Analysis of Usage Patterns in Experiential Multiple Perspective Web Search

Rahul Singh, Ya-Wen Hsu
Department of Computer Science, San Francisco State University, San Francisco, CA
rsingh@cs.sfsu.edu, logoin@sfsu.edu

ABSTRACT

With the rapid growth in the volume, complexity, and heterogeneity of information in the World Wide Web (WWW), the role of user-data interaction paradigms is becoming increasingly critical to the success of web-based information retrieval and assimilation. Currently, the paucity of mature paradigms in this problem area contrasts sharply with the advances in design of search techniques that allow indexing large volumes of information and efficiently executing keyword-based search. State of the art research in mediating user-data interactions over large information repositories such as the WWW has seen the proposition of techniques such as page categorization, page summarization, content-based page clustering, as well as algorithms for recognizing semantic correlations between web-pages having heterogeneous content and supporting experiential and unified interactions across them. In this context, a key challenge involves studying and analyzing user behaviors and usage patterns in such information interaction frameworks. In this paper, we present a study that investigates this issue by analyzing both quantitatively and qualitatively how users interact, query, explore, and assimilate information in such an environment. We also investigate how efficacious such interaction paradigms are as compared to the standard approach of presenting results by ordering them in terms of page rank. Results from our investigation provide important insights into user behavior in heterogeneous information organization and interaction frameworks and will be valuable in further development of information presentation, querying, and interaction techniques.

Categories and Subject Descriptors

H.5.3 [**Information Interfaces and Presentation],** H.5.1 [**Multimedia Information Systems**], H.5.2 [**User Interfaces**]

General Terms

Design, Human Factors, Experimentation

1. INTRODUCTION

Given a query, the standard strategy in most search engines is to present the results as a list where each entry is ranked by its putative relevance to the query. Users have to subsequently peruse the list to satisfy their information needs. While such a strategy is perfectly adequate for many types of queries, it is not without limitations; from an operational perspective for instance, it increases the cognitive load on users by forcing them to cherry-pick from a list of results that may include a variety of hits not all of which may be related to the information goal. The problem is exacerbated if the query terms are polysemous or if the results contain multiple topics.

A deeper analysis brings forth further shortcomings, especially assumptions made about the information need underlying the query. To understand this issue, we begin by noting that the semantics associated with complex information is non-unique, user-dependent, and emergent [13]. In such cases, user context and user-data interactions play a critical role in interpretation and assimilation of the information. The conventional approach of presenting the search results in a long list limits such interactions to the minimum in that a user only has the option of examining each link one at a time. Research in *information foraging* theory [11] asserts that users typically navigate towards their information goal by following *link cues*, which are fragments of information within or near hyperlinks and relevant to the user's information goal. From this perspective also, the conventional approach is minimalistic in that it neither presents any cues about correlations within the list of results nor does it provide sufficiently rich link cues (beyond text snippets). Two distinct but interrelated necessities can thus be identified:

- Capturing the variability in the semantics of the information that may be spread across different web pages identified to be of relevance to the query.
- Supporting efficient and effective interactions between users and the information with the ultimate goal of efficiently satisfying the user information need.

A possible solution to the first problem is to cluster search results into different groups, with each group corresponding to a distinct topic. At the state of the art, commercially available solutions to this problem include Vivisimo [15] and Grokker [4]. In other research, [2] propose a combination of color-texture moments, text, and link information to cluster image search results from the web. Other approaches utilize the predominantly textual content of most web pages [8, 12, 16]. Unlike these approaches, where categories are defined by content, the search engine Northern Light [9] organizes results into categories predefined in library sciences.

The second challenge underlines the need for facilitating exploratory, user-centric capabilities rather than pure syntactic query-retrieval. Here, an interesting parallel can be observed with research in the design of "experiential systems" [7, 14] which deals with the challenges of information retrieval and assimilation in settings involving rich heterogeneous information. Experiential systems take advantage of the human-machine synergy and allow users to use their senses and directly interact with the data. Such systems are: (1) direct, in that they do not use complex metaphors or commands, (2) support same query and presentation spaces, (3) maintain user state and context, (4) present information independent of (but not excluding) media type and data sources,

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.
Copyright 2007 ACM 978-1-59593-701-8/07/0009...$5.00.

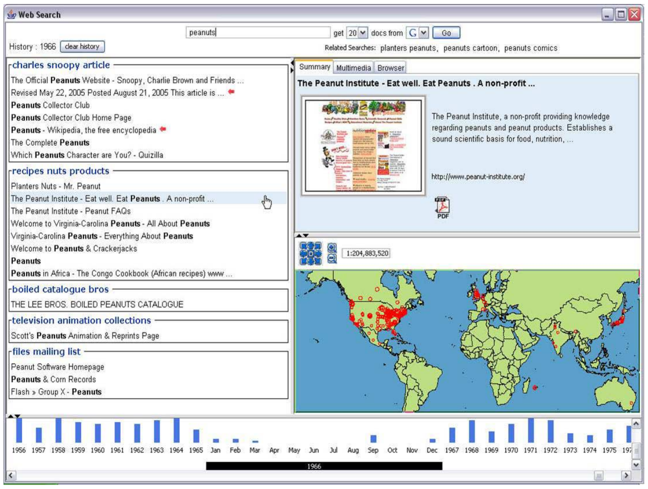
(5) provide multiple semantic perspectives on the data, and (6) promote perceptual analysis and exploration.

In this paper, we present a systematic study of user-behavior and usage patterns in an interaction environment that addresses the aforementioned two challenges for user interface design in web-search. The design of the proposed interface is motivated by the philosophy underlying experiential systems. Our investigations examine efficacy of interactions, usage patterns, and user experience and satisfaction. Furthermore, based on usage tendencies, we identify the importance of specific interaction modalities in information search and assimilation. Results from our research will be valuable not just in the specific contexts of web-search and experiential system design, but also for the development of user-data interaction paradigms in environments with rich heterogeneous data with queries reflecting complex information needs. In the following we briefly outline the user-interface designed and used by us for this study.

2. SYSTEM DESCRIPTION

The primary components of the proposed system include a *web-page retrieval module,* a *data analysis system,* and a *multiple perspective experiential user-data interface*. The interaction process is initiated by directing the (text-based) user query to any search engine of choice. Subsequently the web-page retrieval module is used to obtain the content of the pages as well as a thumbnail of the web page using Alexa [1]. Next, the textual content of the web-page is analyzed to extract any temporal and spatial (geographical) information by cross-referencing the terms in the page with a comprehensive gazetteer of locations. The content of the web-page is also analyzed to extract any media associated with the page including video, audio, Flash, postscript/PDF, MS Word/PowerPoint files. A combination of data-driven and model-based techniques are next applied to the retrieved results to group them into semantically meaningful clusters. *First*, Latent Semantic Indexing (LSA) is used in conjunction with Term Frequency-Inverse Document Frequency (TFIDF) weighting. The basic idea lies in utilizing LSA to reveal semantic correlations between the documents via mapping to a low-dimensional eigenspace and then augmenting the prominence of uncommon words through TFIDF weights derived using the transformed term frequency data. *Second*, ODP [11] based taxonomy information is used to refine the clustering. This allows similar documents to be brought together despite a possible lack of correlating terms. This method is described in detail in [6].

The information derived from the various components is displayed as separate panes in the user interface (Figure 1), with search results presented as clusters of web pages, and related spatial and temporal information shown as points on an interactive map and timeline. As users mouse-over links, a fourth pane presents a thumbnail of the web-page, a summary of its content, and any media files (such as audio or video) that may be associated with it. Additionally, a user can also choose to click on a cluster. In such a case the thumbnails of all the pages in the cluster are displayed along with any associated location and temporal information in the map and the timeline respectively. This allows users to obtain a visual overview of the data in any specific cluster as well as perform *roll-up* and *drill-down* operations. Figure 2 presents an instance, illustrating these capabilities of the interface.



charles snoopy article

The Official **Peanuts** Website - Snoopy, Charlie Brown and Friends ...
Revised May 22, 2005 Posted August 21, 2005 This article is ...
Peanuts Collector Club
Peanuts Collector Club Home Page
Peanuts - Wikipedia, the free encyclopedia
The Complete **Peanuts**
Which **Peanuts** Character are You? - Quizilla

recipes nuts products

Planters Nuts - Mr. Peanut
The Peanut Institute - Eat well. Eat **Peanuts** . A non-profit ...
The Peanut Institute - Peanut FAQs
Welcome to Virginia-Carolina **Peanuts** - All About **Peanuts**
Virginia-Carolina **Peanuts** - Everything About **Peanuts**
Welcome to **Peanuts** & Crackerjacks
Peanuts
Peanuts in Africa - The Congo Cookbook (African recipes) www ...

boiled catalogue bros

THE LEE BROS. BOILED PEANUTS CATALOGUE

television animation collections

Scott's **Peanuts** Animation & Reprints Page

files mailing list

Peanut Software Homepage
Peanuts & Corn Records
Flash » Group X - **Peanuts**

The Peanut Institute - Eat well. Eat Peanuts . A non-profit ...

The Peanut Institute, a non-profit providing knowledge regarding peanuts and peanut products. Establishes a sound scientific basis for food, nutrition, ...

http://www.peanut-institute.org/

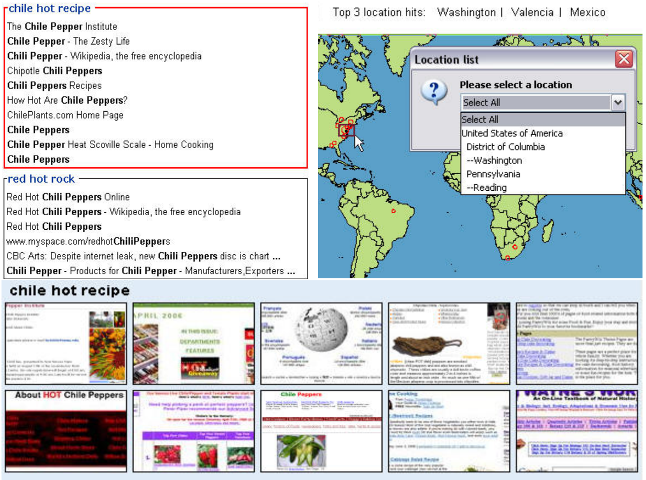
Figure 1: Snapshot of the system. The panel on the top left shows the distinct clusters associated with the query "peanuts". The top right panel displays the summary of a specific page selected by the user. Spatial characteristics of the data are displayed on an interactive map. An interactive timeline at the bottom displays the temporal distribution of the results. All the modules are reflectively interlinked.

Display and interaction with spatial aspects of the information is supported through the interactive map implemented using the OpenMap Java toolkit. Cities appearing in the documents are indicated on the map, where circle size varies logarithmically with the number of documents containing the location. Both point-based and region-based queries are supported through clicking or selecting a region by drawing a rectangle. To disambiguate user intent when a region is selected, a list of locations contained within the selection is displayed (Figure 2) from which the user can make a choice. The temporal information is displayed through a timeline that supports multiple granularities. The timeline supports zooming into a particular time interval to find details or focus on a specific period and zooming out to see the overall temporal distribution.

The design of the interface supports emergent and exploratory user interactions. In it, various semantic perspectives on the data are tightly linked to each other, so that interactions in terms of any one of them are instantaneously reflected in the others. For example, selecting a specific region of time leads all links relevant to it to be highlighted, including in the spatial view. Such coupling is essential for discovering relations hidden in the data. The interface supports direct query and manipulation [5] in that operations are supported using intuitive interactions that are directly executable on the appropriate representation. Finally, combining the query and presentation space alleviates the cognitive load from the user's part unlike traditional search interfaces.

3. USAGE PATTERNS AND EVALUATION

Several experiments were performed to analyze the usage patterns and to compare the efficacy of the proposed approach. The experiments involved both quantitative and qualitative studies. A total of 20 participants from different backgrounds were involved in the studies.



chile hot recipe

The **Chile Pepper** Institute
Chile Pepper - The Zesty Life
Chili Pepper - Wikipedia, the free encyclopedia
Chipotle **Chili Peppers**
Chili Peppers Recipes
How Hot Are **Chile Peppers**?
ChilePlants.com Home Page
Chile Peppers
Chile Pepper Heat Scoville Scale - Home Cooking
Chile Peppers

red hot rock

Red Hot **Chili Peppers** Online
Red Hot **Chili Peppers** - Wikipedia, the free encyclopedia
Red Hot **Chili Peppers**
www.myspace.com/redhot**ChiliPepper**s
CBC Arts: Despite internet leak, new **Chili Peppers** disc is chart **...**
Chili Pepper - Products for **Chili Pepper** - Manufacturers,Exporters **...**

Top 3 location hits: Washington | Valencia | Mexico

Figure 2: Examples of interactions: selection of a cluster (top left) corresponding to the query "chili pepper" leads to thumbnail display of all pages contained in it (bottom). Top right: A region on the US east coast is selected. The drop-down menu is automatically initiated and contains a list of choices for location disambiguation.

The goal of the first study was to evaluate the efficacy of the proposed paradigm in helping users search and explore information as compared to the commonly used commercial search systems *Vivisimo*, *Grokker*, and *Google*. The study consisted of two sections separated by a week, during which the participants were expected to continue to using the systems in order gain expertise. During each section participants were given 20 information goals and asked to find the corresponding information. In Table 1, we list some of the queries and the specific aspects of information they were designed to be most related to. Overall the experiment design anticipated 36% of the queries to be text oriented, 21% to be spatial in nature, 18% to be temporal, and 25% to be media oriented. It should be noted that information goals could also be satisfied by exploring aspects of information different from those anticipated in the design of the experiment. For instance, the media oriented information goal of finding a PDF brochure on turtle-bay, Hawaii, could alternatively be satisfied using spatial cues and search.

The participants were divided into four groups. Each group ran the queries in a different order on the four applications based on the Latin Square to avoid bias due to the ordering of the tasks. Analysis of Variance (ANOVA) was used to determine the statistical significance of the difference among the mean scores of two or more variables. The time and number of clicks needed to find the desired information was recorded for each query executed on each of the four systems. As shown in Figure 3 (top row), across both measurements, our system was more efficacious than the three commercial systems. The reader may also note the high variance observed in this experiment. The effect of participants on the time needed to satisfy the information goal was significant for each system tested: *Google* (F=6.43, p<<0.001), *Vivisimo* (F=4.35, p<<0.001), *Grokker* (F=2.425, p<<0.001), and the proposed approach (F=1.82, p=0.018). On the other hand, the affect of participants was not significant in terms of number of clicks (p << 0.004) across all the applications investigated. It may be noted that our observations are similar to those reported in [3], where a similar variability in search times was observed across users.

Information Aspect	Example Information Goals
Media Oriented	Find: (1) video on recent immigration protests in the US, (2) video of the movie *Heaven's Gate*, (3) PDF tutorial on SPSS, (4) PDF brochure on *turtle-bay (Hawaii),* (5) piano music score *Fur Elise* by Beethoven
Temporal	Find the date of (1) IKEA's opening in Taipei (2) release of LINUX, (3) first discovery of dinosaur eggs, (4) opening of the channel tunnel between UK and France (5) the year in which the hot-air balloon was invented
Spatial	Find the location of: (1) places where Pandas can be found in the US, (2) Asian country where mummies were discovered (3) closest golf course from San Francisco with an ocean view (4) City in which Microsoft Inc opened its first Asian division (5) PDF brochure on *turtle-bay (Hawaii)*
Text Oriented	Find: (1) Five hottest varieties of chilli-pepper, (2) Five uses of lavender in cooking and medicine (3) what Bruce Lee studied in Univ. of Washington, Seattle (4) the religion of which the red lotus is a symbol, (5) recipe for fortune cookies

Table 1: Example information goals used as part of the first study

The second experiment studied the efficacy of the proposed approach in aiding information discovery. The notion of information discovery is very intuitive, yet it is hard to formalize and is arguably subjective. For the purposes of this experiment, we understood this notion to indicate (a) the discovery of any information related to the query and (b) discovery of semantically related groups of information to the query. For instance, the query "*treasure island*" could lead to information about the movie and the resort/casino sharing the same name as the book. This experiment focused exclusively on the information clustering/grouping capabilities of our approach and compared it with the functionally similar facilities provided in Vivisimo. Each participant was given six specific queries, with a varying number of information goals associated with each query, for example, the query *Frida Kahlo*, had the following information goals associated with it*: (1) find four cities where she lived; (2) what happened to her in 1925?* For each such task, the participants rated the two systems on a scale of 1 (very difficult) to 5 (very easy). The results from this experiment are shown in Figure 3 (bottom left); the proposed system scored an average of 4.33 on ease of both information-discovery and understanding of information structure while Vivisimo got scores of 3.2 and 3.3 respectively.

In the third experiment, we tracked the importance of each interaction modality in the information retrieval/exploration tasks. The responses were text-clustering (47%), spatial-display/interaction (19%), temporal-display/interaction (12%), and media-display/interaction (22%). In Figure 3 (bottom right) we compare these numbers with the expected values stemming from the design of the study. The biggest difference was observed in terms of a significantly higher than expected use of the text clustering and interaction modality. We postulate that this is due to the fact that historically web search has been text oriented. Therefore users were more familiar with this functionality and consequently used it more than the other interaction modalities supported in the system.

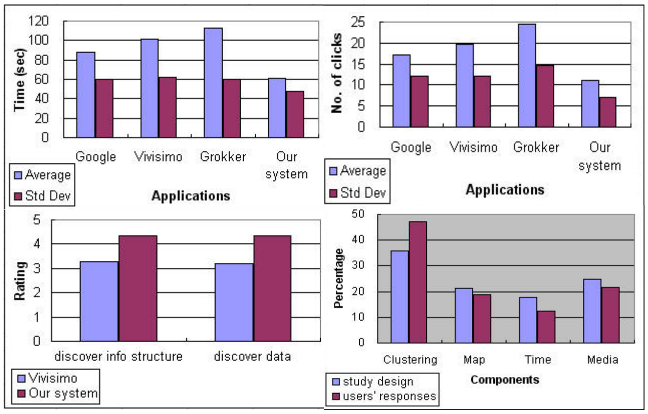


Figure 3: (Top row) Average time and number of clicks required to satisfy the information goals on each system. Bottom row (left): user ratings for Vivisimo and the proposed system for information/data discovery tasks (right): expected and actual usage of the specific interaction modalities during information search and exploration.

In the fourth experiment, a brief longitudinal study was conducted to determine the improvements (if any) in user performance over the period of one week between the first and second sections of the study. We remind the reader that the participants were expected to use all the systems during this period and become familiar with their features. Our hypothesis therefore was that the time required to complete information retrieval/exploration tasks would show a marked decrease, especially for the proposed system, Vivisimo, and Grokker. These results are presented in Figure 4. Briefly, the reduction in time for the proposed system was the most (13.6%). The reduction in time for the other systems was Vivisimo (5.7%), Google (2.2%), and Grokker (1.75%).

As part of the final study, a survey of the users was conducted, where they provided feedback on a Likert scale of 1 (strongly disagree) to 5 (strongly agree). The proposed system received the highest rating when compared with the other three systems involved, in terms of (1) helpfulness for finding desired information and (2) helpfulness for exploring relevant topics.

4. CONCLUSIONS

This paper outlines our research on designing an experiential user-data interaction environment for web search. The underlying design paradigm seeks to support human-machine synergy by identifying semantic correlations in the retrieved data and facilitating direct interactions between the users and the data. Further, mechanisms such as spatio-temporal displays, semantic clustering, and tight coupling between different views of the data help maintain user state and context and provide insights into the semantic structure of the information. This main focus of this paper has been on presenting results from investigations involving detailed user-studies and evaluations conducted in comparative settings with commercially available solutions. These results underline the efficacy and value of the proposed paradigm both in information retrieval and information exploration tasks. While the results in this paper were obtained in the context of web search, they are expected to be of relevance to a wide class of problems in information retrieval involving multifarious heterogeneous data and complex information needs.

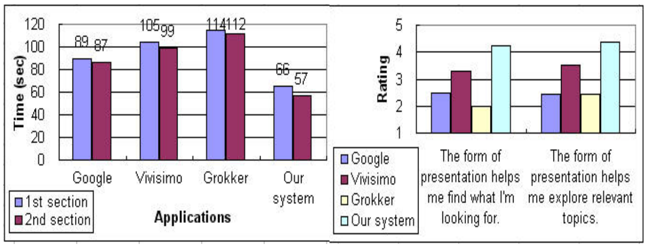


Figure 4: Time to reach the information goal in the first and second sections of the study which were separated by a week (left). Ratings assigned to the four systems by users (right).

ACKNOWLEDGEMENTS

The authors thank Naureen Moon for help in the user studies and Dil Chiture for co-designing the spatial interaction interface.

REFERENCES

[1] Alexa, http://www.alexa.com

[2] Cai D., He Xiaofei, Li Zhiwei, Ma W-Y, and Wen J-R, “Hierarchical Clustering of WWW Image Search Results Using Visual, Textual, and Link Information”, ACM Multimedia, 2004

[3] S. Dumais, E. Cutrell, and H. Chen, “Optimizing search by showing results in context,” ACM SIGCHI, 2001

[4] Grokker, http://www.grokker.com.

[5] Hutchins E.L., J. D. Hollan, D. A. Norman. "Direct Manipulation Interfaces," User Centered System Design. Lawrence Erlbaum Associates. 1986

[6] Ya-W Hsu, N. Moon, and R. Singh, “Designing Interaction Paradigms for Web-Information Search and Retrieval”, *IEEE/WIC/ACM Conference on Web-Intelligence (WI)*, 2006, pp. 815-822, 2006

[7] Jain, R. “Experiential Computing”, Communications of the ACM, Vol. 46, No. 7, July 2003

[8] P. Kruse et al, “Clever Search: A WordNet Based Wrapper for Internet Search Engines,” Proc. 2nd GermaNet Workshop, 2005.

[9] G. Notess, “Review of Northern Light,” 2002

[10] Open Directory Project, http://www.dmoz.org.

[11] P.L. Pirolli, and S. K. Card, Information foraging. Psychological Review. 106: p. 643-675, 1999

[12] D. Radev and W. Fan, “Automatic Summarization of Search Engine Hit Lists,” ACL Workshop on Recent Advancements in NLP and IR, Hong Kong, 2000.

[13] Santini S., A. Gupta, and R. Jain, “Emergent Semantics Through Interaction in Image Databases”, IEEE Trans. On Knowledge and Data Engineering, Vol. 13, No. 3, 2001

[14] Singh R. and Jain R., “From Information Centric to Experiential Environments”, Interactive Computation: The New Paradigm, D. Goldin, S. Smolka, and P. Wegner, eds., Springer Verlag, pp. 323 – 351, 2006

[15] Vivisimo, http://www.vivisimo.com

[16] Zamir O and Etzioni O., “Grouper: A Dynamic Clustering Interface to Web Search Results”, World Wide Web, 1999