,أ,ن,ي,و,ك,VW,كا,X,ه,Z,ه,X,او,_W,ZW, Vgc ن,ZW}
   P stands for the set of prefixes
   P={ل ب,ن,ث ي,لأ,Zo, Po,Vo Z,أZq,Zr} n stands for the total number of character in the Arabic word

Step1: Remove any diacritic in T

Step2: If the length of T is >3 characters then, Remove

The prefixes Waw "و" in position L1

Step3: Normalize},ا, of T to (plain alif)

Step4: Normalize ى, in Ln of T to ى

Replace the sequence of ى in Ln-1,in Ln to ئ

Normalize P in Ln of T to s

Step 5: For all variations of D (لأ) do,
Locate the definite article Di in T
If Di in T matches Di = Di + Characters in T ahead of Di
Stem = T – Di
Step 6: If the length of Stem is > 3 characters then,
For all variations of S, obtain the most frequent suffix,
Match the region of Si to longest suffix in Stem
If length of (Stem -Si) >= to 3 char then,
Stem = Stem – Si
Step 7: If the length of Stem is > 3 characters then,
For all variations of P do
Match the region of Pi in Stem
If the length of (Stem -Pi) > 3 characters then,
Stem = Stem – Pi
Step 8: Return the Stem

4. Collection of index terms from the filtered terms list. [5] Show that the inverted folder is a word-oriented method for indexing a text set to speed up the search process. Index terms can be individual words, group of words or phrases, but most of them are single words [13]. That's why we choose a single word (i.e. a single word).

5. CONSTRUCTION OF SIMILAR THEESAURUS

1-In this stage, 2 important decisions are included in the process of building the thesauri: What is the law used to find "Term Similarity"/"Term Relationship" between the various terms to create a thesauri? For different thesaurus, which formula should be used? Jaccard, Cosine or Dice Internal Product? And what's better? Would they deliver the same results?

2. What is the extent of threshold similarity /what's the relationship that is used as a synonym between the words in the thesauri?

Here we use Cosine equation. This equation is the most commonly used equation in the construction of the similarity thesaurus. Regarding the similarity threshold, it is a parameter to be entered while the system was operating.

$$\text{Cosine similarity } S_{j,k} = \frac{\sum_{i=1}^{N} (w_{i,j} * w_{i,k})}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2 * \sum_{i=1}^{n} w_{i,k}^2}} \quad (2)$$

All the results were between 0 and 1 as ($0 <= W_{i,k} <= 1$) & ($0 <= W_{i,j} <= 1$)

6. RESULTS

The present study aimed to strengthen the IRS based on 242 abstract Arabic documents used by (Hmeidi&Kanaan, 1997) [2]. It is purposed to recognize the importance of using stemmed words in these systems instead of full words. All of these abstracts are included the information system and computer science. The researcher has built and designed an automated system for information retrieval. That was done for handling Arabic text from scratch in order to achieve the intended goal. We obtained after the application of fifty nine queries from the Relevance Judgments documents started and the results were evaluated using the criteria for precision and recall. After doing that, the average precision and recall are calculated. The researcher used inverted file technique to create automatic stemmed words. The researcher built 2 information retrieval systems. That was done based on these indexing words. Through the 1st system, the

researcher used a traditional system for the retrieval of information. That was done through using the term frequency-inverse document frequency (tf-idf) for index term weights. The Similar Thesaurus was used through the second system. After stemming in the conventional information retrieval and using thesaurus. The results of the recovery systems are checked by section with the words: first, second and third section: In the first section: Effect of using thesaurus with stemming than traditional retrieval is shown in Table (1), and the figure (1) has shown the values of the average recall precision when user use Traditional and thesaurus stemming. Table (2) shows the number of the retrieved documents. It shows how many of them are Relevant and Irrelevant, using thesaurus and with Traditional case. In table (3) shows the percentage of the relevant retrieved documents from all the relevant documents in the collection, using thesaurus, and with traditional case. In the table (4) shows the percentage of all the cases together. In the second, Table (5) shows the effect of using similarity thesaurus over traditional retrieving (without using thesauri) by using stemmed words. Figure (2) shows a comparison between the values by using Similarity thesures and Traditional retriveing when using stemmed words. And finally in the third Table (6) shows the effect of using thesauri is much better than traditional information retrieval.

First:

Table 1: Comparison between Traditional and thesaurus stemming

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
Traditional with Stemming	0.92	0.88	0.79	0.7	0.63	0.52	0.44	0.29	0.2	0.08	0.
thesaurus with Stemming	0.874	0.874	0.81	0.71	0.66	0.54	0.44	0.33	0.25	0.14	0.

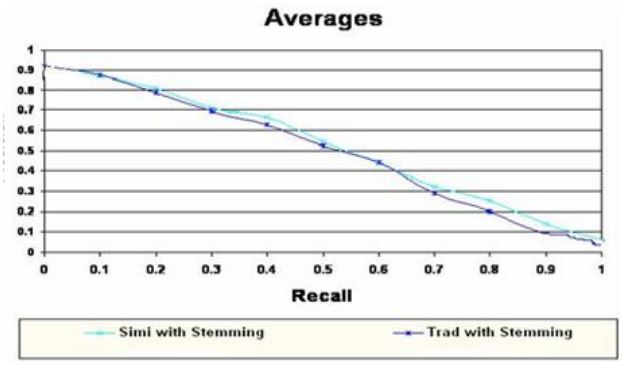

Figure 1: The Recall average between Traditional and thesaurus stemming

Table 2: Number of Retrieved, Relevant, and Irrelevant word when use stemmed

Table (2)			
	Retrieved	Relevant	Irrelevant
Traditional-Stemmed words	2399	1022	1377
thesaurus -Stemmed words	2029	991	1038

Table 3: Percentage of Relevent Document Retrived

	% of Relevant Docs that Retrieved
Traditional-Stemmed words	61.71497585
thesaurus -Stemmed words	62.681159

Table 4: Comparison between Traditional and Stemmed

Table (4)		
	Traditional	Similarity
Stemmed words	61.71497585	62.68115942

Second:

Table 5: The Average Recall Precision

Average Recall Precision			
Recall	Roots with using Similarity Thesaurus	Roots with using Traditional retrieving	% of Improvement for using Association Thesaurus over Traditional retrieving
0	0.908	0.917966102	-1.00%
0.1	0.87	0.875762712	-0.58%
0.2	0.810178571	0.785762712	2.44%
0.3	0.709464286	0.695254237	1.42%
0.4	0.664821429	0.626237288	3.86%
0.5	0.541428571	0.523389831	1.80%
0.6	0.438571429	0.442542373	-0.40%
0.7	0.325357143	0.290847458	3.45%
0.8	0.251428571	0.198305085	5.31%
0.9	0.13875	0.084745763	5.40%
1	0.056428571	0.047288136	0.91%

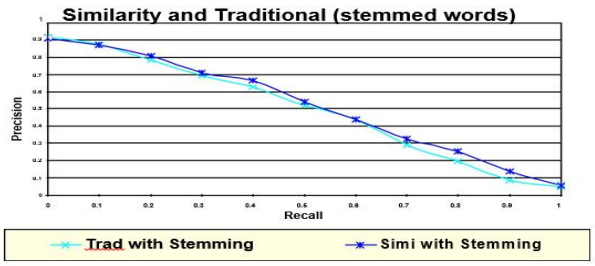

Figure 2: Comparison between the values by using Similarity thesures and Traditional retriveing when using stemmed words

Third :

Table 6: Comparison between Traditional and thesaurus stemming by Recall values

	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
thesaurus with Stemming	0.91	0.87	0.81	0.71	0.66	0.54	0.44	0.33	0.25	0.14
Traditional with Stemming	0.92	0.88	0.79	0.7	0.63	0.52	0.44	0.29	0.2	0.08

7. CONCLUSION

Thesaurus copes very well with these issues. Thus, it can be concluded that this tool plays a vital tool in database retrieval. The main problem that remains is the comprehension of limited thesauri users. If we were able to correct and reform these issues, thesauri might be more useful to users during IR processes. In Arabic, it is much better to use stemmed words with similar thesaurus than using traditional stemmed words. Automatic indexing may be used in Arabic language and its formulas. Using thesauri enforces IRS. Most researchers agreed on this issue. Stemming strengthens and assists IRS by using Arabic words. This applies to others as well.
IRS. Most researchers agreed on this issue. Stemming strengthens and assists IRS by using Arabic words. This applies to others as well.

8. FUTURE WORK

The researcher used stemmed word mechanism in the present paper. In the future, the researcher shall conduct
A study with planning to use all the mechanisms in traditional retrieval and use thesaurus with stemmed word a nd full word.

Acknowledgements

This research is funded by the Deanship of Research and Graduate Studies at Zarqa University, Jordan

REFERENCES

[1] J.Asian, “Effective Techniques for Indonesian Text Retrieval”. PhD Thesis, School of Computer Science and Information Technology, RMIT University Melbourne, Victoria, Australia, 30th March, 2007.

[2]E.Hanandeh, "Building an automatic thesaurus to enhance information retrieval", International Journal of Computer Science Issues, 10 (3). 2013, pp 676-686. https://doi.org/10.5121/ijcsea.2013.3201

[3]G.Kanaan, G. Ghassan and M.Wedyan, "Constructing an Automatic Thesaurus to Enhance Arabic Information Retrieval System". The 2nd Jordanian International Conference on Computer Science and Engineering, JICCSE 2006, Salt, Jordan.2006, pp 89-97.

[4] C. Soanes, A. Stevenson, and S. Hawker. "Concise Oxford English Dictionary". Oxford University
Press, New York, eleventh edition, 2004.

[5] K. Sparck Jones, S. Walker, and S. E. Robertson. "A probabilistic model of information Retrieval: Development and comparative experiments". Information Processing & Manage-
ment, 36(6). 2000, pp 779–840.

[6] A. Spink, D.Wolfram, M. B. J. Jansen, and T. Saracevic. "Searching theWeb: The public and
Their queries", Journal of the American Society for Information Science and Technology,
52(3).2001, pp 226–234.

[7] K. Takeda,"The application of bioinformatics to network intrusion detection", In Proceedings of
The 39th Annual IEEE International Carnahan Conference on Security Technology, pages
130–132, Canary Islands, Spain, 2005. IEEE

[8] B. Baresch, , L.Knight, , D.Harp, , &C.Yaschur,"An analysis of content links on Facebook", The Official Research Journal of International Symposium on Online Journalism, Austin, TX (Vol. 1, No. 2, pp. 1-24), 2011.

[9] A. Hermida, F.Fletcher, D.Korell, &D.Logan, Share, like, recommend: Decoding the social media news consumer. Journalism Studies, 13(5-6), 815-824, 2012.

[10] H. Djoerd, “Information Retrieval Models”, published in Goker, A., and Davies, J. Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, 2009.

[11] S. Memar, L. Affendey, N. Mustapha, S. Doraisamy, and M. Ektefa, “An integrated semantic-based approach in concept based video retrieval,” Multimed. Tools Appl., vol. 64, no. 1, pp. 77–95, Aug. 2011. https://doi.org/10.1007/s11042-011-0848-4

[12] B. Wided, P.Bellot, M.Neji, "A logical representation of Arabic questions toward automatic passage extraction from the Web", *International Journal of Speech Technology*, 2017.

[13] H. Jamil, H. Jagadish,"A structured query model for the deep relational web", In: Proceedings of the 24th ACM

International on Conference on Information and Knowledge Management, pp. 1679–1682. ACM 2015. https://doi.org/10.1145/2806416.2806589

[14] J. Amadaka,V. Salaka, B. Johnson, T. King, "Query expansion classifier for e-commerce", US Patent 9,135,330, 2015.

[15] A. Khwileh, G.Jones, "Investigating segment-based query expansion for user-generated spoken content retrieval In: Content-Based Multimedia Indexing", 14th International Workshop on, pp. 1–6. IEEE, 2016.

[16] S.Kuzi, A.Shtok, O.Kurland,"Query expansion using word embedding". In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 1929–1932. ACM, 2016. https://doi.org/10.1145/2983323.2983876

[17] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: a survey",CoRR, abs/1708.00247, 2017.

[18] W. Ge, W. Huang, D. Dong, and M. R. Scott, "Deep metric learning with hierarchical triplet loss", In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, ECCV (6), volume 11210 of Lecture Notes in Computer Science, pages 272–288. Springer, 2018. https://doi.org/10.1007/978-3-030-01231-1_17

[19] M. Dahab, S. Alnofaie, and M. Kamel,"A tutorial on information retrieval using query expansion", In Intelligent Natural Language Processing: Trends and Applications, pages 761–776. Springer, 2018 https://doi.org/10.1007/978-3-319-67056-0_35

[20] H. Azad, and A. Deepak, "Query expansion techniques for information retrieval: a survey",arXiv preprint arXiv: 1708.00247, 2017.