Leveraging User Search Behavior to Design Personalized Browsing Interfaces for Healthcare Web Sites

Malika Mahoui, Josette F. Jones, Derek Zollinger, and Kanitha Andersen

Indiana University – Purdue University Indianapolis, School of Informatics, USA
{mmahoui,jofjones,dzolling,kpa}@iupui.edu

Abstract. Understanding and leveraging user search behavior is increasingly becoming a key component towards improving web sites functionality for the health care consumer and provider. Hence, the development and improvement of any interactive browser-based information system, such as those used by digital libraries, requires consideration of the type of individuals utilizing the system, an understanding of available content and inclusion of a way to measure user interactivity. Information systems not only need to provide useful content, they must also present content in a way that results in an efficient, effective and satisfying user experience. Functional interface design is assumed to take in consideration the overall environment of the user to support users in their search tasks. Web logs – access logs and search logs - record user interactions with the interface, and as thus provide insight in user search behavior in a natural environment. The present study measures the usability of a digital library through an in depth analysis of the web logs. The study also leverages user interaction with the digital library to propose a use driven browsing interface to improve user interaction with the system.

Keywords: Log data analysis, web usage mining, search term clustering.

1 Introduction

Web-based access is becoming a standard for organization to expose valuable information to interested users. Content of data collections is made available through browsing interfaces supported by searching capabilities. Browsing interfaces also called sitemaps are supported by taxonomies that guide the user progressively into locating his/her information needs. Despite the efforts made in information architecture to facilitate user experience with web sites, it remains a fact that users often find themselves using the search capabilities as the main tool for locating relevant health information. One of the main reasons identified to explain this search behavior is that often web site design is driven by the content of the web site and the requirements set by the web site owner without too much consideration to how the user perceive his way to find the information in the web site. Formative usability evaluations enable the detection of a certain number of usability defects such as difficulty of learning and using the system, high error rate, etc. and the estimation of the degree of seriousness of the defect [1]. The usability evaluation generally is a onetime process and is only

M. Kurosu (Ed.): Human Centered Design, HCII 2009, LNCS 5619, pp. 511–520, 2009.

geared at detecting those aspects of a user interface that may cause the resulting system to have reduced usability. These techniques though, do not focus on how well the interface supports users in their search tasks nor do they take in consideration the overall environment in which the search task is performed [2, 3].

In the present study, the human computer interactions and search behaviors are studied in their natural environment. Web logs are used to detect pattern of cognitive activities as well as behaviors which occur within human-computer interactions.

Research Question
The research questions to be answered by the research team are:

- Can we observe differences in search behavior before and after a website redesign?
- Do people from the same domain name, or similar domain names search for information similarly across a website redesign?
- Do people who search for specific topics search for information similarly across a website redesign?

2 Background

The original 1979 resolution for Sigma Theta Tau International's (STTI) Virginia Henderson International Nursing Library (VHINL) called for "a national nursing library resource offering services to nurses and those interested in nursing" and soon after that an additional call for "a national clearinghouse for information regarding nurse researchers and nursing research". Ten years later, the first computer was purchased for the Library, enabling the beginnings of an electronic library. With that development, and the establishment of a database that stored findings of research studies, nursing knowledge was made available in an electronic format [5]. In 2001 a "re-visioning the library" meeting was convened and participants framed their vision in the context of the dramatic technology changes that have occurred in the 10 years since the VHINL purchased its' first computer. The call is now to extend the VHINL functions to provide access to global knowledge resources and connections made possible by the World Wide Web, as well as retaining and enhancing the rich legacy of knowledge modeling.

The development and improvement of any interactive browser-based information system, such as those used by digital libraries, requires consideration of the type of individuals utilizing the system, an understanding of available content and inclusion of a way to measure user interactivity. Information systems not only need to provide useful content, they must also present content in a way that results in an efficient, effective and satisfying user experience. These considerations guided the redesign of user-centered interfaces to extend the functionality of the Virginia Henderson International Nursing Library (VHINL) from that of a library to that of a Web-based portal for accessing nursing knowledge resources.

Formal usability evaluations of the previous version of the VHINL detected certain inadequacies that hindered navigation of the system and use of its features. A survey collected information to assess the error rate and to estimate the seriousness of identified defects. One defect contributed to what is called "low stickiness." Most user sessions (74.24 percent) lasted only a few seconds, during which page views totaled

one or less. Often, as a result of ineffective strategy, search strings were poorly constructed.

In response to these findings, consideration was given to a tell-and-ask functional interface in which the user—usually a nurse—communicates with the knowledge base by making logical assertions—tell—and posing questions—ask. A prototype of this interface was also developed. Feedback was obtained from key user groups. The formal knowledge structure that currently exists within the library was examined and features needing expansion were identified along with opportunities for a controlled terminology. These analyses led to revisions of the data model and specification of an information model. The goal was that the knowledge structure and data model be complementary and reflect the search behavior of users.

A new interface and search engines have been installed and capturing of objective and subjective user data will continue. Analysis through data mining is planned to evaluate whether the revised knowledge model is consistent with the search behavior model, without unduly restricting a user's logical assertions or reasoning process in making queries.

3 Methodology

The aim of this study is to leverage web log data to improve the design of Sigma Theta Tau International's (STTI) Virgina Henderson International Nursing Library (VHINL) library. To analyze the user search behavior, two main approaches were undertaken: (1) analyze the user interaction with the web site and (2) study the search terms used to query the library.

To perform the former analysis, we leverage the fact that the VHINL web site had already undergone a previous redesign; and therefore the focus was to assess the impact of the redesign on the web site and suggest additional means to improve user interaction with the site.

To perform the latter study, the search terms used by the library visitors were the focus of a clustering analysis. The aim is to generate a browsing interface derived from the search terms deployed by the users to support the web site search capabilities and therefore better target users needs from the web site (see section 3.4).

3.1 Data Collection

Data was collected from STTI- VHINL website including both search logs and access logs. Pretest data was collected from October 1, 2002 until April 30, 2003 (7 months) while the website was operating under an Apache web server. Posttest data was collected from July 1, 2005 until September 4, 2005 (65 days) while the website was operating under an Internet Information Server (IIS) web server. In addition, the posttest data also included random popup questionnaires that complement the web log data.

3.2 Data Preparation

As in any information discovery process, a data preparation or preprocessing phase is needed in web log analysis. One aspect of data preparation is data cleaning. The main challenges faced in this study have to deal with two sets of data produced by two

different applications needing to be studied using similar criteria. Furthermore, the pretest data had a high signal to noise ratio concerning the data, thus making it difficult to map web data usage as the original web design was not available to us anymore. Nevertheless it was important to see whether the redesign of the web site and the underlying web server applications had a positive impact on the site.

- Part of data cleaning involves detecting and removing web bot and web crawler' activities in order to leave only "real" user interactions with the digital library. This step resulted in discarding a large percentage of the log data. Techniques deployed to detect Bot activity involved using the browser information available in the User Agent (UA) section of the log data. The goal was to keep only log data associated with a list of compiled known browsers (white list). In the compilation of this list, efforts were made to include browsers that are longer deployed such as the old Netscape browser. Furthermore, as this filtering technique may generate some false positive we also performed a second filtering to eliminate log lines with UA information including words such as "bot" and "crawl". While the same process was attempted for the pretest data, many of the log lines did not include UA information. Therefore, another method was required to clean the data. Initially AnalyzeSpider [6] was used to help identify and remove web robots. However, further analysis showed that false positives still existed in the filtered data. As a further cleaning process, all logs that accessed the "robots.txt" file were identified. Of the identified logs, their corresponding source (IP address, domain name, etc.) were selected and used to remove all log data originating from these sources.
- Another cleaning task consisted of removing lines that are not useful for the analysis such as lines recording downloading of icons, pictures, scripts, and other files which configure the web page to be correctly presented. This cleaning was performed semi automatically using Unix grep command as well as automatically as part of the software tool WUM used for web mining the log data (see section 3).
- Format conversion is further step in data preparation to ensure that the log data can be processed by the log analysis and web mining tools we used. As most of the open source tools for server log analysis have been designed for common log file (CLF) format produced by the Apache server, IIS2Apache [7] was used to convert the posttest data from the default IIS format to CLF format.
- Furthermore to be able to analyze data using domain names, IP addresses were mapped to fully resolved domain names. We used Web utilization monitor preparation (WUMprep) [8] to perform this conversion.
- To perform the *search terms clustering analysis*, a filtering of the log data was performed to keep only records that include search queries. A list of stop keywords (e.g. Virginia Henderson, the) was also compiled and used to clean each search query. Moreover, stemming was used to be able to better identify main topics shared by the library visitors. Porter stemmer was used for this task.

3.3 Data Sampling

Data sampling was also used during the analysis to study the impact of the web site redesign based on the category of users.

- In the first group of sampling the categories selected emphasized the language used (i.e. English vs. non English), heath affiliation, government affiliation, personal utilization, and educational affiliation. Eight matching samples were therefore created from the data sets based on origin.
  - *English Speaking* sample included all visitors from Australia, Canada, New Zealand, United Kingdom, and South Africa.
  - *Non-English speaking* sample included all foreign visitors not listed in English speaking category.
  - *Government* sample included visitors from Veterans Administration and National Institute of Health.
  - *Health-care* sample included visitors from Banner Health, Carolina Health Care, Emory University Systems Health Care, Heritage Valley Health System, Methodist Health System, Partners HealthCare, Sharp HealthCare, and St. John Health Systems Detroit Hospitals.
  - *Hospitals* sample included Bassett HealthCare, Blessing Hospital, City of Hope, Cleveland Clinic, LeHigh Valley Hospital, St. Jude Children's Research Hospital, and Virginia Mason Hospital/Medical Center. ISPs included Alltel, Ameritech, AT&T, Bellsouth, Covad, Earthlink, Mindspring, Pacbell, Qwest, and South-West Bell.
  - *Military* sample included all visitors from the .mil top-level domain (TLD).
  - *Education* sample included Harvard, Iowa, Kansas, Marquette, Maryland, Mayo Clinic, Miami, Monmouth, Purdue, University of California San Francisco, and Texas Medical Center.

The eight samples comprised about 30% of the entire population (Pretest samples 27.58%, Posttest samples 32.66%). In the remainder 70% of the population were (1) either a disproportional sampling in either the pretest or posttest data set; (2) or, the vast majority of the entries were non-existent in one or the other data set making a matched pre and post analysis not possible.

- In the second sampling method, groups were created based on the subjects searched by the users. For a given topic a list of related keywords is compiled using both scientific words and "common" words to cover for the diversity of the user population accessing the digital library. The "diabetes" topic was selected for this type of sampling. 122 keywords were used. This list included synonyms for Diabetes, such as "diabetes mellitus type 2," "non-insulin dependent diabetes mellitus," and "adult-onset diabetes." This list also included precursors and measures indicating diabetes, such as "hyperglycemia," "familial hyperproinsulinemia," "metabolic syndrome," "fasting blood glucose," and "hemoglobin A1C." Additionally, the list included common comorbidities normally associated with diabetes such as "hypertension", "nephropathy", "stroke," as well as known medications indicated for diabetes, such as "Insulin", "Humalog," and "oral antihyperglycemic." These terms were collected by a medical doctor and were as exhaustive as possible.

3.4 Analysis Method

Two main techniques were utilized to analyze the log data: traditional statistical analysis combined with web usage mining, and search term clustering.

- The aim of web usage mining is to extract knowledge from usage data representing user interaction with a web site. This knowledge can be used for several objectives including improving web site design, and supporting recommendation systems [9]. In addition of the statistical information that web usage mining provide (e.g. frequency of web page visitation), knowledge about the navigational paths that users exhibit is an integral part of the WUM process. To perform the mining process, log data is usually fragmented into users' sessions. Several criteria are proposed to define a user session including setting a time limit of a session or a time limit of non-user interaction with the system. In this study we used WUM tool [10] for web usage mining. All samples, as well as the entire population were analyzed using WUM. WUM tool creates user sessions from the raw data and measures user activities within those sessions. It also generates comprehensive reports with statistical data regarding visited pages and sessions. Report information was further post-processing to conduct statistical comparison between pretest and post data using both the whole population as well as the sample data sets. Five key markers were used to perform the comparison: Average Page Accesses Per Day, Average Visitor Sessions Per Day, Average Page Accesses Per Session, Average Unique Visitors Per Day, and Average Sessions per Unique Visitor.
- The main hypothesis behind search terms analysis is that search terms ultimately vehicle user needs from the web site; therefore they can be used to identify topics of interests shared by the users; and the topics can be combined to build a browsing interface that reflect user needs. To achieve this goal, clustering techniques were used as an unsupervised approach to identify main groups of terms shared by the visitors' web site. Experiments were performed with both k-means and hierarchical clustering using several similarity measures including cosine measure, correlation coefficient and Euclidian distance. The gCLUTO toolkit [11] was used to conduct the experiments given the useful tabular and graphical features it provides to visualize the clustered data. To build the similarity measure, one needed to represent the features to be clustered and their vector representation. For that purpose, we used WUM tool to sessionize the log data including only the search terms for both pretest and posttest data. For each session created, the list of keywords deployed during the visitors' search was identified. From there, two representations for the data was proposed: a session representation where each session has a keyword vector representation; and a keyword representation where each keyword has a session representation. With the first representation, clustering sessions is based on keyword similarity. In the second representation, clustering keywords is based on session similarity. As part of pre-filtering process, only stem keywords with a high frequency were selected for clustering. The search terms frequencies were computed, normalized and scaled for the statistical analysis.

4 Results

4.1 Web Site Usage Analysis

Tables 1 and 2 highlight some of the results obtained after comparing the pretest log data with posttest log data using the five criteria defined above:

- Visitors querying information about diabetes tend to navigate the new web page, visiting about the same amount of web pages, however there were fewer visitors interested in this web site.
- Visitor from English speaking countries tend to navigate the new web page, visiting more web pages, however there were fewer visitors interested in this web site.
- Visitors from non-English speaking countries tend to navigate the new web page, visiting more web pages, however there were fewer visitors interested in this web site.
- Visitors from government institutions tend to visit more often and more frequently, visiting more web pages when they navigated the new web site, overall there was much more interest in the web site from this group than the previous web site.
- While visitors from healthcare institutions tend to be the same, fewer institutions were aware of the newly redesigned web site, and those that did visit utilized the web site more than before.
- Visitors from hospitals tend to utilize and navigate the new web site more often, however individuals did not tend to return nearly as often.
- Visitors from Internet service providers tend to navigate the new web page, visiting more web pages, and once an individual had found the newly redesigned web page, tended to return to it more often as a resource, however there were fewer overall visitors interested in this web site.
- Visitors from military institutions tend to navigate the new web page, visiting more web pages, and once an individual had found the newly redesigned web page, tended to return to it more often as a resource, however there were fewer overall visitors interested in this web site.
- Visitors from educational institutions tend to navigate the new web page, visiting more web pages, and once an individual had found the newly redesigned web page, tended to return to it more often as a resource, however there were fewer overall visitors interested in this web site.

Overall, all visitors tend to navigate the new web page, visiting about the same amount of web pages, however there were fewer visitors interested in this web site.

When comparing each metric across the sample populations and also vis-à-vis the entire population. We can make the following observation

- The number of page access per day dropped vis-a-vis the entire population as well as for the sample populations we studied; except for the health and government related sample populations.
- Similar remark generally holds regarding the average sessions per day metric.
- While the average session length dropped for the entire population, the sample populations we studied all showed an increase in the session length. The most noticeable increase is for the hospitals population
- The number of accesses per session for each sample population follows the same trend as for the entire population. Here also the hospitals category shows a significant relative increase due to the substantially long sessions that this population presents.
- Similar remark holds for the unique visitor per day parameter with a noticeable increase also for the government sample population

- Average session per visitor for the sample populations does not follow the general trends; where we notice a noticeable drop for the English, Hospitals and Diabetes sample populations.

Table 1. Comparison of the site usage between pretest and posttest data for the sample data

Comparison metrics	Percentage change from pretest data to posttest data								
	Diabetes	English	Non English	Govern-ment	Healthcare systems	hospitals	ISPs	Military	Schools
Page Accesses Per Day	-69.08	-79.87	-2.04	359.73	50.66	154.42	-44.61	-24.73	-37.75
Average Visitor Sessions Per Day	-69.19	-84.57	-34.19	206.99	-0.48	19.09	-50.88	-50.13	-41.27
Accesses Per Session	0.37	30.39	48.84	49.76	51.38	113.64	12.76	50.93	5.98
Unique Visitors per Day	-21.50	-75.95	-32.78	192.25	-6.11	171.43	-52.59	-53.09	-44.84
Average Sessions per Visitor	-60.76	-35.83	-2.09	5.04	6.00	-56.13	3.61	6.30	6.47

Table 2. Comparison of the site usage between pretest and posttest data for the entire population

Comparison metrics	Percentage change –entire population
Page Accesses Per Day	-48.93
Average Visitor Sessions Per Day	-51.83
Accesses Per Session	6.01
Unique Visitors per Day	-51.16
Average Sessions per Visitor	-1.37

Overall, when analyzing the results of tables 1 and 2, we observe a general retraction of the web site usage in terms of the number of unique visits, how long they spend within the web site, and the number of pages they visits. However when we look at specific populations, we clearly identify that the government and health related populations have increased their interaction with the web site with more unique visitors spending longer sessions visiting more pages. In addition, we observe that the university-oriented population has decreased its activity with the web site aligning with the overall population trend. These observations suggest that the web site is becoming more professional oriented especially towards hospitals population. The results also indicate that for health and education related populations, the web site content is found useful as supported by the increase of the average user visits.

4.2 Search Terms Clustering Result

- Comparison of search terms between protest and posttest data (see Fig. 1) shows that search topics are similar with a slight higher frequency of searching of the current site ($t = -4.32$, $p < .0001$).

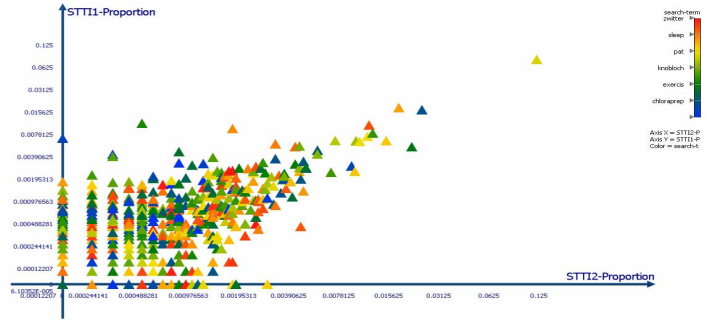

Fig. 1. Statistical comparison of search terms between protest and posttest data

- Clustering experiments using session representation (see section 3.4) allowed isolating the main topics that users are interested in (e.g. pain, nurse practitioner); which are represented in the first column describing a cluster (see fig. 2).
- Potential sub-topics are also generated and represented by the remaining columns in Fig.2. For example, we identified "manage pain" as potential sub-category for "pain" category.

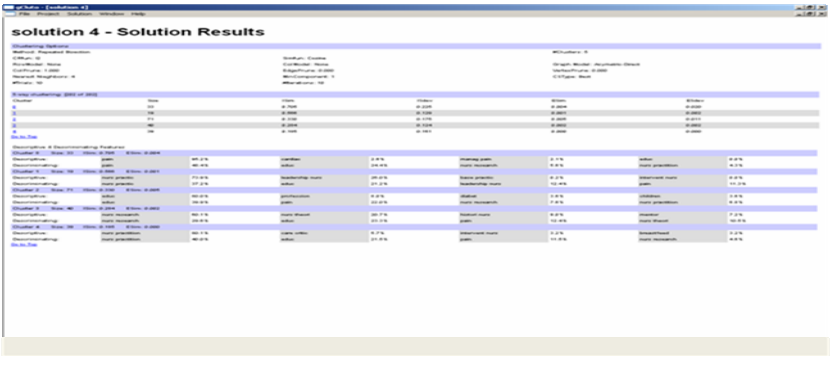

Fig. 2. Sample tabluar representation of the clustering results generated using gCluto software

5 Conclusion

In this study, we described a new approach that combines web usage mining and search terms clustering, leveraging on log data, to improve the design of STTI digital library. While the results of the web site usage mining suggest that the web site is gaining popularity within the health related institutions, it could also reveal a flaw in the current design if the audience targeted by the web site was larger than that. The results of the search terms clustering approach allowed the identification of groups of frequent topics and potentially sub-topics to be used to generate a hierarchical browsing interface for the web site. However, a more elaborated study remains to be performed to assess the utility of the user driven generated hierarchy.

References

1. Shneiderman, B.: Universal Usability. Communications of ACM 43(5), 84–91 (2000)
2. Richardson, J., Ormerod, T.C., Shepherd, A.: The role of task analysis in capturing requirements for interface design. Interacting with Computers 9, 367–384 (1998)
3. Paradowski, M., Fletcher, A.: Using task analysis to improve usability of fatigue modelling software. International Journal of Human-Computer Studies 60(1), 101–115 (2004)
4. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition 5(2), 199–220 (1993)
5. Graves, J.R.: Structuring a knowledge base: the arcs© model. In: C.B. (ed.) Nursing Informatics: Education for practice. Springer Publishing Company, Inc., New York (2000)
6. Jgsoft, Analyse Spyder version 3.01. Downloaded and used 2007-2008 (2004), `http://www.analysespider.com/analysespider.html`
7. Abendschan, J.W.: "iis2apache.pl", Downloaded and used Feburary 27 (2008), `http://www.jammed.com/~jwa/hacks/`
8. WUMprep. Web Usage Mining Preparation Tool, `http://hypknowsys.sourceforge.net/wiki/Web-Log_Preparation_with_WUMprep` (cited downloaded and accessed 2007-2008)
9. Mobasher, B., Cooley, R., Srivastava, J.: Automatic personalization based on web usage mining. Communications of the ACM, 142–151 (2000)
10. WUM. Web Usage Mining tool, `http://hypknowsys.sourceforge.net/wiki/Welcome` (cited downloaded and accessed 2007-2008)
11. gCLUTO. Graphical Clustering Toolkit, `http://glaros.dtc.umn.edu/gkhome/cluto/gcluto/` (cited downloaded and used 2008)