TECHNIQUES OF QUERY REFORMULATION IN INFORMATION RETRIEVAL

[1] MR. D.P.JOSHI, [2]MR. R. D. DOSHI

[1]M.E.[C.E.] Student,V.V.P Engineering College,Rajkot,Gujarat
[2]M.E.[C.E.] Student,V.V.P Engineering College,Rajkot,Gujarat

ddipesh4@gmail.com, er.rushabhdoshi@gmail.com

ABSTRACT: *Query expansion methods have been studied for a long time- with debatable success in many instances. In this paper we present different query expansion technique based on a similarity thesaurus which was constructed automatically. A similarity thesaurus reflects domain knowledge about the particular collection from which it is constructed. We address the two important issues with query expansion: the selection and the weighting of additional search terms in different approach. In contrast to earlier methods, our queries are expanded by adding those terms that are most similar to query terms.*

Keywords—query expansion, query reformulation, information retrieval, relevance feedback.

I: INTRODUCTION

The Web is an innovation that has modified the way we learn, work and live. The novelty lies not only on the freedom to publish, but also in the almost universal communication facilities. It marks the beginning of a new era, of a new society, started by what we may call the information revolution.

In these new times, the volume of information that can be accessed at low cost and high convenience is mind boggling. Volume of information tends to become even larger. As a result, information of critical value appears frequently mixed in with other pieces of information that are not of interest. Because the volume of data is now much larger and frequently poorly organized, finding useful or relevant information might be rather difficult. In fact, this is the case even with modern search engines that take advantage of link analysis to identify popular sources of relevant information.

There is general consensus that web users have poor specifications their data needs in general. Either because they are in a hurry, or because they do not understand well the search process, Web users often specify, short queries with little or no context Information associated with them. They also are the only men at the moment, to decide what is relevant and what is not. The Combination of these two factors, suggesting that the interaction with Users is a crucial step if the precision of the results can be improved. In other words, to improve the user information search we have to ask for more information from him.

In this work, we focus our attention on the problem of improving Query creation process. Our approach provides a high additional reference to the original user query suggestions level for expansion. This is done using information extracted from a log of past queries – an important piece of evidence that is generated in abundance by search engines.

II: QUERY REFORMULATION

Query Reformulation is the process of reformulating a seed query to improve retrieval performance in information retrieval. In the context of Web search engines, including query expansion evaluation of the user's input (which words are entered into the search field and sometimes other types of data) and query expansion match additional documents.

The aim of query reformulation is to reduce this query/document mismatch by expanding the query using words or phrases with a similar meaning or some other statistical relation to the set of relevant documents. This procedure may have even greater importance in spoken document retrieval, since the word mismatch problem is heightened by the presence of errors in the automatic transcription of spoken documents.

Query expansion enables the search appliance to automatically add extra terms to a user's search query, in order to return additional relevant results. When query expansion is enabled, the appliance can expand two types of terms:

1. Words that share the same word stem as the word given by the user. For example, if the user search query includes "engineer," the search appliance could add "engineers" to the query. Query expansion behaviour is context sensitive. The search term "engineer" alone might not be expanded, but "software engineer" is expanded to include "engineers."
2. Terms of one or more space-separated words that are synonymous or closely related to the words

given by the user. For example, if a user searches for "FAQ," the appliance could add "frequently asked questions" to the query, or if a user enters "office building," the query could expand to include "office tower."

III: Techniques of Query Reformulation

Study of query reformulation give the idea of different approach for it and it will expand the query by using synonyms of words, and searching for the synonyms, various morphological forms of words by stemming each word in the search query, Fixing spelling error and automatically searching for the corrected form or suggesting it in the results, Re-weighting the terms in the original query.

1. Simple use of co-occurrence data: The similarities between terms are first calculated based on the association hypothesis and then used to classify terms by setting a similarity threshold value In this way, the set of index terms is subdivided into classes of similar terms. A query is then expanded by adding all the terms of the classes that contain query terms. It turns out that the idea of classifying terms into classes and treating the members of the same class as equivalent is too naive an approach to be useful.

2. Use of document classification. Documents are first classified using a document classification algorithm. Infrequent terms found in a document class are considered similar and clustered in the same term class(thesaurus class). The indexing of documents and queries is enhanced either by replacing a term by a thesaurus class or by adding a thesaurus class to the index data. However, the retrieval effectiveness depends strongly on some parameters that are hard to determine. Furthermore, commercial databases contain millions of documents and are highly dynamic.
The number of documents is much larger than the Number of terms in the database. Consequently, document classification is much more expensive and has to be done more often than the simple term classification mentioned in 1.

3. An automatically derived thesaurus. Thesaurus is developed automatically using co-occurring terms or grammatically related terms. Terms which co-occur quiet frequently in a corpus are more likely to be related. The other approach is to analyze the corpus for grammatical dependencies. For example, entities that grow, walk or move, are more likely to be living organism or more specific, to be humans.

4. Query reformulations based on query log mining. Large amount of user interaction information is available to the web search engines. This information is stored in query logs and can be used to improve the user satisfaction of the users later.

IV: RELEVANCE FEEDBACK

Relevance feedback is a straightforward strategy for reformulating queries. In a relevance feedback cycle, the user is presented with a list of initial results. After examining them, the user marks those documents he or she considers relevant. The original query is expanded according to these relevant documents. The expected result is that the next round of retrieval will move toward the relevant documents and away from no relevant documents.
Typically, expansion terms are extracted from the relevant documents judged by the user. Relevance feedback can achieve very good performance if the user provides sufficient and correct relevance judgments. Unfortunately, in a real search context, users usually are reluctant to provide such relevance feedback.

V: CONCEPTUAL QUERY EXPANSION

In conceptual theories, a concept is expressed by words occurred to a person when he was to express his feeling or idea. There is a one-to-many mapping between the set of concepts and the set of words in WordNet. Even there may not be unique mappings between words and concepts in different expressing process, some mappings must be similar to some other with regard to the vocabulary used in order for communication to occur.
When a user searches for a term, there is more to the "query" than what is actually entered. Humans think in terms of concepts but the search is performed using words. Many times the query terms are ambiguous words, unable to fully represent the concept that the user has in mind. The intended meaning of such words is described by other words commonly occurring in the vicinity or context of these words.
In sentence queries, the sense of words are not only determined by its definitions, but also determined by the relationships with other words in the sentence [6]. Longer queries with punctuations are split into sentences by punctuations and sentences with quotation marks are recovered to its original form. By this pre-processing, sentence queries are transferred to the combination of several individual sentences with each being a complete "clean" sentence with no punctuations in them so that each sentence can be handled further in the later word sense disambiguation procedure.

VI: CONCLUSION

In this paper, we present a query expansion model based on the domain knowledge contained in an

automatically constructed similarity thesaurus. still each technique of query expansion is primarily concerned with the two important problems of query expansion, namely with the selection and with the weighting of additional search terms. The term selection relies on the overall similarity between the query concept and terms of the collection rather than on the similarity between a query term and the terms of the collection. As we illustrate different techniques used for query reformulation with relevance feedback and user personalized approach.

REFRENCES

[1] M.J. Bates, "Search Techniques." Ann. Rev. of Information Science
and Technology,M.E. Williams, ed., pp. 139-169, 1981.
[2] D. Beeferman and A. Berger, "Agglomerative Clustering of a
Search Engine Query Log,"Proc. SIGKDD,pp. 407-416, 2000.
[3] G. Brajnik, S. Mizzaro, and C. Tasso, "Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support,"Proc. 19th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR'96),pp. 128-136, Aug. 1996
[4] C. Buckley, G. Salton, J. Allan, and A. Singhal, "Automatic Query Expansion Using SMART, "Overview of the Third Retrieval Conf.(TREC-3), pp. 69-80, Nov. 1994.
[5]Crouch, C.J., An approach to the automatic construction of global thesauri, Information Processing & Management, 26(5): 629-40, 1990.
[6] Ekmekcioglu, F.C., Robertson, A.M., Willett, P., Effectiveness of query expansion in ranked-output document retrieval systems, J. of Information Science, 18(2): 139-47, 1992.
[7] D. Hull, "Using Statistical Testing in the Evaluation of Retrieval Experiments,"Proc. ACM SIGIR,pp. 329-338, June 1993.
[8]Cliff Goddard.: Semantic Analysis: A Pratical Introduction. Oxford University Press(1998)
[9] R. Richardson, AF Smeaton.: Using WordNet in a Knowledge-Based Approach to Information Retrieval. Proceedings of the BCS-IRSG Colloquium, Crewe(1995)
[10] Smeaton, A.F., van Rijsbergen, C.J., The retrieval effects of query expansion on a feedback document retrieval system, The Computer Journal, 26(3): 239-46, 1983.
[11] Sparck-Jones, K., Barber, E.B., What makes an automatic keyword classification effective? J. of the ASIS, 18: 166-175, 1971.
[12] Hoeber, X.-D. Yang, and Y. Yao.: Conceptual query expansion. In Proceedings of the Atlantic Web Intelligence Conference (2005)
[13]https://developers.google.com/searchappliance/documentation/50/help_gsa/serve_query_expansion
[14] E. N. Efthimiadis, "Query Expansion," Annual Review of Information Systems andTechnology, vol. 31, pp. 121–187, 1996
[15] R. Ozcan and Y. A. Aslandogan, "Concept-based Information Access," in Proceed-ings of the International Conference on Information Technology: Coding and Com-puting, vol. 1, Apr. 4–6, 2005, pp. 794–799.
[16] Text REtrieval Conference (TREC). NIST and ARDA. [Online]. Available:http://trec.nist.gov/
[17]B. J. Jansen and A. Spink, "How are we searching the World Wide Web? A com-parison of nine search engine transaction logs," Information Processing and Man-agement, vol. 42, no. 6, pp. 248–263, 2006.