Library Philosophy and Practice Vol. 7, No. 2 (Spring 2005)
(libr.unl.edu:2000/LPP/lppv7n2.htm)
ISSN 1522-0222

A Model Search Engine Based on Cluster Analysis of User Search Terms

Elaine A. Nowick
Associate Professor, Research and Instructional Services
University of Nebraska--Lincoln Libraries
Lincoln, Nebraska 68588-4100

Kent M. Eskridge
Professor

Daryl A. Travnicek
Computer Specialist

Department of Statistics
University of Nebraska-Lincoln

Xingchun Chen
Senior Software Development Engineer

Jun Li
Graduate Student/Student Programmer

Communications and Information Technology
University of Nebraska-Lincoln

Introduction

Methods of searching for information in electronic collections of documents or records fall into two classes: keyword matching or browsing though a collection arranged in a subject classification scheme. Classification of print documents has traditionally provided users with both intellectual and physical access via shelving of documents in call number order with call numbers corresponding to subject headings. The physical location of the document and the subject classification were inextricably linked together. While the purpose of Library of Congress (LC) Subject Headings is to provide user access to a collection, the subject headings are based on documents rather than on the users' terminology. Studies of controlled vocabularies have indicated that they work well when there is an accepted common terminology describing concepts in the subject area and when users are familiar with the terminology (Voorbij, 1998). Solomon (1991) states, "Classification schemes fail too often because they are not grounded in the language and knowledge of users or in the task or situation of use." With the advent of

electronic information and the Internet, the physical location of the material is of much less importance. This has triggered a reexamination of classification schemes with a greater emphasis placed on intellectual access. With classification freed from the need to shelve one document in one location, subject hierarchies can be made more flexible and there is a greater possibility of customizing classification schemes to fit specific groups of users with particular needs. However, cataloging of information resources is labor intensive. As a practical matter, customized classification of documents will have to be at least partially automated.

One possible technique for examining the user's view of information space with the goal of producing flexible semi-automatic classifications is Cluster Analysis, a statistical technique used for identifying patterns and associations in complex data. Cluster analysis was originally used in analyzing taxonomic data and in creating phylogenetic trees. However cluster analysis has since been used in a wide range of applications from medical image analysis to market research. Cluster analysis based on documents in a collection has been explored as a means of automatic classification (Willett, 1988; McCaffrey,1991).

There has been less research in cluster analysis using user terminology rather than document keywords. Until log files of web sites were made available it was difficult to accumulate enough exact user searches to make a cluster analysis feasible. Another limitation in using searcher terms is that most users of the Internet employ short (one to two word) queries (Jansen et al., 1998). Wu, et al. (2001) used queries as a basis for clustering documents selected by searchers in response to similar queries. This paper reports on an experimental search engine based on a cluster analysis of user free-text for water quality information.

Methods

User queries were collected from a log file analysis of the University of Nebraska(UNL) Agricultural Network Information Center (AgNIC) Water Quality web site, from AgDB, a precursor of the AgNIC system at the National Agricultural Library (NAL), and from users reference questions submitted to the AgNIC web site. A total of 495 queries that included more than one concept or term were included in the cluster analysis. Four distance measures between words x and y, suitable for use with sparse data were calculated: Dice (1- 2A/(2A+B+C)), Kulczynski (1-A/(B+C)), Jaccard (1-A/(A+B+C)), and Ochiai (1- A/SQRT ((A+B)*(A+C)) where A= No.of observations with both words x and y present, B = No. of observations with word x present and word y absent, C= No.of observations with word x absent and word y present (Habalek, 1982). Comparisons were made between clusters produced with terms included in 10 or more queries and clusters produced with terms included in 5 or more queries. Truncated words were included together and some synonymous or similar words were grouped together. For example, names of specific cities were all included in "cityname". Both subject and non-subject query terms were included and some phrases, such as “Best Management Practices” were included rather than separated out as individual words. Clusters were created using the SAS Average Linkage Hierarchical Cluster Procedure (SAS Institute, Inc., 1990).

A consensus cluster map was made using the common elements from the clusters produced by the four methods and a cut-off point of ten clusters. A search program was

incorporated into the Water Quality web site which utilizes the word associations from the consensus cluster map to suggest terms for broadening or narrowing a search.

Results

Word frequency

In this study, the position of terms within the hierarchical trees depended both on the frequencies of words included in the calculations and also on the distance measure. Clusters were calculated for word frequencies of 1, 5, and 10. When terms with a frequency of one or more were included in the calculation, a "chaining" effect was observed. The results showed one very large cluster that included all the terms in the study. When only terms used in five or more searches were considered, there was less chaining and the clusters produced were more logical. When terms used ten or more times were included, the least amount of chaining was produced and the clusters were quite focused on specific subjects. However, considerable information on less frequently made searches was lost. For this reason, it was decided to incorporate terms used by five or more searchers in the analysis.

Distance measures

While clusters varied somewhat depending on the distance measure used in the calculations there were "core" terms that fell into the same cluster regardless of the distance measure used. At the lower levels of the hierarchical tree, the group membership of the terms tended to vary more. Since there are no commonly accepted statistical tests that would point out the best clusters based on mathematical criteria and there were no clear-cut differences in the logical term associations produced, a consensus cluster map was produced incorporating results from all four distance measures (Figure 1.). Terms enclosed by thick lines were found together in the same cluster for all four distance measures. Terms in thinner lines moved between clusters depending on the distance measure used. The consensus map better reflects the "fuzzy" nature of the clusters produced.

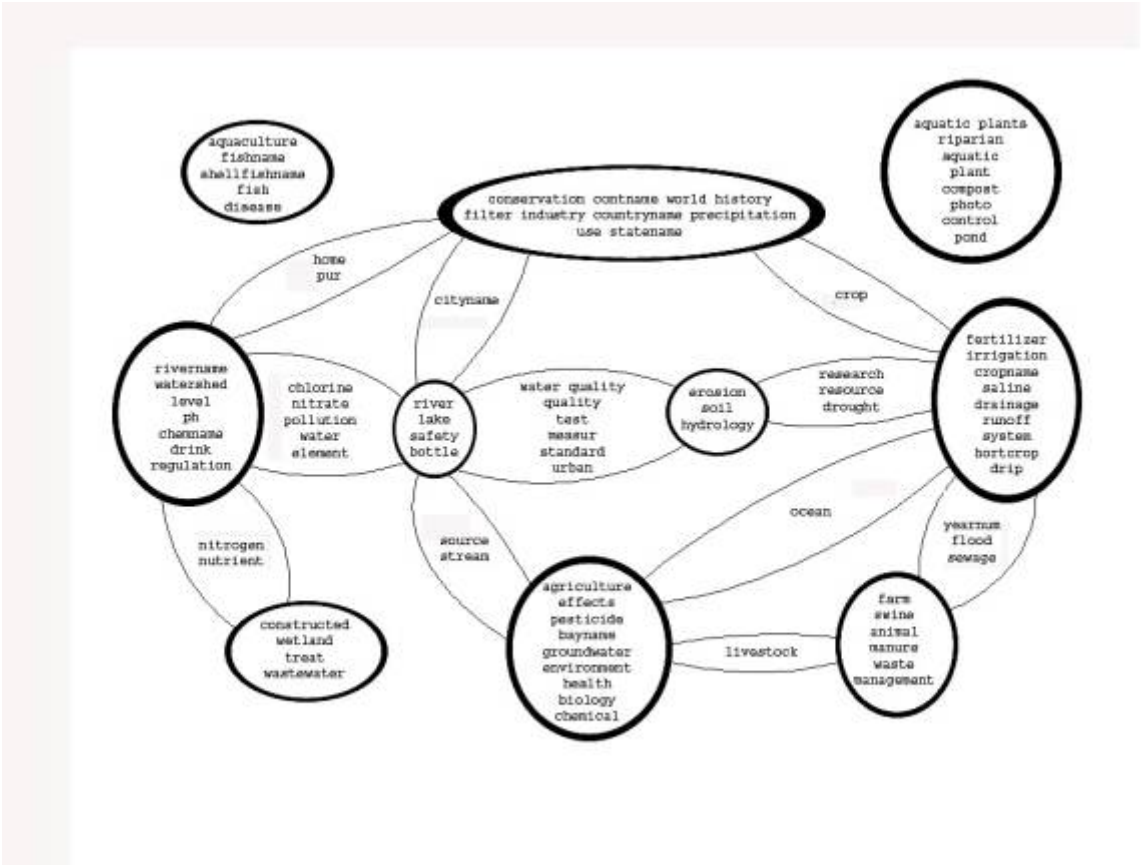

Figure 1 Consensus cluster map.

Search program

The search engine comes into play when the user has a "failed" search, which is defined as having more than 100 or fewer than one hit. For example, a keyword search for "water" would produce more than 100 hits and the user would be offered a suggestion to try the advanced smart search (Figure 2.).

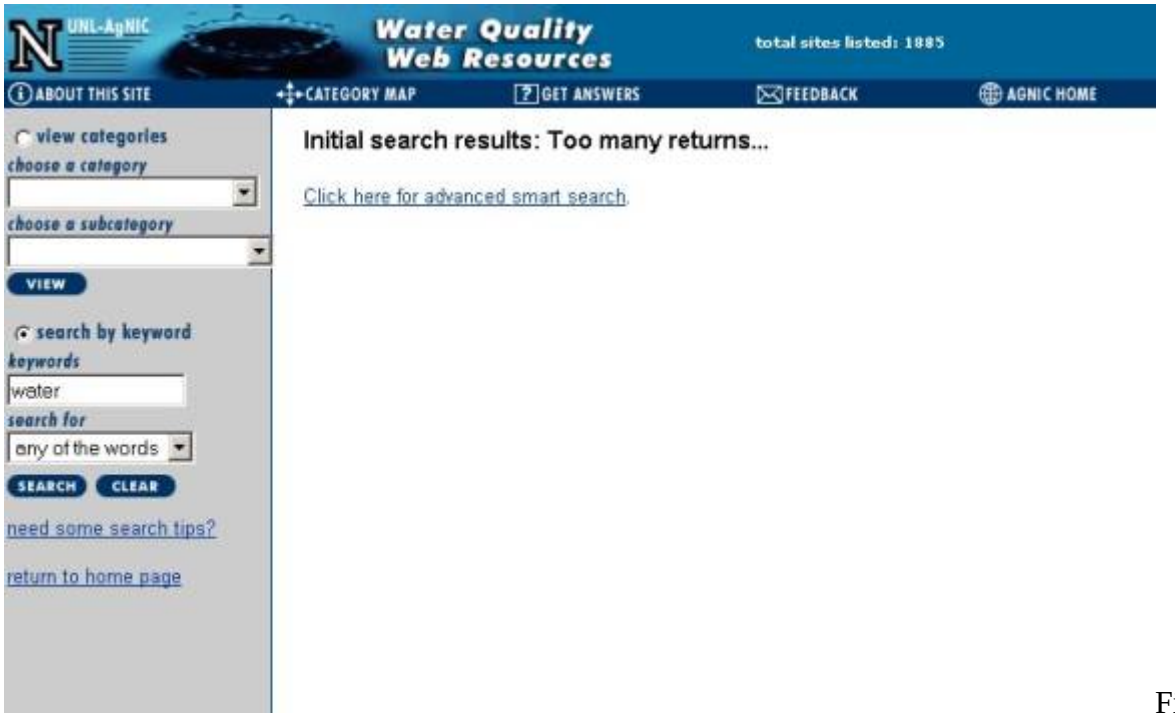

Figure 2. Results of "failed" search.

The advanced search link will take users to a screen with suggested terms to narrow the search (Figure 3.).

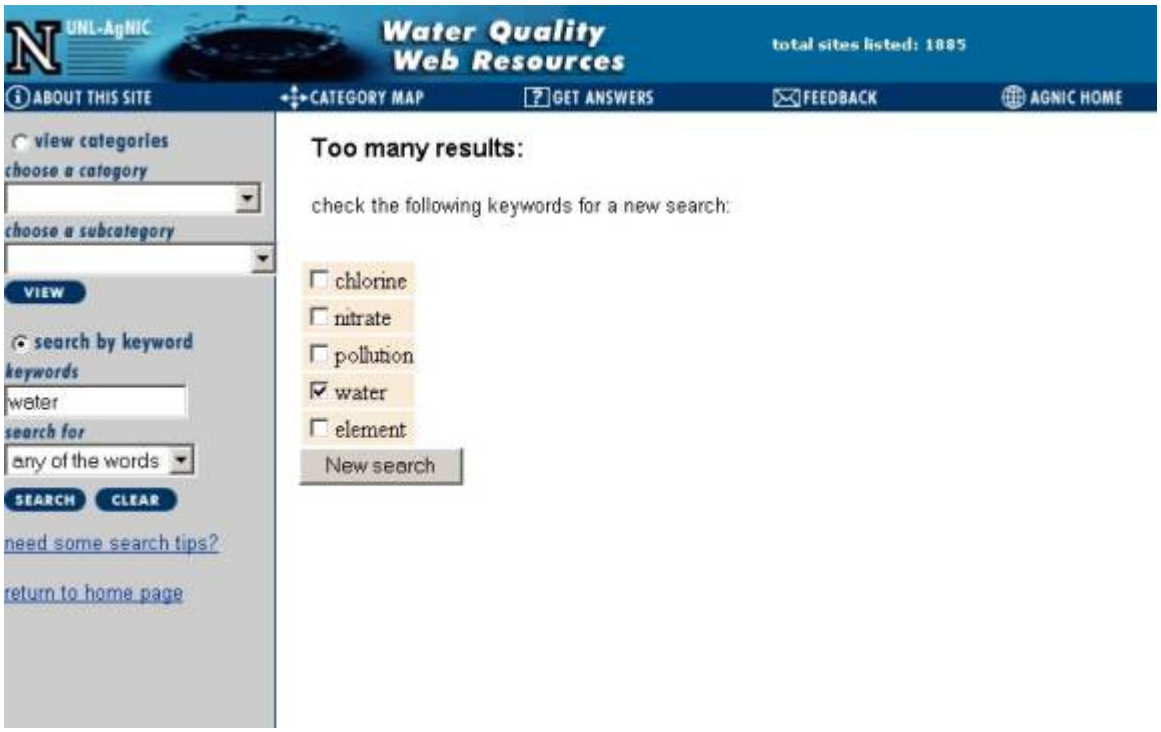

Figure 3. Suggested additional terms based on cluster map

When the user clicks on additional terms and then the search button, an "and" search is performed and the results returned. The additional terms are those that fell into the same cluster as the original term. If the searcher narrows the terms too much, the list of suggested terms will be expanded to include terms farther away in the cluster map. Color highlighting is used to distinguish the group of terms that are in the same core cluster as the original term from the group of terms that are in the "fuzzy" area (Figure 4).

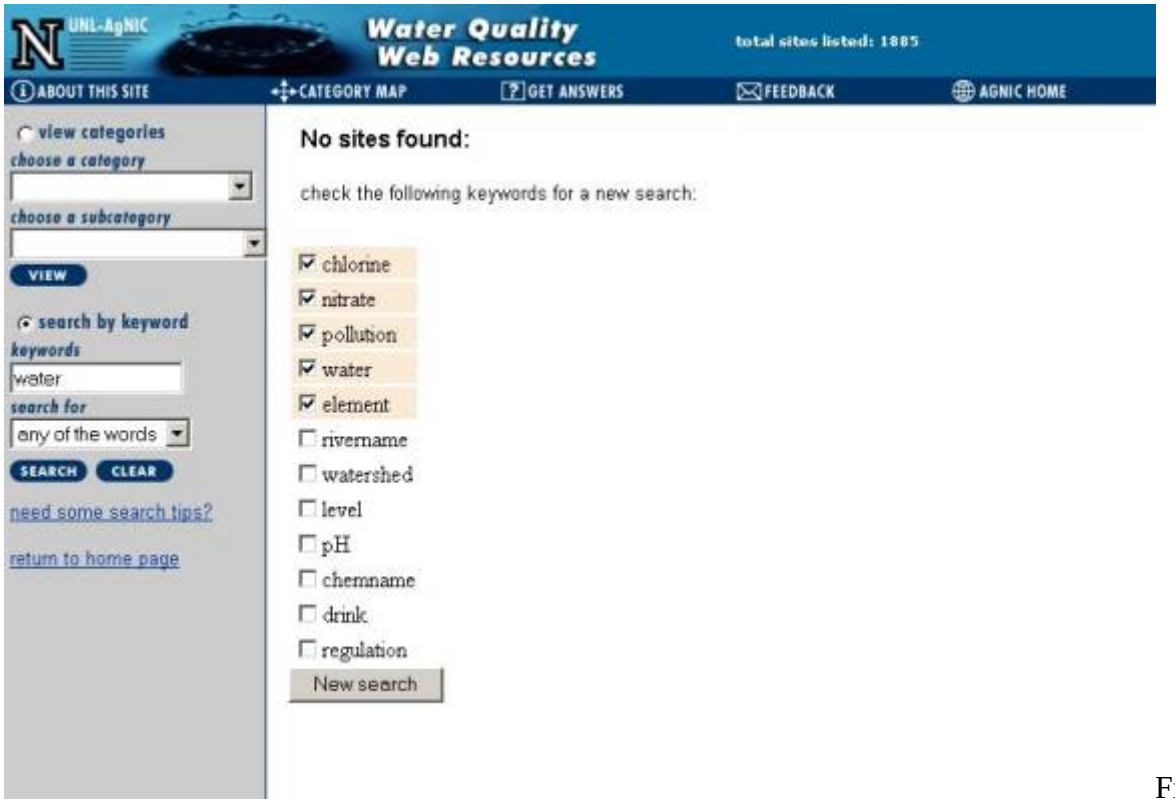

Figure 4. Additional suggested search terms for a too narrow search.

If the searcher's terms are too narrow and there are no results returned an "or" search is performed when the advanced smart search heading is selected.

Discussion

Traditional library cataloging and indexing has been centered on the information source, primarily books and articles. Online catalogs and search engine programming have allowed keyword searching to extend beyond the subject headings, titles, and author fields for searching, but were still centered on the document with the user having to try to fit their search needs, often poorly formulated, to the vocabulary of the author and cataloger or indexer. Programming advances now allow us to collect actual user terminology. Cluster analysis is a technique that enables the researcher to create a picture of the collective users' view of the information space.

The results form this study, serve as a proof of concept and a model of how a user-centric and dynamic classification scheme could work for the method to be effective as a search tool for users it would need to be greatly expanded. While the cluster calculations used in this study

produced slightly different classifications, the same is true for human generated subject hierarchies. As more searches are accumulated and terms added to the database the classification should become more stable. However, one of the strengths of this methodology is its flexibility. Clusters generated through a semiautomatic procedure can be more responsive to changes in search topics of users depending on current issues or scientific breakthroughs and also to changes in vocabulary usage.

A logical extension of the project would be to add controlled vocabulary terms and keywords from keywords indentified from the documents to the cluster analysis. Thus the cluster analysis could provide a logical and semiautomatic link between the user search terms and indexing terminology.

Conclusions

The clusters produced in this study provided intriguing patterns of meaning and hints for more user-friendly organization. However, it is apparent that there is a need for human intervention. Judgments are required on which words to include in stop lists and on whether to include phrases or its separated words in the analysis.

It is possible that query terms from web sites with heavier traffic would produce a more stable clustering pattern. However, a clustering technique that deals better with "fuzzy clusters" may be more appropriate. Future research plans are to explore non-hierarchical and probabilistic clustering methods such as k-means clustering.

References

Hubalek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence) data: An evaluation. *Biological Review* 57 Pp. 669-689.

Jansen, B. J., Spink, A., Bateman, J., Saracevic, T. (1998). Real Life Information Retrieval: A Study of User Queries on the Web. *SIGIR Forum* 32.1 Pp. 5-17.

McCaffrey, A. (1991). Applied cladistics: New methodologies for information classification research. *Proceedings of the 2nd ASIS SIG/CR Classification Research Workshop.* Washington DC, Oct 27, 1991. Pp.85-100.

SAS Institute, Inc. (1990). *SAS/STAT User's Guide,* Version 6, Fourth edition, Vol. 1, Cary, NC: SAS Institute Inc. 943 pp.

Solomon, P. (1991). Use-based methods for classification development. *Proceedings of the 2nd ASIS SIG/CR Classification Research Workshop*. Washington, DC.

Voorbij, H. J. (1998). Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences. *Journal of Documentation* 54 (April). Pp. 466-476.

Willett, P. (1988). Recent trends in hierarchical document clustering: A critical review. *Information Processing & Management* 24(5):577-597.

Wu, Y. H., Chen, Y. C., Chen, A. L. P. (2001). Enabling personalized recommendation on the web based on user interests and behaviors. *Proceedings, Eleventh International Workshop on Research Issues in Data Engineering*. Pp. 17 –24.