SIMILAR THESAURUS BASED ON ARABIC DOCUMENT: AN OVERVIEW AND COMPARISON

Essam S. Hanandeh,

Department of Computer Information System, Zarqa University, Zarqa, Jordan
Hanandeh@zu.edu.jo

ABSTRACT

The massive grow of the modern information retrieval system (IRS), especially in natural languages becomes more difficult. The search in Arabic languages, as natural language, is not good enough yet. This paper will try to build similar thesaurus based on Arabic language in two mechanisms, the first one is full word mechanisms and the other is stemmed mechanisms, and then to compare between them.

The comparison made by this study proves that the similar thesaurus using stemmed mechanisms get more better results than using traditional in the same mechanisms and similar thesaurus improved more the recall and precision than traditional information retrieval system at recall and precision levels.

KEYWORDS

Similarity thesaurus, Recall, Precision, information retrieval, Traditional

1. INTRODUCTION

A thesaurus (plural: thesauri) is a valuable tool in IR, in both indexing and in searching processes. It is used as a controlled vocabulary and as a means for expanding or altering queries (query expansion)[10]. Most thesauri that users encounter are manually constructed by domain experts and/or experts at document description. Manual thesaurus construction is a time-consuming and quite expensive process, and the results are bound to be more or less subjective since the person creating the thesaurus make choices that affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which in addition to the improvements in time and cost aspects can result in more objective thesauri that are easier to update. These developed thesauri have been designed to comply with international and Arabic standards, and is capable of performing all tasks & duties needed in thesauri building. It automates most tasks in the building and maintenance of thesauri and insures integrity of its structures and its relations with database materials.

2. INFORMATION SYSTEM

Information retrieval exhibits similarity to many other areas of information processing. The most important computer-based information systems today are the management information systems (MIS), data base management systems (DBMS), decision support systems (DSS), question-answering systems (QA), as well as information retrieval system (IRS) [8].

Information Retrieval (IR) is best understood if one remembers that the information consists of documents. In that context, information retrieval deals with the representation, storage, and access to documents or representatives of documents [10]. Information Retrieval systems have an important role in the studies of libraries and information, and they are really considered the top of this field. They get use of facts and ideas achieved in all studies, theoretical or practical, such as indexing, classifying, subjective analysis, concluding, computers, bibliography and others.

As we know that half of the science is in its organization, and the main objective of this specialization is to supply and afford the suitable and sufficient information in the suitable time to the suitable person or researcher.

3. REPRESENTATION OF DOCUMENTS AND QUERIES:

The following section describes the steps of representing the documents and queries automatically:

3.1 Filtering:

In filtering the document collection, these filtering processes consist mainly of:

3.1.1 Eliminating Stop words:

Stop word is a word that occurs so frequently in documents in the collection that it is useless for purposes of retrieval [3], Elimination of stop words reduces the size of the indexing structure and thus increases the performance of the system and enables it to retrieve more relevant documents.

Stop words in Arabic include some of grammatical links such as the definite article (ال), attached and separated prepositions, conjunctions, interrogative words, negative words, exclamations , calling letters, adverbs of time and place They also include all the pronouns, demonstratives, subject and object pronouns, the Five Distinctive Nouns, some numbers, additions and verbs. Stop words may be separated or attached ones in a form of prefixes or suffixes [3].

3.1.2 Stemming:

Stemming is the deletion of prefixes and suffixes (getting the root of the word). The root means: the part of the word that is left after deletion of prefixes, suffixes and Infixes [4].

Stems are thought to be useful for improving retrieval performance because they reduce variants of the same root word to a common concept. Furthermore, stemming has the secondary effect of

reducing the size of the indexing structure because the number of distinct index terms is reduced [3].

3.1.3 Indexing:

Selection of index terms from the collection of filtered terms as seen in the construction processes.

Indexing is defined as the process of choosing a term or a number of terms that can represent what the document contains. These terms have been called (Index terms) [8]. Indexing can be performed either manually (Manual Indexing) or through using computers software and programs (automatic Indexing) [6]. To decide the location of accuracy of the keywords in a certain text, the researcher will generate the inverted files. This file will contain all the keywords that the text contains, accompanied by the number of each paragraph containing such words. The paragraphs may be a sentence, a paragraph, a whole page of a document or the complete document [3].

4. ALGORITHM

Phase 1: preparing documents:

1 Use vector space model to put text of documents and query in vectors.
2 Normalization.

- Removing stop words those were collected by Al-Shalabi, et al [2], and they gained 98% success in distinguishing in addition to deleting some signs appeared. (stop words are the words that occur so frequently in documents in the collection that it is useless for purposes of retrieval [9]), Elimination of stop words reduces the size of the indexing structure and thus increases the performance of the system and enables it to retrieve more relevant documents.
- Deleting punctuation marks, commas, follow stops, especial signs, numbers (contents that has no meanings).

3 Stemming : the following stemming algorithm as in [11] with a little bet modification :

Let T denote the set of characters of the Arabic surface Full word
Let Li denote the position of letter i in term T
Let Stem denote the term after stemming in each step
Let D denote the set of definite articles (ال)
Let S denote the set of suffixes

S ={ ت ا، ن، ي، و، ك، ه،ة،
ار، لي، ري، تك، تا، يا، ما، يه، ته، تن، ني، تم، وا، نا، كن، كم، ها، هن، هم، ات، ون، ين، ان،
ية يل، تي، ير،
{لها ينا، رها، رين، مان، رات، يون، يتش، يان، لين،

Let P denote the set of prefixes

P={ ل ب، ن، ، ت ي،
اس فن، ، في فت، لن، ، لت لي، ، با ، فا ، كا سن، ست، سا، سي، لل، ال،
{ الف الك، للم، الع، الس، الا، الم، لال، مال، الح،

Let n is the total number of characters in the Arabic word
Step 1: Remove any diacritic in T

Step2: If the length of T is > 3 characters then,
Remove the prefix Waw " و" in position L1
Step 3: Normalize آ ,إ ,أ of T to ا (plain alif)
Step 4: Normalize ى in Ln of T to ي
Replace the sequence of ى in Ln-1 and ء in Ln to ئ
Replace the sequence of ي in Ln-1 and ء in Ln to ئ
Normalize ه in Ln of T to ة
Step 5: For all variations of D (ال) 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

Phase 2: building a traditional IRS.

4 Selection of index terms from the collection of filtered terms. Ricardo Baeza-Yates and Berthier Ribeiro-Neto in [9], show that the inverted file (or inverted index) is a word oriented mechanism for indexing a text collection in order to speed up the searching task, Index terms can be Individual words, group of words, or phrases, but most of them are single words [9] for this reason researcher choose a single words (i.e., single term) as index terms in this work.

This phase includes an affecting decision, to repeat the required word within the document to be an "Index Term". If it only appears once, can a word be used as an Index Term? Or researcher must use words that repeat many times in the same document?

In this thesis the index terms should be the words that repeated from (2-7) times in the text. After studying some files and taking the average number of occurrence of some words, they found in [7] that the best index terms to be used are those repeated within the average. (not too little, nor too much).

It is expected that the use of a controlled vocabulary leads to an improvement in retrieval performance. So we ignore the terms that appear in most documents in the collection (i.e., has high frequencies), and the terms that appear only once in a document (i.e., terms that low frequencies).

1 Creating the inverted file based on the stemmed words of each documents. (The stemmed words technique used is suffix prefix removal).

2 Compute the frequencies of each index term in each document (tf).

3 Compute the normalized frequencies of terms in each document by using the following formula

$$f_{i,j} = \frac{freq_{i,j}}{\max_{l} freq_{l,j}} \quad (1)$$

1 Compute the inverse document frequency, for each index term K_i in a document d_j, as follows

$$idf_i = \log \frac{N}{n_i} \quad (2)$$

Where N is the total number of documents in the collections, and n_i is the number of documents in which index term k_i appears.

2 Calculate the weight of each term in a document by multiplying the normalized frequencies with inverse document frequencies as follows

$$w_{i,j} = f_{i,j} * idf_i \quad (3)$$

After these steps, we have an inverted file that contains index terms (i.e., words) and terms frequency and the weight of each term in a document.

Phase 3: Building Similar Thesaurus:

In this paper, the researcher uses Cosine equation, as it is the most common equation in building the similarity thesaurus, and the threshold similarity was a variable to be entered while the system working.

$$\text{Cosine similarity } S_{j,k} = \frac{\sum_{i=1}^{n} (w_{i,j} * w_{i,k})}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2 * \sum_{i=1}^{n} w_{i,k}^2}} \quad (4)$$

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

5. EXPERIMENTS AND RESULTS

This study aims to reinforcing IRS depending on 242 Arabic abstract documents that used by (Hmeidi & Kanaan, 1997) in [5], also to realize the importance of using stemmed words in these systems instead of full words. All these abstracts involve computer science and information system.

To achieve this aim, the researcher designed and built an automatic information retrieval system from scratch to handle Arabic text. Work on these results that we got after applying 59 queries

from the Relevance Judgments documents began and results were analyzed using the Recall and Precision criteria. After that, Average of Recall and Precision were calculated.

Researcher has constructed an automatic stemmed words and full word index using inverted file technique. Depending on these indexing words researcher has built three information retrieval systems; in the first system, the researcher has used a Traditional Information Retrieval system using a term frequency-inverse document frequency (tf-idf) for index term weights. In the second one, the researcher used Similar Thesaurus by using Vector Space Model with Cosine formula using a term frequency-inverse document frequency (tf-idf) for index term weights, and compare between the similar measurements to find out the best that will be used in building the Similar thesaurus.

3.1 Results

Table 1 shows the effect of using Traditional (full words, stemmed) and using Similarity thesaurus(full words, stemmed).

Table 1 :Number of Retrieved, Relevant, Irrelevant documents using Traditional and when using similarity thesaurus

	Retrieved	Relevant	Irrelevant
Traditional-Full Words	1706	763	943
Traditional -Stemmed words	2399	1022	1377
Thesaurus -Full Words	1704	771	933
Thesaurus -Stemmed words	2029	991	1038

Figure 1 & Figure 2: Comparison values of the Average Recall Precision when using Traditional and when using similarity thesaurus

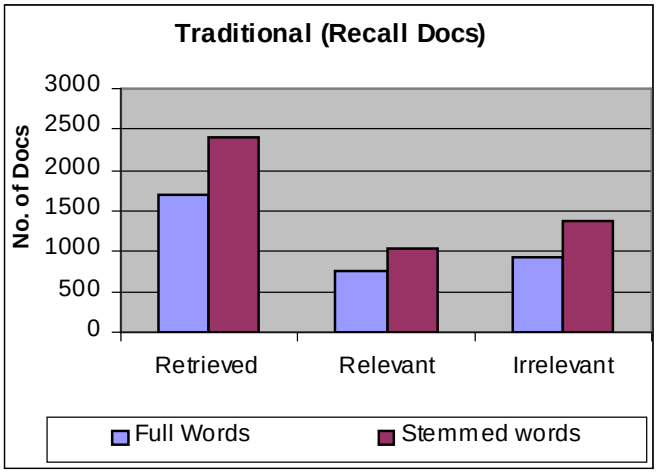

Figure 1: relevant and irrelevant shape from retrieved document in Tradition system

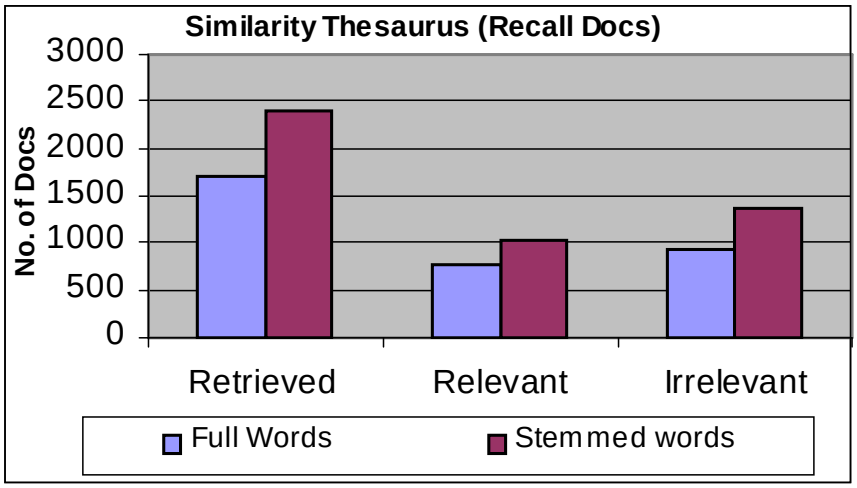

Figure 2: relevant and irrelevant shape from retrieved document in the case of similarity thesaurus retrieving system

Table 2, Figure 3, and Figure 4 shows the percentage of the relevant retrieved documents from all the relevant documents in the collection, when using Traditional and when using similarity thesaurus

.

Table 2: shows the percentage of the relevant retrieved documents	
	% of Relevant Docs that Retrieved
Traditional-Full Words	46.07487923
Traditional - Stemmed words	61.71497585
Thesaurus -Full Words	46.557971
Thesaurus -Stemmed words	62.681159

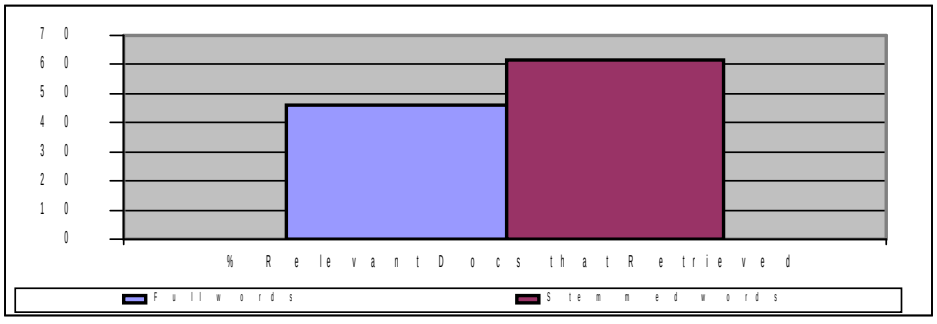

Figure 3: % of retrieved document in Traditional system

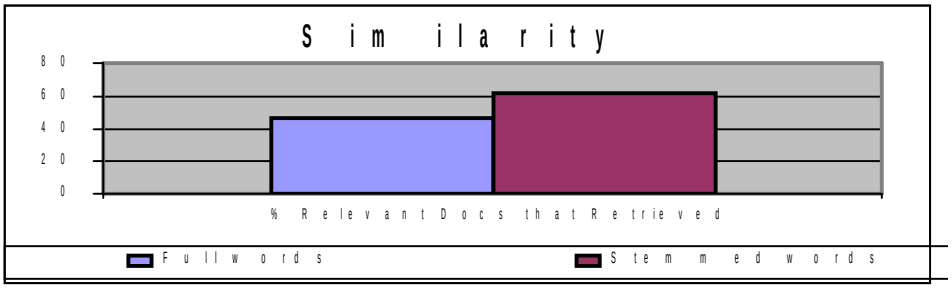

Figure 4: % of retrieved document in thesaurus system

Table 3 shows percentage when using Traditional and when using similarity thesaurus

Table 3: Percentage of using Traditional & similarity thesaurus		
	Traditional	**Similarity**
Full words	46.07487923	46.55797101
Stemmed words	61.71497585	62.68115942

Table 4: shows how much better were the results of using Traditional and when using similarity thesaurus

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%
q1	`2q2	0.047288136	0.91%

Table 4: Effect of using Traditional and when using similarity thesaurus

Table 5 shows the effect of using the stemmed words for information retrieving were always better than using Full words, and ensure that using thesauri is much better than using Traditional information retrieval.

Table 5: Average of all the Relative work

	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Traditional-Full Words	0.91	0.87	0.81	0.71	0.66	0.54	0.44	0.33	0.25	0.14
Traditional - Stemmed words	0.92	0.88	0.79	0.7	0.63	0.52	0.44	0.29	0.2	0.08
Thesaurus - Full Words	0.92	0.8	0.68	0.54	0.4	0.33	0.21	0.13	0.11	0.07
Thesaurus - Stemmed words	0.9	0.82	0.65	0.49	0.39	0.26	0.21	0.12	0.08	0.04

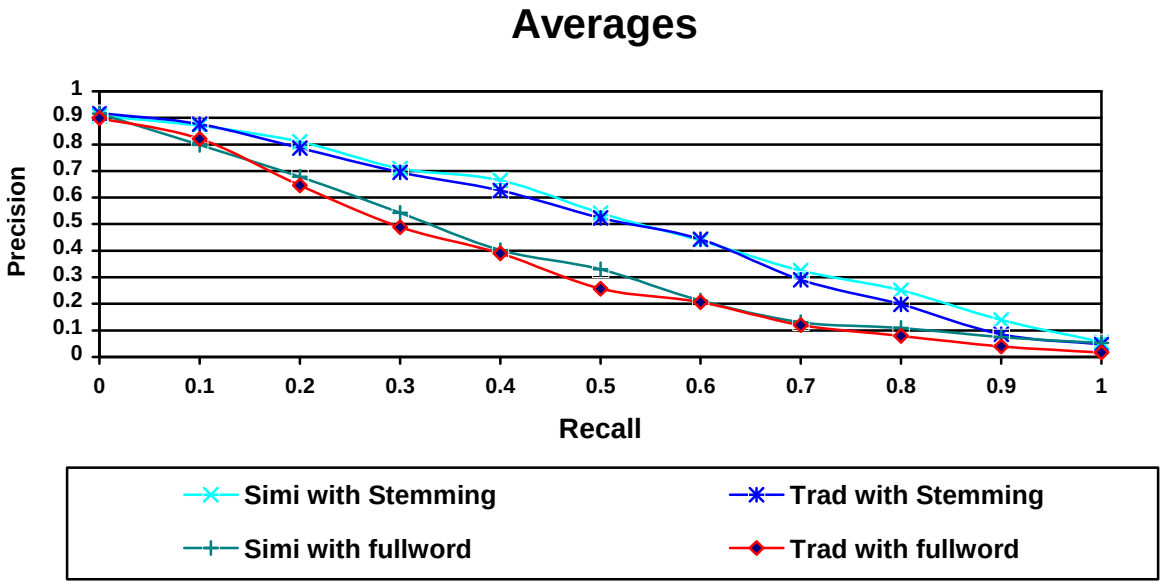

Figure 5: A comparison between the values of average Recall Precision with all cases

6. **CONCLUSION:-**

In this paper, researcher built similar thesaurus in tow mechanisms (full word, and stemmed) , researcher found out that the results for retrieval information in used stemmed word get better result than using full word in case of used traditional search, but when researcher used the similar thesaurus in both mechanisms, got better result of stemmed than full word.. Finally, the results when using stemmed in similar thesaurus got better result over than using stemmed in traditional search.

7. FUTURE WORK:-

In this study i built the similar thesaurus in both mechanisms full word and stemming word. I hope in the future to build an associative thesaurus and compare between them to know what is better for retrieval information.

REFERENCES

[1] Adriani, M. and Croft, W. "Retrieval Effectiveness of Various Indexing Techniques on Indonesian News Articles", 1997.

[2] Al-Shalabi, R. Kannan, G., Al-Jaam, J., Hasnah A., and Helat, E., "Stop-word Removal Algorithm for Arabic Language", processing of the 1st International Conference on Information & Communication Technologies: from theory to Applications-ICTTA, Damascus, 2004.

[3] baeza-yates R.,and Rierio-neto B., "Modern Information Retrieval" , Addison-Wesley,New-York,1999.

[4] Darwish, K., "Building a Shallow Arabic Morphological Analyzer in one Day", Acl Workshop on Computational Approaches to Semitic Language, PP 47-57, 2002.

[5] Kanaan, G. "Comparing Automatic Statistical and Syntactic Phrase Indexing for Arabic Information Retrieval", Ph.D.Thesis, University of Illinois, Chicago, USA, 1997.

[6] Monica Lassi "Automatic Thesaurus Construction", A paper written within the GSLT course Linguistic Resource, autumn 2002.

[7]Al-Shalabi, R. Kannan, G., Al-Jaam, J., Hasnah A., and Helat, E., "Stop-word Removal Algorithm for Arabic Language", processing of the 1st International Conference on Information & Communication Technologies: from theory to Applications-ICTTA, Damascus, 2004.

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

[9]Smeaton, A.F., Van Rijsbergen, C.J., The Retrieval Effects of Query Expansion on Feedback Document Retrieval System, The Computer Journal, 26(3), p239-46, 1983.

[10]T. R. Addis, Machine Understanding of Natural Language, International Journal of Man-Machine Studies, Vol. 9, No. 2, March 1977, pp. 207-222.

[11]Aljlayl, M, and Frieder, O, "on Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach" ACM Conference on Information and Knowledge Management, Mcelean, VA , November, 2002

AUTHORS

Assistant Prof. in Zarqa University , jordan