Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management xxx (2007) xxx–xxx
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

An alternative approach to natural language query expansion in search engines: Text analysis of non-topical terms in Web documents

Rahmatollah Fattahi [a,b,*], Concepción S. Wilson [a], Fletcher Cole [a]

[a] School of Information Systems, Technology and Management, University of New South Wales, Sydney, Australia
[b] Ferdowsi University of Mashhad, Iran and Visiting Research Fellow, University of New South Wales, Sydney, Australia

Received 25 June 2006; received in revised form 20 September 2007; accepted 25 September 2007

Abstract

This paper presents a new approach to query expansion in search engines through the use of general non-topical terms (NTTs) and domain-specific semi-topical terms (STTs). NTTs and STTs can be used in conjunction with topical terms (TTs) to improve precision in retrieval results. In Phase I, 20 topical queries in two domains (Health and the Social Sciences) were carried out in Google and from the results of the queries, 800 pages were textually analysed. Of 1442 NTTs and STTs identified, 15% were shared between the two domains; 62% were NTTs and 38% were STTs; and approximately 64% occurred *before* while 36% occurred *after* their respective topical terms (TTs). Findings of Phase II showed that query expansion through NTTs (or STTs) particularly in the 'exact title' and URL search options resulted in more precise and manageable results. Statistically significant differences were found between Health and the Social Sciences vis-à-vis keyword and 'exact phrase' search results; however there were no significant differences in exact title and URL search results. The ratio of exact phrase, exact title, and URL search result frequencies to keyword search result frequencies also showed statistically significant differences between the two domains. Our findings suggest that web searching could be greatly enhanced combining NTTs (and STTs) with TTs in an initial query. Additionally, search results would improve if queries are restricted to the exact title or URL search options. Finally, we suggest the development and implementation of knowledge-based lists of NTTs (and STTs) by both general and specialized search engines to aid query expansion.

Keywords: Query expansion; Natural language; Non-topical terms; Semi-topical terms; Search engines

* Corresponding author. Address: Ferdowsi University of Mashhad, Iran and Visiting Research Fellow, University of New South Wales, Sydney, Australia. Tel.: +98 511 8783010; fax: +98 511 8783012.
E-mail addresses: fattahi@ferdowsi.um.ac.ir (R. Fattahi), c.wilson@unsw.edu.au (C.S. Wilson), f.cole@unsw.edu.au (F. Cole).

0306-4573/$ - see front matter © 2007 Published by Elsevier Ltd.
doi:10.1016/j.ipm.2007.09.009

1. Introduction

1.1. Current problems in information retrieval on the Web

The World Wide Web has now become one of the most used sources for accessing public as well as scholarly information. However, retrieving the relevant information on the Web and identifying the most useful items are not an easy task for many users. The number of Web documents generally retrieved in response to a search is huge and it is hard and unrealistic for the user to scan many. For example, an undergraduate student who needs a few simple and short documents defining 'globalization', enters the term 'globalization' in one of the search engines and is provided with thousands of Web documents, many of which may not be simple or short. The student may look through a number of sites appearing on the first few pages but still may not be happy with the results because either the document type or the approach of the content of the documents to the topic is not what is expected.

There are many intelligent search engines which, with their advanced search features, can help users fulfil their information needs. While a majority of users are too easily satisfied with what they retrieve, many remain dissatisfied with the results as often the retrieved documents are generally irrelevant to their specific needs (Casasola & Gauch, 1997; Chowdhury & Soboroff, 2002; Pokorny, 2004; Soboroff, 2004; Sugiura & Etzioni, 2000).

Some of the major problems with information retrieval in search engines are:

1.1.1. Choice of search term(s)

The formulation of queries is difficult for many Web searchers (Baeza-Yates, Hurtado, & Mendoza, 2004; Chowdhury, 1999; Chowdhury & Chowdhury, 1999Doan et al., Doan, Plaisant, Shneiderman, & Bruns, 1997; Filman & Pant, 1998; Lawrence & Giles, 1998; Lykke & Ingwersen, 1999; Voorbij, 1999). It is also very difficult for the average searcher to carry out phrase searching since they may not be familiar with the right set of terms which constitute the phrase.

1.1.2. General keyword searching as the default

Recall is high and precision and relevance are low (Chowdhury & Soboroff, 2002; Soboroff, 2004; Sugiura & Etzioni, 2000; Tillett, 2001).

1.1.3. Navigating and browsing retrieved pages

Time consuming and sometimes frustrating tasks; may mislead many searchers (Dias, Gomes, & Correia, 1999; Oyama, Kokubo, & Ishida, 1999; Spink & Xu, 2000).

1.1.4. Search algorithms

Often with inadequate explanation of how queries are interpreted by the search engines (Ellis, Ford, & Furner, 1998; Spink, Greisdorf, & Bateman, 1998).

1.1.5. Query refinement, reformulation and/or expansion

Generally unknown to many users, or if known, preference is to ignore the functions. Many users do not modify their original query or view subsequent results (Jansen, Spink, & Saracevic, 2000). They often approach IR systems with a query formulated from words that come to mind (Lykke & Ingwersen, 1999).

1.1.6. Facilities for query expansion

A few search engines (e.g., Lycos, Altavista and AskJeeves) have limited facilities for query expansion through proposing related topical terms (Baeza-Yates et al., 2004), but many do not normally provide users with additional, relevant search terms.

In refining a search query, searchers normally add terms which describe further topical aspects of the documents. Iivonen (1995) reports that inexperienced searchers use a surprisingly wide range of words to describe the same thing; that is, they try alternative (synonymous) terms rather than terms designating a particular

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

aspect of the subject of interest. This, as stated earlier, leads to the retrieval of pages which are less relevant or irrelevant to the searcher's needs.

1.2. Aim of the study

The *aim* of this research was to identify and categorise English language non-topical (i.e., general) terms and phrases in Web documents in two subject domains: Health and the Social Sciences. The researchers attempted to analyse the text of Web documents to see what general (i.e., non-topical) terms and phrases do normally occur in conjunction with keywords illustrating the content of the documents. Therefore, identification of the most frequently general terms in each domain and also the possible development of a dictionary list were the main motivations for the research being undertaken. The outcome would facilitate natural language query expansion in two ways: (1) it would help the searcher to specify his/her query more precise, and (2) the search engine designer can develop an intelligent tool to provide the searcher with most frequently general terms occurring with keywords in a given domain.

1.3. Definitions of major concepts

1.3.1. Query expansion (QE)

The process of refining a query retrieving too many or too few relevant items. QE occurs when users modify, amplify or further specify their search queries by typing a variety of additional terms. The intention of QE is to improve precision in topic search results, through specifying aspects of what is needed such as, document type, intended audience, readership level, and depth of content.

1.3.2. Topical terms (TTs)

Topical terms represent the subject content of documents. TTs are typically the terms which web searchers use to find relevant sources of information. Terms such as 'globalization', 'child abuse', 'Skin care', and 'Cosmetic plastic surgery' are examples of topical terms. Lists of subject headings (e.g., Library of Congress Subject Headings) and thesauri cover topical terms.

1.3.3. Non-topical terms (NTTs)

In most subject searches, these general NTTs are not used independently. NTTs usually occur in conjunction with (before or after) topical terms (expressions or concepts, for example) to represent a specific aspect of the subject (i.e., the nature of the document such as readership level, approach to the content, type of document, and so on). Using NTTs in queries can improve precision in retrieval. Examples of general NTTs and their uses are: '∼ for beginners' (e.g. 'Internet for beginners'), 'introduction to ∼' (e.g. 'introduction to globalization), '∼ Websites' (e.g. 'skin care Websites'), '∼ surveys' (e.g. 'child abuse surveys'), 'about ∼' (e.g. 'about breast cancer') and so on.

1.3.4. Semi-topical terms (STTs)

Like non-topical terms, these terms do not normally 'stand alone' and are not normally used for searching by themselves. STTs are used in conjunction with topical terms to narrow or further specify the subject aspect of the TTs. Thus, 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. Note that domain-specific STTs can occur in multiple domains as in, for example, 'risk of globalization' or 'risk of lung cancer'.

1.4. The semantics of non-topical terms (NTTs)

The distinction between TTs and NTTs is based on differences of meaning and function, which are matters of linguistic, and particularly semantic interest. Lyons (1995) points out that meaning is simultaneously determined at several levels; by the words that are used (**lexical** meaning), by the sentences in which they appear (**sentence** meaning), and by the wider **context** and the use to which a text is put (**utterance** meaning).

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

The typical document information retrieval (IR) system exploits lexemes and phrasal expressions freely occurring in full-text documents, or more precisely, exploits the specific instances (string tokens) that are evident. Some of these will be lexically simple (single words, irreducible phrases), but some will be less tightly-linked (composite phrasal expressions). TTs and NTTs appear in both these classes. Besides drawing on naturally occurring expressions, some IR systems also use pre-assigned subject terms to provide additional lexical handles to aid IR.

However they are analysed linguistically, NTTs stand in a qualifying relationship to the TTs with which they are associated. This relationship appears to function at both the lexical and the sentence level, for an aspect of their meaning is derived from their **grammatical** construction also. In practice, distinguishing between TTs and NTTs also draws on resources in the reading process which point to the ways language is actually used and understood (i.e. the **pragmatics** of language) in different knowledge domains. Those resources involve semantics at a comprehensive level, namely, language-meaning within a specific social and cultural setting. All this suggests that an exclusive reliance on the attributes of string tokens to differentiate TTs from NTTs has intrinsic limitations, but they are ones that are worth exploring, as in this research.

2. Literature review

Query expansion and related issues have long been of interest in IR research. The literature on how searchers formulate and reformulate their queries to improve precision and/or recall is extensive (see for example Anick & Tipirneni, 1999; Billerbeck & Zobel, 2[illegible] Bruza, McArthur, & Dennis, 2000; Efthimiadis, 1995, 2000;Harman, 1988; McArthur & Bruza, 2000).[1] Many research projects have investigated different approaches to the use of thesauri (manual or online) for query expansion. However, it should be noted that, in many cases, subject clustering of results and thesaurus-based expansion of queries do not improve precision without considering aspects of the documents and the users' queries. While information sources may differ with respect to document types, intended audience, readership level, depth of content, etc.[illegible]h [illegible]rences have not been dealt with adequately in many thesauri and in query enhancement features.

Little research has been undertaken on the use of general and non-topical terms for query expansion. A research project, somewhat in line with this study, was carried out by Sugiura and Etzioni (2000). With their prototype Q-Pilot routing system they found that each query can be routed to an appropriate specialized search engine by identifying the appropriate query category. This 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. Their research focus, however, was on the architecture of routing the queries to specialized search engines. Q-Pilot does not deal specifically with the identification and categorisation of non-topical terms based on their frequency, location and linguistic attributes.

Another study, which aimed at the use of non-topical terms in query expansion, was conducted by Chan, Childress, Dean, O'Neill, and Vizine-Goetz (2001) 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. However, FAST is restricted to non-topical terms only to pre-determined sub-divisions (such as geographical, chronological and form) appearing in LCSH.

An approach to query expansion similar to the present research has been implemented in the search engine 'Ask Jeeves' (www.askjeeves.com) from May 2005. In its 'Zoom' query refinement tool, Ask Jeeves provides a list of suggested phrases related to the user's query. While the concept has technically been implemented well by Ask Jeeves, the scope and range of suggested terms and phrases are limited and do not cover many of the non-topical phrases frequently being used in Web documents. Furthermore, there is no distinct arrangement or categorisation in the list of suggested phrases and in many cases there are phrases which have no contextual or topical relation to the query. For instance, in the query 'child abuse' (carried out on the 23rd October, 2005), Ask Jeeves variously displayed phrases like 'drug abuse', 'animal abuse' and 'alcohol abuse' for query

[1] A list of major research in this area can be found at: http://wotan.liu.edu/docis/search?query=query+expansion.

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

expansion. Many of the phrases suggested by Ask Jeeves are not extracted from the actual Web documents. These suggested phrases, which may belong to Ask Jeeves' own dictionary/thesaurus, are like 'blind references' where query phrases in the retrieved documents are not present. Hence, questions about the relevance of the results arise.

Our approach to query expansion in this study is different since we focus on the use of non-topical terms in 'exact phrase' searching using the 'Advanced search' options in Google; we mainly investigated the most frequently occurring non-topical and semi-topical terms in conjunction with topical terms in Web documents. As will be discussed below, the use of non-topical terms for query expansion can improve precision (to varying degrees) in information retrieval on the Web.

3. Research questions

From our primary research aim, the following research questions arose:

1. What are the most frequently occurring non-topical terms in Health and Social Sciences subjects?
2. What are the most frequently occurring non-topical terms used with topical terms in Health and the Social Sciences?
3. What are the most frequently occurring non-topical terms appearing either *before* or *after* the topical terms in the two subject areas?
4. Is there a significant difference between Health and Social Sciences subjects vis-à-vis the frequency of retrieved pages in different search options expanded with non-topical terms: keyword, exact phrase, exact title, exact URL?

4. Design of the study

This research used textual analysis (TA) generally referred to as content analysis: words or phrases of the text are taken to signify their meaning. Words in the text were counted; their semantic relationships identified; and their co-location noted. Two techniques of content analysis, 'elemental analysis' and 'structural analysis', were used. Elemental analysis deals with the identification of words, word groups and word frequencies, while structural analysis is concerned with the identification of elements (words) and the relationships between them (Hicks, Rush, & Strong, 1977, p. 90). This procedure primarily aims to identify meaning at a lexical level, and not so explicitly identify it at a sentence or contextual level (c.f. Section 1.3).

4.1. Data collection

The research was carried out in two phases

Phase I: Twenty simple queries in two broad subject domains (10 queries in the Social Sciences and 10 in Health) were searched in Google (www.google.com). The subjects or topical terms (see below in Tables 2 and 3 of the Section 5) of the queries were selected from those of current interest to Web searchers such as undergraduate students or the general public. We did not use the terms identified as most highly used in 2004 or 2005 by the different Websites, as most were either proper names or terms outside the scope of our study.

We chose Google as it is one of the most highly used search engines (Griffiths & Brophy, 2005). For each topical term searched in Google, ten websites were selected from the retrieved lists on the first, second and/or third pages. We excluded commercial Web sites as most had sparse or irrelevant text. In total, 200 web sites were selected for textual analysis to identify and extract non-topical terms occurring in conjunction with topical terms.

Since web searchers tend to browse only the first few pages of the sites retrieved in response to their searches (Spink et al., 1998; Jansen et al., 2000; Henzinger, Motwani, & Silverstein, 2002), only the first four pages retrieved were analysed. Thus over 800 web documents were analysed. Only the phrases/sentences which included the query term (i.e., the topical term) were selected for textual analysis and extraction. Furthermore, only the terms occurring immediately *before* or *after* the topical terms were extracted for further analysis. This

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

procedure is based on the notion that the average user normally enters one or two terms in his/her query and does not formulate a long phrase or sentence (Jansen et al., 2000).

The non-topical terms thus identified were extracted and entered into SPSS to determine: (1) the frequency of each NTT in the two subject areas, (2) the frequency of NTTs occurring before and/or after the TTs, and (3) the frequency of NTTs shared between the two domains. Altogether 1442 terms were identified as non-topical terms, but those which were either proper names (e.g., acronyms for associations and organisations) or occurred less than three times were excluded. Thus 1071 NTTs (387 in Health and 684 in the Social Sciences) were ranked based on their frequencies in the retrieved Web documents of both subject domains.

Phase II: In the second phase of the study,[2] each of the 20 topical terms examined in the first phase 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 NTTs identified in Phase I. In all, 4284 searches (1071 NTTs in 4 search options) were performed and the number of items retrieved was recorded in SPSS. Relevant statistical analyses, such as frequency, percentage, mean, *T*-test, and χ^2 test were carried out on the data gathered.

4.2. Data analysis

To analyse the data collected we used two different statistical methods. Descriptive statistics, such as frequency and percentage, were used to find out about the frequency of NTTs and STTs for different search options in the two domains. Also, *T*-test and χ^2 were applied to show the difference between the results of the frequencies in the two domains.

5. Results and discussion

5.1. Frequencies of NTTs and STTs in Health and Social Sciences

The 1071 terms and phrases identified as NTTs or STTs were ranked based on their frequency of occurrence in the 800 pages analysed. The top 50 NTT and STT terms/phrases in the two domains are shown in Appendix A. Table 1 summarizes the complete frequency data for NTTs (including STTs) for the Health and Social Sciences domains separately, for both domains, and for shared (or overlapping) numbers and percentages. The locations of NTTs (including STTs), either before or after the topical terms (TTs), and the numbers and percentages separately for NTTs and for STTs are also given.

5.2. Difference in frequencies of NTTs between the Social Sciences and Health

Based on the NTTs (including STTs) with frequencies over three, the average frequency was nearly five for Health and slightly over six for the Social Sciences. Hence, it is likely that topics in the Social Sciences domain on the web will use, on average, more NTTs than topics relating to Health. This may be due in part to the older (more mature) terminology of topical subheadings used in conjunction with Medical Subject Headings (MeSH) in the Health Domain – e.g., analysis, diagnosis, methods, prevention, trends, etc. (National Library of Medicine, 2006; HealthLink, 2006).

5.3. Most frequent non-topical terms shared between the Social Sciences and Health

Of the 1071 NTTs (including STTs), only about 15% ($n = 156$) were shared between the documents of the topics searched in the two domains. The overwhelming majority of the NTTs (including STTs) were unique to each of the subject domains and perhaps, even to topics within the same domain. This finding may have considerable implications in developing intelligent tools especially for specialized search engines. Since even fewer STTs ($n = 20$) overlap in the two domains, this implies that each major subject domain has its own separate terminology especially in providing STTs to qualify aspects of a topical term. For example, the STT 'economic

[2] Both Phases of this study took place during August–September, 2005.

Table 1
Frequencies, locations, and overlaps of non-topical terms and phrases (NTTs and/or STTs) in two domains

Combinations of subject domains	Frequencies of NTTs (incl. STTs)		Location of NTTs (incl. STTs)				Separate frequencies			
			Before TTs		*After* TTs		NTTs		STTs	
	No.	%	No.	%	No.	%	No.	%	No.	%
Health	387	36.1	232	59.9	155	40.1	235	60.7	152	39.3
Social Sciences	684	63.9	464	67.8	220	32.2	434	63.5	250	36.5
Total for both domains	1071	100.0	696	65.0	375	35.0	669	62.5	402	37.5
Shared (overlap)	156	14.6	86	55.1	70	44.9	136	87.2	20	12.8

248 ∼‘ occurring *before* a Social Sciences topic was not found in the Health domain; conversely ‘diagnosed with ∼
249 ‘ occurred only in the Health domain (see Appendix A, Ranks 20 and 29, respectively).

250 ### 5.4. Location of NTTs (including STTs)

251 The location of terms in a search query is very important, since algorithms for relevance ranking of retrieval
252 results on the web generally consider term proximity and co-occurrence of words/terms. Table 1 shows that
253 nearly two-thirds of NTTs (including STTs) occur *before* the topical terms, somewhat higher in the Social Sci-
254 ences (67.8%) than in Health (59.9%). A χ^2-test showed that there is a significant difference ($p = .001$) between
255 the two domains.[3] Our study therefore shows that the discourse in the Social Sciences domain is more likely to
256 place qualifying terms (NTTs or STTs) before topical terms as in, for example, ‘Definition(s) of ∼ ‘ where the
257 topical terms can be ‘domestic violence’, ‘terriorism’, etc. Interestingly, the NTT, ‘Definition(s) of ∼’ did not
258 appear with any of the Health domain topical terms searched (see Appendix A, Ranks 13 and 48, respectively).
259 A further interpretation leads to the more likely occurrence of NTTs (including STTs) before topical terms as
260 more ‘natural’ in normal English discourse in the Social Sciences than in Health domains.

261 ### 5.5. Separate NTTs and STTs Frequencies

262 With respect to the type of non-topical terms (general or domain-specific) found in our study, a χ^2-test
263 again showed that there is a significant difference ($p = .000$) between the two domains with general terms
264 (NTTs) occurring nearly two-thirds of the time in both domains and domain-specific terms (STTs) only in
265 about one-third of the time. These findings suggest that STTs should be considered (with caution) for query
266 expansion in certain sub-domains, despite the fact that STTs tend to be more ‘specific’ than NTTs and there-
267 fore more likely to improve the precision of a search on the web. Additionally, the selection of STTs for a
268 subject domain (or sub-domain) may be too time consuming and therefore too costly to implement.

269 ### 5.6. Differences in search results between the Social Sciences and Health

270 As stated above, all of the 20 topical terms were searched in Google’s four search options twice: first alone
271 and then in conjunction with each of their related NTTs and STTs. The results are shown in Tables 2 and 3.
272 The average number of retrieved pages in response to topical searches using Google’s four search options is
273 much more than what an average searcher can browse through. Even the progressive reduction in search
274 results from ‘Keyword search’ to ‘Exact phrase’, to ‘exact title’ and finally to ‘exact URL’ option would still
275 retrieve too many pages. As stated elsewhere (e.g. Chowdhury & Soboroff, 2002; Soboroff, 2004; Sugiura &
276 Etzioni, 2000; Tillett, 2001), this is a serious problem for search engine users. Performing topical queries with-
277 out further expanding them with other relevant terms results in an overwhelming number of hits.
278 However, as an alternative, query expansion through non-topical terms (adding the NTT to the TTs in nat-
279 ural language phrases) has much greater success in finding more manageable results (i.e., low recall, high pre-

[3] However, When considering only the top-50 occurring NTTs (including STTs), their locations either before or after the topical terms are equally divided: 25 *before* and 25 *after* (see the Appendix).

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

Table 2
Frequencies of search results for topical terms in health

Topical terms/phrases	Keyword	Exact phrase	Exact title	Exact URL
Suicide	93,600,000	93,600,000	1,550,000	465,000
SARS	15,000,000	15,000,000	879,000	837,000
Occupational health	66,200,000	16,600,000	315,000	44,700
Eating disorder	9,510,000	2,090,000	85,000	41,500
Breast cancer	59,600,000	44,600,000	1,400,000	695,000
Skin care	54,600,000	13,800,000	2,470,000	879,000
Drug abuse	49,900,000	18,700,000	228,000	254,000
Cosmetic plastic surgery	10,500,000	525,000	32,300	2,550
Yoga	28,000,000	27,800,000	1,840,000	2,430,000
Mental health	142,000,000	93,500,000	1,660,000	2,340,000
Average (mean)	44,921,000	23,438,500	895,970	752,457

Table 3
Frequencies of search results for topical terms in the Social Sciences

Topical terms/phrases	Keyword	Exact phrase	Exact title	Exact URL
Child abuse	55,700,000	14,500,000	296,000	47,400
Abortion	12,500,000	12,600,000	360,000	164,000
Human cloning	13,900,000	1,770,000	50,400	819
Social security	273,000,000	110,000,000	1,840,000	384,000
Domestic violence	38,000,000	22,300,000	428,000	98,900
Globalization	51,000,000	51,000,000	563,000	287,000
Human rights	440,000,000	126,000,000	3,170,000	177,000
Terrorism	139,000,000	139,000,000	1,770,000	635,000
Adoption	127,000,000	127,000,000	1,660,000	2,300,000
Feminism	13,800,000	13,800,000	238,000	153,000
Average (mean)	116,390,000	61,797,000	1,037,540	424,712

cision). Findings based on the strategies for query expansion in this research showed that the frequencies of retrieved pages were reduced considerably and led to more precise and more manageable results to browse. Table 4 illustrates the difference between the average frequencies of retrieved pages in response to 'TT searches' and 'NTT + TT searches' in the two domains.

As Table 4 shows, there are significant differences between Health and the Social Sciences with respect to keyword ($p = 0.001$) and exact phrase ($p = .006$) searches. More items are retrieved in the Social Sciences through these two search options. However, there is no significant difference between the two domains with respect to Exact title ($p = 0.851$) and Exact URL ($p = 0.170$) searches. It can be said that query expansion through NTTs in Exact title and Exact URL searches in the two domains perform equally well.

Also to illustrate the value of query expansion though NTTs we divided the average frequency of 'NTTs + TTs' to 'TTs' in each of the four search options and came to the ratios shown in Tables 5 and 6.

Table 4
T-test for retrieval results in different search options (queries expanded through NTTs)

Search options	Domain	Number of Web pages	Mean	Standard deviation	|*t*-test|	Degree of freedom	Probability value
Keyword	Health	387	9361203	13926204			
	Social Sciences	683	20298788	33517754	6.126	1068	0.001
Exact phrase	Health	387	148476	690518			
	Social Sciences	683	324080	1149376	2.738	1068	0.006
Exact title	Health	387	3236	19175			
	Social Sciences	683	3483	21496	0.188	1069	0.851
Exact URL	Health	387	754	8672			
	Social Sciences	683	279	1940	1.374	1069	0.170

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

Table 5
Average frequency and ratio of results for 'topical search' and 'topical search expanded through NTTs

Subject area	Mean	Keyword	Exact phrase	Exact title	Exact URL
Health	Average (of TTs)	44,921,000	23,438,500	895,970	752,457
	Average (of TTs expanded through NTTs)	9,332,489	148,094	3,227	752
	Ratio (of 'TTs + NTTs to TTs)	20.77%	0.63%	0.36%	0.1%
Social Sciences	Average (of TTs)	116,390,000	61,797,000	1,037,540	424,712
	Average (of TTs expanded through NTTs)	20,258,444	382,811	3,505	279
	Ratio (of 'TTs + NTTs to TT)	17.40%	0.61%	0.33%	0.06%

Table 6
T-test for means of results (TTs expanded through NTTs) in different search options

Search options	Domain	Number of Web pages	Mean	Standard deviation	Probability value
Keyword	Health	387	9332488.82	13905474.23	
	Social Sciences	684	20258444.24	33501509.51	.001
Exact phrase	Health	387	148094.29	689665.99	
	Social Sciences	684	382811.20	1200271.72	.001
Exact title	Health	387	3227.36	19150.41	
	Social Sciences	684	3505.43	21488.72	.832
Exact URL	Health	387	752.47	8661.31	
	Social Sciences	684	279.20	1940.15	.171

$$\text{Ratio} = \frac{\text{Average frequency of search results for 'NTTs + TTs}}{\text{Average frequency of search results for 'TTs}} \times 100$$

T-test analysis shows that there are significant differences between the two domains with respect to the means of the pages retrieved in 'Keyword' and 'Exact phrase' search options. However, regarding the 'Exact title' and 'Exact URL' searches, there is no significant difference in the two domains. This finding is in agreement with the previous findings on the average 'NTT + TT' retrieval results (Table 4). This suggests that query expansion through general terms in exact title and exact URL search options is equally successful in the two domains and, therefore, is recommended to all searchers.

The ratios of pages retrieved in response to query expansions through NTTs can also be calculated from a different perspective. The average frequencies of retrieved Web documents in the 'Keyword' search options can be divided by other options. The aims would be: (1) to see how the number of retrieved documents are reduced and results are more precise when moving from a more general search option to a narrower one, and (2) to see if there is any significant difference between the two domains. Tables 7 and 8 show the findings.

As can be seen, there is a significant difference between the two domains with respect to the three ratios. The *P*-value for the ratio of 'Exact phrase to Keyword', 'Exact title to Keyword' and 'exact URL to Keyword' searches are (.024), (.006) and (.017), respectively. This finding implies that searching through NTTs and narrowing the searches from 'Keyword' to other options in Health led to more precise results. As the ratio of retrieved documents using an 'Exact title search' and 'Exact URL search' in Health is relatively less than the respective ratios in the Social Sciences (see Table 7), there is a greater chance of retrieving more precise results in Health through query expansion using NTTs in Exact title and URL search options. The possible

Table 7
Ratios of retrieval results in different search options to one another (based on query expansions through NTTs)

Domain	Type of terms/phrases	'Exact phrase' to 'Keyword' (%)	'Exact title' to 'Keyword' (%)	'Exact URL' to 'Keyword' (%)
Health	TTs (without QE)	52	1.99	1.67
Social Sciences	TTs (without QE)	1.58	0.34	0.008
Health	TTs + NTTs	52	0.89	0.36
Social Sciences	TTs + NTTs	1.88	0.017	0.001

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

Table 8
T-test for means of retrieval results in different search options in Health and the Social Sciences

Ratios	Domain	*N*	Mean	SD	|*t*|	d*f*	*P*
Keyword/ Exact phrase	Health	387	90849.34	985444.11			
	Social Sciences	682	5248.05	38419.35	2.266	1067	.024
Keyword/ Exact title	Health	323	636036.74	1643141.78			
	Social Sciences	579	1480592.04	5366069.39	−2.757	900	.006
Keyword/ Exact URL	Health	227	1367462.68	4174828.74			
	Social Sciences	414	4904793.23	22079173.32	−2.390	639	.017

reason for this is that, although the range of NTTs in Health is limited in comparison to the Social Sciences (see Table 1), such terms have a wider usage in the title and URL of Web documents in Health.

6. Conclusions

Language is one of the most important factors in creating, representing and retrieving documents with different features. It is through the words and phrases that the topical as well as non-topical aspects of documents can be represented to the readers. It is therefore vital for IR systems to do word-related analysis of documents for more effective indexing and retrieval. This paper aimed to extend the options for information storage and retrieval through the identification and categorisation of a range of non-topical terms/phrases which normally occur in conjunction with topical terms in Web documents. Our findings show that the retrieval results would be more precise if the searcher enhanced a topical query with one or more non-topical terms or phrases.

One important semantic issue which should be taken into account by search engine developers is that the indexing of Web documents should not only take topical terms into consideration but also general and domain-specific non-topical terms which usually come in conjunction with topical terms in the natural language word order. In the indexing of Web documents, search engines can give more weight to NTTs in exact titles and URLs. Therefore, retrieval results would be more precise and more manageable if the queries are carried out in the 'exact title' or 'URL' search options. Thus the problem in retrieving irrelevant information on the Web would be solved considerably through adding non-topical terms to queries in exact title and exact URL searches. Also the default search option in search engines can be set on 'exact title' or 'exact URL' search to retrieve more precise and more manageable results.

Based on our findings, Web site developers should be encouraged to assign more meaningful terms in the title and URL of Web documents to help search engines give more weight to such elements in the indexing of the Web. Our findings could also be used to improve retrieval effectiveness by means of developing an intelligent interface providing searchers with a list of such terms (e.g. Appendix B). Such a tool could be integrated into these search options so that searchers can browse and choose the non-topical terms/phrases that most suit their information needs. Since many of the non-topical terms in the Social Sciences and Health are domain-specific, specialized search engines can develop their own list of non-topical terms which relate particularly to their speciality.

This line of research is a fruitful area of study, particularly because it contributes to the development of methods which match the language of queries to the actual language of Web documents. Web searchers always look forward to seeing more powerful and more intelligent aids which enable them to do that match.

7. Uncited references

Kelly and Fu (2006), Ruthven et al. (2002), Tombros, Jose, and Ruthven (2003), White, Jose, and Ruthven (2005).

Acknowledgement

This work was funded, in part, by the John Metcalfe Visitor's Grant awarded to the first author, Rahmatollah Fattahi.

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

Appendix 1

Top 50 non-topical terms/phrases in the Health and Social Sciences Domains

Rank	Non-topical terms/phrases	NTT	STT	Location: before/after	Total frequency	Frequency in Social Science	Frequency in health	Shared terms
1	about ∼	X		B	78	40	38	X
2	∼ is	X		A	72	47	25	X
3	∼ and	X		A	63	44	17	X
4	∼ prevention		X	A	49	11	38	X
5	And ∼	X		B	40	15	25	X
6	National ∼		X	B	40	22	18	X
7	∼ Information	X		A	25	9	14	X
8	information about ∼	X		B	22	4	18	X
9	∼ in	X		A	20	7	13	X
10	International ∼		X	B	20	20	0	
11	∼ research	X		A	15	2	13	X
12	∼ issues	X		A	15	15	0	
13	Definition of ∼	X		B	15	15	0	
14	Teen ∼		X	B	15	0	15	
15	∼ foundation	X		A	14	12	2	X
16	∼ reform		X	A	14	12	2	X
17	risk of ∼		X	B	14	0	14	
18	∼ of	X		A	13	10	3	X
19	∼ cases		X	A	13	0	13	
20	Commission on ∼		X	B	13	13	0	
21	Economic ∼		X	B	13	13	0	
22	∼ system		X	A	12	7	5	X
23	What is ∼	X		B	12	10	2	X
24	∼ products		X	A	12	0	12	
25	Against ∼		X	B	11	10	1	X
26	Types of ∼	X		B	11	10	1	X
27	∼ Awards		X	A	11	11	0	
28	diagnosed in ∼		X	B	11	0	11	
29	Prescription ∼		X	B	11	0	11	
30	To prevent ∼		X	B	10	3	7	X
31	Nuclear ∼		X	B	10	10	0	
32	∼ statistics		X	A	9	7	2	X
33	Information on ∼	X		B	9	3	6	X
34	∼ program		X	A	9	9	0	
35	Diagnosed with ∼		X	B	9	0	9	
36	The term ∼	X		B	9	9	0	
37	∼ articles	X		A	8	2	6	X
38	∼ laws		X	A	8	6	2	X
39	∼ news	X		A	8	6	2	X
40	∼ survivors		X	A	8	1	7	X
41	Prevention of ∼		X	B	8	6	2	X
42	Survivors of ∼		X	B	8	2	6	X
43	∼ attempt		X	A	7	0	7	
44	∼ law		X	A	7	7	0	

(*continued on next page*)

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

Appendix 1 (*continued*)

Rank	Non-topical terms/phrases	NTT	STT	Location: before/after	Total frequency	Frequency in Social Science	Frequency in health	Shared terms
45	~ outbreak		X	A	7	0	7	
46	~ patients		X	A	7	0	7	
47	~ tips	X		A	7	0	7	
48	Definitions of ~	X		B	7	7	0	
49	Domestic ~		X	B	7	7	0	
50	Radical ~		X	B	7	7	0	

Appendix 2

A sample list of non-topical and semi-topical terms occurring with the topical term 'globalization'

About **globalization**	**globalization** articles
Against **globalization**	**globalization** attempt
Articles on **globalization**	**globalization** awareness
Aspects of **globalization**	**globalization** basics
Basic information about **globalization**	**globalization** benefits
Benefits of **globalization**	**globalization** campaign
Books on **globalization**	**globalization** center
Combating **globalization**	**globalization** FAQs
Commission on **globalization**	**globalization** foundation
Counter **globalization**	**globalization** Guide
Definition of **globalization**	**globalization** information
Definitions of **globalization**	**globalization** information center
Economic **globalization**	**globalization** issues
Effects of **globalization**	**globalization** laws
FAQs about **globalization**	**globalization** links
History of **globalization**	**globalization** news
information about **globalization**	**globalization** principles
Information on **globalization**	**globalization** problems
International **globalization**	**globalization** process
national **globalization**	**globalization** programs
Prevention of **globalization**	**globalization** refers to
Radical **globalization**	**globalization** reform
Resources on **globalization**	**globalization** research
Risk of **globalization**	**globalization** statistics
Survivors of **globalization**	**globalization** stories
To prevent **globalization**	**globalization** system
Types of **globalization**	**globalization** taxes
Understanding **globalization**	
Victim of **globalization**	
War on **globalization**	
What is **globalization**	

References

Anick, P. G., & Tipirneni, S. (1999). *The paraphrase search assistant: Terminological feedback for iterative information seeking. Proceedings of the 22nd annual international Berkeley, California*. New York: ACM Press.

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

Baeza-Yates, R., Hurtado, C. & Mendoza, M. (2004). *Query recommendation using query logs in search engines edbt workshops* (pp. 588–596). http://www.dcc.uchile.cl/churtado/clustwebLNCS.pdf.

Billerbeck, B., & Zobel, J. (2003). *When query expansion fails. Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval*. New York: ACM.

Bruza, P., McArthur, R., & Dennis, S. (2000). *Interactive Internet search: Keyword directory and query reformulation mechanisms compared. Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, July 24–28, 2000, Athens, Greece*. New York: ACM Press.

Casasola, E. & Gauch, S. (1997). Intelligent information agents for the World Wide Web. Technical Report ITTC-FY97-11100-1, Information and Telecommunication Technology Center, The University of Kansas.

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.

Chowdhury, G. G. (1999). The Internet and information retrieval research: A brief review. *Journal of Documentation, 55*(2), 209–225.

Chowdhury, A. & Soboroff, I. (2002). Automatic evaluation of World Wide Web search services, SIGIR'02 (pp. 421–422). http://citeseer.ist.psu.edu/chowdhury02automatic.html.

Chowdhury, G. G., & Chowdhury, S. (1999). Digital library research: Major issues and trends. *Journal of Documentation, 55*(4), 409–448.

Dias, P., Gomes, M. J., & Correia, A. P. (1999). Disorientation in hypermedia environments: Mechanisms to support navigation. *Journal of Educational Computing Research, 20*(2), 93–117.

Doan, K., Plaisant, C., Shneiderman, B., & Bruns, T. (1997). Query previews for networked information systems: A case study with NASA environmental data. SIGMOD Record (ACM Special Interest Group on Management of Data).

Efthimiadis, E. N. (1995). User choices: A new yardstick for the evaluation of ranking algorithms for interactive query expansion. *Information Processing and Management: an International Journal, 31*(4), 605–620.

Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science, 51*(11), 989–1003.

Ellis, D., Ford, N., & Furner, J. (1998). In search of the unknown user: Indexing and hypertext and the World Wide Web. *Journal of Documentation, 54*(1), 28–47.

Filman, R. E., & Pant, S. (1998). Searching the internet. *IEEE Internet Computing, 2*(4), 21–23.

Griffiths, J., & Brophy, P. (2005). Student searching behavior and the web: Use of academic resources and Google. *Library Trends, 53*(4), 539–578.

Harman, D. (1988). Towards interactive query expansion, *Proceedings of the 11th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 321–331). Grenoble, France.

HealthLink. (2006). MeSH Subheadings and Families of Subheadings. http://healthlinks.washington.edu/howto/meshsubh. html#alpha.

Henzinger, M. R., Motwani, R., & Silverstein, C. (2002). Challenges in Web search engines, Colloquium Papers, Right Now Technologies (pp. 1–12). http://ai.rightnow.com/colloquium/papers/henzinger.pdf.

Hicks, C., Rush, J., & Strong, S. (1977). Content analysis. *Encyclopedia of Computer Science and Technology* (Vol. 6). New York: Springer.

Iivonen, M. (1995). Searchers and searchers: Differences between the most and least consistent searchers. In E. A. Fox, P. Ingwersen, & R. Fidel (Eds.), *Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 149–157). New York: ACM Press.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). Real life, real users and real needs: A study and analysis of users queries on the Web. *Information Processing and Management, 36*(2), 207–227.

Kelly, D., & Fu, X. (2006). Elicitation of term relevance feedback: An investigation of term source and context. *Proceedings of the 29th annual acm international conference on research and development in information retrieval (SIGIR'06)*, Seattle, WA.

Lawrence, S., & Giles, C. L. (1998). Context and page analysis for improved web search. *IEEE Internet Computing, 2*(4), 38–46.

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.

Lyons, J. (1995). *Linguistic semantics: An introduction*. Cambridge University Press.

McArthur, R. & Bruza, P. D. (2000). The Ranking of Query Refinements of Interactive Web-based Retrieval. In: *Proceedings of the information doors workshop (held in conjunction with the ACM hypertext and digital libraries conferences)*.

National Library of Medicine. (2006). Medical Subject Headings. http://www.nlm.nih.gov/mesh/topsubscope2006.html.

Oyama, S., Kokubo, T. & Ishida, T. (1999). Keyword spices: A new method for building domain-specific web search engines. http://www.lab7.kuis.kyoto-u.ac.jp/ishida/pdf/ijcai01.pdf.

Pokorny, J. (2004). Web searching and information retrieval. *IEEE Computer Software, 6*(4), 43–48.

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).

Soboroff, I. (2004). On evaluating web search with very few relevant documents. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. New York: ACM Press.

Spink, A., Greisdorf, H., & Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. *Information Processing and Management, 34*(2/3), 257–274http://www.informatik.uni-trier.de/ley/db/indices/a-tree/s/Spink:Amanda.html.

Spink, A., & Xu, J. L. (2000). Selected results from a large study of Web searching: the Excite study. *Information Research, 6*(1)http://informationr.net/ir/6-1/paper90.html.

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009

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.

Tillett, B. B. (2001). Authority control on the Web. http://www.loc.gov/catdir/bibcontrol/tillett_paper.html.

Tombros, A.; Jose, J. M. & Ruthven, I. (2003). Clustering top-ranking sentences for information access. In: Tombros, A., Jose, J., Ruthven, I. (Eds.), *Proceedings of the 7th european conference on digital Libraries*, ECDL 2003 (pp. 523–528).

Voorbij, H. J. (1999). Searching scientific information on the Internet: A Dutch academic user survey. *Journal of the American Society for Information Science, 50*(7), 598–615.

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), 1113–1125.

Please cite this article in press as: Fattahi, R. et al., An alternative approach to natural language query expansion ..., *Information Processing and Management* (2007), doi:10.1016/j.ipm.2007.09.009