Thesaurus-assisted search term selection and query expansion:
A review of user-centred studies

Ali Asghar Shiri, Crawford Revie, Gobinda Chowdhury

Department of Computer and Information Sciences, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH Scotland, UK

Abstract

This paper provides a review of the literature related to the application of domain-specific thesauri in the search and retrieval process. Focusing on studies which adopt a user-centred approach, the review presents a survey of the methodologies and results from empirical studies undertaken on the use of thesauri as sources of term selection for query formulation and expansion during the search process. It summaries the ways in which domain-specific thesauri from different disciplines have been used by various types of users and how these tools aid users in the selection of search terms. The review consists of two main sections covering, firstly studies on thesaurus-aided search term selection and secondly those dealing with query expansion using thesauri. Both sections are illustrated with case studies that have adopted a user-centred approach.

1. Introduction

The selection of search terms for query formulation and expansion is a challenging task within the information search and retrieval process. Two general approaches have been adopted in studies on search term selection: system-centred and user-centred. The system-centred approach is represented by work on algorithms and evaluation based on the traditional IR model, a model that fundamentally ignores the users and their interaction with the system. In contrast, the user-centred approach focuses on the cognitive, interactive, and contextual aspects of IR and considers users, use, situations,

context, and interaction with system (1). The user-centred approach has been developed to address a range of poorly understood issues relating to behavioural and cognitive aspects of the IR process. Spink (2) and Spink and Saracevic (3) have emphasised the need for further research into the user-centred approach to search term selection and query expansion in order to improve the use of and interaction with IR systems. This approach is concerned with the ways in which users of IR systems select their search terms for formulating and/or expanding their queries. It also considers factors and variables that cognitively and behaviourally affect the user's decision-making in the search term selection process.

Knowledge structures in general and domain-specific thesauri in particular are potential sources of search terms for query formulation and expansion. Several studies have evaluated the use of thesauri by different types of users. While all of these studies have adopted user-oriented approaches, their treatment of the use of thesauri, types of users and methodologies vary.

The main objective of this review is to provide a survey of the methodological issues and main findings of a relatively comprehensive set of user-centred studies on thesaurus-aided search term selection and query expansion. Two criteria were taken into consideration when defining the scope of this review. First, the review focuses on studies that have considered thesauri as sources of term selection and query expansion. Second, as there are different types of thesauri, only studies which have applied or evaluated domain-specific thesauri with standard relationships (hierarchical, equivalence and associative) are included in this review. Thus, for example, studies that have evaluated the role of automatically constructed thesauri in search term selection and query expansion are excluded.

This review is structured into two major sections. The first section deals with the use of thesauri for search term selection. The second section is concerned with the application of thesauri for query expansion purposes to improve search results. A short summary of the main issues examined is given for each of these sections while an overall conclusion is provided at the end of the review.

2. Approaches to search term selection

The selection of search terms for query formulation and expansion within the information retrieval process has been studied from a range of perspectives. Spink and Saracevic (3) identified two general types of search term selection research, namely the algorithmic and human approaches.

The focus of the algorithmic approach is to develop and evaluate different types of algorithms for selecting, weighting and/or ranking search terms in the process of query formulation or expansion to improve information retrieval. Examples of research of this type include (4-10) as well as much of the research documented in the Text REtrieval Conference (TREC) proceedings (11).

The human approach, in contrast, is concerned with studying and evaluating the ways in which users choose terms for formulating, expanding or modifying their queries during the search process. It deals with cognitive and behavioural models and issues that affect the selection of search terms by users. Research has focused on user-centred variables such as those relating to information needs, user intentions, personal characteristics, and different user information seeking profiles, and investigates their relationship to term selection in the search process.

2.1 Sources of search term selection

A number of sources for search term selection in query formulation and expansion have been suggested in the literature. For instance, Fidel (12) analyses search term selection based on two types of source: free-text and controlled vocabulary descriptors. Efthimiadis (13) categorises sources for term selection in query expansion into two types: those based on the relevance feedback process to initial search results and those which use some form of knowledge structure. Such structures can be either collection-dependent (corpus based) or collection-independent such as thesauri and dictionaries. Spink and Saracevic (3) identify five sources of search terms in a study investigating their effectiveness during mediated online searching. These sources are: the question statement, user interaction with intermediary, thesaurus, the human intermediary, and term relevance feedback. In these studies and those reviewed in the following sections thesauri were recognised as tools suited to the provision of search terms at either the query formulation or expansion phases of the information retrieval process.

2.2 Thesaurus-aided search term selection: user studies

Studies within this area have investigated information searching behaviours, tactics and strategies to better understand term selection and the ways in which thesauri are utilised by users during this phase of the search process.

Bates (14) in a discussion on search strategy, defines four categories of search tactics, two of which are relevant in the present context; these she labels search formulation tactics and term tactics. Search formulation tactics refer to the process of designing or redesigning the search formulation and ways of analysing information request and query elements. This includes making the search formulation precise, broad or specific. Term

tactics relate to the selection and revision of specific terms within the context of search formulation. These tactics focus on moving upward, downward, or sideways within a hierarchical structural to find broader or narrower terms during the selection of search terms. These tactics attempt to capture the complexity of the human search term selection process and the sophisticated decision-making effort involved.

Several studies have investigated search term strategies and the term selection behaviour of users from a wide variety of disciplines, backgrounds and environments. Although treated differently, thesauri were considered or evaluated as search term sources in all of the studies discussed below.   These studies can be categorised according to the user population involved:

- professional searchers only

- professional searchers and end-users in an mediated environment

- professional and novice search behaviour compared

- end-users only

2.2.1 *Search term selection behaviour of professional searchers*

While investigating the searching behaviour of professional online searchers is a well-developed research theme, the major studies concerned with the use of thesauri are due to Raya Fidel. In an early investigation Fidel (15) studied the search formulation, reformulation, and search term selection behaviour of five experienced searchers. She identified two types of information searching behaviours which she named the *operationalist* and *conceptualist* searching styles. Operatioanlist searchers try to formulate a query by identifying the related descriptors for each component of the

request. They look for the descriptors and check their categories and locations in the hierarchical structure not only to find permitted entry term but also to gain a better understanding of the request. In situations where they cannot find a descriptor to represent a concept they search using the free-text mode. In contrast, conceptualist searchers focus on the structure of the vocabulary for conceptual analysis of the request and query formulation. They look for relationships between the facets of the request and the structure of the controlled vocabulary. They tend to use free-text terms for the initial search formulation in cases where they are confident that there is no controlled way to express the concept, or where terms are very specific and well-defined, and very little has been written about the subject. Both searching styles represent the term selection styles of professional searchers using thesauri and the ways in which they cognitively process terms in the request and translate them into queries accepted by the retrieval system.

In another study Fidel (16) observed the online searching behaviour of eight searchers in order to provide a set of rules for search term selection by an intermediary expert system. The search keys selected by the searchers and the reasons for choosing these terms were examined. Using this evidence a "selection routine" detailing the conditions for the selection of search keys from both free-text and controlled vocabularies was developed. The selection routine describes different conditions under which searchers try to map a search term to a descriptor in the thesaurus or controlled vocabulary, and those which lead to a decision to use the term as a free-text key. Mapping to descriptors can be carried out through an exact match, a partial match, or the key might be mapped to a broader descriptor or to narrower terms. By analysing the patterns of selection routine, Fidel illustrated the significance of decisions made during search key selection to the success of a search. Her model also showed that a "good" search term is a single-meaning term that can be mapped to a descriptor, while a term is "not adequate" if it is a common term

and/or cannot be mapped to a descriptor. She refers to the lack of a general typology for requests to support the selection of different ways of choosing terms and suggests that more investigation should be undertaken on searching behaviour to reveal under which conditions each of the term selection options is chosen. The "selection routine" approach identifies the problematic points in the search term selection process and provides guidelines for research into the searching behaviour of online searchers.

In a larger investigation of the search term selection process, Fidel (17) studied 47 professional online searchers performing job-related searches. Two distinct types of search keys: words used in free-text searching, and descriptors taken from a controlled vocabulary, were compared. In the first part of this study she built on her previous work (16) on the validation and expansion of a search term selection routine. She defined *single-meaning terms,* as those which are good for free-text searching, and *common terms* for those which are not. In order to build a formal model of search term selection, different conditions of selecting a term were identified. For example, whether a common or single-meaning term can or can not be mapped to a descriptor and if it can be mapped, is this an exact or partial match. Different strategies for choosing broader, narrower or synonymous terms were identified. Searchers who participated in the study were asked to provide reasons for the selection of search keys. These reasons were categorised as: request-related, database-related and searcher-related. The first dealt with the characteristics of the requests, the second was associated with facilities and characteristics of the databases used by the searchers (for example, their use of thesauri or other features), and the third related to the individual searching behaviours and habits of the searchers. These factors influenced the way in which searchers chose free text or thesaurus descriptors. For instance, if a request were very specific, the searcher would

tend to use free text terms; or, if a database did not have a thesaurus, the searcher would be more inclined to enter free text terms.

In the second part of the study Fidel (12) observed how the 47 professional searchers used descriptors and text words as search terms. A number of variables were defined to measure the factors affecting the selection of search keys, including: institutional setting, subject area, databases used, number of search keys, number of moves, ratio of free-text keys to total number of keys, and *thesaurus neglect* ratio. The results showed that of the 3,200 search keys selected 50% were descriptors and 50% were free-text terms. In an analysis of all the search terms used it was revealed that searchers consulted a thesaurus for 75% of the case selections. Characteristics of databases, thesauri and requests were the major factors affecting search key selection. One of the main results of this part of the study was that searchers used a thesaurus when it was of satisfactory quality and was easily available to them. In contrast the non-availability of a thesaurus tended to increase the number of search keys and the number of moves in a search, and thus increased the effort necessary to perform a search. Fidel also suggested that database designers and search-system vendors should encourage the use of thesauri by designing easy-to-use and flexible thesauri, particularly as reliable sources of synonyms.

The third part of Fidel's study (18) explored the searching styles of the professional searchers specifically in relation to their modification of search strategies. She identified different moves for increasing and reducing the size of retrieved set and moves to increase both recall and precision with regard to the *operationalist* and *conceptualist* searching models (15). She stated that the association between the number of search keys and the total number of moves was a significant pattern in online searching behaviour. It was revealed that the searchers who made more moves were likely to use more search keys than were searchers making fewer moves. Searchers selected on average around 13 keys

per search but the average number of keys varied greatly between searcher, ranging from just under three to almost seventy. The results also demonstrated that the average number of keys per search is typical for a searcher, and that a person's searching style will thus determine how extensive will be their use of terms. It was found that searchers considered recall to be the most important factor when they select search keys and also when they modify search strategies.

2.2.2 *Search term selection behaviour of professional searchers and end-users in a mediated environment*

Studies considered in this section examine the search term selection and interaction behaviour of professional searchers and end-users in a mediated retrieval environment.

In a large-scale empirical investigation to characterise the elements involved in information seeking and retrieving, particularly from the cognitive and human decision making perspective, Saracevic et.al. (19-21) examined the selection of search terms by different searchers for the same questions. User, question, searcher, search, and items retrieved were the main variables observed in this study. Forty users, thirty-nine searchers and forty questions constituted the experimental environment and different techniques including: questionnaires, interviews, transaction logs and videotape were used for data collection. This study made a significant effort to explore the different cognitive structures involved in the information retrieval interaction process, including: users, intermediaries (searchers), information retrieval systems, questions and their interplay. In order to analyse and evaluate all variables influencing the search and retrieval process, the study defined four sources of search terms: a) terms derived from an oral statement about the problem recorded by the user, but without any reference to the written question; b)

terms extracted from the recorded oral statement and the written question submitted by the user; c) terms from the written question using only the words in the question as search terms without any further elaboration and; d) terms derived from written question plus terms from an appropriate thesaurus for elaboration. The results indicated that searches based on the user's written question plus the use of a thesaurus were rated as the second best searches in terms of recall and precision, outperformed only by the relatively intensive approach that used oral and written user statements. The research demonstrated that given the same question, different searchers tend to select a few common terms and a considerable number of terms that are different. Searches based on different search term selection sources produced a significant difference in recall but no significant difference in precision. While end-users were involved in the research, this study examined primarily the process of search term selection by professional searchers and the ways in which they evaluated the search requests and selected terms based on the four term sources noted above.

In another study by Spink and Saracevic (3) investigating the selection and effectiveness of search terms. The searches were again performed by professional searchers but end-users played an active role in selecting search terms during online sessions. The data consisted of the interactions of 40 faculty and doctoral students with 4 professional searchers, searching 40 questions provided by the participants. Interviews, questionnaires, and transaction logs together with video recording were the main data gathering techniques. The variables defined in the study were as follows: user satisfaction ratings, search outcome variables (including, number of relevant and non-relevant items retrieved and precision measures), search process variables (including, number of cycles and moves), and user characteristics such as domain knowledge. They identified and classified sources of search terms as follows:

Question statements: search terms derived from the user's written request

User interaction: search terms suggested by the user prior to and/or during the online search, but not included in the user's question statement

Thesaurus: search terms derived from a thesaurus associated with the database

Intermediary: search terms suggested by the search professional prior to and/or during the online search

Term relevance feedback: search terms suggested either by the user or professional searcher taken from the retrieved items identified by the user as relevant

They evaluated the effectiveness of each source and their contribution to the search results. Question statements and interaction with the user were responsible for 38% and 23% of the selected terms respectively, with thesauri contributed an additional 19% of the terms. A further 11% of the search terms came from term relevance feedback, while professional searchers were responsible for the remaining 9%. In addition to supplying the largest proportion of terms question statements were also the most productive in terms of retrieving relevant items. User-interaction terms were slightly less effective with around 50% resulting in relevant retrieved items. Terms derived from thesauri were less effective again, a fact which caused Spink and Saracevic to conclude that thesaurus terms prove most effective when combined in search statements with user terms. This finding emphasises the significance of interaction between users' terms and the terms taken from the thesaurus. Although thesauri were one of the sources of terms used during the search process, the interaction between the end-user and thesauri was not examined in this study.

In addition to the above studies, there are a number of investigations that have explored the interaction between users, professional searchers and IR systems in a mediated

environment. These studies have considered the process of search term selection in less detail, and have not specifically examined thesauri as sources of term selection (22-23).

2.2.3 *Search term selection behaviour of professional and novice searchers compared*

To compare the search term selection and information searching behaviour of users with different levels of experience, a number of studies have employed professional and novice searchers. In an investigation of the behaviours associated with the process of online bibliographic searching, Fenichel (24) examined the differences among users searching online systems who had different levels of experience. Five groups of searchers with different levels of online searching experience performed searches using ONTAP, a subset of the ERIC database. The main variables studied included: environmental, searcher, search process, and search outcome factors. Variables associated with the search process were: the number of commands used, free-text and thesaurus terms chosen, sets viewed, cycles, search modifications, and connect time. The results showed that all five groups used more thesaurus terms than free-text terms. The most experienced group used a significantly higher proportion of descriptors taken from the thesaurus than did the other groups. In addition searchers with ERIC experience used significantly more thesaurus terms than subjects without such experience. The study showed that having experience of databases equipped with thesauri affects the ways in which searcher select terms. However, the finding that novice searchers also make use of thesauri indicates the importance of recognising that thesauri can be useful sources of search terms for users with varying levels of experience.

Hsieh-Yee (25) investigated the effects of subject knowledge and search experience on novice and experienced searchers' use of search tactics in online searches. Using

transaction logs, on-site observation and think aloud techniques, she studied the online searching process of 33 professional searchers and 30 novice searchers. Based on previous search tactics, she defined a number of variables related to search term selection, such as use of searcher's own terms, searcher's reliance on the thesaurus structure, off-line term selection efforts, online usage of search terms, inclusion of similar concepts and synonyms, and the searcher's combination of terms. To evaluate the effect of subject knowledge the study defined two familiar and two unfamiliar questions for both novice and professional searchers to consider. The results showed that when searching a topic of which searchers had some knowledge, experienced searchers included more synonyms and tried more combinations of search terms than did novice searchers. Experienced searchers looked to the thesaurus for term suggestion and tended to formulate a more comprehensive search, while novice searchers tended to rely on their own terms. Novice searchers consulted the thesaurus much less frequently than did experienced searchers. While the use of thesauri as a term selection tactic was considered in this study, the interaction of searchers with the thesaurus interface was not examined.

In a study of differences in search term selection between the most and the least consistent searchers, Iivonen (26) evaluated the inconsistencies among 32 subjects, 24 experienced searchers and 8 undergraduate students of information studies. The sibjects were given the option of using both descriptors from a thesaurus and free-text terms to perform specified search tasks. Three factors relating to the terminological style of the searchers were identified: the number of search terms per search request, the number of search terms per search concept, and the proportion of descriptors among search terms. The results suggested that as the number of search terms increases the term consistency is adversely affected. However, increasing the proportion of descriptors will lead to improved term consistency. Those searchers who chose only a few search terms per

request and per search concept, and attempted to ensure that these terms were controlled vocabulary descriptors, achieved higher inter-searcher consistency than the average. These results are in line with prior expectations in that by selecting from the controlled vocabulary the number of potential search terms is already limited. It was also shown that differences in the searchers' experience resulted in the use of different terminological styles.

Based on an empirical study of searchers during the pre-online stage, Iivonen and Sonnewald (27) proposed a cognitive model of the search term selection process. Once again their study population was made up of 24 professional searchers, who had backgrounds in special, university and public libraries, together with 8 students. Each searcher was presented with 12 requests to formulate query statements prior to the search process. Searchers used a Finnish database to select descriptors but also had the option to incorporate free-text terms in formulating their queries. The results of the study revealed six different sources of term selection, referred to by the researchers as six "discourses". By discourse they meant a specific way of thinking and talking about a certain topic within a community. These six discourses, which constituted the elements of their proposed model, were: controlled vocabularies, documents and domain, indexing practice, client's search request, databases, and the prior search experience of the searcher. Documents and domain refer to the titles and abstracts of records, which represent the way the topic is discussed within a community of authors and publishers. Indexing practice relates to a searcher's perception of indexing rules and the practices adopted by indexers in the use of particular terms and concepts. Database discourse refers generally to the content and structure of a database and specifically implies knowledge of the subject categories and fields available in a given database. The study revealed that all but two of the searchers used controlled vocabularies. Irrespective of the searcher's

background the only discourse which was frequently cited as a source of search terms was the controlled vocabulary. This illustrates the strong influence which the discourse of controlled vocabularies exerts upon professional searchers. Controlled vocabularies provide a mechanism to describe a topic and to aid navigation by showing the relationships between topics. The results of this study illustrated the multidimensionality and complexity of the search term selection process and provided insights into the ways in which searchers navigate different discourses. The study suggested that further research was needed to fully identify and validate the characteristics of each discourse and to explore those aspects which facilitate the search term selection process.

2.2.4 *Search term selection behaviour of end-users searching independently*

While there have been a number of studies on end-user searching behaviour, few studies have focused on search term selection.

Among the studies which consider discipline-oriented search terminology is the work of Bates et.al. (28) and Siegfried et.al. (29) which provide insights into the search terms used by humanities scholars as end-users. Their research examined the search techniques, queries and search terms of twenty-seven scholars searching DIALOG databases. The researchers defined three major terminological categories in order to identify the vocabulary used by humanities researchers, namely: type of search need, such as works of an author or works on a subject; bibliographic features, such as date or form of publication; and types of subject, including individuals, geographic locations, date or period. While there were cases of using controlled vocabulary descriptors during search formulation, the study did not examine how and to what extent the searchers chose such descriptors or their impact on search term use. The results of the study showed that the

terminology used by humanities researchers was remarkably different from the vocabulary used in other fields, as were aspects of the information seeking and online searching behaviour. The humanities scholars searched for more named individuals, geographical terms, chronological terms and discipline terms. This finding has significant implications for developing thesauri and online search aids for the humanities, suggesting that thesauri developed for the humanities should incorporate more comprehensive sets of geographic and chronological terms as well as proper names.

Sutcliffe et.al. (30) investigated end-user behaviour by studying the performance of 17 medical students searching the Medline database. The subjects were categorised as expert or novice searchers based on their knowledge of the search system. Search performance, query pattern, search strategy, query construction, term use, and system facility usage were the variables studied. The study found that more than eighty percent of the expert searchers used the thesaurus or term suggestion facilitates to explore concepts. However no reference is made as to how the subjects interacted with the thesaurus or what difference this made to search term selection. Expert searchers used more terms in their queries, constructed more complex queries, performed more iterations and used more system facilities. Novices, in contrast, used simple queries and fewer search terms. They made less use of system facilities and carried out fewer search iterations. The research concluded that although there were behavioural differences between novice and expert searchers, no simple correlations between behaviour and performance were found. A number of usability problems with the system thesaurus and term suggestion facilities were found. Suggestions were made as to how access to alternative terms might be enhanced and the effort required to use the thesaurus or other term suggestion facilities reduced.

Vakkari (31-32) investigated changes in search term usage and tactics among eleven students carrying out a research proposal as part of their Master's theses. Using survey questionnaires, interviews, the think-aloud technique and transaction logs, he studied the search behaviour of students as they performed this task. He examined the number and types of search terms and tactics used and their relationship to the students' prior knowledge of their topics. He analysed different types of new terms introduced by the students and classified them as broader, narrower, related and synonymous terms. The results demonstrated that a growing focus and clearer understanding of the task led students to choose narrower and synonymous terms, to discard broader terms from their search formulation, and to use simpler search tactics. He suggested that for novices in a domain, structured terminological support would not only improve search results by encouraging the use of an increased number of narrower terms, but might also support the user in deconstructing the topic and interrelating its constituent parts.

A number of additional studies on end-user behaviour in different environments have commented on the specific issue of term selection (33-35). Belkin et. al. (36) have reported the use and effectiveness of term suggestion facilities for supporting end-users in an interactive and relevance feedback environment. Another line of investigation adopted by researches has been the study of search term selection patterns used by children as part of their information seeking and retrieval behaviour (37-38). While these studies have not evaluated thesauri as sources of term selection as such, their results indicate the need to incorporate thesauri and terminological support in the information retrieval interface to facilitate query formulation by end-users.

In addition to the above studies, research examining search term use among end-users of World Wide Web search engines has been carried out (39-42). These studies have investigated a number of variables related to search terms used by the public, including

the number and types of search terms, the number of terms per query, search term categorisation, search term frequencies and co-occurrences, as well as search strategies and tactics. The results reveal that most people use few search terms, view few Web pages, and rarely employ advanced search features such as Boolean operators or relevance feedback.

2.3 Summary of search term selection research

Table 1 illustrates key characteristics of the major search term selection studies discussed above.

Research adopting a user-centred approach to search term selection demonstrates the complexity and importance of human decision making in the information retrieval process. There is a growing interest in carrying out research into cognitive and behavioural aspects of the users' search term selection and into the factors and variables that affect the process. Searching behaviour and the search process have been studied to explore issues associated with different types of user-system interaction and to discover how variables such as the user's initial request, search experience, environment, and domain knowledge affect the term selection process.

Since the search process in general and search term selection in particular involve a range of cognitive and behavioural characteristics, most studies have employed a combination of data gathering techniques to provide sufficient qualitative as well as quantitative data. The qualitative techniques derive mainly from disciplines within the social sciences, human-computer interaction and psychology. Data gathering tools utilised in these studies have included questionnaires (pre-search, during-the-search and post-search), interviews

(both audio and video taped), the think aloud technique, transaction logging, and observation.

Most of the studies reviewed above have considered in varying degrees of detail the role and influence of thesauri as one source for term selection. While some regarded thesauri as marginal or peripheral tools, others considered them to be substantial sources of terms which users can take advantage of during their online searching.

Format and presentation of thesauri appear to affect their use. Most of the thesauri used by searchers and users in the studies noted were in printed format and were thus considered to be external sources. Even in those studies where thesauri were available in electronic format, no specific attempt was made to examine and evaluate the user-thesaurus interaction as a part of the term selection process.

Most of the studies reviewed focus on the term selection behaviour of professional searchers. Those studies specifically dealing with end-users give little attention to the role of thesauri as aids to search term selection. Rather their results are limited to suggestions on methods to facilitate the use of thesauri by end-users during the online search process. For instance Fenichel (24) suggests the need for facilities such as hierarchical display of descriptors or "exploding" options to include all narrower terms of a descriptor. Hsieh-Yee (25) proposed that front-end software could incorporate features that would encourage searchers to survey possible search terms, such as prompting the user as to whether synonyms should be included as search terms. Vakkari (31) points to the need for system-provided synonyms and narrower terms as a means of facilitating search term selection and query reformulation.

As the number of electronic thesauri attached to information retrieval systems has grown several interface features and facilities have been developed to aid users in the selection

of search terms. However, almost no research has been carried out into how end-users interact with these types of interfaces or into the ways in which these integrated thesauri affect search term selection.

Research is required to shed light on the search term selection behaviour of various end-user communities who make use of IR systems that provide thesaurus and other terminological support for improving search performance. The interaction between end-users and thesauri and the ways in which thesauri affect their search term selection is an area clearly in need of research. This research should not only address thesauri as term sources but also evaluate the impact of different types of user interface used to provide thesaurus-aided search facilities.

3. Query expansion

Query expansion is defined as a stage of the information retrieval process during which a user's initial query statement is enhanced by additional search terms in order to improve retrieval performance. As studies addressed specific stages of the search process the application of thesauri in this expansion and reformulation became an area of increasing interest. Query expansion is rationalised by the fact that initial query formulation does not always reflect the exact information need and request of the user, therefore, the query may often be enhanced by the addition of search terms in a manner that results in improved information retrieval.

Three types of query expansion are discussed in the literature, namely: manual, automatic and interactive (also known as semi-automatic, user-mediated or user-assisted). These approaches use different sources of search terms and a variety of expansion techniques (13). Beaulieu and Robertson (43) have argued that the distinction between manual and

interactive methods of query formulation is problematic, since both involve human intervention. The difference is that the manual approach does not include any consultation of the collection, while in the interactive approach the query is modified through a feedback process. In both cases, however, assistance can be sought from other sources, including a dictionary or thesaurus.

Spink (2) has commented that extensive research has been carried out on automatic and semi-automatic query expansion techniques. Automatic query expansion techniques exploit the text of a user's question and/or retrieved documents found to be relevant by the user, as input for techniques to derive a set of search terms to retrieve additional relevant documents. In interactive query expansion users are responsible for selecting from candidate search terms suggested by the retrieval system. Several studies have been conducted on interactive query expansion to evaluate the ranking algorithms based on users' relevance judgement of candidate search terms (44, 8), to study user interaction through graphical user interfaces (45-46), and to investigate the simulated users' term selection within the context of interactive query expansion (47,10).

The human approach to query expansion research, "investigates the user's representation of their question and whatever tools e.g. thesaurus or experiences they use to extract or modify a set of search terms during query expansion" (2). This approach stresses the importance of decision-making as well as behavioural and cognitive characteristics of users in reformulating and expanding their search statements.

3.1 Query expansion using thesauri

Several studies have reported the construction and use of different types of thesauri as aids to the query expansion process. In general thesauri within information retrieval

systems can be categorised as belonging to one of three main types: standard manually constructed thesauri, searching thesauri, and automatically constructed thesauri.

Standard thesauri with hierarchical, equivalence and associative relationships have been widely used for search term selection and query expansion purposes. Much of the research in this area has focused on comparing the performance and effectiveness of controlled vocabularies versus free text terms in information retrieval (48-54). These types of thesauri have also been incorporated as knowledge bases or interface components in several prototype expert and intelligent systems to assist users in the process of search terms selection and query expansion (13).

Searching thesauri, also referred to as end-user thesauri, are defined as a category of thesauri enhanced with a large number of entry terms that are synonyms, quasi-synonyms or term variants which assist end-users to find alternative terms to add to their search queries (49, 55-57). A number of searching thesauri have been designed and developed (58-59) and have been evaluated in query expansion research (60-62).

The design and testing of several types of automatically constructed thesauri has also been extensively reported in the literature. A number of researchers have constructed co-occurrence based thesauri to evaluate the performance of thesaurus-based query expansion (63-64). Using a laboratory environment and the TREC test collections, these studies resulted in a slight improvement in retrieval performance. General-purpose thesauri such as WordNet have also been evaluated in the query expansion process but have demonstrated little difference in retrieval effectiveness (65). Thesauri constructed automatically using a linguistic approach have also demonstrated a marginal improvement in retrieval performance (66). Combining different types of thesauri for

query expansion has shown better retrieval results than using only one type of thesaurus (67).

Automatically constructed thesauri have also been evaluated in user-oriented environments (68-70). In addition some researchers have found that the integration of automatically and manually constructed thesauri has a positive effect on the query expansion process (71-74).

The research reviewed in the following section focuses on those studies that have investigated thesaurus-based query expansion using standard domain-specific thesauri in a user-oriented environment.

3.2 Thesaurus-aided query expansion: user-centred studies

Research on user-assisted query expansion using domain-specific thesauri can be divided into three categories depending on the extent to which users are involved in the search and retrieval process:

- users involved only in providing requests and relevance judgements;

- user involvement mediated by professional searchers;

- users operating in an interactive environment.

The following sections review studies conducted in each of these categories.

3.2.1 *Users involved only in providing requests and relevance judgements*

In the studies reported in this section users were not involved in the actual search process, which was typically performed by a professional searcher, who also carried out any thesaurus-aided query expansion. Kristensen and Jarvelin (60) studied the effectiveness of a small searching thesaurus, with 328 terms, on recall and precision in a full-text

database. The test environment was an operational database with around 34,000 newspaper articles relating to economic issues and the queries were elicited from five journalists. The researchers performed the searches while relevance judgements were undertaken by the journalists. Each query was searched in three distinct modes: basic search, synonym search, and related term search. The basic search included only the journalists' initial query statements. Expanding original queries with synonyms using a searching thesaurus significantly increased the number of new relevant records. Query expansion using related terms also increased the recall rate but led to a marked decline in precision. Thus related terms can be used as query expansion terms if high recall is required. They concluded that a searching thesaurus could improve results in free-text searching of a full-text database.

Following this study, Kristensen (62) investigated the effects of the searching thesaurus on recall and precision in a full-text database using an expanded set of five distinct search modes: basic search, synonym search, narrower term search, related term search, and union of all these searches. In this study, an operational database of 227,000 newspaper articles, a test thesaurus of 1573 terms, and 30 queries elicited from journalists constituted the test environment. The researcher carried out all modes of search, with relevance judgement again being supplied by the journalists. The study concluded that the effect of a search-aid thesaurus was substantial and improved retrieval with twice as many relevant documents being found using the union search as compared to the basic search, with only a 10% decrease in precision. Each of the three expanded search modes retrieved several unique articles but it was found that the expansions using synonyms and related terms performed better in terms of relevance than those using the narrower terms. The findings showed that a searching thesaurus was clearly a recall-enhancing tool and the author

suggests that the active involvement of end-users in the term selection process could enhance the levels of precision.

3.2.2 *User involvement mediated by professional searchers*

To evaluate thesaurus-aided query expansion in a more user-oriented setting, some researchers have studied the selection of search terms for query expansion by users in mediated search environments. Spink (2) for instance examined the selection and effectiveness of search term sources while observing used-based query expansion to provide guidelines for IR system enhancement. The users' written question statements and pre-online interviews as well as the search logs created during online interaction between 40 users and professional searchers were collected. The study identified five different sources of term selection for query expansion: user question statement, user interaction (suggested by user during interaction with intermediary or online search), intermediary, thesaurus, and term relevance feedback (TRF) based on terms extracted by users after examination of retrieved documents. Spink also evaluated the effectiveness of these sources in retrieving relevant items with a special focus on term relevance feedback. The thesaurus was rated as the third most effective source of search term after the user's question statement and user interaction with the intermediary. The results showed that the TRF search terms were more effective in retrieving relevant items than search terms suggested by the professional searchers or those selected from the thesaurus. TRF terms that proved effective in retrieving relevant records were largely selected from the title and descriptor fields of the records viewed. This finding suggests that descriptors, which will normally be terms in the thesaurus, contributed significantly to the retrieval of relevant records. Thus, due to the fact that end-users did not have any direct interaction with the

thesaurus as a main source of term selection, the value of thesaurus-based terms may have been underestimated. The study concluded that users' written question statements, terms derived during the interaction between users and professional searchers, and terms selected from the title and descriptor fields of the retrieved records were the most effective sources of search terms for query expansion.

Greenberg (75-76) investigated the effectiveness of different thesaural relationships for automatic and interactive query expansion in an operational environment. She aimed to explore how end-user search terms can be mapped to a thesaurus and if there are additional thesaurus terms representing some type of semantic relationship to investigate which are good candidate terms for interactive and automatic query expansion. The test environment consisted of the *Proquest* controlled vocabulary, ABI/Inform (a business periodical database), and forty-two queries obtained from business administration students. The study collected data from user profile questionnaires, relevance judgements, and a post-evaluation questionnaire. In evaluating automatic query expansion using thesauri, different query treatments were reported in the study. These 'treatments' differed in terms of the sources used to expand the query. They included the participants' initial search statement and five treatments involving the thesaurus: mapped terms, narrower, broader, related, and synonymous terms. In the interactive query expansion process, the participants were presented with lists of thesaurus terms related to their initial queries and were asked to select search terms that they thought would have been useful. In both the automatic and interactive cases the researcher developed the search strategy and performed all the searches while end-users were mainly responsible for providing real information requests together with relevance judgement. The findings suggested that synonyms and narrower terms yielded more precise results and are good candidates for automatic query expansion because they increase recall with a minimum loss in precision.

All semantic relationships increase recall if applied in automatic query expansion. If precision is required, narrower terms can be good candidates for interactive query expansion. Related terms can be better candidate terms for interactive query expansion than for automatic query expansion. The results also demonstrated that end-users can have a significant impact on the precision of the results when selecting terms via interactive query expansion. This study indicated that semantic relationships in standard thesauri can have an impact on retrieval performance and can be used as a source for query expansion.

3.2.3 *Users operating in an interactive environment*

The research reviewed in the previous two sections covers studies with only partial user involvement in thesaurus-aided search term selection for query expansion. There are few studies that have investigated such selection behaviour by end-users in an interactive environment. Jones et.al. (77) for instance investigated user interaction behaviour with a thesaurus as a source of query expansion terms to identify strong patterns for possible rule- or weight-based systems for term expansion and to compare the effectiveness of queries enhanced by thesaurus use with that of the original query terms. The INSPEC thesaurus, 39 users with real information needs and OKAPI, a ranked output information retrieval system, were the experimental environment within which the users' thesaurus navigation and term selection behaviour was recorded. The interface was implemented in such a way that following the entry of the original query users were shown exact or partially matched thesaurus terms and encouraged to browse and select thesaurus terms for expansion. The researchers examined the number and types of terms seen and selected by users. The analysis showed that the majority of terms retrieved by thesaurus navigation

came through the association relationship, perhaps reflecting the fact that INSPEC had more associative links than hierarchical ones. The focus of the study was on the performance of thesaurus terms in relevant retrieval. A comparison of the results generated by original queries with those generated by the selected thesaurus terms was made based on the users' relevance judgements (using only author, title, and journal) and showed no significant differences in the retrieval performance. The study also evaluated the retrieval performance of hybrid searches, which included both the original and controlled terms. The hybrid search retrieved more records than the original search although it did not show a marked improvement in overall performance. The researchers concluded that thesaurus-based query expansion may increase recall and uncover additional relevant documents, but that is will not improve queries which are already quite fully specified. It was also suggested that users obtain good results if they have a large number of terms from which to select, and that since their choice of terms will differ it is infeasible to design a generic automatic expansion procedure. Rather tools should be provided to aid the end-user in thesaurus navigation. User feedback also indicated that thesaurus navigation was a useful and informative activity. It was also found that depth and coverage of the thesaurus and the quality of the user interface were important issues in determining the degree of enhancement to the query expansion process.

Beaulieu (46) reported on experiments conducted to evaluate different interfaces which support query expansion, again based on the Okapi search engine. The experiments were carried out using an online library catalogue and also the INSPEC database and thesaurus. Observation, talk-aloud, online pre-search and post-search questionnaires were used together with transaction logs to gather data. Three types of interfaces supporting automatic and interactive query expansion were tested by examining user interaction behaviour. Thesauri were used in different ways within the interface design, as tools for

supporting interactive query expansion. The results showed that both the explicit and implicit use of a thesaurus (using interactive or automatic query expansion respectively) can be beneficial. It was found that the overall number and specific presentation of candidate terms for query expansion was important. For instance the way that thesaurus terms and terms extracted from the documents were displayed affected the ultimate set of search terms selected by the user. The study also suggested that the different cognitive styles adopted by users when seeking information were an important issue in designing interfaces that could effectively support query expansion.

3.3 Summary of thesaurus-aided query expansion research

Table 2 provides an overview of some key characteristics of the query expansion studies discussed above. This research has focused on the utilisation of domain-specific thesauri for query expansion in user-oriented settings. While some studies employed laboratory-type experiments to examine the retrieval performance of thesaurus-aided query expansion, the majority adopted a user-centred approach and analysed the user - thesaurus interaction in an operational setting. The results of these studies demonstrate the usefulness of thesauri both in terms of providing users with alternative search terms for query expansion and in improved retrieval performance. However, these studies focused primarily on search outcomes rather than the search process. Interactive query expansion research (46) has shown that user interface design can play a major role in encouraging the use of thesauri as query expansion term sources. This has implications for the design and evaluation of interfaces enhanced with thesauri to support query expansion. Efthimiadis (78) has suggested that during query expansion, a thesaurus could be used to display the relationships of the selected terms to other terms. This could be achieved, for example, by displaying the hierarchical tree to which a term belongs (as in the INSPEC or MESH tree displays) or by presenting broader, narrower or related terms on screen for

users to browse and make selections from. As end-user searching becomes more prevalent, research on user - thesaurus interaction within the context of the information searching process is also needed to evaluate the ways in which various end-user communities use and interact with thesauri and how this affects their term selection for query expansion. Such research must take into account the user's attitudes, cognitive aspects of the search process, and the mechanisms by which the user may select terms, issues which are little investigated in the reviewed literature. Research also suggests that the coverage and richness of thesauri play a significant role in their contribution to the users' term selection for query expansion (77, 79). Given the fact that few domain-specific thesauri have been evaluated in terms of their coverage and performance for query expansion, research needs to be carried out to evaluate thesaurus-aided query expansion in a range of subject domains.

4. Conclusion

This review has provided a survey of user-centred studies on search term selection and query expansion. In particular the use of domain-specific thesauri as sources of term selection has been examined. The studies reviewed were classified according to the type of user involved and it was seen that most of the search term selection studies investigated the behaviour of professional searchers. This was due to the fact that a number of the studies were carried out at a time when mediated online searching was common practice. In addition, most information retrieval systems did not have integrated online thesauri and so professional searchers were the main users of these tools. Research demonstrated that user characteristics, including cognitive and behavioural factors, are important in analysing end-user search term selection.

Query expansion studies by contrast have been more concerned with the end-user, although in most cases end-users were not involved in the actual search process. Moreover the focus of thesaurus-based query expansion research has been on retrieval performance rather than the search process and its cognitive aspects. The results of these studies demonstrated that thesauri have the potential for improving retrieval performance and can be used for both automatic and interactive query expansion.

Two major trends in modern information retrieval justify additional research into thesaurus-assisted search term selection and query expansion by end-users. First, end-user searching is widely becoming universal owing to the developments associated with the Internet and the World Wide Web. To provide better systems, research which explores the behaviour and attitudes of different types of end-users who access thesaurus-enhanced information retrieval systems is needed. Second, a growing number of commercial retrieval systems have recently incorporated thesauri into their search interfaces to encourage users to enhance their queries. This development opens avenues for more research into the usefulness and usability of these systems and the extent to which they do in fact assist end-users in selecting search terms and in expanding queries. Studying end-users' interaction with online thesauri will extend our knowledge of user behaviour and provide guidance for the design and implementation of better systems.

References

(1) Saracevic, T.: Information Science. *Journal of the American Society for Information Science*, 50 (1999) No 12, p. 1051-1063.

(2) Spink, A.: Term relevance feedback and mediated database searching: implications for information retrieval practice and systems design. *Information Processing and Management*, 31 (1994) No 2, p.161-171.

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

(4) Spark Jones, K.: Search term relevance weighting given little relevance information. *Journal of Documentation*, 35 (1979) No 1, p. 30-48.

(5) Van Rijsbergen, C.J., Harper, D..J., and Porter, M.F.: The selection of good search terms. *Information Processing and Management*, 17(1981) No 2, p.77-91.

(6) Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24 (1988) No5, p. 513-523.

(7) Robertson, S.E.: On term selection for query expansion. *Journal of Documentation*, 46(1990) No4, p. 359-364.

(8) Efthimiadis, Efthimis N. A user-centred evaluation of ranking algorithms for interactive query expansion. Korfhage et.al. SIGIR93: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 16th Annual International Conference on Research and Development in Information Retrieval. 146-159. 93. New York, ACM, 1993.

(9) Robertson, S., Walker, S., and Beaulieu, M.: Laboratory experiments with Okapi: Participation in the TREC programme. *Journal of Documentation*, 53 (1997) No1, p. 20-32.

(10) Magennis, M. and Van Rijsbergen, C. J. The potential and actual effectiveness of interactive query expansion. SIGIR97: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 20th Annual International Conference on Research and Development in Information Retrieval. 27-31 July, Philadelphia, 1997.

(11) Text REtrieval Conference (TREC) proceedings: Available at: http://trec.nist.gov

(12) Fidel, R.: Searchers' selection of search keys: II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science*, 42(1991) No 7, p. 501-514.

(13) Efthimiadis, E.N. Query expansion. In: *Annual Review of Information Science and Technology*, edited by M.E. Williams, pp. 121-187. Information Today, Medford, NJ. 1996.

(14) Bates, M.J.: Information search tactics. *Journal of the American Society for Information Science*, 30 (1979) No 4, p. 205-214.

(15) Fidel, R.: Online searching styles: a case-study-based model of searching behaviour. *Journal of the American Society for Information Science*, 35 (1984) No 4 , p. 211-221.

(16) Fidel, R.: Towards expert systems for selection of search keys. *Journal of the American Society for Information Science*, 37(1986) No 1, p. 37-44.

(17) Fidel, R.: Searcher's selection of search keys: I. The selection routine. *Journal of the American Society for Information Science*, 42: (1991) No 7, p. 490-500.

(18) Fidel, R.: Searchers' selection of search keys: III. Searching styles. *Journal of the American Society for Information Science*, 42(1991) No 7, p. 515-527.

(19) Saracevic, T., Kantor, P., Chamis, A., and Trivison, D.: A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information science*, 39(1988) No 3 , p. 161-176.

(20) Saracevic, T. and Kantor, P.: A study of information seeking and retrieving: Users, questions and effectiveness. *Journal of the American Society for Information Science*, 39(1988) No 3, p. 177-196.

(21) Saracevic, T. and Kantor, P.: A study of information seeking and retrieving .III. Searchers, searches, and overlap. *Journal of the American Society for Information Science*, 39(1988) No 3, p. 197-216.

(22) Spink, A.: Multiple search sessions model of end-user behaviour: an exploratory study. *Journal of the American Society for Information Science*, 47(1996) No 8, p. 603-609.

(23) Spink, Amanda, Goodrum, Abby, Robins, David, and Wu, Mei Mei. Elicitations during information retrieval: implications for IR system design. SIGIR' 96: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 19th Annual International Conference on Research and Development in Information Retrieval. 120-127. 1996.

(24 Fenichel, C.H.: Online searching: measures that discriminate among users with different types of experiences. *Journal of the American Society for Information Science*, 32(1981) No 1, p. 23-32.

(25) Hsieh-Yee, I.: Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*, 44 (1993) No 3 , p. 161-174.

(26) Iivonen, Mirja. Searchers and searchers: Differences between the most and least consistent searchers. Fox, Edward, Ingwersen, Peter, and Fidel, Raya. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 149-157. 95. New York, ACM. 1995.

(27) Iivonen, M. and Sonnenwald, D.H.: From translation to navigation of different discourses: A model of search term selection during pre-online stage of the search process. *Journal of the American Society for Information Science*, 49(1998) No 4, p. 312-326.

(28) Bates, M.J., Wilde, D.N., and Siegfried, S.L.: An analysis of search terminology used by humanities searchers: The Getty online searching project report No 1. *Library Quarterly*, 63 (1993) No 1, p. 1-39.

(29) Siegfried, S., Bates, M.J., and Wilde, D.N.: A profile of end-user searching behaviour by humanities scholar: The Getty Online Searching Project Report No 2. *Journal of the American Society for Information Science*, 44 (1993) No 5, p. 273-291.

(30) Sutcliffe, A.G., Ennis, M., and Watkinson, S.J.: Empirical studies of end-user information searching. *Journal of the American Society for Information Science*, 51(2000) No 13, p. 1211-1231.

(31) Vakkari, Pertti.: Cognition and changes of search terms and tactics during task performance: a longitudinal case study. RIAO 2000: Content-based Multimedia Information Access.

(32) Vakkari, P.: A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study. *Journal of Documentation*, 57(2001) No 1, p. 44-60.

(33) Marchionini, G. et.al. : A study of user interaction with information retrieval interfaces: progress report. *The Canadian Journal of Information Science*, 16 (1991) No 4, p. 42-59.

(34) Meadow, C., Wang, J., and Yuan, W.: A study of user performance and attitudes with information retrieval interfaces. *Journal of the American Society for Information Science*, 46(1995) No 7, p. 490-505.

(35) Brajnik, Giorgio, Mizzaro, Stefano, and Tasso, Carlo. Evaluating user interfaces to information retrieval systems: a case study on user support. Proceedings of the 19th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval. 128-136. 96. Germany, Hartung-Gorre Verlag Konstanz, 1996.

(36) Belkin, N.J .et.al. : Iterative exploration, design and evaluation of support for query reformulation in interactive information retrieval. *Information Processing and Management*, 37 (2001) No 3, p. 403-434.

(37) Marchionini, G.: Information-seeking strategies of novices using a full-text electronic encyclopaedia. *Journal of the American Society for Information Science*, 40 (1989) No1, p. 54-66.

(38) Solomon, P.: Children's information retrieval behaviour: a case analysis of an OPAC. *Journal of the American Society for Information Science*, 44 (1993) No5, p. 245-264.

(39) Spink, A., Bateman, J., and Jansen, B.J.: Searching the web: a survey of Excite users. *Internet Research: Electronic Networking Applications and Policy*, 9 (1999) No 2, p. 117-128.

(40) Spink, A., Jansen, B.J., and Ozmultu, H.C.: Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic networking Applications and Policy*, 10 (2000) No 4, p. 317-328.

(41) Spink, A., Wolfram, D., Jansen, B.J., and Saracevic, T.: Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52 (2001) No 3, p. 226-234.

(42) Jansen, B.J., Spink, A., and Saracevic, T.: Real life, real users, Real needs: A study and analysis of user queries on the web. *Information Processing and Management*, 36 (2000) No2, p. 207-227.

(43) Beaulieu, M., Robertson, S., and Rasmussen, E.: Evaluating interactive systems in TREC. *Journal of the American Society for Information Science*, 47(1996) No 1, p. 85-94.

(44) Efthimiadis, E.N.: User choices: a new yardstick for the evaluation of ranking algorithms for interactive query expansion. *Information Processing and Management*, 31 (1995) No4, p. 605-620.

(45) Hancock-Beaulieu, M., Fieldhaouse, M., and Do, T.: An evaluation of interactive query expansion in an online library catalogue with a graphical user interface. *Journal of Documentation*, 51 (1995) No 3, p. 225-243.

(46) Beaulieu, M.: Experiments of interfaces to support query expansion. *Journal of Documentation*, 53 (1997) No1, p. 8-19.

(47) Harman, Donna K. Toward interactive query expansion. Chiaramella, Yves ed. Proceedings of the Association for Computing Machinery Special Interest group on Information Retrieval (ACM/SIGIR) 11th International Conference on Research and Development in Information Retrieval. 321-331. 88. Grenoble, France , Presses Universitaires de Grenoble, 1988.

(48) Markey, K., Atherton, P., and Newton, C.: An analysis of controlled and free-text search statements in online searches. *Online Review*, 4 (1980) No 3, p. 225-236.

(49) Perez, E.: Text enhancement: controlled vocabulary vs. free text. *Special Libraries*, 73 (1982) 183-192.

(50) Svenonius, E.: Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science*, 37(1986) No 5, p. 331-340.

(51) Dubois, C.P.R.: Free text vs controlled vocabulary: a reassessment. *Online Review*, 11(1987) No 4, p. 243-253.

(52) Cousins, S.A.: Enhancing access to OPACs: controlled vs natural language. *Journal of Documentation*, 48(1992) No 3, p. 291-309.

(53) Rowley, J.: The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research. *Journal of Information Science*, 20(1994) No 2 , p. 108-119.

(54) Muddamalle, M.R.: Natural language versus controlled vocabulary in information retrieval: a case study in soil mechanics. *Journal of the American Society for Information Science*, 49 (1998) No 10. P. 881-887.

(55) Piternick, A.: Searching vocabularies: a developing category of online searching tools. *Online Review*, 8 (1984) No 5, p. 441-449.

(56) Bates, M.J.: Subject access in online catalogs: a design model. *Journal of the American Society for Information Science,* 37 (1986) No 6, p. 357-376.

(57) Cochrane, P.A. Indexing and searching thesauri, the Janus or Proteus of information retrieval. In: *Classification Research for Knowledge Organization*, edited by N.J. Williamson, et al, pp. 161-178. FID, 1992.

(58) Anderson, James D. and Rowley, Fredrick A. Building end-user thesauri from full-text. Barbara H. Kwasink and Raya Fidel. Advances in Classification Research, Proceedings of the 2nd ASIS SIG/CR Classification Research Workshop. 1-13. 91. Medford, NJ, Learned Information, 1991.

(59) Knapp, S.D., Cohen, L.B., and Judes, D.R.: A natural language thesaurus for humanities. *Library Quarterly*, 68 (1998) No 4, p. 406-430.

(60) Kristensen, J. and Jarvelin, K.: The effectiveness of a searching thesaurus in free text searching of a full-text database. *International Classification*, 17(1990) No 2, p. 77-84

(61) Kristensen, Jaana and Jarvelin, K. The impact of query structure and query expansion on retrieval performance. Croft, W. B. & Moffat A. & van Rijsbergen C. J. & Wilkinson R. & and Zobel, J. Eds. SIGIR'98: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 130-137. 98. Melbourne, New York, NY, ACM Press, 1998.

(62) Kristensen, J.: Expanding end-users' query statements for free text searching with a search-aid thesaurus. *Information Processing and Management*, 29 (1993) No 6, p. 733-744.

(63) Qiu, Yoggang and Frei, H. P. Concept-based query expansion. SIGIR'93: Proceedings of the Association for computing Machinery special Interest Group on Information Retrieval (ACM/SIGIR) 16th International ACM SIGIR Conference on Research and Development in Information Retrieval. 27 June- 1 July, Pittsburgh, PA, 1993. 160-169.

 (64) Schutze, H. and Pedersen, J. A co-occurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 33 (1997) 3, p. 307-318.

(65) Voorhees, Ellen M. Query expansion using lexical-semantic relations. Croft, W. Bruce and Rijsbergen, C. J. eds. SIGIR94 Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR), 17th Annual International Conference on Research and Development in Information Retrieval. 61-69. 94. Berlin, Germany, Springer-Verlag, 1994.

(66) Jing, Y. and Croft, W. Bruce. The association thesaurus for information retrieval. RIAO '94: Intelligent Multimedia Information Retrieval Systems and Management. 146-160. 94. Paris, France, CID. 1994.

(67) Mandala, Rila, Tokunaga, Takenobu, and Tanaka, Hozumi. Combining multiple evidence from different types of thesaurus for query expansion. SIGIR' 99: Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 22nd Annual International Conference on Research and Development in Information Retrieval. 191-197. New York, ACM, 1999.

(68) Chen, H. and Dhar, V.: Cognitive process as a basis for intelligent retrieval systems design. *Information Processing and Management*, 27(1991) No 7, p. 405-432.

(69) Chen, H., Yim, T., Fye, D., and Schatz, B.: Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science,* 46(1995) No 3, p. 175-193.

(70) Chen, H. and Ng, T.D.: A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the worm community system. *Journal of the American Society for Information Science*, 48 (1997) No 1, p. 17-31.

(71) Chen, H. and Ng, T.: An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): Symbolic Branch-and-Bound search vs. connectionist Hopfield net activation. *Journal of the American Society for Information Science*, 46 (1995) No 5, p. 348-369.

(72) Schatz, Bruce R., Johnson, Eric H., Cochrane, Pauline A., and Chen, Hsinchun. Interactive term suggestion for users of digital libraries: using subject thesauri and co-occurrence lists for information retrieval. Proceedings of the 1st ACM International Conference on Digital libraries, 126-133. 1996.

(73) Chen, H., Martinez, J., Kirchhoff, A., Ng, T.D., and Schatz, B.R.: Alleviating search uncertainty through concept associations: Automatic indexing, co-occurrence analysis, and parallel computing. *Journal of the American Society for Information Science*, 49(1998) No 3, p. 206-216.

(74) Ding, Y., Ghowdhury, G.G., and Foo, S.: Incorporating the results of co-word analyses to increase search variety for information retrieval. *Journal of Information Science*, 26 (2000) No 6, p. 429-451.

(75) Greenberg, J.: Automatic query expansion via lexical-semantic relationships. *Journal of the American Society for Information Science and Technology*, 52 (2001) No 5, p. 402-415.

(76) Greenberg, J.: Optimal query expansion (QE) processing methods with semantically encoded structures thesauri terminology. *Journal of the American Society for Information Science and Technology*, 52 (2001) No 6, p. 487-498.

(77) Jones, S., Gatford, M., Hancock-Beaulieu, M., Robertson, S.E., Walker, W., and Secker, J.: Interactive thesaurus navigation: intelligence rules Ok? *Journal of the American Society for Information Science*, 46 (1995) No 1, p. 52-59.

(78) Efthimiadis, E.N.: Interactive query expansion: a user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science*, 51(2000) No 12, p. 1-25.

(79) Fidel, R.: Who needs controlled vocabulary? *Special Libraries*, 83 (1992) No 1, p. 1-9.

Table 1. Major studies which have adopted a user-based approach to thesaurus-assisted search term selection

Author(s)	Research population	Subject domain	Data collection techniques	Variables	Sources of search terms
Fenichel (1981)	Professional searchers & LIS students	Education	Questionnaire, search transcriptions	Search, searcher, search outcome, environmental	Thesaurus terms & free text terms
Fidel (1984)	Professional searchers	Health sciences	Observation, interview, think aloud, search protocol	Exploratory study – no specific variables reported	Thesaurus terms & free text terms
Fidel (1986)	Professional searchers	Life sciences	Observation, interview, think aloud, search protocol	Exploratory study – no specific variables reported	Thesaurus terms & free text terms
Saracevic et.al. (1988)	Professional searchers & end-users	Medicine, sciences, social sciences & humanities	Video-taped interview, questionnaire, search records	Users, questions, searchers, searches, retrieved items	Oral & written problem statement, thesaurus
Fidel (1991)	Professional searchers	Medicine, sciences, social sciences	Observation, interview, think aloud, search protocol	Text-word ratio, thesaurus-neglect ratio, search keys, moves, subject area, environment	Thesaurus terms & free text terms
Bates et.al. (1993)	End-users	The humanities	Transaction logs, interview	Database, search features, commands, terms and vocabulary	Free text statements and term relevance feedback
Hsieh-Yee (1993)	Experienced & novice searchers	Education	Transaction logs, think aloud, observation	Search experience, subject knowledge, search tactics	Thesaurus terms & free text terms
Iivonen (1995)	Professional searchers	Social sciences	Interview	Education, experience, environment, search term selection and strategy	Thesaurus terms & free text terms
Spink & Saracevic (1997)	Professional searchers & end-users	Medicine, social science, physical sciences and the humanities	Transaction logs, video-taped interview	Search terminology, user characteristics, user satisfaction, search process and outcome	Questions, user interaction, searcher, thesaurus, term relevance feedback
Iivonen & Sonnenwald (1998)	Professional searchers	Social sciences	Interview	Exploratory study – no specific variables reported	Thesaurus, domain and documents, indexing practice, search request, database, search experience
Sutcliffe et.al. (2000)	End-users	Medicine	Audio-video recording, think aloud	Query terms, query syntax, search strategies and effectiveness	Thesaurus terms & free text terms
Vakkari (2001)	End-users	Library & information science	Transaction logs, think aloud, interview, questionnaire	Knowledge of the topic, number and types of search terms, operators and tactics used	Free text

Table 2. Key studies which have considered the role of users in thesaurus-based query expansion

Author(s)	Research population	Subject domain	Database	Thesaurus
Kristensen & Jarvelin (1990)	End-users	Economics	Finnish newspaper articles	In-house search-aid thesaurus
Kristensen (1993)	End-users	Economics	Newspaper articles	In-house search-aid thesaurus
Spink (1994)	Professional searchers and end-users	Various subjects	Dialog bibliographic	Not specified
Jones et.al. (1995)	End-users	Computer, electronics, & information science	INSPEC	INSPEC
Beaulieu (1997)	End-users	Computer, electronics, & information science	INSPEC	INSPEC
Greenberg (2001)	End-users	Business	ABI/Inform	Proquest controlled vocabulary

Correspondence to: Ali Asghar Shiri
Department of Computer and Information Sciences, University of Strathclyde,
26 Richmond Street, Glasgow G1 1XH Scotland, UK

Tel: +44-141- 548 4065
Fax: +44-141- 553 1393
Shiri@dis.strath.ac.uk