THESAURUS AND QUERY EXPANSION

Hazra Imran[1] and Aditi Sharan[2]

[1]Department of Computer Science, Jamia Hamdard ,New Delhi ,India
himran@jamiahamdard.ac.in
[2]School of Computers and System Sciences, Jawaharlal Nehru University, New Delhi, India
aditisharan@mail.jnu.ac.in

ABSTRACT

The explosive growth of the World Wide Web is making it difficult for a user to locate information that is relevant to his/her interest. Though existing search engines work well to a certain extent but they still face problems like word mismatch which arises because the majority of information retrieval systems compare query and document terms on lexical level rather than on semantic level and short query: the average length of queries by the user is less than two words. Short queries and the incompatibility between the terms in user queries and documents strongly affect the retrieval of relevant document. Query expansion has long been suggested as a technique to increase the effectiveness of the information retrieval. ***Query expansion is the process of supplementing additional terms or phrases to the original query to improve the retrieval performance.*** *The central problem of query expansion is the selection of the expansion terms based on which user's original query is expanded. Thesaurus helps to solve this problem. Thesaurus have frequently been incorporated in information retrieval system for identifying the synonymous expressions and linguistic entities that are semantically similar. Thesaurus has been widely used in many applications, including information retrieval and natural language processing.*

.

KEYWORDS

Network Protocols, Wireless Network, Mobile Network, Virus, Worms &Trojon Thesaurus, Automatic query expansion, Local context analysis, Information retrieval

1. INTRODUCTION

Millions of user's use web search engines to meet their information needs. Advances in web search effectiveness are perhaps the most important outcome of information retrieval research. The conventional role of an information retrieval system is to inform users about the existence of documents related to their request. Though existing search engines work well to a certain extent but they still face problems like word mismatch [1], which arises because the majority of information retrieval systems compare query and document terms on lexical level rather than on semantic level and short query: the average length of queries by the user is less than two words [2]. However search engines commonly return many irrelevant documents especially when user's queries are not specific enough. According to the Spink and Saracevic[3] there are five sources of search terms selection in query formulation and expansion. These sources are: the question statement, user interaction intermediary, thesaurus, the human intermediary and term relevance feedback. In this paper we will discuss one of the sources of term selection "thesaurus" in detail. In its simplest form **the treasury consists of:** ***i) a precompiled list of important words in a given domain of knowledge and ii) for each word in this list ,a set of related words.*** Related words are, in its most common variation, derived from a synonymity relationship.The paper is divided into 5 sections. Section 2 addresses some basic issues related to query expansion. In Section 3 we will discusses some important strategies that have been used for automatic query expansion. Section 4 highlights the types and the construction of thesaurus and finally we conclude in the Section VI.

2. BASIC ISSUES IN QUERY EXPANSION

There are three different ways of expanding the query: Manual [4], Interactive and Automatic.

Manual and Interactive query expansion requires users involvement. Sometime user may not be able to provide sufficient information for query expansion, therefore query expansion methods are needed which do not require user's involvement. **Automatic query expansion is the process of supplementing additional terms or phrases to the original query to improve the retrieval performance without user's intervention.** It seemed more effective to users for simpler search tasks whereas interactive query expansion appeared more productive for more complex search tasks [5].No matter which method is used, the key point is to get the best refinement words that are used to expand the query.We identified following as the basic issues dealing with query expansion: ***Source of term selection, criteria for term selection, construction of thesaurus and weighting/re weighting of the query terms.*** Therefore in this section we focus on these aspects.

2.1. Source of Term Selection

Sources used for query expansion can be grouped into 3 categories: Approaches based on the information derived from set of documents initially retrieved(Local),Approaches based on global information based on document category(Global),Approaches based on user's feedback. Further some of the query expansion methods rank the passages instead of the document [6]

2.2. Criteria of Term Selection

Query expansion is done by selecting those terms that are related to the query terms and help in increasing retrieval efficiency. Such terms might be synonyms, stemming variations, terms that co-occur with query terms or terms which are close to query terms on the text. Some of the important criteria for term selection include: *Simple use of co-occurrence data, Document classification, and syntactic context[7] and Relevance feedback.* [8]

2.3. Construction of Thesaurus

A thesaurus is a classification system compiled of words or phrases organized with the objective of facilitating the user's idea .In section 4 we highlight the construction of thesaurus in detail.

2.4. Term Weightage

In some cases all the terms in the query may contain an equal weight i.e. all terms are of equal importance. However many times it happens that different weights have to be assigned to various terms depending on its importance. There are different ways of assigning weights to query terms [8].

3. STRATEGIES FOR AUTOMATIC QUERY EXPANSION

Automatic query expansion has been a target of research for decades and a lot of methods have been proposed. This section presents and analyzes some important strategies that have been used for automatic query expansion. We are discussing the strategy considering the source of selection. (Local Vs Global)

3.1. Global Vs Local Automatic Query Expansion

Automatic query expansion has been a target of research for decades and a lot of methods have been proposed. This section presents and analyzes some important strategies that have been used for automatic query expansion. We are discussing these strategies considering two points of views: One dealing with source of selection (Local Vs Global) and other dealing with criteria for selecting terms and constructing thesaurus of similarity terms. A query expansion method based on global analysis usually builds a thesaurus using the word co-occurrence and the relationship in the corpus as a whole to assist users to reformulate their queries. One of the earliest global analysis techniques is term clustering[7,9].Queries are simply expanded by adding similar terms in the same cluster formed according to term co-occurrences in documents. Their work suggested better performance for query expansion with term clustering than with unclustered terms. A query expansion model using a global similarity thesaurus is presented by Qui and Frei[10]. Another work based on a global statistical thesaurus is [11], which first clusters documents and then selects low frequency terms to represent each cluster. Although the global analysis techniques are relatively robust, the corpus-wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it focuses only on the document side and does not take into account the query side, global analysis only provide a partial solution to the term mismatch problem.

Another group of techniques for query expansion is called local analysis, which extracts expansion terms from the top ranked documents retrieved by the original query[12,13].Local techniques have shown to be more effective than global technique but existing local techniques are not robust and can seriously hurt retrieval when few of retrieved documents are relevant.

Xu and Croft[13][14] have suggested a new technique ***Local Context Analysis*** which applies a co-occurrence measure of terms to local feedback and takes advantages of both global and local analysis. Local Context Analysis ranks the concepts (it can be simple term, phrases) according to their co-occurrences within the top ranked documents with the query terms and uses the top ranked concepts for query expansion. Following are the steps by which query is expanded using local Context Analysis

1. Top n ranked passages are retrieved using the information retrieval system.
2. Concepts in retrieved passages are ranked according to the formula

$$bel(Q,c) = \prod_{t_i \in Q} \left(\partial + \log\left(af\left(c,t_i\right)\right) idf_c / \log(n)\right)^{idf_i}$$

Where

Q is query

c is concept

$$af\left(c,t_i\right) = \sum_{j=1}^{j=n} ft_{ij} fc_j$$

$$idf_i = \max\left(1.0, \log 10\left(N / N_i\right) / 5.0\right)$$

$$idf_c = \max\left(1.0, \log 10\left(N / N_c\right) / 5.0\right)$$

ft_{ij} is the number of co-occurences of t_i in passage p_j

fc_j is the number of co-occurences of c in passage p_j

N is the number of passages in the collection

N_i is the number of passages containing t_i

N_c is the number of passages containing c

∂ is 0.1 to avoid zero bel value.

The above formula is a variant of the tf-idf measure used by most information retrieval system.

3) m top ranked passages are added to the query

Advantage of local context analysis is that it is computationally practical and once the top ranked passages are available query expansion is fast. But, it is based on the hypothesis that a frequent term from the top-ranked relevant documents will tend to co-occur with all query terms within the top-ranked documents which is not always true.

The role of thesaurus is very important in query expansion[15][16][17] so in next section we will discuss different type of thesaurus and their construction.

4. TYPES OF THESAURUS

4.1. Global Vs Local Thesaurus

In a global approach, thesaurus classes are constructed based on word co-occurrence and their relationship in the corpus as a whole and these classes are *used to index both documents and queries whereas the local thesaurus [12,18] uses information obtained from the top rank documents retrieved in response to a particular query Thus a global thesaurus is constructed prior to the indexing process whereas a local thesaurus is constructed dynamically during query processing and uses information retrieved in response to a specific query to modify only that query [12].*Although the global analysis techniques are relatively robust, the corpus-wide statistical analysis consumes a considerable amount of computing resources. Moreover, since it focuses only on the document side and does not take into account the query side, global analysis only provide a partial solution to the word mismatch problem.

4.2. Manual Vs Automatic thesaurus

The approaches for constructing thesaurus can be broadly classified into four groups: General purpose Hand crafted thesaurus (Manual thesaurus), Co-occurrence based automatically constructed thesaurus, Similarity based automatically constructed thesaurus, Head Modifier based automatically constructed thesaurus. Construction of manual thesaurus is very labor intensive work and it lacks domain specific terms as it generally cover general terms.Hand-crafted thesaurus describe the synonymous relationship between words[19].Voorhees [20] performed query expansion using Wordnet, a manually constructed network of lexical relationships, and founds that expansion helps only for very short queries. Therefore there is need of automatically generated thesaurus. Automatically constructed thesaurus has shown

improvements in retrieval performance [21]. They are based on the co-occurrence information, linguistic information and relevance judgment information.

In *Co-occurrence based automatically constructed thesaurus*, similarity between terms are first calculated based on association hypothesis and then used to classify terms by selecting a similarity threshold value. In this way the set of index term is divided into classes of similar term. The query is then expanded by adding the terms of classes that contain query term. Such strategies are based on local clustering of terms

A similarity based automatically constructed thesaurus is built considering the term to term relationship rather than simple co-occurrence data. Terms for expansion are selected based on similarity to the whole query rather than their similarity to individual term.

In Head –modifier based automatically constructed thesaurus term relations are gathered on the basis of linguistic relations [22].

4.2.1. Wordnet

Many researchers in area of information retrieval have long used WordNet. The system was developed at Princeton by a group led by Miller [23]. It has power of both an on-line thesaurus and on-line dictionary. The basic building block in WordNet is synset (set of synonyms). WordNet partitions the lexicon into Nouns, Verbs, adjectives and adverbs, all are organized into form of synsets. A synset represents a concept in which all words in a synset are interchangeable in some syntax. Knowledge in a synset includes the definition of these words (glosses) as well as pointers to other related synsets. The purpose is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. WordNet has been used for a number of different purposes in information systems, including word sense disambiguation, information retrieval, automatic text classification, automatic text summarization.[24]

We will discuss the construction of Wordnet and Concept based Thesaurus in detail.

1. Words are organized into taxonomy where each node is the set of synonyms. Following are the four taxonomies

1. Hyponym/Hypernym(IS-A/HAS A)
2. Meronym/Holonym(Part-of/Has-part)
3. Meronym/Holonym(Member-of/Has-Member)
4. Meronym/Holonym(Substance-of/Has-substance)

2. The similarity between words a and b can be defined as the shortest path from each sense of a to each sense of b, as below

$$sim_{ab} = \max\left[-\log\left(\frac{N_p}{2D}\right)\right]$$

Where N_p is the number of nodes in path p from a and b

D is the maximum depth of the taxonomy.

3. Similarity values are normalized.

4.2.2. Concept based Thesaurus

The Similarity based thesaurus makes it possible to expand the complete query concept rather than the individual terms separately [25]. *A similarity thesaurus* is built considering the term to term relationship rather than simple co-occurrence data. Terms for expansion are selected based on similarity to the whole query rather than their similarity to individual term. To construct this thesaurus the terms of collection are considered documents and the documents are used as index terms. To understand the idea of concept based thesaurus let us consider, T as the set of indexing terms and q as the query given by the user containing terms t and t`. In Fig 1 the pair wise similarity is represented as fine lines. The closer two linked terms are to each other, the more similar they are. The query concept q_c can be obtained by calculating the centroid of q.

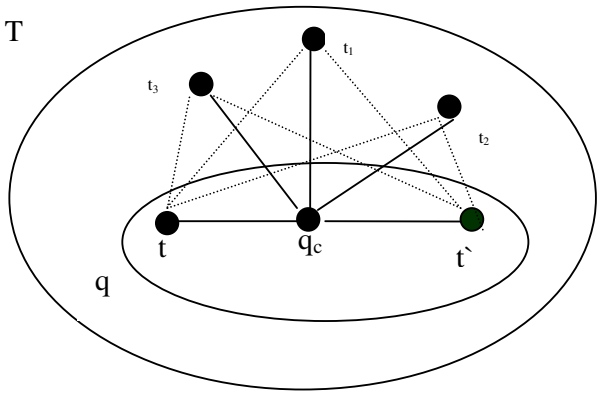

Fig. 1

In construction of similarity thesaurus, a mechanism is used by which each term in the collection is represented as a vector in document vector space.Each term t_i in the collection of m terms will be represented by a vector $\vec{t_i} = (d_{i1}, d_{i2} \mathrm{K})^T$ where d_{ik} is the weight of index document d_k in the representation of the term t_i . tf-idf scheme is used to calculate the value of d_{ik} The weights are normalized [24]. Now we discuss how to construct Similarity thesaurus:

Each term t_i in the collection of m terms is represented as vector of N components in the vector space of documents, $\vec{t_i} = (p_{i1,} p_{i2} \ldots p_{iN})$ where p_{ij} denotes the weight of the index document d_j in the representation of the term t_i .Use tf-idf to compute the value of p_{ij} .

$$p_{ij} = \frac{\left(0.5 + 0.5\frac{f_{ij}}{\max_k(f_{ik})}\right).itf_j}{\sqrt{\sum_{u=1}^{N}(0.5 + 0.5\frac{f_{iu}}{\max_k(f_{ik})})^2.itf_u^2}}$$

Where

f_{ij} is the number of times term t_i appears in document d; $\max_k(f_{ik})$ is the maximum of the frequency values for the term t_i in document collection; $itf_j = \log \frac{m}{|d_j|}$ is the inverse term frequency for document d_j.

Calculation of inverse document frequency shows that a short document plays more important role than a large one. Then Query q is represented in term of terms present in query as $\left(\sum_{t_i \in q} q_i, \vec{t}_i \right)$ where q_i represent the weight of the i^{th} term. Similarity between two terms can be calculated as Scalar product of t_i and t_j :

$$SIM(t_i, t_j) = \vec{t}_i^T * \vec{t}_j = \sum_{k=1}^{N} p_{ik} \cdot p_{jk}$$

Calculation for all term pairs in the collection produces similarity thesaurus. It is a symmetric matrix with values between 0 and 1. Its construction is computationally costly although it is done only once. If new documents are added to the collection the values for terms included in new document must be updated

Now following are the steps to expand the query using similarity thesaurus

1. Represent each term t in document vector space as discussed above.
2. Represent query q as discussed above
3. Calculate the similarity between query concept and each term using following formula

$$sim(q,t) = \vec{q}^T * \vec{t} = (\sum_{t_i \in q} q_i \vec{t}_i)^T * \vec{t} = \sum_{t_i \in q} q_i \cdot (\vec{t}_i^T * \vec{t}) = \sum_{t_i \in q} q_i \cdot SIM(t_i, t)$$

In fact the values have already been calculated and can be obtained form similarity thesaurus constructed earlier. The first r terms (according to decreasing order of similarity) of the collection or terms crossing a similarity threshold can be used for query expansion.

Research have shown that combination of manually and automatically constructed thesauri has a positive effects on the query expansion process[26]. Salton's works on the automatic thesaurus construction and query expansion [18],Rijsrergen's work on co-occurrence[27] and Minker [7] came to the conclusion that automatic term classification with relevance judgment can produce significant improvements. Crouch and Yang[11,19] build a term-vs-term thesaurus that produced significant improvements.

5. CONCLUSION

A well-constructed thesaurus has been recognized as a valuable tool in the effective operation of an information retrieval system as expansion of queries with related terms using thesaurus can improve performance. Some of the suggested directions include: extending the concept of similarity based thesaurus where instead of frequency based representation both the query and the documents can be represented in terms of concept using some knowledge based representation, using different types of similarity measures for finding the similarity between terms such as : dice coefficient; Jaccard coefficient; use of semantic network and ontology for thesaurus construction, use of soft computing methods like Genetic Algorithm and Neural Network for thesaurus construction. In this paper we have tried to address basic issues related to the query expansion along with some important strategies that have been used for automatic query expansion. We have tried to present the concept behind automatic thesaurus, identifying important types of thesaurus, providing an outline of steps for their construction and their use in query expansion. We think that there is a large scope for improving the technique for automatic query expansion. Considering the source of selection we think local context analysis is an important technique that combines global analysis with local feedback. The idea can further be extended to increase the efficiency of local context analysis. Further passage based query expansion, query expansion using N grams and query expansion using heterogeneous thesaurus are important directions to target automatic query expansion.

REFERENCES

[1] Jinxi Xu., "Solving the word mismatch problem through automatic text analysis.", Ph.D. Thesis, Department of Computer Science, University of Massachusetts, Amherst, MA, USA, May 1997.

[2] J.R.Wen, J.Y.Nie and H.J. Zhang, "Clustering *User Queries of a Search Engine",*. In Proc. of WWW10, pp. 587-596, 2001

[3] Spink,A and Saracevic,T., "Interaction in information retrieval:selection and effectiveness of search terms" , Journal of the American Society for Information Science 48 ,No 8,p.741-761,1997

[4] Harter, Stephen P, "Online Information Retrieval: Concepts, Principles, and Techniques", Orlando: Academic Press, 1986.

[5] H. Fowkes and M. Beaulieu., "Interactive searching behavior: Okapi experiment for TREC-8.", Proceedings of 22nd BCS-IRSG European Colloquium on IR Research, Electronic Workshops in Computing. Cambridge. 2000.

[6] Callan J, "Passage level evidence in document retrieval", In Proceedings of the Seventeeth Annual International ACM SIGIR Conference on Research and Development in Information retrieval,pages 302-310,1994 .

[7] Minker, J., Wilson, G.A., Zimmerman, B.H., "An evaluation of query expansion by the addition of clustered terms for a document retrieval system. Information Storage and Retrieval, 8:329--348, 1972

[8] Grossman, D.A. and Frieder, O., "Information Retrieval: Algorithms and Heuristics.", Kluwer,1998

[9] Ruge, G.,"Experiments on linguistically-based term associations", Information Processing & Management, 28(3): 317-32, 1992.

[10] Y. Qiu and H.P. Frei, "Concept-based query expansion", in SIGIR ,1993.

[11] C. J. Crouch and Bokyung Yang , "*Experiments in automatic statistical thesaurus construction",* SIGIR 92,77-88,1992.

[12] Baeza-Yates, R. and Berthier Ribiero—Neto,"Modern Information Retrieval.", Addison Wesley,1999

[13] Jinxi Xu and W. Bruce Croft., "Query expansion using local and global document analysis.", In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 4--11, 1996.

[14] J. Xu, W. B. Croft, Improving the effectiveness of information retrieval with local context analysis, ACM Transactions on Information Systems,2000.

[15] Blocks, D., Binding, C., Cunliffe, D., Tudhope, D., "Qualitative evaluation of thesaurus-based retrieval", in Agosti, M., Thanos, C. (Eds),Proceedings of 6th European Conference on Research and Advanced Technology for Digital Libraries, Springer, Berlin, Rome, 16-18 September (Lecture Notes in Computer Science), pp.346-61,2002.

[16] Sihvonen, A., Vakkari, P., "Subject knowledge improves interactive query expansion assisted by a thesaurus", Journal of Documentation, Vol. 60 No.6, pp.673-90., 2004
[17] Shiri, A.A., Revie, C., "Query expansion behaviour within a thesaurus-enhanced search environment: a user-centred evaluation", Journal of the American Society for Information Science, Vol. 57 No.2,2006
[18] G. Salton, Automatic Information Organization and Retrieval, *McGraw-Hill Book Company*, 1968.
[19] Wang,Y.,Vandendorpe,J.,and Evens ,M., "Relational thesauri in information retrieval", Journal of the American Society for Information Science,36(1):15-27,1985.
[20] E.M. Voorhees., "Query expansion using lexical-semantic relations", In Proceedings of the 17th ACM-SIGIR Conference, pp. 61-69, 1994
[21] Jing .Y and Croft,w.Bruce, " The association thesaurus for information retrieval", RIAO'94,Intelligent Multimedia Information Retrieval Systems and Managemnet,146-160,Paris France,CID,1994.
[22] D. Hindle.,"Noun classification from predicate-argument structures.", In Proceedings of 28th Annual Meeting of the ACL, pp. 268-275, 1990.
[23] Miller, G.A., Beckwith, R.T., Fellbaum, C.D., Gross, D., and Miller, K." *WordNet: An On-line Lexical Database",* International Journal of Lexicography, 3(4):235-244,1990
[24] Francisco Joao Pinto et al, "Joining automatic query expansion based on thesaurus and word sense disambiguation using WordNet", International Journal of Computer Applications in Technology, Volume 33, Pages 271-279 ,2009
[25] Y. Qiu and H.P. Frei, "Concept-based query expansion", in SIGIR ,1993.
[26] Ding,Y.,Ghowdhury,G.G. and Foo,S., " Incorporating the results of co-word analyses to increase search variety for information retrieval",Journal of Information Science,26,429-451,2000 .
[27] C.J. Rijsbergen, D. J. Harper, and , M. F. Porter , "*The selection of good search terms*.", Information Processing and Management 17: 77 – 91 , 1981

Authors

Hazra Imran is a Ph.D. scholar at the School of Computer and Systems Sciences (SC & SS), Jawaharlal Nehru University (JNU), New Delhi, India, and also working as lecturer in Department of Computer Science, JamiaHamdard, New Delhi, India (e-mail:hazrabano@gmail.com).

Aditi Sharan is an assistant professor at the School of Computer and Systems Sciences (SC & SS), Jawaharlal Nehru University (JNU), New Delhi, India. Her research areas include information retrieval, text mining, Web mining; email:aditisharan@mail.jnu.ac.in).