Using Clusters on the Vivisimo Web Search Engine

Sherry Koshman and Amanda Spink

School of Information Sciences
University of Pittsburgh
135 N. Bellefield Ave., Pittsburgh, PA 15237
skoshman@sis.pitt.edu, aspink@sis.pitt.edu

Bernard J. Jansen

School of Information Sciences and Technology
The Pennsylvania State University
329F IST Building, University Park PA 16802
jjansen@ist.psu.edu

Abstract

Web search engine interfaces typically return search results as lists of Web links accompanied by brief textual summaries. An improvement on this results interface design has been introduced with the use of clusters. Clusters represent groups of similar items and are used to aggregate search results. Operational Web interfaces containing cluster representations are relatively new. This paper presents research from a transaction log analysis of Vivisimo.com, which is a Web search engine that dynamically clusters the user's search results. The primary research question is: what are the characteristics of cluster use among Vivisimo searchers? One week's worth of data was analyzed from April 25 to May 2, 2004. The findings show that searchers frequently clicked on cluster labels although seldom manipulated the cluster tree. This research has implications for interface design utilizing clustering technology and Web retrieval.

1 Introduction

An increasing amount of people turn to the Web to resolve their information needs, however standard Web search engine interfaces do not utilize sophisticated interface design principles to enhance the retrieval process. The search and display functions have remained constant. A search box is used for query entry and corresponding results lists are presented to the user. This is apparent with Google, which has a simplistic interface and is a frequently used Web search engine.

The introduction of clustering technology has changed they way in which searchers can view search results since cluster labels are used to identify a category of related items. This is an attempt to improve the navigation of cumbersome results page lists by identifying groups of similar documents to aid user selection in discovering items most relevant to their query. This investigation reports on query log research to identify specific characteristics associated with cluster use in a novel clustering Web search engine. The overall goal of this research is to extend further the line of user interaction research in Web retrieval and specifically with the Vivisimo search engine.

2 Related Work

Investigations into user interaction with clusters for information retrieval generally follow a methodology where direct observation with the interface is employed. Frequently a comparison is made between using an experimental system representing the clustering interface and a traditional list-oriented system, which acts as the baseline. Clusters are generally represented by text labels and their use in interface design in Web retrieval is shown to be favoured by users in several studies.

Hearst and Pedersen (1996) showed that users were successful in interacting with the clustering mode in their experimental system and could select the cluster containing the most relevant documents. Chen and Dumais (2000) studied 18 users with a categorization interface and a traditional text-based interface for organizing Web results. They reported that the categorization interface excelled in objective and subjective measurements after the comparison was made using the same search tasks, Web search engine, and results.

Web-based visualization systems utilize clustering techniques to represent document groupings. Osdin, Ouni and White (2002) examined 16 users who conducted eight tasks with HuddleSearch, an experimental clustering-based visualization tool and Panoptic, a traditional list-based search engine. The tasks adhered to the Text Retrieval Conference (TREC) guidelines, and the researchers collected data on task completion, task times, and user satisfaction. Their results showed that users preferred the clustering tool for Web retrieval.

Rivadeniera and Bederson (2003) tested 15 participants with factual retrieval questions using an operational visualization interface, Grokker, for Web searching. The visualization was compared it to a version of Grokker's textual interface and Vivisimo's textual clustering interface. The objective measures showed no significant differences, however subjective satisfaction measures favoured the textual interfaces.

This investigation uses a transaction log analysis to study Vivisimo user search characteristics. An advantage of this method is that a large-scale view of user interaction with the system is provided. A transaction log analysis was used to evaluate user interaction with Grouper, a clustering interface for Web search engine results (Zamir & Etzoni, 1999). The findings showed that users tended to examine more clusters than hypothesized by the investigators. The logs were also analyzed to compare Grouper with a traditional text-based interface, HuskySearch, in order to determine the number of documents clicked-on by the users. Their results showed that users followed more multiple documents using the Grouper clustering interface and more single documents using HuskySearch.

Transaction log analyses offer an unobtrusive method to study user interactions with traditional Web search engines (Jansen, Spink, & Saracevic, 2000; Spink & Jansen, 2004; Spink, Wolfram, Jansen, & Saracevic, 2001). This study extends the research of Web search engines to a cluster-based environment

3 Research Questions

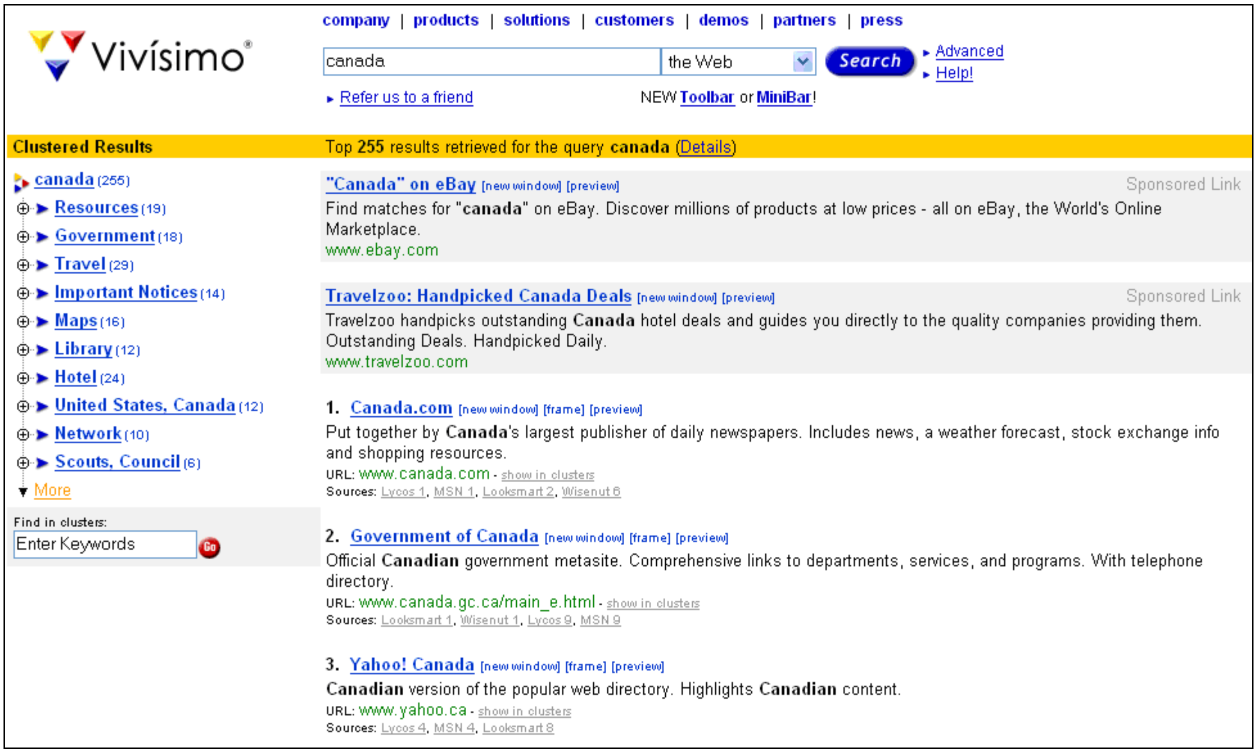
The research questions asked are: 1) how is the interface's cluster tree used by Vivisimo searchers?, and 2) what is the frequency of use for cluster labels? The data resulting from user cluster interactions were captured in the log file and analyzed according to extent and frequency for this investigation.

4 Research Design

4.1 System

The Vivisimo interface contains a dialog box for inputting queries and supports Boolean and exact phrase matching. The default search source is the Web and a drop down menu provides options for additional source selection (e.g. CBC, CNN, Wisenut). Searches can be limited by domain or host name, by link content, Web page or Uniform Resource Locater (URL) information.

Vivisimo offers an "Advanced" search form containing options for source and language selection, defining the number and display of search results, deciding how links should be opened, and whether or not the content filter is applied. After a user submits a query, Vivisimo presents the clusters using a tree metaphor, which is similar to that used for viewing folders in Windows Explorer. The clusters appear on the left side of the page and the results pages are featured on the right of the main search page (Figure 1).



Top **255** results retrieved for the query **canada** (Details)

"Canada" on eBay [new window] [preview] Sponsored Link
Find matches for "**canada**" on eBay. Discover millions of products at low prices - all on eBay, the World's Online Marketplace.
www.ebay.com

Travelzoo: Handpicked Canada Deals [new window] [preview] Sponsored Link
Travelzoo handpicks outstanding **Canada** hotel deals and guides you directly to the quality companies providing them. Outstanding Deals. Handpicked Daily.
www.travelzoo.com

1. Canada.com [new window] [frame] [preview]
Put together by **Canada**'s largest publisher of daily newspapers. Includes news, a weather forecast, stock exchange info and shopping resources.
URL: www.canada.com - show in clusters
Sources: Lycos 1, MSN 1, Looksmart 2, Wisenut 6

2. Government of Canada [new window] [frame] [preview]
Official **Canadian** government metasite. Comprehensive links to departments, services, and programs. With telephone directory.
URL: www.canada.gc.ca/main_e.html - show in clusters
Sources: Looksmart 1, Wisenut 1, Lycos 9, MSN 9

3. Yahoo! Canada [new window] [frame] [preview]
Canadian version of the popular web directory. Highlights **Canadian** content.
URL: www.yahoo.ca - show in clusters
Sources: Lycos 4, MSN 4, Looksmart 8

Figure 1: Vivisimo Interface

Unlike typical Web search engines that present lists of search output, Vivisimo's clustering feature creates dynamic post-search categories in a meta-searching environment. Users can click on cluster labels to retrieve results pages. Clusters can be expanded by clicking on the plus sign to reveal sub-clusters and the cluster tree may be elongated by clicking on the "More" option. Search terms can be entered in the "Find in clusters" search box to search the clusters.

The results pages are initially displayed as a result of the initial search. Results pages are retrieved when the user clicks on the clusters and additional results pages may be selected at the bottom of the window. Hyperlinks may be accessed for individual items and Web pages may be previewed, opened in the results frame, or opened in a new window.

An item on the results pages may be identified within the clusters by clicking on the "show in clusters" option next to the item. This highlights the clusters on the tree that contain the item. The "Details" feature shows the number of results for the sources searched.

4.2 Data Collection

The Vivisimo transaction log data used for this study represents a one-week period from April 25 to May 02, 2004. The transaction logs recorded 100% of the traffic on the Vivisimo Web site during this period and contained 4,219,925 records.

4.3 Data Analysis

The log is a flat ASCII file, which was imported into a relational database, and a unique identifier for each record was assigned. Using four fields (*User Identification*, *Date, Time of Day*, and *Query Terms*), the initial query was located and the chronological series of actions on a given day was recreated to represent a user session.

A *term* is any series of characters separated by white space or other separator. A *query* is the entire string of terms submitted by a searcher in a given instance of interaction. A *session* is the entire series of queries submitted by a user during one interaction with the Web search engine on a given day.

Sessions with 100 or fewer queries were separated into an individual transaction log. This cut-off was selected because it is almost 50 times greater than the reported mean search session for human Web searchers and it assured that human searches were not excluded (Silverstein, Henzinger, Marais, & Moricz, 1999).

Vivisimo assigns a unique code to identify a user's multiple interactions with the system. To address the issue of duplicate queries, the Vivisimo transaction log was collapsed by combining all identical queries submitted by the same user to generate unique queries for analyzing sessions, queries and terms, and pages of results viewed. The complete un-collapsed sessions were used to obtain an accurate measure of the session duration and the number of results pages visited. Key fields were extracted for the analysis and a series of UNIX text manipulation commands were used to parse and calculate statistics on the data.

5 Results

The cluster analysis of post-search records showed that almost half of the interactions (48.2%) involved clicking on a cluster label. However, a small percentage of Vivisimo searchers (12.6%) used the cluster expansion feature to manipulate the cluster tree. Clusters were most frequently expanded once and the maximum number of clusters expanded was 26 (Figure 2).

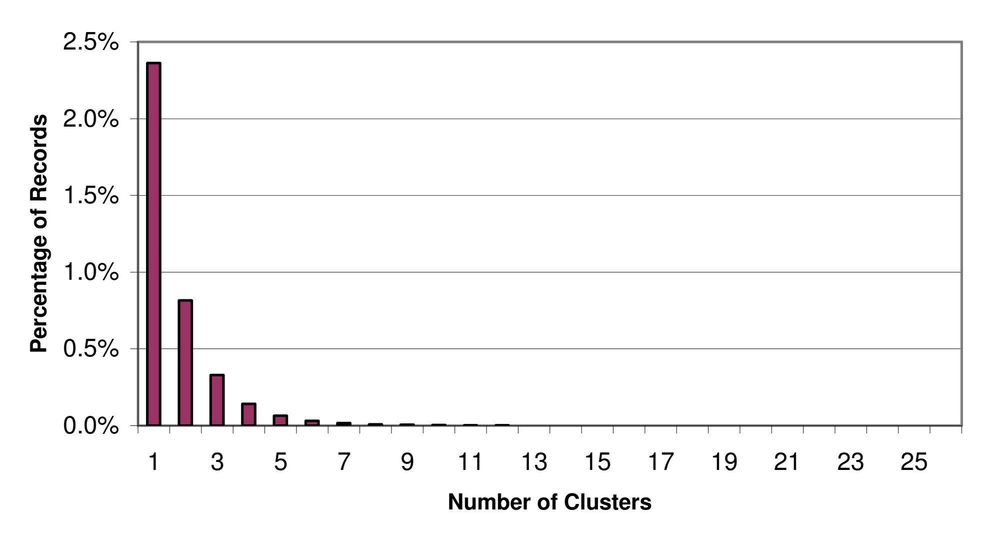


Figure 2. Total Cluster Expansion Distribution

6 Discussion and Implications

Direct manipulation techniques are used to operate the cluster tree, which is Vivisimo's unique interface feature. The tree is presented in a separate frame on the left side of the screen. The result lists change as the user selects different cluster labels. The cluster label text is presented as underlined blue text, which is a familiar hyperlink visual for Web users. The number of items found is located next to the cluster label in parentheses. While clicking on the cluster labels was very successful among users, the expansion and elongation of the cluster tree was not.

Vivisimo employs a familiar hierarchical tree metaphor for navigating the clusters. Users need to click on the plus sign contained in a circle to expand the cluster and to click on the "more" option at the bottom of the tree to view additional cluster labels. Infrequent usage findings may be attributed to the visibility of the plus sign within the circle and understanding its function or because people found what they were looking for. Once the cluster is expanded, the cluster‘s hierarchy is visualized by arrows and lines to depict relationships between root clusters and sub-clusters. The “more” option is presented in underlined light yellow text at the bottom of the cluster tree and may not have been clear to the user.

7 Conclusions

These results will be used to design a formal usability study, which observes searchers’ real time interaction with the Vivisimo interface. The following factors will be considered 1)) the user’s selection of cluster labels and the depth of clusters from which results pages are viewed, 4) the search patterns which incite the user to expand and elongate the cluster tree, and 5) subjective satisfaction measures to determine if searchers resolved their queries through cluster selection.

The log analysis reported in this paper identified Vivisimo interface features used and provided a pattern of cluster usage to fulfil the research questions. The primary goal for future HCI research is to examine in detail the role of Vivisimo’s interface features on Web searching effectiveness.

8 Acknowledgements

We would like to thank Vivisimo for providing the data for this analysis without which we could not have conducted this research.

9 References

Chen, H., & Dumais, S. (2000). Bringing order to the Web: automatically categorizing search results. *ACM SIGCHI Conference on Human Factors in Computing, The Hague, The Netherlands* (pp. 145-152). New York: ACM Press.

Hearst, M., & Pedersen, J. (1996). Re-examining the cluster hypothesis: scatter/gather on retrieval results. *ACM SIGIR'96: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Zurich, Switzerland* (pp. 76-84). New York: ACM Press.

Jansen, B. J., Spink, A., & Saracevic, T. (2000). A study and analysis of user queries on the Web. *Information Processing and Management, 36*(2), 207-227.

Osdin, R., Ounis, I., & White, R. (2002). *Using Hierarchical Clustering and Summarization Approaches for Web Retrieval: Glasgow at the TREC 2002 Interactive Track*. Retrieved October 20, 2004, from citeseer.nj.nec.com/580060.html

Rivadeneira, W., & Bederson, B. (2003). *A Study of Search Results Clustering Interfaces: Comparing Textual and Zoomable User Interfaces*. Retrieved April 20, 2004, from ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2003-36html/2003-36.htm

Sliverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a very large Web search engine query log. *ACM SIGIR Forum, 33*(1), 6-12.

Spink, A., Wolfram, D., Jansen, J., & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226-234.

Zamir, O., & Etzoni, O. (1999). *Grouper: A Dynamic Clustering Interface for Web Search Results*. Retrieved January 10, 2004, from http://www.cs.washington.edu/research/projects/WebWare1/etzioni/www/papers/www8.pdf