Thesaurus-aided search

Denisa Popescu
Graduate School of Business
George Washington University
Washington, DC

Introduction

Among emerging technologies, the Internet is the most prominent. The Internet is usually considered the democratizing medium of shared knowledge because its decentralized structure allows anyone to publish and search for any type of information. Nevertheless, many advocates of the Internet often confuse access to information with access to relevant information. We all agree the Internet provides the ideal platform for global access to information, but important issues must be addressed for the Internet to achieve its full potential as a tool for global access. As Brian Pinkerton states, "The World Wide Web is decentralized, dynamic and diverse; navigation is difficult and finding information can be a challenge" (Pinkerton, 1994). And yet there are few tools to integrate and organize related and relevant information from different sources.

Fifty years ago, Vannevar Bush believed that technology can bring "a new relationship between thinking man and the sum of our knowledge" (Bush, 1945). He was particularly concerned with information structuring, retrieval and transmission, and imagined technologies that would allow scientists to deal with the massive information overload in ways in which people think, not linearly but by association. Bush believed that the most important component in representing knowledge was semantic content of information. In this context, the semantic structures provided by controlled vocabularies can play an important role in both representing and retrieving Web information.

Because of the amazing interconnectivity of the Web, information retrieval must be perceived as a social activity that involves thousands of users in addition to intelligent agents with different cultural, social and economic backgrounds, different languages, and many ways of expressing their own knowledge and

information needs. Typically, information retrieval on the Web consists of browsing and searching activities in which users attempt to locate relevant resources for their information needs. In the first situation, a user who has an interest that is poorly defined might use an interactive interface (i.e., hierarchical) to simply look around in the collection for documents. On the other side, search mechanisms give users more control in performing a retrieval task and locating information in response to a defined query that intends to describe a specific information need. However, the main problems with conventional Web searches are that relevant resources to the users are not identified, usually because different words were used in the document that did not match the search terms directly. This problem refers to the situation in which the search mechanism is based on character string matching and does not make any attempt to understand the meaning of the terms used. Another problem derives from the so-called "vocabulary problem," where users may know what they are looking for, but are unable to clearly articulate the problem in terms recognized by the search system (Cooper and Byrd, 1997).

Borgman (1986) identifies two types of knowledge necessary to search effectively: mechanical aspects of searching including syntax and semantics of the query language and the conceptual aspects of system relating to ways to broaden and narrow searches using vocabularies. To address the conceptual problems of defining an information need, controlled vocabularies, like thesauri, may be used to complement free-text searches and allow searchers to use a standardized search language to better define their information problem and improve retrieval (Fidel, 1990). This research will suggest ways that controlled vocabularies can be used to better organize information and help the users retrieve relevant information.

A thesaurus captures and structures knowledge in a specific domain and, by so doing, captures the meaning of concepts that are specific to that domain. This meaning is then extended to end-users through retrieval tools, such as the search engine that would apply the thesaurus concepts and relationships for better information retrieval.

While free text searching retrieves only the resources containing the words specified in the query, thesaurus-aided searching retrieves resources containing the specified term and any equivalent terms. Terms submitted through the search will be pre-processed through the thesaurus and this process will automatically expand the number of terms in the query, incorporating any synonym terms found in the thesaurus—automatic query expansion. If a user is not happy with the results because they are too general or too specific, it may be necessary to consult a thesaurus to aid in the selection of terms. After the initial query is submitted, the system returns a list of related terms derived from a thesaurus to complement the initially chosen term—interactive query expansion (Harman, 1988).

Thesaurus-aided search is one example of how controlled vocabularies can be used to improve the quality and completeness of answers to user queries on the Web. When thesaurus-aided search is considered for query formulation, refinement and expansion, we argue that it is an effective tool that can help users define their information needs more precisely and therefore contribute to increased relevance and efficiency of information. My final aim is to provide a better understanding of the impact of controlled vocabularies on Web searches, as well as describe and explain users' reactions to the experience of a thesaurus-aided search.

Literature Review

This section reviews the relevant research literature related to application of thesaurus in information retrieval on the Web in the area of information searching behavior, search term selection and query expansion, and information retrieval interface evaluation.

Searching behavior studies

In studying information-seeking strategies of elementary school children searching a full-text encyclopedia, Marchionini (1989) found that users had difficulties in selecting terms for query formulation. Suggestions for

an addition of a thesaurus or usage-sensitive search aid were made towards providing more efficient and effective searches.

Yee (1993) investigated the effect of search experience and subject knowledge on the search tactics of novice and experienced searchers. She pointed out that search experience affected searchers' use of many search tactics, and suggested that subject knowledge became a factor only after searchers have had a certain amount of search experience. She suggested that new interfaces that encourage searchers to view possible search terms for selection and take advantage of an online thesaurus would be desirable.

Vocabulary Problem

The importance of vocabulary problem has been illustrated by the numerous researchers. The major theoretical description of the vocabulary problem came from Furnas et al. (1987). Their work has discussed the situation where users must translate their information problems into query terms, and not knowing which terms will produce the sought information constitutes a significant barrier for all search users. The authors argue that there are few aspects to vocabulary problem. When translating an information need into a query, the query terms might contain ambiguous terms or the concepts user may not be precise enough to ensure an adequate response. There is also the situation where users failed to find what they were looking for only because the content they were searching against contained different representations of the terms they searched for. They concluded that there is a need to recognize the use of vocabularies to address this problem. Several approaches to alleviate the vocabulary problem include search term selection and query expansion and retrieval feedback.

Search term selection and query expansion research

Fidel (1991) studied the search key selection behavior of 47 professional online searchers while performing search tasks. Her research showed that better quality and availability as well as support for cross-database

searching are likely to increase the use of controlled vocabularies. She found the use of the thesaurus terms for query formulation or expansion to be a major characteristic of searcher information searching behavior.

Saracevic (1997) studied the end-user search term selection behavior under real-life circumstances using a database thesaurus, users written question statements, terms derived from relevance feedback, user's domain knowledge, and intermediaries during the interaction as a source for selection of search terms. The authors argue that selection of search terms for query formulation is a dynamic and highly interactive process that calls for enhancing interface features to better support users.

Efthimiadis (2000) in a user-centered investigation of interactive query expansion in a relevance feedback environment concluded that query expansion terms were identified as being synonyms ore related terms to the initial query terms. Based on this finding, he suggested that during query expansion, a thesaurus could be used for displaying the relationships of the selected terms to other terms for example by displaying the hierarchical tree to which the term belongs or by presenting broader, narrower or related terms on the screen for user to browse and select from. He also emphasized the need for more research into the process of term selection by users because of its importance for understanding the users' searching behavior.

Harman (1988) developed techniques that automatically select feedback terms to offer guidance to user wishing to improve their initial queries. Results showed that query expansion by these terms provides significant improvement over no query expansion.

The research undertaken by Jones et.al.(1995) look into the thesaurus searching and browsing behaviors of the users and how they interacted with thesaurus as a source of term expansion. The thesaurus-aided query expansion was examined in a relevance feedback environment with real users in an attempt to identify some patterns to incorporate into automatic procedures.

IR Interface evaluation

Bates (1986) has discussed the improvement of both interfaces and vocabularies of information retrieval systems to support users in the selection of search terms. Her model of an end-user thesaurus and a front-end system mind (FSM) provides end-users with a wide range of search terms, alternative term displays and various approaches to term selection.

Borgman et al. (1989) studied the information needs and information-seeking habits of researchers in the energy domain and designed and evaluated a set of microcomputer-based training and assistance programs to support end user access to the energy database on the DOE information retrieval systems. The study showed that there was a need to provide search term support for end-users as they had considerable difficulties in selecting search terms. They concluded that some of the vocabulary problems could be attributed to the lack of online access to the energy database subject thesaurus to find out the alternative broader, narrower, and related terms. However, they also commented that a significant proportion of the vocabulary problems could be resolved through having index browsing capabilities. There was no indication as to whether having an index browsing facility in place would eliminate the need for thesauri.

Research Hypothesis

The research will examine whether the thesaurus-aided web searching will improve the relevance and efficiency of web searching for users.

H1. The use of thesaurus terms for query formulation will improve the relevance of searches for the users.

H2. The use of thesaurus terms for automatic query expansion will improve the relevance of searches for the users.

Due to the sheer volume of unstructured information on the Web, it is becoming increasingly difficult to locate useful information and there is an immediate need for technologies that give a well-defined meaning to information. A potential substitute to represent meaning in a Web environment is the thesaurus that the indexer uses to interpret and represent the themes, concepts, and language of the content and that the searcher uses to interpret and represent sometimes vague expressions. Previous research has shown that the thesaurus will shape the terminology of the search queries and increase users' ability to find what they need with better results. Evidence was found to support this hypothesis.

Thesaurus aided web searching is defined as:

- exact phrase searching, where the query phrase is validated against a thesaurus before it is matched against search indexes
- where the query match includes transparent search for all synonyms
- where the results display includes presentation of
  - broader concepts to broaden the concept scope of the query
  - narrower concepts to narrow the concept scope of the query
  - related concepts

Relevance as a measure of effectiveness of information retrieval refers to users judgments of the retrieved documents to his information needs. (Greisdorf, 2000) For example, if a user clicks on a document, it is likely that the document is relevant to the query, or at least related to it to some extent. Therefore, for our purposes, we do consider a clicked document to be relevant to the query.

Methodology

The study will use search logs analysis as part of the research framework. We investigated a large sample of searches, represented by logs of queries from Gateway search including variables such as query

sessions that lead to document clicks (user-supplied queries and query refinement), term selection characteristics and composition of query refinement from thesaurus relationships terms.

Sample

In studying actual thesaurus aided web searching by the public at large, we analyzed queries by users of the Development Gateway search engine. DG searches are based on the exact terms a user enters in a query. An online thesaurus is used for automatic query expansion and to find related terms for the terms entered. Terms submitted through the search will be pre-processed through the thesaurus and this process will automatically expand the number of terms in the query, incorporating any synonym terms found in the thesaurus. If a user is not happy with the results because they are too general, the initial search will also return a list of narrower terms related to the original search terms which the user might use to narrow the search on a secondary try. If the initial search has fewer results, then the user can select among the broader relationships terms that are relevant to his query and perform the search again. The search results page will include the narrower, broader or related terms only if there is an exact match between the user query and the thesaurus list of terms. Search response is in result pages, listing URLs and a short description of sites that match the query.

The data we analyzed consisted of a log of transaction record of 193,727 user queries submitted during a period of three months. User activities are time-stamped and include query terms, query options, and thesaurus use. The log records the number of returned documents (results) as well as the number of resulting documents that the user chooses to view (click) for each query.

Data Collection

The variables examined are divided into three conceptual clusters and are evaluated by the search logs. The categories include: request characteristics, thesaurus usage characteristics, and retrieval effectiveness.

Each request contains one query, a set of documents which the user clicked on (clicked documents), and set of results provided by the search engine. Thesaurus usage characteristics provide information about the query expansion terms that were selected by the system, and query refinement terms that were selected by the user. Retrieval effectiveness is based on the assumption that the clicked documents are relevant to the query. Table 1 identifies each of the variables measured in each of the concept clusters, as well as its definition.

Table 1. Table of Measures

s_time	Start time of user query request
f_time	End time of user query request
query	User query request terms
results	Number of results
clicks	Number of clicked documents
no_terms	Number of thesaurus terms displayed for query refinement (RT, BT, NT)
no_syn	Number of synonyms used for automatic query expansion
term_p	Identifies if the request is initiated from the thesaurus terms (0,1)
clicksr	Clicks recoded (0,1)
thes	User-supplied queries that match a term in a thesaurus (0, 1)
resultsr	Results recoded (0,1)

Results

The data is summarized in Table 2. For the majority of queries (63.1%), the search returned a set of results comprising one or more documents. Disappointingly, the majority of queries (94.7%) did not lead to users viewing document content, with 3.3% of queries resulting in the viewing of only one document. Therefore, one can argue that overall the returned documents are not relevant to user queries.

Table 2. Clicked documents and results set

Results recoded	Valid Percent	Clicks recoded	Valid Percent
No results	36.9%	No clicks	94.7%
One or more results	63.1%	One or more clicks	5.3%
Total	100%		100%
	N= 193727		N= 193727

For example, in 65% of all requests, the DG search engine consulted the thesaurus terms either for automatic query expansion or to present the results as an aide for user query refinement. Of the thesaurus-associated queries, only 3% were initiated as a query refinement task from the thesaurus list of terms.

Table 3. Use of Thesaurus

	Valid Percent
No use of thesaurus	34.1%
Use of thesaurus either for query expansion or query refin.	65.9%
Total	100%
	N= 193727

The number of results is positively correlated with both the number of synonyms used in automatic query expansion and the number of clicked documents. On the one hand, the use of thesaurus terms (SYN)

improves the number of results, and on the other hand more results are associated with more clicks. Thus one can argue that the use of synonyms for automatic query expansion has an indirect effect on the relevance of information retrieved through its positive impact on the results set.

Table 4. Correlations Results * No_syn, Clicks

	No_syn	Clicks
Results Pearson Correlation	.235**	.027**
Sig. (2-tailed)	.000	.000

Correlation is significant at the 0.01 level (2-tailed).

To investigate the main hypothesis, a cross tabulation was conducted to provide proof that the above statement was true. Table 5 indicates that query refinement tasks using thesaurus terms are related to more instances of documents clicked. Of the searches initiated from the thesaurus list, 9.1% led to one or more clicked documents compared to 5.2% of the searches not initiated with the help of the thesaurus. When used for query formulation, and refinement, the thesaurus contributed to increased relevance of information retrieved, here defined by the number of clicked documents.

Table 5. Cross Tabulation between Clicks and Term_p

	uery	Refinement	requests
		Not initiated from thesaurus list	**Initiated from thesaurus list**
Number Of Clicked documents recoded	**No clicks**	179939	3576
		94.8%	**90.9%**
	One or more clicks	9853	359
		5.2%	**9.1%**
	Total	189792	3935
		100%	**100%**

chi-square = 119.345 p<.001

Conclusions and future research

This exploratory study has provided useful results regarding thesaurus-aided web searching. We investigated a sample of searches on the Web, represented by logs of queries from Development Gateway. The results demonstrated that most user queries do not result in documents being viewed. While a percentage of users did go on to formulate their original query with thesaurus-proposed terms, those queries initiated from the thesaurus list were more likely to result in clicked documents than those that were not. The results also provide evidence for the effectiveness of automatic query expansion. The use of a thesaurus to expand a request by adding synonymous terms is positively correlated with the number of resulting documents from a search request. Acknowledging the limitations of the data set, one can argue that the initial results of this study provide some support for embedding a thesaurus in the Web searching mechanism.

The conclusions from this review are suggestive of many directions for future research on methods for ranking algorithms and of other methods for thesaurus usage in the search, user studies on conceptual search with emphasis on how searchers select terms for query refinement, how do they perceive term relations, and how these affect their understanding of the subject.

Reference

Bates M.J. Subject access in online catalogs: A design model. Journal of the American Society for Information Science 37 (6): 357–376. 1986.

Borgman, CL. Why are online catalogs hard to use? Lessons learned from information-retrieval studies. Journal of the American Society for Information Science 37(6): 387-400. 1996.

Borgman. C.L., Case. D.O. and Meadow. C.T. The design and evaluation of a front-end user interface for energy
researchers, Journal of the American Society for Information Science 40(2): 99–109. 1989

Bush, Vannevar. As We May Think. The Atlantic Monthly. July 1945

Cooper, James W. and Byrd, Roy J. "Lexical navigation: visually prompted query expansion and refinement". Proceedings of the second ACM international conference on Digital libraries. July 1997.

Efthimiadis, E.N. Interactive query expansion: a user-based evaluation in a relevance feedback environment. Journal of the American Society for Information Science 51(11): 1–25. 2000.

Fidel, Raya. Searchers' Selection of Search Keys I. The Selection Routine; II. Controlled Vocabulary or Free-Text Searching; III. Searching Styles. Journal of the American Society for Information Science 42(7):490-527. 1991.

Furnas. G.W., Laundauer. T.K, Gomez. L.M., and Dumais, S. T. The Vocabulary problem in human-system communication. Commun, ACM 30: 964-971. 1987.

Jones S., Gatford M., Hancock-Bealieu M., Robertson S.E., Walker W., and Secker J., Interactive thesaurus navigation: intelligence rules Ok? Journal of the American Society for Information Science 46(1): 52–59. 1995.

Greisdorf, Howard. Relevance: An Interdisciplinary and Information Science Perspective". Informing Science Journal 3(2): p. 67-71. 2000.

Harman, Donna. Towards Interactive Query Expansion. In Proceedings ACM SIGIR'88: 321-331. 1988.

Marchionini. G. Information-seeking strategies of novices using a full-text electronic encyclopaedia. Journal of the American Society for Information Science. 40(1): 54–66. 1989.

Pinkerton, Brian. Finding What People Want: Experiences with Webcrawler. http://www.thinkpink.com/bp/WebCrawler/WWW94.html. 1994.

Yee. I. Hsieh. 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(3): 161–174. 1993.

Saracevic, T., Kantor. P. et al. A Study of Information Seeking and Retrieving. Journal of the American Society for Information Science 39(3): 161-176, 177-196, and 197-216. 1988.

Saracevic, T., Spink. A. Interaction in Information Retrieval: Selection and Effectiveness of Search Terms. Journal of the American Society for Information Science 48(8): 741-761. 1997.

Discussion Points

81. What does the author mean by classification?

82. Does the research involve creation or implementation of a classification scheme?

83. How does the researcher use classification to improve the automated approach?

84. How do these methods compare to current human-generated approaches to classification?

85. How does the reported research expand our understanding of classification?

86. Does the research suggest an improvement over human-generated classification?

87. What do you think are the most important lessons learned in this research?

88. What do you think are the best practices reported in this research?

89. What would you recommend to the researcher as the next step in this approach?

90. Is there other related research that you would recommend the researcher become acquainted with?