Exploring Health Website Users by Web Mining

Wei Kong and Josette Jones

Health Informatics, School of Informatics, Indiana University,
Indianapolis, Indiana, Usa
weikong@umail.iu.edu, jofjones@iupui.edu

Abstract. With the continuous growth of health information and users of the Internet, how to build an easy-to-use interface for different users is one fundamental desideratum when constructing a health website. The goal of this paper is to explore different information needs by examining the search terms of different Internet user groups. In order to deeply investigate the information, five months' daily access weblog files from one particular health provider's website are collected and the log data is analyzed by Web Mining technique. Based on the mining results, the paper also gives some suggestions of how to redesign the interface to be more intuitive to users.

Keywords: Web Mining; Web Log; Internet; User-Computer Interface; Search Term; Information Seeking Behavior.

1 Introduction

With the rapid development of the Internet and technologies used in the health area, people have more opportunities than ever to resort the Internet for health information seeking. Surveys [1-2] have shown that more than half of patients have used the Internet to access health information. In addition, more than 70% of Internet users prefer to use search engines rather than medical portals or libraries to start the information searching[3]. Several studies [4-7] have also described the importance of the use of the World Wide Web (WWW) as a source for health information and demonstrated that individuals who seek health information on line for decision-making promoted disease management and thus improving quality of life. It's very clear that the Web is progressively playing a significant role in patients' healthcare, and the impact of the Internet cannot be overlooked.

However, the information on the Web is huge and overwhelming. Although general search engines, such as Google and Yahoo are good starting points for users, the precision of the information retrieval results still need to be improved [8-10]. Obtaining the maximal benefits from the Internet must be built on understanding the users' interests, characteristics, and preference first. Understanding the users' preference is the first step to provide the tailored health service and user-friendly Web interfaces. The purpose of this study is to examine users' information seeking behaviors based on different user groups by extracting their search terms with Web mining technology.

C. Stephanidis (Ed.): Universal Access in HCI, Part IV, HCII 2011, LNCS 6768, pp. 376–383, 2011.

Like Lambert & Loiselle mentioned, "*Seeking information about one's health is increasingly documented as a key coping strategy in health-promotive activities*"[11]. Some studies have shown some progress of Web design by mining the user behaviors in health domain. Chen and Cimino[12] analyzed a Web-based clinical information system's (New York Presbyterian Hospital) logs to discover patient's pattern of usage. The result of mining data indicated that users usually view radiology and laboratory results in one session. Hence, they suggested adding "shortcuts" in these Web pages to provide patients a quicker access to the information. Graham and Tony[13] did navigation research on a consumer health website (ClincialTrials.gov) and one of their findings showed that majority of the users were referred by general search engines to access the webpage. Therefore, they suggested increasing the use of general search engines like Google. Rozic-Hristovsk and Hristovski[14] investigated the usage of the central medical library of their University by exploring weblog files and decided to reconstruct more intuitive reference pages to fulfill the increasing visitors. These applications have represented that Web log analysis is a powerful tool for researchers to explore the users' usage patterns and correlate these characteristics to website construction.

However, most of the applications focused on all Web users instead of specific users. As Rozic-Hristovsk and Hristovski[14] stated in the limitation discussion in their study that "*the analyse adequaletly reveal overall usage patterns but can only provided estimated of individual user characteristics*". Although it is hard to predict every single user's preference, it is useful that users can be divided into groups to examine their information needs.

In past studies, some researchers investigated health information seeking behavior from either a patient's or a physician's perspective. A study[15] of cancer patients' information needs showed that all participated patients just want basic information rather than every detail information at all stages of their illness, while other study[16] has shown that primary physicians would like to spend more time gathering information focused on the diagnose and treatment. Also, people would like to assume that there is a different preference between the Web users when they seek information. For example, HON.ch [17] provides different search options and the result are different even we search a same term. For patients, it provides some consumer health website links to other websites like MedlinePlus, WebMD and family doctor. While for health professionals, the result is more focused on professional articles from journals and knowledge base. For example, it provides some articles from eMedicine, which is an online clinical medical knowledge maintained by WebMD. Although users are "assumed" differently, limited research and comparison was done to prove this assumption quantitatively. In this study, we will investigate different information needs of patient and doctor group based on the Web query analysis.

2 Method

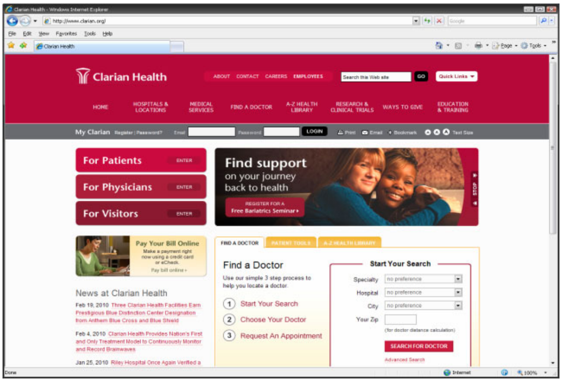
The website of Clarian Health [18] was used for this study. Clarian Health, now renamed as "IU Heath", was first formed in 1997 and it is a private, nonprofit organization owns more than 20 hospitals and health centers throughout Indiana.

Daily access log files of this website were collected and analyzed for pattern detection in search behaviors of different user groups such as patients, providers and occasional visitors. As is suggested from the literature and communication with users, their needs are complementary but different.

Web mining technology was used for data analysis. WUM-prep and Perl script were used to clean the original data; descriptive statistics described the features of the logs and user groups; association rule was employed to discover the frequency and patterns of the users; auto-classification categorized the users even without logged information.

2.1 Data Collection

The study was based on five months' weblog collected in 2007. The usage data of this website is considered sufficient to provide some trends, and also the website has already built up the navigation bars for patients, physicians and the visitors, which would greatly facilitate the user classification. Figure1 is a screenshot of this website.



Make a payment right now using a credit card or eCheck.

News at Clarian Health

Feb 19, 2010 Three Clarian Health Facilities Earn Prestigious Blue Distinction Center Designation from Anthem Blue Cross and Blue Shield

Feb 4, 2010 Clarian Health Provides Nation's First and Only Treatment Model to Continuously Monitor and Record Brainwaves

Jan 25, 2010 Riley Hospital Once Again Verified a

Use our simple 3 step process to help you locate a doctor.

(for doctor distance calculation)

Fig. 1. Screenshot of Clarian's website

2.2 Data Preparation

Raw log data was processed to reduce the noisy. After removing the identification information of the users, we got rid of the Web spiders, irrelevant and duplicate records. Web spiders, also called Web robots or Web crawlers, are programs that automatically collect relevant contents from Web pages, so the search queries generated by these spiders do not represent the actual information needs of the real users. And then user sessions were made by cookie and 30 minutes time constrain. The user groups of patients, doctors and visitors were separated by URL. The last step was to extract the search terms of the users. Figure 2 describes the whole process.

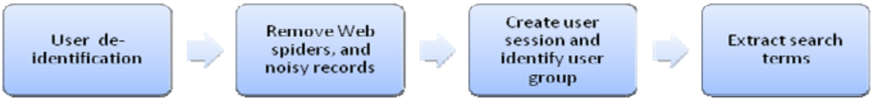

User de-identification
Remove Web spiders, and noisy records
Create user session and identify user group
Extract search terms

Fig. 2. Process of Data Preparation

3 Result

3.1 Log File Statistics

Figure 3 describes the statistics of the log files. There were totally 11 million original log records, but after the cleaning process, there were only 58% left to be used to process the study. And during the five months' time period, April and May have the top visit amount. This may be due to the season's reason or some event happened during that time period.

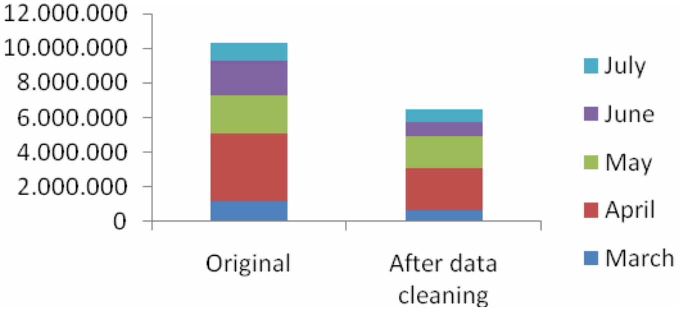


Fig. 3. Log file statistics

3.2 User Session

Patients have around 200 thousands user sessions, which are almost 10 times of doctors. No user logged as a visitor during five months period. 73% percentage of the users did not log as any of the user groups when they were surfing the website. Although this website has built the log in button, majority of the users still didn't logged, so they might not have the special service provided based on user groups.

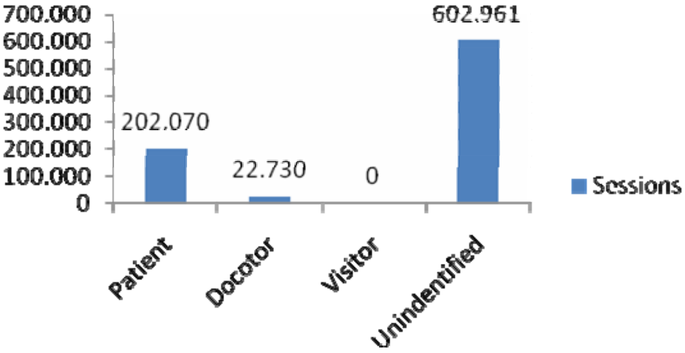


Fig. 4. User sessions

3.3 Search Engine Distribution

We examined the query request from four most popular search engines, Google, Yahoo, MSN, now is Bing, and the Clarian's site search. The result has shown that 30% of the users employ search engine to look for health information and google.com is the most popular search engine for both users logged as patients or doctors. It can be seen that intranet is well used through Web and we recommend that this website could consider increase the server support ability and optimize the website to Google.

Another finding we can see from the result is that doctors relied on site search engines much more than patients. It can be predicted that when doctors were browsing this website, they usually cared about specific topics rather than general ideas. So they preferred to search directly in Clarian's site.

Figure 5 give the search engine distribution chart for both patient and doctor group.

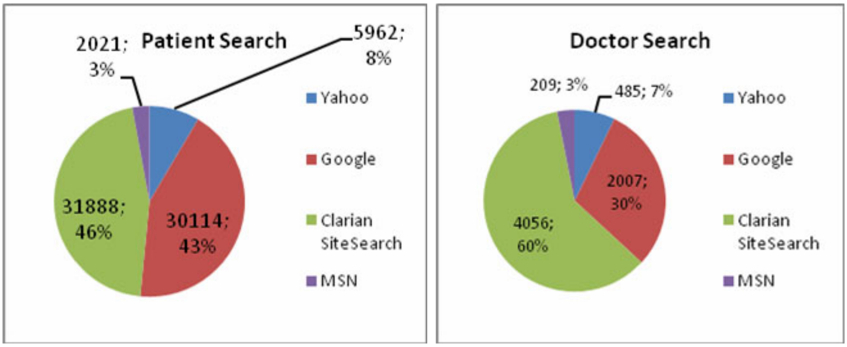


Fig. 5. Search engine distribution

3.4 Search Term

We extracted the top 20 search terms from Clarian's site search engine according to patients and doctors groups.

For patient group, it is surprise to see that more than half of the terms are related to employment and education information. In another word, users logged as patient cared about jobs and training rather than health information. This prediction may be true, but there may be another reason that the users logged as patients are not the real "patients". They might be some "seekers" looking for employment or training. As they didn't know where to go through the home page, they clicked on "patients" to start. In this case, we can see that the homepage interface is not intuitive for these users. They were confused and mislead by the webpage.

Compared to patient group, doctors used more medical terms to search, like pathology, pain. However, they were more likely to use this website as a handy tool to search auxiliary information, as doctor detail information (by search Dr.name), patient medical record (by careweb) and lab or surgery data.

Table 1. Top 20 Clarian Search Terms with Term Frequency

Patient		Doctor	
job	814	dr.	214
clarian	666	center	75
center	641	clarian	69
methodist	616	medical	68
employment	598	care	58
health	582	methodist	55
medical	571	health	41
care	480	*pulse*	40
human	409	john	39
patient	387	physician	38
program	355	*surgery*	35
resource	349	clinic	34
address	323	doctor	33
hospital	294	careweb	33
employee	293	*laboratory*	31
nurse	289	*lab*	28
class	284	*cancer*	27
career	241	*pathology*	26
dr.	282	*transplant*	26
service	254	*pain*	23

3.5 Association Discovery

In this study, we also applied market basket analysis[19] to find the associations of the popular search terms. As we know which terms patients or doctors are most likely to search together, we can better understand the users' needs, and also provide dynamic site search suggestions to users to promote their search. Table 2 lists some of the associations found with high confidence rate.

Table 2. Associations for patient and doctor groups

Patient			Doctor		
Term1	**Term2**	**Confidence**	**Term1**	**Term2**	**Confidence**
human	resources	92.83%	order	sets	95.24 %
therapy	physical	84.80%	women's	health	84.62 %
phone	number	83.33%	west	clarian	84.62 %
life	child	93.52%	group	medical	70.00 %
information	patient	56.57%			
community	plunge	91.30%			
records	medical	92.31%			
people	mover	97.44%			
financial	assistance	72.97%			
occupational	health	69.23%			

3.6 User Group Classification

As we seen before, 73% of the users did not log as any group when browsing, so any "tailored" service would not be available to them. Nevertheless, if a classifier can be built to automatically identify the user role based on the search terms users input, majority of the users can still get benefit even they don't log in.

In this study, we tried several popular classifiers, like naïve Bayes and neural network. Among these, we found the binary linear classifier, Support Vector Machine (SVM), has the best F-score. The classifier was tested based on 600 patient queries and 600 doctor queries randomly selected from the five month data. With this classifier, we are able to categorize the users. So when people search information, they can be suggested or directed according to their user roles, no matter whether they remembered to log in or not.

Table 3. Performance of SVM classifier

		Patient **(F = 74.57%)**	**Doctor** **(F = 84.00%)**	
Predict	Patient	True Positive = 447	False Positive = 96	Precision=82.32%
	Doctor	False Negative = 153	True Negative = 504	Precision = 76.71%
		Recall:= 74.50%	Recall:= 84.00%	

4 Conclusion

The previous results prove that users logged as patients or doctors have different information preferences. Web mining technology can help us understand what information the users are really interested. Based on the results we found, we would purpose some suggestions to redesign the website and build more user-friendly interface for different users. The suggestions are summarized as below.

- For the homepage, remove the "visitor" log portal, and instead of it, build a log portal for employment careers, like "employer" or "future employee".
- Differentiate the entry pages for different user groups. For patient group, build friendly links to training, education programs and general information.
- For doctor group, build intuitive links to people contact directory, knowledgebase, and auxiliary medical data access.
- For search engine, provide dynamic searching suggestions and implement the SVM classifier to promote the search criteria.

5 Limitations and Future Study

This study only represents the users' seeking pattern from one website and the results can only be used as an estimate data for other health websites. The user separation is based on the log in information, so we cannot say the user groups of patients or doctors are the real patients and doctors in life. As we have pointed out that the users logged as patient may be some "employment seeker" who are just look this website for jobs. Only we can get rid of these noisy data from patient user group, can we future investigate more topics of the patient group. Another study may be more focus on the navigation pattern of the different groups, like what is the path to find the same topic, are there any waste steps in the process to get the final information. Get to know that, we can redesign and refine the website for more convenient use.

References

1. Ayantunde, A.A., Welch, N.T., Parsons, S.L.: A survey of patient satisfaction and use of the Internet for health information. Int. J. Clin. Pract. 61(3), 458–462 (2007)
2. Trotter, M.I., Morgan, D.W.: Patients' use of the Internet for health related matters: a study of Internet usage in 2000 and 2006. Health Informatics J. 14(3), 175–181 (2008)
3. Eysenbach, G., Köhler, C.: How Do Consumers Search For And Appraise Health Information On The World Wide Web? Qualitative Study Using Focus Groups, Usability Tests, And In-Depth Interviews. BMJ: British Medical Journal 324(7337), 573–577 (2002)
4. Eysenbach, G.: The Impact of the Internet on Cancer Outcomes. CA Cancer J. Clin. 53(6), 356–371 (2003)
5. Fox, S., Fallows, D.: Internet Health Resources. Internet & American Life Project (July 16, 2003) [cited August 30, 2003)]; report/survey], `http://www.pewinternet.org/`
6. Fox, S.: Health Information Online. PEW Internet & American Life Project (2005)
7. Rice, R.E.: Influences, usage, and outcomes of Internet health information searching: Multivariate results from the Pew surveys. International Journal of Medical Informatics 75(1), 8–28 (2006)
8. Chang, P., et al.: Are Google or Yahoo a good portal for getting quality healthcare web information? In: Proc. of AMIA Annu. Symp., 2006, p. 878 (2006)
9. Morita, T., et al.: A study of cancer information for cancer patients on the internet. Int. J. Clin. Oncol. 12(6), 440–447 (2007)
10. Wu, A.S., et al.: Evaluation of Negation and Uncertainty Detection and its Impact on Precision and Recall in Search. J. Digit Imaging (2009)
11. Lambert, S.D., Loiselle, C.G.: Health information seeking behavior. Qual. Health Res. 17(8), 1006–1019 (2007)
12. Chen, E.S., Cimino, J.J.: Automated discovery of patient-specific clinician information needs using clinical information system log files. In: Proc. of AMIA Annu. Symp., 2003, pp. 145–149 (2003)
13. Graham, L., Tse, T., Keselman, A.: Exploring user navigation during online health information seeking. In: Proc. of AMIA Annu. Symp., 2006, pp. 299–303 (2006)
14. Rozic-Hristovsk, A., Hristovski, D., Todorovski, L.: Users' information-seeking behavior on a medical library Website. J. Med. Libr. Assoc. 90(2), 210–217 (2002)
15. Leydon, G.M., et al.: Cancer patients' information needs and information seeking behaviour: in depth interview study. BMJ 320(7239), 909–913 (2000)
16. Gonzalez-Gonzalez, A.I., et al.: Information needs and information-seeking behavior of primary care physicians. Ann. Fam. Med. 5(4), 345–352 (2007)
17. Health On the Net Foundation, `http://www.hon.ch/`
18. Clarian Health, `http://www.clarian.org`
19. Agrawal, R., et al.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM, Washington, D.C (1993)