Sholeh Arastoopoor - Regional Information Center for Science and Technology, Shiraz, Iran
Rahmatollah Fattahi - Ferdowsi University of Mashhad, Iran

A More Effective Web Search through Developing a Small Thesaurus of Non-Topical Terms: A Proposed Model to Improve Pertinence and Retrieval Relevance

Abstract
Purpose: A higher level of retrieval relevance along with pertinence is what information systems are required to provide if they are to gain more user satisfaction. Although in almost every IR system, general and non-topical terms are not considered to play a major role in indexing, the underlying assumption of current study stems from the fact that these terms could be effective in better searching by end-users. Thus this paper aims at proposing a new approach in this regard using a small thesaurus of general and non-topical terms which can be used along with queries (i.e., topical keywords). This would help making users' queries closer to natural language.
Method: In the first phase of this study, a set of 669 non-topical terms, which were identified through a previous study, was assumed as the test bed. Based on the main goal of the present paper, the list was analyzed and divided into major categories. As for the second phase, a preferred label representing each category was selected and assigned based on its "use warrant"; and to this end Google Trends was applied for determining the most-frequent general and non-topical terms among users' web searches. At last the developed thesaurus was tested in searching and the retrieved results were evaluated in terms of relevance and pertinence.
Results: The findings of this study show that, although there is rather a diverse range of general and non-topical terms appearing before or after topical keywords in Web documents, users are still using no or only a small fracture of them in their search queries. The findings also show that these terms would be of great assistance in providing more relevant results and a meaningful display of the results. This paper then proposes a conceptual model for applying such a thesaurus in searching the Web.

Introduction

The nature of free text searching is different from controlled searching. Although sometimes free text searching is considered the same as keyword search, there is a slight difference between them; at least from the very nature of the queries point of view. Keyword is usually considered to be a significant word or phrase which has been extracted from title, abstract or even the fulltext of a document. Thus it has to contain or convey an important concept, subject or topic. This means the keyword search in most of the IR systems (i.e. databases) has to be topical. Yet free text search is not supposed to. The queries submitted to web search engines in this form of searching can contain also non-topical terms for the benefit of refining or expanding a query. Non-topical terms are those terms that have no particular topicality. They only come along with the keywords to give certain meanings or features to them. This would help the user to conduct some kind of faceted search. Since according to Ceri et al. (2013) user request is ambiguous, retrieved results are sometimes similar to each other, and because search results are often enriched with representative metadata, it is possible to combine text search with query refinement which is the essence of the faceted search. Thus it is possible to perform such a search using non-topical terms. For example, in phrases such as "meaning of royalties" and "more on royalties", both "meaning of and "more on" are non-topical. They have particular meaning but this meaning has no direct relation to any specific subject or topic. Yet each of them reflects a different feature of

the same topic. Other examples are "FAQ", "Statement", "articles", "report", "anniversary" and etc. Walker and Jane (1999) state that, although one-concept searches are widely used, there is always at least another concept lingering in user's mind. Walker and Jane do not mention non-topical terms in a distinctive group, but they mention terms such as "materials" or "controversy". Thus, search refinement through using non-topical terms, based on the situation, could be used for both limiting and expanding search results (query expansion). Although in almost every IR system, general and non-topical terms are not considered to play a major role in indexing, the underlying assumption of current study stems from the fact that these terms are effective, additional help in better searching by end-users (clarifying a specific aspect of the subject being searched). Thus this paper aims at proposing a new approach in this regard using a small thesaurus of general and non-topical terms which can be used along with queries (i.e., topical keywords). This would help making users' queries closer to natural language searching and thus improve pertinence and retrieval relevance. According to Soergel (1994) relevance and pertinence are defined as follows: a result is topically relevant if it can help to answer a user's question and it would be pertinent if it is topically relevant and also appropriate for the user (i.e. the user can understand it and use its information).

Literature Review

Little research has been done on the use of general and non-topical terms in searching especially in using thesauri for retrieving more relevant results. Sugiura and Etzioni (2000) prepared a prototype Q-Pilot routing system that worked based on extracting phrases and clustering them into topical and non-topical. Non-topical terms were used to get new topical terms. Another study, which aimed at use of non-topical terms in query expansion, was conducted by Chan et al. (2001) who developed a new approach to subject vocabulary for Web searching using FAST and its non-topical terms. Aside from the ASK Jeeves attempt in phrase suggestion in searching, Fattahi, Wilson and Cole (2008) in a study regarding natural language query expansion, focused on using non-topical terms based on analysis of 800pages of Web documents in two fields of health and social sciences. They found that using such terms would increase relevance and pertinence. Other more recent studies regarding this matter are carried out by Badie, Tayefeh Mahmoudi & Ghaderi (2010) and Wan etal. (2012).

Methodology

The corpus generated based on Fattahi, Wilson and Cole's study consisted of a 1071 general terms (684 from the social sciences and 387 from health domain). For the purpose of current study this corpus was used but from 1071 identified terms 402 terms were considered to be semi-topical. These terms were also omitted from the test bed. At last 669 terms were analyzed in order to prepare the small thesaurus required for this study. Thus, based on the main goal of the present paper, the list was analyzed and divided into major categories. As for the second phase, a preferred label representing each category was selected and assigned to each minor sub-category based on its "use warrant"; and to this end Google Trends was applied for determining the most-frequent general and non-topical terms among users' web searches. The obscure or less-frequent terms were also crossed off the list. Then the remaining terms along with their categories were organized based on a thesaurus structure. Finally the developed

thesaurus was tested in searching the Web and the retrieved results were evaluated in terms of relevance and pertinence according to the opinion of actual users (30 web users were interviewed).

Findings and Discussion

In order to discuss the issue in a logical and step by step manner the findings of this study are presented here in the form of answering three major questions.

1) Is it Possible to create a small thesaurus based on the non-topical terms?

As previously mentioned, 669 terms extracted from the corpus developed through Fattahi, Wilson and Cole (2008) study were considered for constructing the thesaurus in first place. For the purpose of this study and for the first phase of a more comprehensive one, the basic relationships of "use" and "related term" were assumed for studying the feasibility of thesaurus construction. In this phase the verbs and verb phrases were also excluded from the initial term list, which reduced the number of real non-topical terms to 609. Through a content analysis the terms were divided in to 8 general categories. Categorizing the terms was done through a reciprocal process through which the terms from one family were grouped together (Singulars, Plurals, different forms or speeches of the same term were put together). Then those families of terms that shared the same meaning were again grouped with each other. Each family in each group was labeled with a different color. This procedure was followed until there was not possible to merge any family in other categories. Table 1 shows the distribution of terms among these 8 categories.

Table1. Major categories and their sub-categories along with their term distribution

	Main Sub-categories	Minor Sub-categories	Preferred Terms	Total Terms
Informative Category	9	37	54	225
Information Types	4	27	26	87
Directories	7	28	17	96
Events	5	13	10	55
Monetary Features	--	5	4	24
Categorical Terms	3	7	10	37
Categories of People	--	12	35	9
Views, Debates and Arguments	3	12	9	35

As for selecting the top terms, Google trends' statistics of word usage in search queries was used as criteria. For grouping the terms researchers made the decision since the terms were general no field specialty was required. But at last the developed thesaurus was presented to a linguist.

In each category there were some minor sub-categories which were related and formed the main sub-categories on table 1. Those two categories with only one main sub-category are "monetary features" and "categories of people". Informative category is the biggest and most enriched category of terms identified. This category includes terms that deal with giving certain information about a topic. Terms ranging from

"information", "info", "examples" to "meaning", "more on", "facts and figures" and etc. come in this category. Information types category deals with those terms regarding different information based on their types, yet disregarding the very specific topic. The terms ranging from "research", "survey", "articles, "resources" to "glossary", "report" and "book" are considered as members of this category. Directories deal with different listings that might be available in different subjects. The Event terms category includes those terms such as "anniversary", "day", "month", "awards" and etc. Monetary terms category encompasses those money-related terms in different fields. Categorical terms and categories of people largely deal with different general terms such as "types", "fields" and etc. which are used mainly for categorizing groups. Views, debates and arguments have their own group of general non-topical terms which shape a different category.

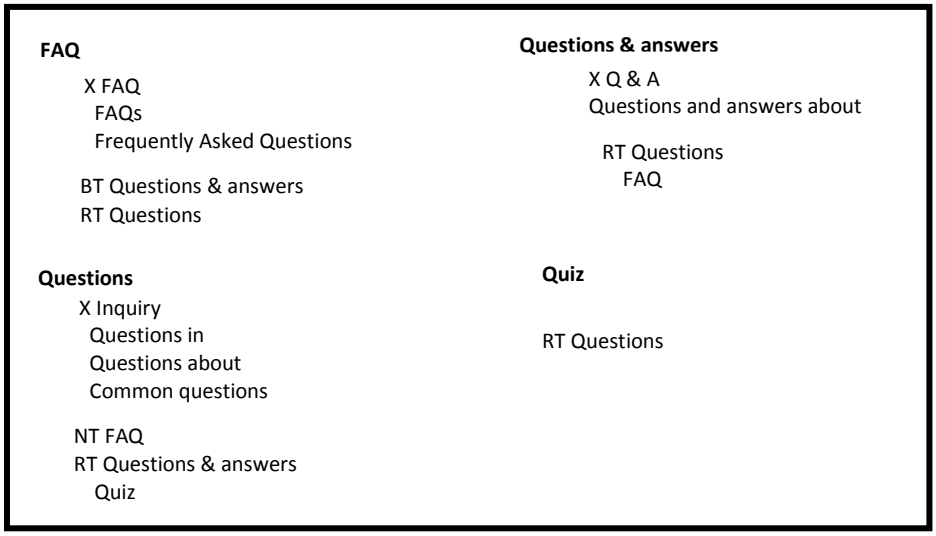

Figure1. The thesaurus relationships defined for three minor sub-categories[1]

As for checking the reliability and viability of the thesaurus and its basic structure, it was presented to a specialist[2] for control and editing. Figure 1 demonstrates three minor sub-categories of one of the main sub-categories in "informative terms".

2) Is the developed thesaurus effective in refining the search results?

During the previous phase 141 minor sub-categories were identified. For testing the developed thesaurus, 10 minor sub-categories were selected randomly (two minor sub-categories from the first two rows of the table 1 and one from other rows of that table). For each category all of the main and related terms were searched in Google combined with 3 different topical terms used by users of the RICeST[3]. These terms were "psychology", "science history" and "feminism". The minor sub-categories were

[1] The Thesaurus is currently arranged in an xls file and authors can send a copy through email upon request.
[2] A Ph.D. in General Linguistics who has worked on at least two research projects regarding thesauri and ontologies.
[3] Regional Information Center for Science and Technology, Shiraz, Iran

"issues", "about", "articles", "report", "experts", "award", "costs", "fields of", "debate", and "national". 168 search queries were submitted to Google search engine. The first page of the search results of the topic itself was retained. The first page of the search results of the topic with the preferred non-topical terms of the thesaurus was also retained. The search results of the topical term with the other not-preferred or related non-topical terms were analyzed, the redundancies were omitted and the first 10 retrieved documents were retained.

Then for the purpose of testing the effectiveness of the search, the three sets of results were presented to 30 users. They were selected from those who come to the RICeST in order to find information resources relevant to their needs. Most of them were university students. Thus 30 interview sessions were arranged with them and they were asked to compare those three sets of results with each other from the relevance and pertinence point of view. Each session took at least 30 minutes. The results show that the entire users share the same notion about that the third set of results is the most complete and also refined one. This third set was the cumulative one from searching the topic along with different non-topical terms, In other words, the list presents some good documents that are not retrieved as the first 10 most relevant results of the other two result sets. As for the second set of results (i.e., search results from preferred non-topical terms along with the topics) about 66% (20 interviewees) were positive that the relevance of the search results would increase if non-topical terms are added to the topical term. Analyzing the results deeper would show that conducting a phrase search using non-topical terms and the topic in quotation marks produce better results. In other words, "science history articles" or "essays on science history[1]" generate better results than (articles + science history) or ("essays on" + "science history"). But generating automatic phrase search is not an easy task, since some of the non-topical terms are to be used before the topic and some has to be placed after the topical term. But there are some terms like "information", which also happens to be a preferred non-topical term that cannot be added to the topical term on its own in a phrasal combination. Thus if the thesaurus terms are to be used in automatic phrase search, certain rules based on grammar have to be embedded in conjunction to the thesaurus terms.

The most significant finding of this study is that, based on the users' opinion, although searching topical terms along with non-topical preferred terms is beneficial, using all the related terms in the form of not preferred, related or narrower terms produces better results. This finding is enforced with the previous finding that 66% of the users were satisfied with the second set of results but all of them were satisfied with the third set of results.

3) What is the users' impression about searching topical terms along with the non-topical terms?

Users were asked to answer whether they have previously used non-topical terms in their search or not. Only two of them replied that they sometimes do but they only use one non-topical term and since usually this procedure does not help them they prefer to use the topical terms only. All of them affirmed that they do not see any importance or informational value in non-topical terms. For example, one of them said: "for the one who wants to conduct a research about a subject such as feminism the documents regarding its meaning, dimensions, even costs, debates and all these sort of things are

[1] Based on the constructed thesaurus "essays" and "essays on" are two other terms for the preferred term "articles". During the search using the thesaurus all of the results of these were also included.

important; so I think when I search only for the term “feminism” I suppose I would retrieve all of these types of information. When a system cannot give me that then it is not a successful one”. Another user said that: “I didn’t know that using such terms with my search term could be of any use […]. I would certainly try that for other searches, but I think it is a bit hard to think about some of them. I mean I was not aware of the difference between results of searching a term in plural or singular form”. Based on the users’ interviews, most of them were unaware of the benefits of searching with non-topical terms; and as it is indicated in both of the above quotes from the interview, they do not know exactly how to use these terms along with their topical terms. Moreover, even if they rarely use some general non-topical terms they usually fail to search more with similar non-topical terms. For example, even when someone searches for “feminism definition” they would not search a second time for “feminism meaning”.

On the other hand the researchers conducted two other interviews with the librarians who performed the searches for the users. They also were not fully aware of the benefit of searching in this way. As one of them mentioned: “As LIS students, we had the opportunity to become familiar with learned and standard search strategies. We learn that keywords are those terms or phrases that contain significant information and they have topicality. Thus we try to adhere to these rules for better results. It is not considered normal to use non-topical terms in search queries, especially in databases and since we normally run database searches, we do not use such terms.” Conducting a good and successful search could be considered as an art. This ability evolves through time and creativity is very important in this regard. But analyzing the librarians’ responses reveals that their performance is largely based on the search procedures they are accustomed to; and since they usually search in databases, their web search is also the same, as it is not normal to use non-topical terms in databases. This result is not promising since it shows no difference between the attitudes of the users and librarians as professionals.

Concluding Remarks and the Model

This study is to be considered as the first phase of a broader and more exhaustive project. But it clearly indicates that a thesaurus of non-topical terms could help in web search; especially regarding the relevance and also the pertinence of retrieved documents. Since non-topical terms, put the topical terms in context; and this context stems from user's needs. This context especially when searched in a phrase mode, encapsulates the topic along with that particular aspect which is needed; and since the essence of relevance lies on meeting the information needs of the user according to the context, it seems the proposed model would be of help. Figure 2 demonstrates a simplified conceptual model for implementing and incorporating the thesaurus into the search process. In this model the thesaurus would be available in the second phase of the search, when the preliminary results are present. The role of the thesaurus here is important in both search refinement and display of the results. The search results could be filtered through choosing special aspects featured in the thesaurus and search refinement options using the thesaurus would conduct a whole new search using the network of non-topical terms, based on the users' need. Both of these phases (Simple and Thesaural searches) are demonstrated in Figure 2.

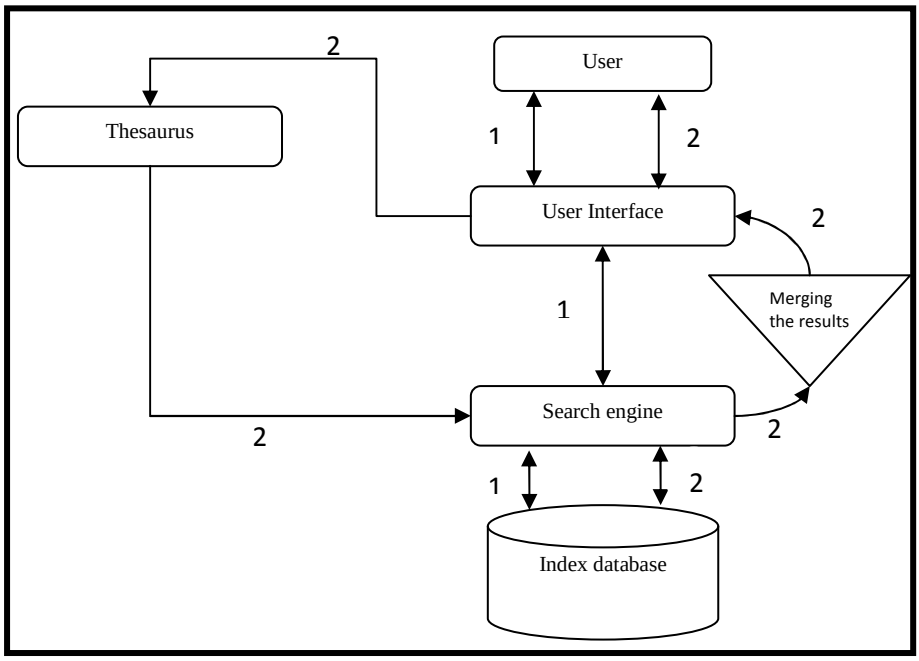

Figure 2. The simplified model for incorporating thesaurus of non-topical terms in web search

What is important to know is that this field of study is very young and it needs further research to ascertain the findings of current study. Yet the results presented here are promising and shows that through this approach higher search efficacy will be possible.

References

Badie, K., Tayefeh Mahmoudi, M., & Ghaderi, M. (2010). A Framework for Query Expansion Based on Viewpoint-Oriented Manipulation of the Related Concepts. In AMS '10 Proceedings of the 2010 Fourth Asia International Conference on Mathematical/Analytical Modeling and Computer Simulation (pp. 112-117). Washington: IEEE Computer Society.

Ceri, S. (2013). Web Information Retrieval. Berlin: Springer.

Chan, L., et al. (2001). A Faceted Approach to Subject Data in the Dublin Core Metadata Record. Journal of Internet Cataloging, 4(1/2): 35-47.

Fattahi, R.; Wilson, C. & Cole, F. (2008). An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents. Information Processing and Management, Vol. 44 (4): 1503-1516.

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. Journal of the American Society for Information Science, Vol. 45 (8): 589-599.

Sugiura, A., & Etzioni, O. (2000). Query Routing For Web Search Engines: Architectures And Experiments. In Proceedings of the 9th international World Wide Web conference on computer networks: The international journal of computer and telecommunications networking (pp. 417-429). Amsterdam: North-Holland.

Walker, G., & Janes, J. (1999). Online retrieval: A dialogue of theory and practice. 2nd ed. Littleton, Colo.: Libraries Unlimited

Wan, J. et al. (2012). Query Expansion Approach Based On Ontology and Local Context Analysis. Research Journal of Applied Sciences, Engineering and Technology 4(16): 2839-2843.