IX CONGRESS CONGRESO
ISKO-SPAIN ISKO-ESPAÑA
Valencia 11th, 12th, 13th March 2009 11, 12 y 13 de Marzo de 2009
New Perspectives for the organisation and dissemination of knowledge
Nuevas perspectivas para la difusión y organización del conocimiento

# INTEGRATING GENERAL, NON-TOPICAL TERMS INTO ONLINE THESAURI: A NEW APPROACH TOWARDS IMPROVING RETRIEVAL RELEVANCE ON THE WEB THROUGH NATURAL LANGUAGE TERMS

***Rahmatollah Fattahi***
*PhD., Professor, Department of Library and Information Studies, Ferdowsi University of Mashhad, Iran (fattahi@ferdowsi.um.ac.ir)*

***Mehri Parirokh***
*PhD., Associate Professor, Department of Library and Information Studies, Ferdowsi University of Mashhad, Iran (parirokh@ferdowsi.um.ac.ir)*

***Marie Rahimi***
*Master of Library and Information Science, Ferdowsi University of Mashhad (marie.rahimi@gmail.com)*

## ABSTRACT

Thesauri facilitate searching through controlled vocabularies which cover domain specific (topical) terms in a hierarchical or tree structure. One major limitation of thesauri is that they do not cover general terms representing particular aspects of the documents (such as context, approach, readership level, form, etc.) being important to authors as well as to readers. Documents on the same topic and with the same index term(s) differ from one another in some respects. Descriptors assigned to similar documents do not represent the non-topical aspects of documents. General terms can help searchers who are looking for documents with particular approach to content, structure, and/or format. Such terms can be used by the searcher or the system in query expansions through a natural language approach to retrieve more relevant documents. Research was undertaken to identify and study the availability of general, non-topical terms (GNTs) in Web documents as well as in two different thesauri (Hasset and ERIC). Findings show that, while GNTs are frequently present in the titles and URLs of Web documents (before or after topical terms), there are a small number of such terms in the thesauri studied. Searching through a combination of GNTs and descriptors in a natural language string would help users retrieve more relevant documents. Suggestions are provided on how to incorporate GNTs in all levels of the hierarchy (BTs, RTs, NTs) in thesauri based on their appearance in titles and URLs of Web documents. A sample list of GNTs is developed to be used in conjunction with index terms for searching.

## KEYWORDS

Thesauri, Online thesauri, General, non-topical terms, relevance performance, Web documents

## INTRODUCTION

Thesauri are extensively used in indexing and retrieval of documents irrespective of their physical formats. They focus on the semantic relationship between and among index terms to provide better access to the content of documents. With the increasing access to online and on-disc databases and with the exponential availability of information on the Web the use of thesauri for searching by end-users has also increased considerably. It is very important for the end-users to find more relevant information when they use online thesauri for searching.

One of the approaches thesauri developers take to help end-users to find more relevant information is to provide them with access to narrower terms. This is automatically done when a searcher looks for a term the hierarchical structure of the thesaurus leads him/her to narrower terms to continue the search. In many cases the searcher may be satisfied with the retrieval results indicating more specific information. However, in many other cases the documents may not satisfy the searcher because of the different approaches in their content. Documents even on the same topic differ from one another. While they may differ with respect to document context, intended audience, readership level, depth of content, etc., such differences have not adequately been dealt with in many thesauri and in query enhancement features (Fattahi, Wilson and Cole, 2008). Regarding the expansion terms, some researchers (such as White et al., 2005; Tombros et al., 2003; Ruthven et al., 2002) examined users' queries to identify top-ranking sentences based on deeper examination of the content of documents retrieved. They concluded that by reformulating of users search terms, retrieval results would be significantly effective and efficient. Reformulation can be done through query expansion.

## APPROACHES TO QUERY EXPANSION

Query expansion is the process of refining a query retrieving too many or too few relevant items. It occurs when the user modifies, amplifies or further specifies his/her search queries by typing a variety of additional terms. The intention is to improve precision in search results, through specifying aspects of the information needed.

There are two approaches to do query expansion to get more precision: 1) through replacing the query with narrower topical terms, or 2) through adding general, non-topical or semi-topical terms (GNTs) which indicate specific aspects of the topic (information) being sought. GNTs cover the following:

*1. **Non-topical terms (NTTs)*** are general terms which are not used independently for search. They usually occur in conjunction with (before or after) topical terms (TTs = expressions or concepts, for example) to represent a specific aspect of the subject (i.e., the nature of the document such as readership level, depth of content, approach to the content, context of document, and so on). Examples of NTTs and their uses are:

'~ **for beginners**' (e.g. 'Internet for beginners'), '**introduction to** ~' (e.g. 'introduction to globalization),

'~ **surveys**' (e.g. 'child abuse surveys'),

'**about** ~' (e.g. 'about breast cancer') and so on.

*2. **Semi-topical terms (STTs)***: Like non-topical terms, these terms are not normally used for searching by themselves. However, the difference between NTTs and STTs is that the latter are normally domain-specific. Terms such as '**prevention**', '**risk of**' ~, '~ **commission**', '~ **incidents**' and so on belong to this category. Domain-specific STTs can occur in multiple domains as in, for example, 'risk of globalization' or 'risk of lung cancer'.

Using GNTs in queries can improve precision in retrieval (Fattahi, Wilson and Cole, 2008).

## RESEARCH PROBLEM

The different advantages of the application of thesauri in information storage and retrieval, such the improvement of retrieval relevance have extensively been addressed in the literature. However, in terms of the structure and functions of thesauri in the online environment we need to identify new approaches to make them more responsive to searchers information seeking behavior. While thesauri are controlled vocabulary tools many users use natural language queries including both topical and general terms. Soergel (1999) states that users are losing their interest in using thesauri because thesauri are not responsive to the new environment.

In terms of helping the searcher to formulate queries some of the shortcomings of thesauri are:

Facilities for query expansion: Thesauri do not usually provide help to the searcher other than hierarchical controlled vocabulary terms (Narrower, Related, and/or Broader terms which represent topical aspects of documents). Furthermore, index terms in thesauri do not represent non-topical aspects of documents which the searcher may be interested in. Lykke & Ingwersen (1999) point out that most searchers often approach IR systems with a query formulated from words that come to mind. Among such words may be general or non-topical terms.

One major shortcoming of thesauri is lack of taking the readership level and the content approach of documents into account. As it was pointed out earlier, while many Internet users are ordinary/average individuals with different levels of knowledge, thesauri have not taken this into consideration. In other words, different documents being indexed based on thesauri are treated similarly irrespective of the differences in approaches to the contents.

The present research aims at investigating the availability of GNTs in online thesauri. It mainly examines the most frequently occurring GNTs in conjunction with topical terms in Web documents. It also attempts to provide solutions by which thesauri can help the searcher specify the scope and approach of his/her query. Thus the present paper attempts to answer the following questions:

1. How and to what extent non-topical terms are available in online thesauri?
2. To what extent general terms are addressed in online thesauri?
3. To what extent general terms added to index terms can improve precision and relevance in retrieval?

## REVIEW OF THE LITERATURE

Little research has been undertaken on the identification and use of general and non-topical terms in thesauri and for query expansion. The existing body of literature focuses mainly on the use of non-topical terms for query expansion. For example, the prototype Q-Pilot routing system developed by Sugiura & Etzioni (2000) was based on extracting phrases and then clustering terms into two groups: topical and non-topical. Non-topical terms were then added to the original user query to get new topical terms and the revised queries were then re-routed to relevant search engines. Chan et al., (2001) investigated the use of non-topical terms in query expansion on the Dublin Core metadata record to develop a new approach to subject vocabulary for Web searching. Their research on FAST (*Faceted Application of Subject Terminology*) is based on the LCSH (*Library of Congress Subject Headings*). In FAST, non-topical terms are separate from topical terms and placed in different elements provided in the Dublin Core

metadata record. Findings of a recent research on the use of NTTs for query expansion in search engines (Fattahi, Wilson and Cole, 2008) proved that such terms improved precision and relevance considerably in retrieval results.

## DESIGN OF THE STUDY

Textual analysis method generally referred to as content analysis was used for this research. Words in the texts of Web documents and in two thesauri were counted; their semantic relationships and their co-location were identified. The research was carried out in two phases:

***Phase I:*** Based on the list of 1071 non-topical terms developed by Fattahi, Wilson and Cole (2008) a list of 70 GNTs in the Social Sciences with a frequency of 3 or more was developed (see Appendix 1). These terms are among the most frequent GNTs occurring with topical terms in the texts of web documents. This list was then checked against two online thesauri (i.e., ERIC and Hasset) to examine to what extent and how GNTs are presented in them, that is to determine: (1) the frequency of each GNTs in the two thesauri, and (2) the frequency of GNTs occurring before and/or after the TTs.

***Phase II:*** In this phase, each of the 10 topical terms examined in the first phase were combined with GNTs and were searched in ERIC database to examine if they can improve precision in search results. Findings showed that almost all searches resulted in no hits as ERIC did not use such approach in developing descriptors. In comparison, to prove that the use of GNTs in conjunction with TTs would improve search results on the Web, the same 10 topical terms were searched in Google's four search options (i.e., keyword search, exact phrase search, exact title search, and exact URL search), this time in conjunction with their respective GNTs identified in Phase I. Altogether, 210 searches were carried out. The aim of this phase was to see to what extent GNTs were able to improve precision and relevance on the Web in comparison to thesauri which lack such approach.

## FINDINGS AND DISCUSSIONS

1. **The presence of general, non-topical terms in online thesauri.** The number of GNTs being available and where they appear in conjunction with topical terms were examined in each thesaurus. Table 1 shows the findings.

**Table 1:** Number, percent and location of GNTs in both thesauri

| Location of GNTs | Number | Percent |
|---|---|---|
| Before the TTs | 37 | 52.85 |
| After the TTs | 33 | 47.15 |
| Total | 70 | 100 |

The location of terms (i.e., term proximity and co-occurrence of words/terms) in a search query is very important. As can be seen, %52.85 of GNTs appears before and %47.15 appears after the TTs. This is in line with what Fattahi, Wilson and Cole (2008) found in their research. Regarding the frequency of GNTs most of them occur 3 to 10 times in both thesauri. Table 2 shows the findings.

**Table 2:** Frequency of GNTs in both thesauri

| Frequency | Number | Percent |
|---|---|---|
| 31-40 | 1 | 78.58 |
| 21-30 | 2 | 17.14 |
| 11-20 | 12 | 2.85 |
| 4-10 | 55 | 1.42 |
| Total | 70 | 100 |

With regard to the main aim of the research, that is to examine the availability of GNTs in the two thesauri (i.e., ERIC and Hasset), it was revealed that there is not much difference between the two thesauri in this regard (%48.57 in ERIC an%52.65 in Hasset). Tables 3 and 4 show the findings:

**Table 3:** Availability of GNTs in ERIC Thesaurus

| Number of NTTs | Availability | Location | Number | Percent |
|---|---|---|---|---|
| 70 | Available 36 | NTT alone | 14 | 20 |
| | | Before TTs | 7 | 10 |
| | | After TTs | 6 | 8.57 |
| | | Before and after TTs | 9 | 12.85 |
| | Not available 34 | - | 34 | 48.57 |

**Table 4:** Availability of GNTs in Hasset Thesaurus

| Number of NTTs | Availability | Location | Number | Percent |
|---|---|---|---|---|
| 70 | Available 34 | NTT alone | 13 | 18.57 |
| | | Before TTs | 7 | 10 |
| | | After TTs | 8 | 11.42 |
| | | Before and after TTs | 6 | 8.57 |
| | Not available 36 | | 36 | 52.65 |

GNTs in the two thesauri have appeared in two forms: either in the same form as they appear in the list 70 terms developed for this research (based actually on the text of web documents), or as a part of compound words/phrases. The location of the latter differs according to its relation to the TTs (i.e., before, after or in between the topical terms).

Findings show that about %50 of GNTs selected for this research are available in the two thesauri. This would indicate the importance of such terms for indexing. However, as tables 3 and 4 show, most of the GNTs appear as a part (before or after) of compound words indicating GNTs do not usually appear standalone in text and are not used alone by searchers to find information. Some of the GNTs identified in the two thesauri are as follow:

"~**System** "as in "***Metric system*** or in "***Solar system***"

"~**anti**" as in" ***Anti-social behavior*** or in "**Anti-Poverty programs** "

"~**Administration**" as in "***Educational Administration***" or in "***Public Administration Education***"

"**History of ~**" as in "***History of education***" or in "***history of Language***"

## Query expansion using non-topical terms in Google

As pointed out earlier, in the second phase of the research we examined the effectiveness of GNTs in retrieving relevant information on the Web. Ten (10) topical terms from the Social Sciences domain were searched in four different strategies in Google (i.e., simple search, exact phrase, exact title search, and exact URL search). Table 5 shows the 10 terms and the retrieval results in the 4 searching ways.

**Table 5:** Number and mean of retrieval results in four searching ways in Google (topical terms only)

| Keywords | keyword | Exact phrase | Exact title | Exact URL |
|---|---|---|---|---|
| Violence | 159,000,000 | 159,000,000 | 3,670,000 | 1,370,000 |
| Depression | 105,000,000 | 105,000,000 | 2,790,000 | 1,380,000 |
| Behavior Problems | 17,900,000 | 2,270,000 | 57,100 | 9,960 |
| Social Services | 134,000,000 | 38,700,000 | 960,000 | 709,000 |
| Child Neglect | 92,4000 | 58,400 | 13,900 | 3,290 |
| Educational Aims | 1,290,000 | 133,000 | 6,240 | 307 |
| Social Identity | 168,000,000 | 691,000 | 46,100 | 3,690 |
| Death | 652,000,000 | 652,000,000 | 19,000,000 | 5,650,000 |
| Mental Disorders | 10,100,000 | 5,760,000 | 136,000 | 30,900 |
| Foreign Relations | 30,000,000 | 10,400,000 | 472,000 | 97,800 |
| Mean | 127,821,400 | 97,453,800 | 2,384,834 | 845 |

Search was further carried out in Google this time with queries consisting of the 10 topical terms (TTs) with the 10 GNTs identified in the first phase of the research. These GNTs were as follows:

~ws / ~principles / ~statistics / ~news / against~ / about~ / effects of ~/ what is ~ / history of ~/ theory of ~

Each of the queries was performed in the 4 different search strategies in Google. Retrieval results were analyzed to examine the number and mean of retrieved Web documents before and after the query expansions. Findings are shown in table 6.

**Table 6:** Number, mean and ratio of retrieved documents (in Social Sciences) before and after query expansion in Google

| Search status | keyword | Exact phrase | Exact tile | Exact URL |
|---|---|---|---|---|
| **Before query expansion** | 127,821,400 | 97,453,800 | 2,848,834 | 8,564,576 |
| **After query expansion** | 25,690,880 | 164,415 | 100,419 | 632 |
| **ratio** | %20.09 | %0.16 | %4.21 | %0.07 |

Findings in Table 6 show that the means of the retrieval results in all the 4 search strategies decrease considerably using GNTs in conjunction with TTs. Ratio analysis for all the different retrieval results approves this conclusion.

The ratios of retrieval hits for simple keyword searching, exact phrase search, exact title search, and exact URl search decrease to 0.20.09, 0.16, 0.4.21, and 0,.07 respectively. This indicates that, while using GNTs would decrease the number of retrieved documents to a manageable number, the documents are more relevant to users' needs. The reason for this is that, technically, when the searcher narrows the scope of search using GNTs, the search engine only retrieves those documents which match the expanded query (i.e., a string of ...). In other words, only those documents are retrieved which represent the aspects that the searcher specifies in his/her expanded search query.

## CONCLUSIONS

Online thesauri are supposed to help the searcher in finding information with more relevance and precision. However, they have limitations in this regard as they do not provide the indexer or the searcher with the sufficient and appropriate additional terms which would represent the specific aspects of the documents such as the content approach, readership level, depth of content, and form. While a considerable number of GNTs are present in the title of documents indexed in ERIC database they are not treated appropriately in the structure of the thesaurus. The capacity of GNTs to represent different aspects of topics is considerable. Thus their significance should not be overlooked.

It can be concluded that the descriptors alone in thesauri are not appropriate for the search and retrieval of relevant information since they do not take concepts such as the approach, readership level, context, and form of documents. Thus identification of the most frequently general terms and also the possible development of a dictionary list which can be incorporated into the search options of thesauri would facilitate query expansion in natural language approach. This would have two advantages: 1) to specify the query in a more precise way, and 2) to design and develop an intelligent thesauri capable of providing the searcher with most frequently general terms occurring with index terms (topical terms).

GNTs can be incorporated in all levels of the hierarchy (BTs, RTs, NTs) in thesauri based on the frequency of their appearance in titles and URLs of Web documents. For example, thesauri can add the relevant GNTs under "See also" references for descriptors (broader, narrower and/or related terms). An actual entry and a sample entry based on such approach are presented as follows:

**School administration**
**BT Educational administration**
**NT School based management**
**RT School supervision**
SA Aids in **School supervision**
Approaches to **school supervision**
Certification in **School supervision**
Handbook of **School supervision**
Introduction to **School supervision**
Quality of **School supervision**
Theory of **school supervision**
**School supervision** essays
**School supervision** jobs
**School supervision** program
**School supervision** research
**School supervision** services

**Figure 2.** Sample entry for ERIC extended with GNTs

Another possible enhancement to thesauri is to incorporate a dictionary list of GNTs (see Appendix 1, for example)in them or as attachment (e.g., a small relational database) so that the searcher uses the appropriate GNTs for query expansion. A scope note under the main descriptor may explain how the searcher can use GNTs in combination of his/her query. Then a link to the list of most frequent GNTs would automatically lead him/her to the relevant general, non-topical terms most appearing with the query term(s). A sample entry would like such as the following:

**School administration**
**BT Educational administration**
**NT School based management**
**RT School supervision**
Also expand your search using any of the above terms in conjunction with the following general terms such: Aids in ~, Approaches to ~, Certification in ~, Definition of ~, Handbook of ~, Introduction to ~, ~ essays, ~ program, ~ research, services, and MORE...

**Figure 3.** Sample entry for ERIC extended with scope note for GNTs

Further research in this area is needed to identify the implications of implementing such approach for the structure and functions of thesauri. Research also would provide useful results because new findings would contribute to the development of methods which match the language of searchers (i.e., queries including non-topical and semi-topical terms) to the actual language of documents. Online searchers are eager to see and use more intelligent thesauri which enable them to find information with more precision and relevance.

The findings of this research showed that %20 of NTTs identified in Web documents are available in the two thesauri in the same form and %30 appear with slight change.

## REFERENCES

CHAN, L., CHILDRESS, E., DEAN, R., O'NEILL, E. T., & VIZINE-GOETZ, D. (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. AND 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 & Management*. Vol. 44 (4): 1503-1516.

LYKKE, M. & INGWERSEN, P. (1999). The word association methodology - a gateway to work-task based retrieval.
http://66.102.7.104/search?q=cache:d4V9LACbPeQJ:ewic.bcs.org/conferences/1999/mira99/papers/paper6.pdf+%22subject+searching%22+non-topical&hl=en

RUTHVEN, I. LALMAS, M. & VAN RIJSBERGEN, C. J. (2002). Ranking expansion terms using partial and ostensive evidence. *Proceedings of the 4th International Conference on Conceptions of Library and Information Science. CoLIS 4.* pp. 199-220.

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.

TOMBROS, A.; JOSE, J. M. & RUTHVEN, I. (2003). Clustering top-ranking sentences for information access. A. Tombros, J. Jose, and I. Ruthven. *Proceedings of the 7th European Conference on Digital Libraries. ECDL 2003.* pp. 523-528.

WHITE, R. W.; JOSE, J. M. & RUTHVEN, I. (2005). Using top-ranking sentences to facilitate effective information access. *Journal of the American Society for Information Science and Technology*. 56 (10), pp. 1113-1125.