Building an Automatic Thesaurus to Enhance Information Retrieval

Essam Hanandeh[1]

[1] **Computer Information System,**
Zarqa University, Zarqa, Jordan

Abstract

One of the major problems of modern Information Retrieval (IR) systems is the vocabulary Problem that concerns with the discrepancies between terms used for describing documents and the terms used by the researcher to describe their information need. We have implemented an automatic thesurs, the system was built using Vector Space Model (VSM). In this model, we used Cosine measure similarity. In this paper we use selected 242 Arabic abstract documents. All these abstracts involve computer science and information system. The main goal of this paper is to design and build automatic Arabic thesauri using term-term similarity that can be used in any special field or domain to improve the expansion process and to get more relevance documents for the user's query**.** The study concluded that the similarl thesaurus improved the recall and precision more than traditional information retrieval system in terms of recall and precision level**.**

Keywords: *Information* ***Retrieval,*** *similarity thesaurus****,*** *Vector Space Model****,*** *Query Expansion****.***

1.Introduction

Information retrieval (IR) can be defined broadly as the study of how to determine and retrieve from a corpus of stored information the 'portions' which are responsive to particular information needs (Tengku, 1989). IR is also concerned with text representation, text storage, text organization, and the retrieval of stored information items that are similar in some sense to information requests received from users. The term IR covers a wide range of disciplines, and have some similarities with many other areas of information processing, e.g., management information systems, database management systems, decision support systems, question-answering systems, natural language processing, as well as document retrieval systems.

A thesaurus (plural: thesauri) is a valuable tool in IR, both in the indexing process and in the searching process, used as a controlled vocabulary and as a means for expanding or altering queries (query expansion)[10]. Domain experts and/or experts at document description manually construct most thesauri that users encounter. 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 can affect the structure of the thesaurus. There is a need for methods of automatically construct thesauri, which can in addition to the improvements in time and cost aspects can result in more objective thesauri that are easier to update.

2. Vector Space Model

The vector space model uses non-binary weights that are assigned for the documents and queries index terms[13]. This will suggest a partial matching retrieval instead of the relevant / non-relevant matching. The non-binary weights assigned for both the queries and documents are ultimately used to measure the degree of similarity(equation 1) between each of the documents in store in the system and the user query. Hence, the vector model will also take into consideration documents which match the query terms partially.

The vector model uses the t-dimensional vectors to represent both document and query. For a document **dj** (where j is the document number) and a query **q**, their t-dimensional representations are **dj** and **q** as follows:

The query q representations is :

$$\overrightarrow{q} = (w_{1,q}, w_{2.q}, \ldots, w_{t,q})$$

and the document dj representation is :

$$\overrightarrow{dj} = (w_{1,j}, w_{2.j}, \ldots, w_{t,j})$$

where $\mathbf{w_{i,q} \geq 0}$ and t is a total number of index terms in the system.

The vector model proposes to evaluate the degree of similarity of the document **dj** with regard to the query **q** as the correlation between the vectors **dj** and **q**. This correlation can be quantified, for instance, by the cosine of the angle between these two vector [13], That is,

$$sim(d_j, q) = \frac{\sum_{i-1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}} \quad (1)$$

Vector model can uses different similarity measures other than cosine similarity as shown in Table 1 [13]:

Table 1: Similarity Measures

Similarity Measure	Evaluation for Binary Term Vector	Evaluation for Weighted Term Vector
Cosine	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert^{1/2} \bullet \lvert q \rvert^{1/}}$	$sim\ (d_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i.j}^2} \times \sqrt{\sum_{j=1}^{t} w_{i,q}^2}}$
Dice	$sim(d,q) = 2\frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert}$	$sim(d_j, q) = \frac{2\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1}^{t} w_{i,j}^2 + \sum_{i=1}^{t} w_{i,q}^2}$
Jaccard	$sim(d,q) = \frac{\lvert d \cap q \rvert}{\lvert d \rvert + \lvert q \rvert - \lvert d \cap q}$	$sim(d_j,q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sum_{i=1} w_{i,j}^2 + \sum_{i=1} w_{i,q}^2 - \sum_{i=1}^{t} w_{i,j} \times w_{i,}}$
Inner	$\lvert d_i \cap q_k \rvert$	$Sim = \sum_{k=1}^{t} (d_{ik} \bullet q_k)$

Note $|d|$ is the number of term in document d.

3. Query Expansion:

Information retrieval deals with the representation, storage, organization , and access to information items. It is important that this representation provides users with easy access to the information in which they are interested [15] [18].

Short queries submitted to search engines are behind the missing of important words or terms from the user's queries. To solve this problem, the researcher of information retrieval have been investigated the Query expansion as a method to help the user in formulating a better queries [22].

Many users find it difficult to formulate queries that are well-designed for effective retrieval, and they often use a great variety of words to refer to what they want. Expansion or modification of the user's query can lead to a considerable improvement in the retrieval results [19].

Information retrieval refers to the processing of user requests, commonly referred to as queries to obtain relevant information from a collection of documents [19].

An obvious approach to solve this problem is Query Expansion.

Ricardo Baeza-Yates & Berthier Ribeiro-neto point out that [18] that they examine a variety of approaches for improving the initial query formulation through query expansion and term

reweighing. These approaches are grouped in three categories: (a) approaches based on feedback information from the user; (b) approaches based on information derived from the set of documents initially retrieved (called the local set of documents); and (c) approaches based on global information derived from the document collection, which is the objective of this paper.

4. Previous work

Many techniques and algorithms for Information Retrieval Systems (IRS), building thesauri and query expansion have been devised and proposed in the literature. There have been different approaches for building a thesaurus, some of them based on finding the similar terms for the query term in the documents, while others based on mapping both the query and the documents to some kind of a thesaurus, while others used to expand their queries finding synonyms or build a hierarchy relation between terms. Below are some of the results of using or building thesaurus in information retrieval and query expansion.

David Walker, (2001) in his paper talks about the different types of query expansion, he divides them into three types as (1) human and computer generated thesauri, (2) relevance feedback, and (3) automatic query expansion with taking into account the strengths and weaknesses of each. In the conclusion, he has shown that automatically expanded queries via pseudo-relevance feedback and computationally generated thesauri address the needs of users, but have not improved effectiveness of search engines beyond that encountered in relevance feedback [24].

Abu Salem, (1992) studies the IR in Arabic Language. His study based on 120 documents that he received from the Saudi Arabian National Computer Conference and on 32 queries. In his research, he studies indexing by using full words and by using the roots only. He finds out that using the roots are superior to other ways. He also built a manual thesaurus using the relation between expressions to test the possibility of supporting an IRS through this thesaurus. He finds out that the thesaurus makes IR much better [9].

Al-Shalabi et al, (2004) suggests an algorithm for determining and deleting stop words in Arabic texts. This method depends on Finite State Machine. They tested the system using the 242 documents that were presented in the Saudi Arabian National Computer Conference in addition to some verses taken from the Holy Quran. They reached to an accuracy of 98% [11].

Kanaan et al, (2006) Construct an Automatic Thesaurus to enhance Arabic Information Retrieval System. This study was based on 242 documents taken from Saudi Arabian National Computer Conference and they used 24 queries. Their study find out that using Similar thesaurus will make the efficiency of the Arabic IRS better when using roots of the words **[37]**.

5. Experiments Procedure

We do the following steps:

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 [11], 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 [54]), 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 [67] 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،
ير، 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،
{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

4 Selection of index terms from the collection of filtered terms. Ricardo Baeza-Yates and Berthier Ribeiro-Neto in [54], show that the inverted file 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 [54] for this reason we choose a single words (i.e., single term) as index terms in this work.

5 Building Similar Thesaurus

In this step, the process of building the thesauri includes two important decisions:

1 What is the law used in finding "Term Similarity"/"Term relationship" between the different terms to build a thesauri? Which formula should be used in similar thesaurus? Inner product, Cosine, Jaccard or Dice? And which is better? Will they give the same results?

2 what is the degree of threshold similarity/relationship used between the expressions in the thesauri to be used as a synonym?

We use here Cosine equation(equation 2) , 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.

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

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

6. Results

This study aims to reinforcing IRS depending on 242 Arabic abstract documents that used by (Hmeidi & Kanaan, 1997) in [36], 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 objective, the researcher designed and built an automatic information retrieval system from scratch to handle Arabic text. Working 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 using inverted file technique. Depending on these indexing words, researcher has built two information retrieval systems; in the first system, 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, researcher used Similar Thesaurus by using Vector Space Model with four similarity measurements (Cosine, Dice, Inner product and jaccard) using a term frequency-inverse document frequency (tf-idf) for index term weights, and compare between the similarity measurements to find out the best that will be use in building the Similarity thesaurus.

The results of the retrieval systems can browse using the words after stemming in the Traditional retrieval of information and in using thesaurus:

First：

Table (1): Effect of using thesaurus with Stemming than Traditional retrieval

Table(1)

	0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1**
Traditional with Stemming	0.92	0.88	0.79	0.7	0.63	0.52	0.44	0.29	0.2	0.08	0.05
thesaurus with Stemming	0.874	0.874	0.81	0.71	0.66	0.54	0.44	0.33	0.25	0.14	0.06

Figure (1): Comparison values of the Average Recall Precision when use thesaurus than traditional.

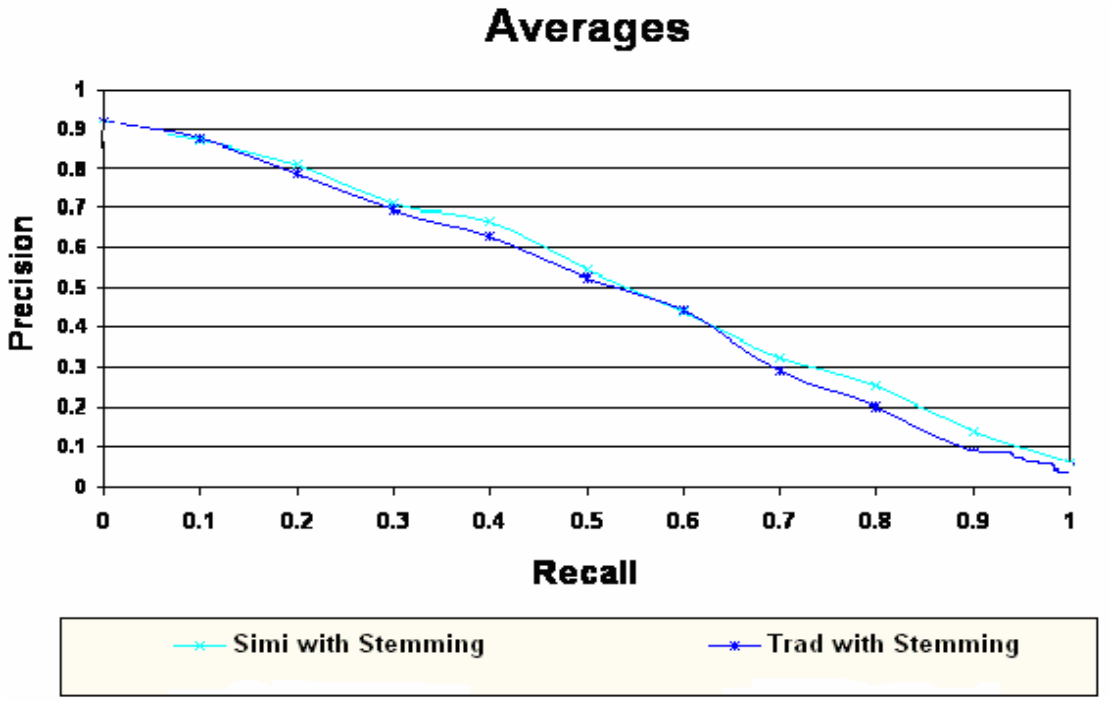

Table (2) shows the number of retrieved documents and how many of them are Relevant and Irrelevant, using thesaurus and with Traditional case.

Table (2)

	Retrieved	Relevant	Irrelevant
Traditional-Stemmed words	2399	1022	1377
thesaurus -Stemmed words	2029	991	1038

Table (3) shows the percentage of the relevant retrieved documents from all the relevant documents in the collection, using thesaurus and with Traditional case

Table (3)

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

Table (4) shows percentage of all the cases together

Table (4)

	Traditional	Similarity
Stemmed words	61.71497585	62.68115942

Second：

Table (5) Effect of using Similarity thesaurus over traditional retrieving (with out using thesauri) by using stemmed words.

Table (5)

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%

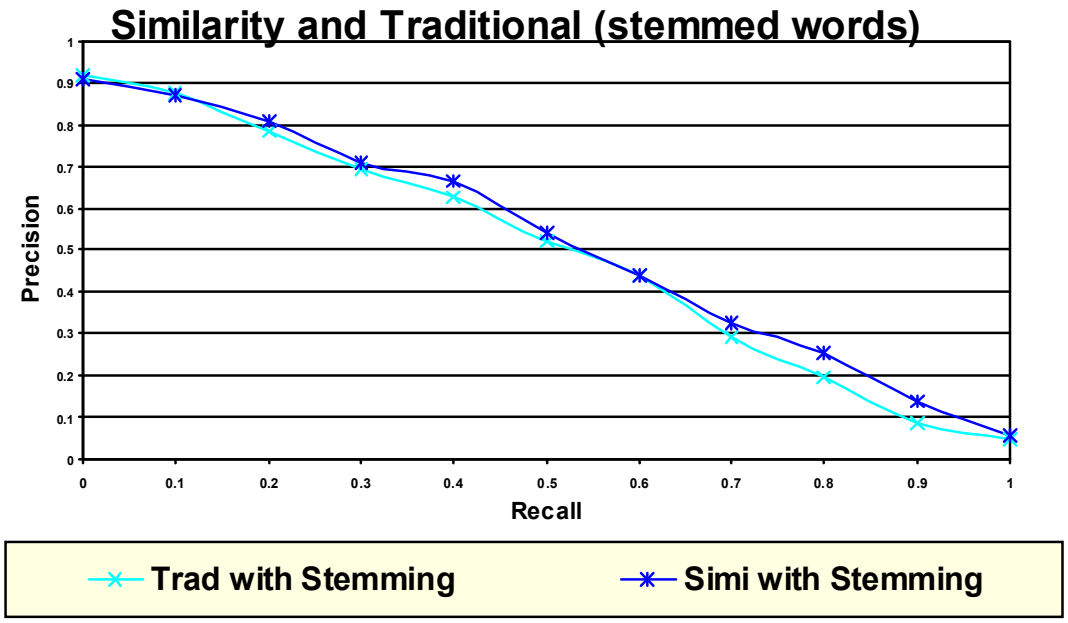

Figure (2) Comparison between the values by using Similarity thesaurus and Traditional retrieving when using stemmed words

Third:

Table (6) shows the effect of using thesauri is much better than using Traditional information retrieval.

Table (6): 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
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

Conclusion

Using stemmed words with similar thesaurus in Arabic language retrieving system is much better than using stemmed words with traditional. [9]. There is a possibility of applying Automatic Indexing and its equations in the Arabic language. Using thesauri enforces IRS. Most researcher agreed on this point. Using the stemming of Arabic words reinforces and supports IRS. This is also true for other languages.

Future work

In this paper, we use stemmed word mechanism. In future, we plan to use all mechanism with stemmed word and full word in traditional retrieval and use thesaurus.

References

[1] Abu Salem, H., A Microcomputer Based Arabic Bibliographic Information Retrieval system With Relation Thesaurus, Ph.D. thesis, University of Illinois, Chicago, USA, 1992.

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

[3] 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.

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

[5] C. J. van Rijsbergen "information retrieval, Butterworth",1979

[6] Chengfeng Han, Hideo Fujii, and W. Bruce Croft, "Automatic Query Expansion For Japanese Text Retrieval", Umass Technical Report, 1994.

[7] Cui H. Wen J. Nie. J., Ma W., "Query Expansion by Mining User Logs," IEEE Transaction on Knowledge and Data Engineering, 2003; 15(4); 829-839.

[8] David Walker, "Query Expansion using Thesauri: Previous Approaches and Possible New Directions", 2001.

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

[10] 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.

[11] 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.

[12] 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

[13] Salton,G., and McGill,M., Introduction to Modern Information Retrieval,McGraw-Hill,New-York, 1983.