BUILDING ARABIC AUTOMATIC THESAURUS USING CO-OCCURANCE TECHNIQUE

Assist.Prof.Dr. Hayel Khafajeh
Faculty of Computing and Information Technology
Zarqa University, Zarqa, Jordan
E-mail: hayelkh@zu.edu.jo
Mobile: 0777721311

Assist.Prof.Dr. Mohamad Refai,
Faculty of Computing and Information Technology
Zarqa University, Zarqa, Jordan
refai@zu.edu.jo

Assist.Prof.Dr. Nidal Yousef
Faculty of Computing and Information Technology.
Al-Isra University, AMMAN, Jordan
nidal.yousef@ipu.edu.jo

Abstract
One of the major problems of modern Information Retrieval (IR) systems is the word mismatch word mismatch that concerns the discrepancies between terms used for describing documents and the terms used by the researchers to describe their information need. One way of handling the Word mismatch is by using a thesaurus, that shows (usually semantic) the relationships between terms. The main goal of this study is to design and build an automatic Arabic thesaurus using Co-occurrence technique 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. Results from this study were compared with the traditional information retrieval system.

Two hundred and forty two Arabic documents and 59 Arabic queries were used for building the requirements of the thesaurus, such as inverted File, indexing, term-term co-occurrence matrix, etc. All of these documents involve computer science and information system vocabulary.

The system was implemented in ORACLE 10 g environment and run on Pentium-4 laptop with 2.13GHz speed, 2.86MB RAM memory, and hard disk capacity of 500GB.

Building this technique can be used in any special field or domain to improve the expansion process and to get more relevant documents for the user's query.

In this paper, we concluded that the Co-Occurrence thesaurus improved the recall. However, it has many limitations over the traditional information retrieval system in terms of recall and precision level.

Keywords: Query Expansion, Co-Occurrence thesaurus, Similarity thesaurus, Thesaurus, Indexing, Natural language (NL), Synonyms.

Introduction

Information retrieval (IR) deals with the representation, storage, organization and access of information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested.

Unfortunately, characterization of the user information-need is not a simple task because of the language of the user. [13]

The word thesaurus has Greek and Latin origins and is used as a reference to a treasury of words. [7]

The Thesaurus involves some normalization of the vocabulary and includes a structure much more complex than a simple list of words and their synonyms, the popular thesaurus published by Peter Roget [23].

A thesaurus (plural: thesauri) is a valuable tool in Information Retrieval (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)[8] . 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 besides from the improvements in time and cost aspects can result in more objective thesauri that are easier to update.

Is a statistical approach where the occurrences of terms in documents, chapters or some other unit are computed? The closer the words occur, the more significant is the co-occurrence. Many automatic indexing methods do not consider how closely words occur, just if they occur in the same document [6].

Related Work

Many researchers discussed the co-occurrence analysis of the documents text such as Chen and Lynch [8], Crouch [7], and Salton [24].

The limitation of the popular symmetric similarity functions (such as cosine, Dice, and Jaccard's) have been reported by Peat and Willett [21]. Their research showed that similar terms identified by symmetric co-occurrence function tended to occur very frequently in the database that is being searched and thus did little or nothing to improve the discriminatory power of the original query. They concluded that this can help explain Sparck Jones finding that the best retrieval results were obtained if only the less frequently occurring terms were clustered and if the more frequently occurring terms were left UN clustered.

The co-occurrence analysis used by Schutze and Pedersen in their research was based on number of times a word co-occurs with other words in a document. Schutze and Pedersen described this matrix as a “term-by-term matrix" (Schutze and Pedersen, 1997) [24].

Topical or semantic similarity between two words can then be defined as the cosine between the corresponding columns of the matrix.

The assumption is that words with similar meanings will occur with similar neighbors if enough text material is available. (Schutze & Pedersen 1997, p.311) [26] there are efficiency problems with this approach: the matrix that is used to Compare each word in the vocabulary to all other words in the vocabulary tend to be quite large, and it takes quite a long time to process the word comparisons, depending on the size of the vocabulary.

Although Crouch and Yang (1992) [6] automatically generated thesaurus classes from text keywords, which can subsequently be used to index documents and queries. Crouch's approach is based on Salton's vector space model and the term discrimination theory. Documents are clustered using the complete link clustering algorithm (agglomerative, hierarchical method). Ekmekcioglu et al. [12] tested retrieval performances for 110 queries on a database of 26,280 bibliographic records using four approaches: original queries and query expansion using co-occurrence data, Soundex code (a phonetic code that assigns the same code to words that sound the same), and string similarity measure (based on similar character microstructure), respectively. The four approaches produced 509 (original queries), 526 (term co-occurrence), 518 (Soundex), and 534 (string) documents, respectively. They concluded that there were no significant differences in retrieval effectiveness among these expansion methods and initial queries. However, a close examination of their results

revealed that there was a very small degree of overlap between the retrieved relevant documents generated by the initial queries and those produced by the co-occurrence approach (19% overlap using the Dice coefficient). This suggests that search performance may be greatly improved if a searcher can select and use the terms suggested by a co-occurrence thesaurus in addition to the terms he/she has generated.

Several research groups have experimented with an algorithmic approach to cross-domain term switching recently. Chen et al. experimented extensively in generating, integrating, and activating multiple thesauri (some were existing thesauri, others automatically generated, all in computing-related areas) [9] [11]. Both Kim and Kim [18] and Chen et al. [9] proposed treating (automatic and manually-created) thesauri as a neural network or semantic network and applying spreading activation algorithms for term-switching. Despite questions about the usefulness of automatic thesaurus browsing heuristics [15], our recent experiment revealed that activation-based term suggestion was comparable to the manual thesaurus browsing process in document recall and precision, but that the manual browsing process was much more laborious and cognitively demanding [11].

Collecting Terms

Thesaurus construction requires collecting a set of terms. Some of these will end up becoming preferred terms and others may not appear in the thesaurus at all in their original form, but they may suggest concepts that need to be covered in some way.

In a global strategy as in [27] the query expansion technique presented explored the lexical-semantic links in Wordnet in order to expand hierarchically related terms to the original query. In a local strategy, the top-ranked documents retrieved for a given query are examined to determine terms for query expansion.

Apart from this expansion has been carried out by replacing or adding thesaurus words or synonyms to the existing query. Research pioneer Voorhees [Voorhees, 1994] has shown that this mechanism decreases the IR performance. However, her research points out that a manually built corpus specific thesaurus can give better results.

Arabic Language Problems

The problems of Arabic language that are related to our project are:

1) A word may take several meanings, depending on it position on the text and if the text is pointed or not, so that it makes an ambiguous view.
2) several words "حاسوب" (computer), "حاسبات" (Computers), "حوسبة" (Computing),"حساب"(Computations) and "محاسبة" (accounting), have the same root "حسب" (Compute), in spit of that there meaning is differ, and our calculation are based on root only.
3) Some words may have more than one root."مخافة",(Fear) it has two roots "خفى",(Hid), and "أخفى" (Cached).
4) When we deal with pointed text is a big problem?

Co-Occurance Analysis

Co-occurrence analysis is a statistical approach, where the occurrences of terms in documents Term co-occurrence analysis is one of the approaches used in IR research for forming multi-phrase terms. Local Context Analysis, implemented as term−suggestion devices. The closer the words occur, the more significant is the co-occurrence.

Any IR system performs the following tasks [5]:

1- Deleting the stop word from the documents.
2- Extracting Stems for each term in the documents.
3- Creating the inverted file based on the root of each documents. (The root technique used is suffix prefix removal).

Two hundred and forty two Arabic documents were used to build the database of the thesaurus. These documents contain 2499 distinct terms. An inverted file of nearly size 22478 record was build. The problems faced in building the thesaurus were:

1- Compute the weight of each term in each document.
2- Compute the weight of each two terms in the same document.
3- Compute the similarity between each two terms (Compute the cluster weights).

After terms were identified in each document, we first computed the term frequency and the document frequency for each term in a document. Term frequency, tf_{ij}, represents the Number of occurrences of term j in document i. Document frequency, df_j, represents the Number of documents in a collection of n documents in which term j occurs. A few Changes were made to the standard term frequency and inverse document frequency measures.

Usually terms identified from the title of a document are more descriptive than terms identified from the abstract of the document. In addition, terms identified by the user Filters are usually more accurate than terms generated by automatic indexing. This is due To the fact that terms generated by automatic indexing are relatively noisy [10].

We then computed the combined weight of term j in document i, d_{ij}, based on the product of "term frequency" and "inverse document frequency" as follows:

$$d_{ij} = tf_{ij} * \log \frac{N}{df_j}$$

Where N: represents the total number of documents the collection.

We then performed term co-occurrence analysis based on the asymmetric "Cluster Function" developed by Chen and Lynch [8]. We have shown that this asymmetric Similarity function represents term association better than the popular cosine function.

The weighting-factor appearing in the equations below is a further improvement of our Cluster algorithm.

$$ClusterWeight(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(T_k)$$

These equations indicate the similarity weights from term T_j to term T_k, d_{ij} and d_{ik} were calculated based on the equation in the previous step. d_{ijk} represents the combined weight of both Terms T_j and T_k in document $_i$. d_{ijk} is defined similarly as follows:

$$d_{ijk} = tf_{ijk} * \log \frac{N}{df_{jk}}$$

Where tf_{ijk} represents the number of occurrences of both term j and term k in document i (The smaller number of occurrences between the terms was chosen).

df_{jk} represents the Number of documents (in a collection of N documents) in which terms $_j$ and $_k$ occur together.

In order to penalize general terms (terms which appeared in many places) in the co-occurrence analysis, we developed the following weighting schemes which are similar to the inverse document frequency function:

$$WeightingFactor(T_k) = \frac{\log \frac{N}{df_k}}{\log N}$$

Terms with a higher df_k value (more general terms) had a smaller weighting factor value, this caused the co-occurrence probability to become smaller. [10]

So here weight cluster is like the similarity in similarity thesaurus, applying the Co-Occurrence analysis and finding the weight factor between each two terms.

Expansion Process

The co-occurrence analysis started with computations of each term's document frequency (the number of documents in a collection in which a word occurs) and term frequency (the frequency of occurrence of a word in a document). Terms appearing in the title of a document were assigned higher weights than terms in the abstract or other parts of the document. Terms that had been identified by the object filters in the first step were also assigned higher weights than those identified in the automatic indexing process. The inverse document frequency was then computed with some extra features. Multiple-word terms were assigned higher weights than single word terms since the former usually convey more precise semantic meaning than the latter.

Our Co-Occurrence thesaurus was based on all the documents in the collection (Global analysis). In our research we expanded the greater 10 terms associated with greater weight Cluster and we consider its terms as an expanded term and we expand the original Query, and after we retrieve the documents we rank it, (Lu et al., 2008) suggest that ranking by relevance can result in better retrieval performance. Thus, we computed TF-IDF scores for retrieved documents (Kim and Wilbur, 2005; Lu et al., 2008) and then ranked them based on these scores. A document with a higher TF-IDF score is returned earlier in a list.

Discussion

One of the major problems of the modern IR systems is the word mismatch that concerns the discrepancies between terms used for describing documents and the terms used by the searchers to describe an information need. A way of handling the word mismatch is by using a thesaurus, which shows (usually semantic) relationships between terms. Thesauri can aid the indexer or the indexing system in choosing the correct terms to describe the contents of documents, and in normalizing the terms so that all terms are e.g. presented in singular form. In the searching process, thesauri can help the searcher to find terms to refine a query, by expansion of the original query.

Some of the relationships between terms that are handled by thesauri are narrower term (NT), broader term (BT), and related term (RT). There are some obvious problems with manually constructing thesauri. It is an expensive and time-consuming process that requires a domain-expert or an expert at document description. In domains where new research fields develop frequently, thesauri become out of date, and need to be updated, which again is time-consuming and expensive. By using documents published in the domain in question as a corpus, a thesaurus can be created and updated automatically. The terminology of the researchers of the field will be the basis of the indexing process and the assignment of index terms. There are a number of different approaches available for automatically creating thesauri, among others different kinds of statistical co-occurrence analyses. A way of following up this paper would be to go in deeper on the different approaches, and/or select the one most interesting for my future thesis project.

This study is implemented on Oracle 8i, and the project was tested on the 242 Arabic documents that were used by Hmeidi and Kanaan (1997) [14]. The user query was 59 Arabic queries in many general and scientific fields (mostly were related to computer science field) [7].

The following results were found from the study:-

1- The recall is better when using the co-occurrence thesaurus than using the traditional IR system. This result is also reported by Qiu and Frei [22].
2- The precision is almost better when using traditional IR system than using co-occurrence thesaurus in small range.
3- On average recall/precision levels, the co-occurrence thesaurus makes a good effect on the last five levels (0.5 to 1). While it has limitation on the first 5 level (0 to 0.4) this mean that traditional IR is better in the first 5 level (Figure 3)
4- Many researchers concluded that the effective of the retrieval process when we using a thesaurus will increase, when we increase the number of documents in the collection.

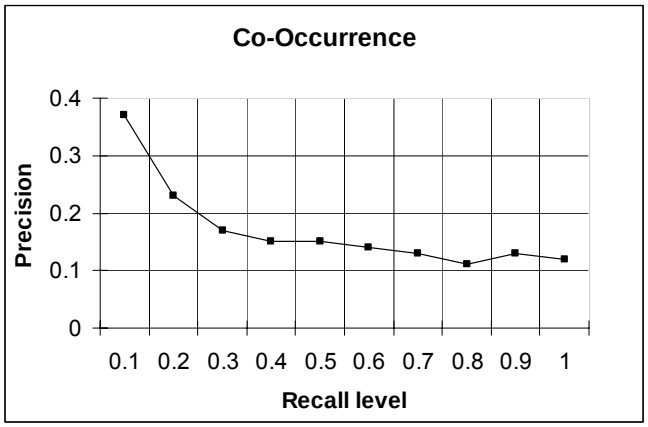

Figure 1: Using co-occurrence model

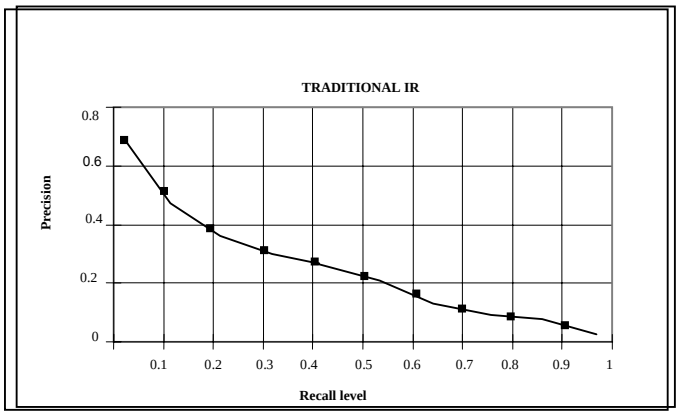

Figure 2: Traditional information retrieval

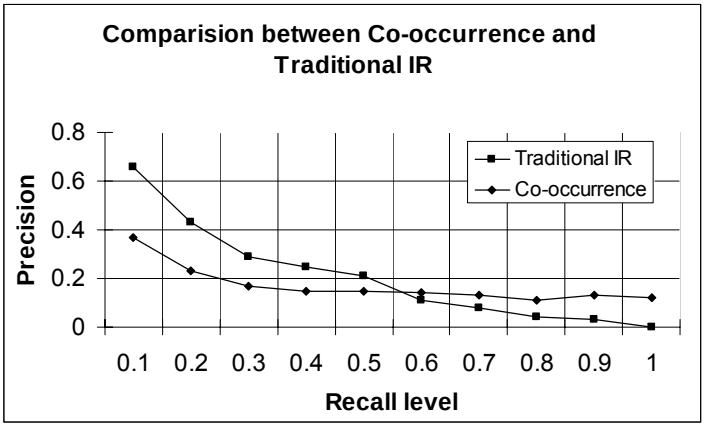

Figure 3: Comparison between co-occurrence and traditional IR model

Conclusions

In a world of increasing facing information overload, where the issue is not how many documents can be found in a particular research subject, but rather how to weed throw thousands of documents on a topic to find the most relevant ones. Based on the results of this study, the following conclusions may be drawn:

1) The co-occurrence improves the recall in a good manner
2) The co-occurrence affects the precision in a negative form
3) The co-occurrence thesaurus based on recall/precision level does not improve the effective of the retrieval task of the system. Qui and Frei [22] support this conclusion. They reported that most of query expansion methods (including co-occurrence) failed to improve the retrieval process. But on anther hand Khafajeh [16] showed that using Association thesaurus in Arabic language retrieving system has been improved the effective of the retrieval task of the system.
4) The experiments results showed that using the stemmed words improved the retrieval process when they were used by co-occurrence analysis. While when the full words were used in the traditional system, the system's performance was the worst in the continuous retrieval process, because the precision values decreased in a remarkable way. When the recall values increased, mostly the precision values reached to zero. But, in the same system with using the stemmed words, its performance degraded less sharply.
5) Finally, we present some of the future works that can be achieved. These works are related to anther techniques for using query expansion. Especially, there are many query expansion methods that are not applied on the Arabic corpuses. Continuing our program of studying different methods of query expansion in Arabic information retrieval (AIR), we may examine the effects of varying methods of term suggestion for user−controlled query expansion such as Relevance Feedback, and improving automatic method to build Arabic corpus.

References

[1] Abdelali, A. Localization in Modern Standard Arabic. Journal of the American Society for Information Science and technology Volume 55, Number 1, 2004.
[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] Amir Hossein Jadidinejad, Hadi Amiri," Local Cluster Analysis as a Basis for High-Precision Information Retrieval ,"In Proceeding of INFOS2008 International Conference on Informatics and Systems ,Egypt, 2008
[4] Baeza-Yates, Ricardo And Berthier Ribeiro-Neto. "Modern Information Retrieval". Addison-Wesley, New York City, NY, ACM Press, 1999.
[5] Baeza-Yates, R. and Ribeiro-Neto, B. (1999) "Modern Information Retrieval". Addison Wesley.
[6] Crouch, C. and Yang, B. (1992) " Experiments in Automatic Statistical Thesaurus Construction" In Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp 77-88, Copenhagen, Denmark, June 21-24.
[7] Crouch, C. J. (1990) "An approach to the automatic construction of global thesauri, Information Processing and Management," Vol. 26, No. 5, pp. 629-640

[8] Chen, H. and Lynch, K. J. (1992) "Automatic construction of networks of concepts characterizing document databases" IEEE Transactions on Systems, Man and Cybernetics, Vol. 22, No. 5, pp. 885-902.
[9] Chen, H., and Lynch, K., et al.(1993) "Generating, integrating, and activating thesauri for concept-based document retrieval," IEEE EXPERT, Special Series on Artificial Intelligence in Text-based Information Systems, Vol. 8, No. 2, PP. 25-34.
[10] Chen, H., Ng, T. D., et al. (1997). "A Concept Space Approach to addressing the word mismatch in scientific Information Retrieval: An experiment on the Worm Community System." Journal of the American Society for Information Science, Vol. 48, No. 1, pp. 17-31.
[11] Chen, H., and Ng, D., (1995) "An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound vs. connectionist Hopfield net activation," Journal of the American Society for Information Science, Vol. 46, No. 5, PP. 348-369.
[12] Ekmekcioglu, F., Robertson, A., et al. (1992) "Effectiveness of query expansion in ranked-output document retrieval systems," Journal of Information Science, Vol. 18, PP. 139-147.
[13] http\\www.ai.bpa.arizona.edu/papers.
[14] Hmeidi, I., kanaan, G., Evans, M. (1997) "Design and Implementation of Automatic Indexing for information Retrieval with Arabic Documents. Journal of the American Society of information science. Vol. 48, No. 10, pp. 867-881.
[15] Jones, S., et al. (1995) " Interactive thesaurus navigation: intelligent rules OK?" Journal of the American Society for Information Science, Vol. 46, No. 1, PP. 52-59.
[16] Khafajeh, H. et al. Automatic Query Expansion for Arabic text retrieval based on Association and Similarity thesaurus, European Mediterranean & Middle Eastern Conference on Information Systems (EMCIS 2010), April 12-13, 2010,
[17] Kim W, Wilbur WJ. A strategy for assigning new concepts in the MEDLINE database. AMIA Annu, Symp Proc 2005:395–399. [PubMed: 16779069], Lindberg D, Humphreys B, Mccray A. The unified medical language system. Methods Inf Med 1993;32, (4):281–291. [PubMed: 8412823]
[18] Kim, Y., and Kim, J. (1990) "A model of knowledge based information retrieval with hierarchical concept graph," Journal of Documentation, Vol. 46, PP. 113-116.
[19] Lu Z, Kim W, Wilbur WJ. Evaluating relevance ranking strategies for medline retrieval. in press, 2008
[20] Miller, U. (1997) "Thesaurus construction: problems and their roots." Information Processing and Management Vol. 33, No. 4, pp. 481-493.
[21] Peat, H. J. and Willett, P. (1991) "The limitations of term co-occurrence data for query expansion in document retrieval systems," Journal of the American Society for Information Science, Vol. 42, No. 5, pp. 378-383.
[22] Qiu, Y. and Frei, H. P. (1993) "Concept based query expansion". In Proc. 16th Int'l Conference on R&D in IR (SIGIR), pp. 160-169.
[23] Roget, P. (1988) Roget's II the new Thesaurus, Houghton Mifflin Company, Boston, USA.
[24] Salton, G. (1989) Automatic Text Processing. Addison-Wesley Publishing Company, Inc., Reading, MA, USA.
[25] S. Battiato, G. M. Farinella, G. Impoco, O. Garretto, and C. Privitera "Cortical Bone Classification by Local Context Analysis", Conference: MIRAGE - MIRAGE , pp. 567-578, 2007.
[26] Schutze, H., and Pedersen, J. (1994) "A Co occurrence-based Thesaurus and Two Applications to Information Retrieval," In Proceedings of RIAO, pp. 266-274.
[27] Voorhees, E. (1994). Query Expansion using lexical-semantic relations. In Proceedings of ACM SIGIR International Conference on research and development in Information Retrieval, pp. 61-69.