2000

Use of Clustering and Information Visualization for Managing Information Overload in the Web Environment

Ozgur Turetken
Oklahoma State University, turetke@okstate.edu

Use of Clustering and Information Visualization for Managing Information Overload in the Web Environment

Ozgur Turetken, Oklahoma State University, Department of Management, Stillwater OK 74078, turetke@okstate.edu

Abstract

This study addresses the information overload problem that one faces while browsing the results of a web search query. The increasing indexing capabilities of the commercial web search engines makes it common for a broadly formulated search query to result in thousands of web documents. This causes a nontrivial overload problem. Our study proposes the use of clustering and information visualization as a remedy to this problem.

The proposed prototype system groups the search results according to their contents. The information seeker is presented a visual overview of these clusters to identify the general characteristics of the document collection. Based on such an understanding of the information space, the user of the system can focus on document groups of more interest in order to reach the information sought for. We will incorporate two different zooming methods for this purpose: the traditional full zoom (strict-filtering), and a fisheye zoom which provides details in context. We will empirically test the success of the visualization based presentation system in comparison to its text based counterpart in an experimental setting. The success of the fisheye zoom approach in comparison to the full zoom approach will also be tested by means of the same experiment.

Introduction

The promise of the information age is for the information user to access the highest quality information in the right form, at the right time, and right place. Information technologies are progressing at an unprecedented rate and have the potential of making this promise come true. However, this fast progress has brought the problem of information overload along with it. Information overload occurs when an information user is exposed to more information than (s)he needs, and more importantly, is able to process. This phenomenon has adverse impacts on the use of information and the decisions made thereof. This research focuses on a specific information overload problem, one that is encountered while searching the web by means of a commercial search engine.

Typically, web search engines present their results as a ranked list. For some rather broad search queries, such a list may contain thousands of documents. Research has suggested that the users of these systems are not likely to go beyond the top 20 to 30 documents on these lists before they get bored or frustrated (Roussinov, 1999). Consequently, the chance of reaching the relevant information is reduced when the searcher is overloaded with the irrelevant documents at the top of the list. We are in the process of developing a prototype system that aims to address this issue by using clustering (grouping) and information visualization. Clustering is a well-known technique commonly used to identify patterns in an unstructured group of objects. On the other hand, information visualization is the common name for a group of techniques that use the idea of supporting the cognitive system by means of visual cues for better and quicker understanding of information (Shneiderman, 1996). Our methodology is based on the combined use of clustering and information visualization for the treatment of the aforementioned problem. In the next section, we briefly describe the details of our prototype development effort.

Methodology

The proposed visual system presents an overview of search result groups instead of a linear ranked list of each individual result. These groups are formed according to the content of documents as follows: first the significant terms in the document collection (i.e. the search result list) are identified, and the documents are represented as vectors of the term frequencies, then the clustering algorithm groups these vectors.

The visual overview of the document clusters summarizes the information space and lets its users recognize certain patterns. Based on this recognition, the information seeker can focus (zoom) on the document groups of more interest with the purpose of finding the information (s)he is looking for. In this study, we will use and compare two different zooming methods: full zoom and fisheye zoom. The full zoom method strictly filters out the groups outside of the immediate interest area while the fisheye zoom examines the regions of immediate interest in full detail while keeping in view a summary of the remaining groups to provide context. We believe that it may be desirable to examine local detail without losing awareness of global context especially in the visual exploration of search result clusters where the boundaries between the clusters in the visual space are rather imposed than natural. Past research (e.g. Furnas, 1986; Schafer et. al., 1998) supports this observation, and suggests that a fisheye view is a plausible alternative to

the full zoom approach for exploring information-intensive structures. Accordingly, one major contribution of our research will be the application of the fisheye view idea in the visualization of web search results.

Research Questions and Hypotheses

We propose that the use of clustering and visualization as explained above will not only help in finding the relevant information within search results, but also in finding this information faster. In other words, this approach will provide better information access with less overload. In this respect, we aim to achieve high search success by means of increasing search effectiveness and search efficiency. This use of the "success" concept refers to the better organization of information rendering it possible to display more information without causing overload. We should also clarify that our use of the term "search" assumes a user rather than a system perspective. Although the system we are describing does not improve the available search methods per se, it aims to enhance the outcomes of the user's search efforts, hence the terms effectiveness and efficiency are used regarding these outcomes.

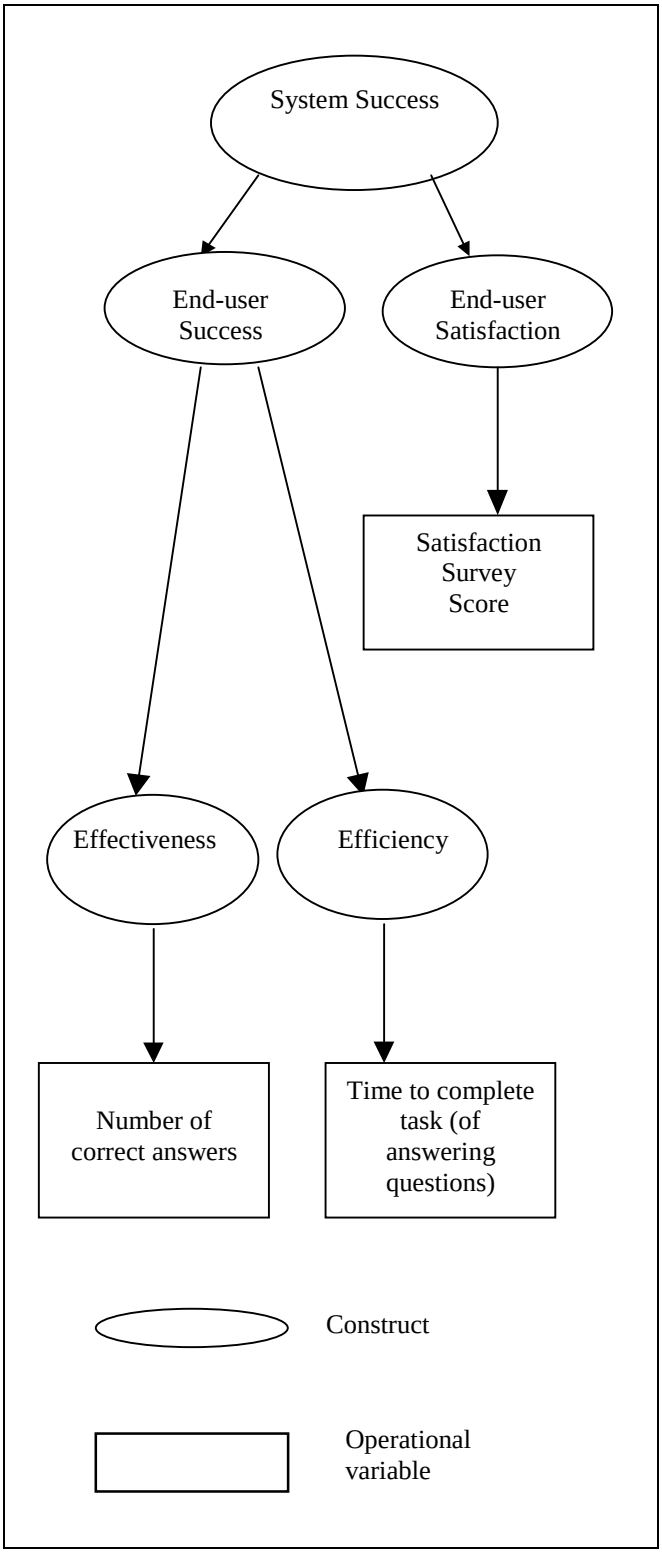
It has been observed that the scarcity of empirical studies on the usefulness of information visualization has been a weakness of knowledge management research in this area (Zamir 1998; Roussinov 1999). Hence, an important contribution of this study will be the empirical testing of the prototype to discover whether our approach fulfills the aim of high search success. Specifically, we will seek answers to the following research questions:

1. *Can a (clustering-based) visual presentation system improve search success over one without such a support?*
2. *Can a (clustering-based) visual presentation system supporting fisheye zoom improve search success over one with full zoom only?*

Figure 1 displays the factor structure of the "System success" variable (construct). According to this structure, system success has two dimensions: "success of the end-users" using the system, and their "satisfaction" with the system. "End-user success" in turn can be approximated by the user's effectiveness and efficiency. The operationalization of the "effectiveness", "efficiency", and "satisfaction" variables is shown in the rectangular boxes in Figure 1. Based on this operationalization, we formulate the following hypotheses:

H1a: Existence of (clustering-based) visualization affects the number of correct answers given to an objective set of questions that have their answers in the search results.

Figure 1 Factor Structure and Operationalization of the Dependent Construct



H1b: Existence of visualization affects the time to complete the task of answering these questions.

H1c: There is a significant difference between the visual systems and the text-based system with respect to user satisfaction.

The hypotheses in this first group aim to measure the success of visualization. The second group hypotheses aim to measure the success of the fisheye zoom method in comparison to the full zoom method, and are formulated as follows:

H2a: Use of the fisheye zoom instead of the full zoom affects the number of correct answers.

H2b: Use of the fisheye zoom instead of the full zoom affects the time to complete the task of answering the questions.

H2c: There is a significant difference between the full zoom and fisheye zoom systems with respect to user satisfaction.

Experimental Design and Data Analysis

For testing the above hypotheses, we will collect data by means of a lab experiment based on a repeated-measures design as depicted in Table 1. The experiment will consist of three phases, each of which will contain a different task (answering questions on a different search topic). Three groups of subjects will answer these questions using different types of supporting presentation methods. In each phase of the experiment, the number of correctly answered questions will be recorded ,and the time it takes for the subjects to finish the task of answering these questions will be measured. The subjects will also complete a satisfaction survey at the end of each phase.

Table 1. The Repeated Measures Design

Phase	Task/Support		
	Group 1	Group 2	Group 3
Phase 1	Question Set 1 No visualization	Question Set 1 Full zoom	Question Set 1 Fisheye zoom
Phase 2	Question Set 2 Fisheye zoom	Question Set 2 No visualization	Question Set 2 Full zoom
Phase 3	Question Set 3 Full zoom	Question Set 3 Fisheye zoom	Question Set 3 No visualization

The data analysis methods (parametric vs. non-parametric) will be determined depending on the distributional characteristics of the data.

References

Furnas, G.W. "Generalized Fisheye Views," *Proceedings of the ACM SIG-CHI 86 Conference on Human Factors in Computing Systems*, 1986, pp.16-23.

Roussinov, D. "Information Foraging Through Automatic Clustering and Summarization: A Self-Organizing Approach," Doctoral Dissertation, University of Arizona, 1999.

Schafer, D., Zuo, Z., Greenberg, S., Bartram L., Dill J., Dubs, S., and Roseman, M. "Navigating Hierarchically Clustered Networks Through Fisheye and Full-Zoom Methods," *ACM Transactions on Computer-Human Interaction* (13:2), 1998, pp. 162-188.

Shneiderman, B. , "The eyes have it: A task by data type taxonomy of information visualizations," *Proceedings of the IEEE Symposium on Visual Languages*, 1996, pp. 336-343.

Zamir, O., "Visualization of Search Results in Document Retrieval Systems," General Examination Report, Department of Computer Science and Engineering, University of Washington, 1998.