EXPLORING HEALTH WEBSITE USERS BY WEB MINING

Wei Kong

Submitted to the faculty of the School of Informatics

in partial fulfillment of the requirements

for the degree of

Master of Science in Health Informatics,

Indiana University

May. 2012

Accepted by the Faculty of Indiana University,

in partial fulfillment of the requirements for the degree of Master of Science

in Health Informatics

Master's Thesis

Committee

_____

Josette Jones, Ph.D., Associate Professor, Chair

_____

Malika Mahoui, Ph.D., Adjunct Professor

_____

Hadi Kharrazi, M.D., Ph.D, Assistant Professor

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

I would like to thank IU Health for providing us the important raw data and grant us the authorization to analyze them. I sincerely appreciate Prof. Josette Jones's help for providing excellent mentorship and very useful guidelines to complete this research. I also want to thank my other committees, Prof. Malika Mahoui and Hadi Kharrazi for very helpful methodology reviews and precious comments. In addition, I would like to give my thanks to my thesis group classmates. Their good suggestions helped a lot to inspire my thinking. The concepts I learned from this research will benefit a lot in my future study.

ABSTRACT

Wei Kong

EXPLORING HEALTH WEBSITE USERS BY WEB MINING

With the continuous growth of health information on the Internet, providing user-orientated health service online has become a great challenge to health providers. Understanding the information needs of the users is the first step to providing tailored health service. The purpose of this study is to examine the navigation behavior of different user groups by extracting their search terms and to make some suggestions to reconstruct a website for more customized Web service. This study analyzed five months' of daily access weblog files from one local health provider's website, discovered the most popular general topics and health related topics, and compared the information search strategies for both patient/consumer and doctor groups. Our findings show that users are not searching health information as much as was thought. The top two health topics which patients are concerned about are children's health and occupational health. Another topic that both user groups are interested in is medical records. Also, patients and doctors have different search strategies when looking for information on this website. Patients get back to the previous page more often, while doctors usually go to the final page directly and then leave the page without coming back. As a result, some suggestions to redesign and improve the website are discussed; a more intuitive portal and more customized links for both user groups are suggested.

CHAPTER ONE: INTRODUCTION & BACKGROUND

Background

With the rapid development of the Internet and technologies used in the field of health care, people have more opportunities than ever to use the Internet for health information. Surveys (Ayantunde, Welch, & Parsons, 2007; Trotter & Morgan, 2008) 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 searching for information (Eysenbach & Köhler, 2002). Several studies (Eysenbach, 2003; Susannah Fox, 2005; Susannah Fox & Fallows, 2003; Rice, 2006) have also described the importance of the use of the World Wide Web (WWW) as a source of health information, and have demonstrated that individuals who seek health information online for decision-making have promoted disease management, thus improving their quality of life. A recent study (Chiu, 2011) shows that patients like to search health information on the Internet to probe and verify their doctors' competence. The Internet helps them to understand the doctors' jargon and thus pushes doctors to prepare more for patient's questions. Nowadays, users are not only accessing health information on the Web, but are also using an increasing number of Web applications, like search engines or Personal Health Records (PHRs) to improve their perceived knowledge of health problems (Fernandez-Luque, Karlsen, & Bonander, 2011). 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, there are thousands of websites distributing health information, from public government-owned websites to individuals publishing based on their experiences. The quality of information on the Web is diverse. Research indicates that there are big variances existing in some health websites (Greenberg, D'Andrea, & Lorence, 2004), but accepted standards lack the ability to uniformly evaluate them (Morahan-Martin, 2004). Although general search engines, such as Google and Yahoo are good starting points for users, the precision of the information retrieval results still needs to be improved to be useful (Chang, Hou, Hsu, & Lai, 2006; Morita et al., 2007). There are also many studies evaluating the quality of health information on the Internet, but the results demonstrate that the suggestions given online have not been proved beneficial (Hallingbye & Serafini, 2011; Lawrentschuk, Abouassaly, Hackett, Groll, & Fleshner, 2009; Tangri & Chande, 2011). In addition, health information in the Internet environment is inherently generalized, so it cannot fulfill the users' individual needs dynamically according to their own situations (Risk & Dzenowagis, 2001). With the continuous growth of information on the Internet, dealing with information overload and learning how to develop more "dynamic" and "personalized" Web service will be a main challenge to the website builders.

Moreover, in recent years, health applications have been getting more and more popular in social network service, and health consumers are more than ever expecting personalized experience in Web health applications (Fernandez-Luque et al., 2011). In order to obtain the maximum benefits from the Internet, the first step is to understand the users' interests, characteristics, and preferences so that tailored health service, user-friendly Web interfaces, and profitable Web applications can be built.

In this situation, Web mining provides a good method to find what exactly the users want. Bing Liu (Liu, 2007) defines Web mining in his, book: "Web mining aims to discover useful information or knowledge from the Web hyperlink structure, page content, and usage data" (p. 6). Furthermore, users searching for health information usually show specific information-seeking behaviors that are highly individualized (Stavri, 2001). Web usage mining, which is an important branch of Web mining, offers valuable methods for personalized service from a user's individual perspective.

"Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user" (Liu, 2007). By analyzing the weblogs, providers can observe more individual information needs, like what patients are looking for and how they get the information from the websites. It's a significant step toward improving the satisfaction of the potential users.

Purpose of the Study

The purpose of this study is to (1) examine the information-seeking behaviors of providers, patients and visitors using a Midwest health institution's website by extracting their search terms with Web mining software and (2) to make some suggestions to reconstruct the website to increase its functionality for different user groups.

Exploring the users' preference is the first step to providing tailored health service. A thorough examination of users' seeking behavior—such as what they are looking for, what categories of topics they are most concerned about, and how they get the information from the websites—would help us to clarify user preferences. With such

information, health providers could build more attractive websites and provide more efficient search results based on different user groups.

Significance of the Study

For health organizations, the research results will provide preliminary data for reconstructing existing health websites and for building new user-orientated ones.

For particular Internet users, like patients and physicians, the research will contribute to offering a customized health service. Knowing what the users' needs are, therefore, becomes a significant step toward improving their satisfaction.

Furthermore, the results will provide fundamental knowledge for developers of websites and Web applications to establish long-term user profiles and thus lay the foundation for dynamic search filters that would filter the search results according to a user's personal information, such social roles and potential interests.

The target groups in this study will be defined as physicians and patients.

Web Mining Methodology

Because Web mining will be the primary technology used to conduct this research, some background knowledge of Web mining will be introduced in this section.

Information Retrieval and Web Mining

"Retrieving information simply means finding a set of documents that is relevant to the user query" (Liu, 2007)(p.183). Due to its convenience and richness, the Web is increasingly becoming a major source of information. Web mining provides a method of automatically discovering and extracting information from Web documents and services;

thus it is a part of the Web Information retrieval process (Kosala & Blockeel, 2000). In his research, Orland Hoeber (Hoeber, 2008) provided a vision of the opportunities and challenges of the future Web search engines and the "Web information retrieval support systems."

Web Mining Categories

According to Bing Liu (Liu, 2007) and Raymond Kosala (Kosala & Blockeel, 2000), web mining tasks can be categorized into three types based on the research interest. They are Web Content Mining, Web Structure Mining and Web Usage Mining. Web content mining is used to extract textual information from documents on the Web; Web structure mining is used to discover the structural information behind the hyperlinks; and Web usage mining tries to discover the sessions or behaviors of the users by referring to the log analysis and clickstreams. Table 1 (Liu, 2007) gives an overview of Web mining categories.

Table 1.

Web Mining Categories and Applications

	Web Mining		
	Web content mining	**Web structure mining**	**Web usage mining**
Data Source	Hyperlinks	Content documents	Usage data (ex. logs)
Application Categories	-Categorization -Clustering -Finding extraction rules -Finding patterns in text	-Categorization -Clustering	-Site construction, adaptation, and management -Marketing -User modeling

CHAPTER TWO: LITERATURE REVIEW

Summary of Literature Review

As Lambert and Loiselle (Lambert & Loiselle, 2007) observed, “Seeking information about one's health is increasingly documented as a key coping strategy in health-promotive activities.” There are many studies that have stressed the issue of personalized website service. Yet to design tailored information services requires an understanding of the user’s behavior and search approaches. With Web mining technology we can provide patients health information individually and improve the design of websites so that they can get the information quickly.

Web Mining Process and Web Usage Mining

Web Mining Process

There are three stages of the Web mining process, which follows the general data mining process (Liu, 2007). They are collection and pre-processing, pattern discovery, and pattern analysis. Figure 1 (Liu, 2007) graphically summarizes the above process.

Collection and pre-processing. The first step involves not only the collection of suitable target data, like access logs and server logs, but also the cleaning and partitioning of these raw data. Although this is the most difficult and time consuming stage in the process, there is no doubt that the result in this step is very critical to the success of the application. It is the crucial precondition, and the final result greatly relies on this task. This stage often requires special algorithms and heuristics not commonly employed in

other domains. Robert Cooley (Cooley & Srivastava, 1999) presents several data preparation techniques that are necessary for performing Web mining.

Pattern discovery. In the pattern discovery stage, data mining, machine learning, and statistical operations are performed to obtain hidden patterns that reflect the typical behavior of users. The users are automatically segmented and classified based on their similar behavior, and then the adaptive user model is developed. This type of model represents a collection of personal data associated with specific users, such as preferences, interests, and skills. In this research, descriptive statistics were used to describe the features of the logs and user groups; association rules were employed to discover the frequency and patterns of the users; and auto-classification categorized the users even without logged information.

Pattern analysis. In the last stage of the process, the discovered patterns and statistics are further processed and filtered in order to meet the different representation requirements.

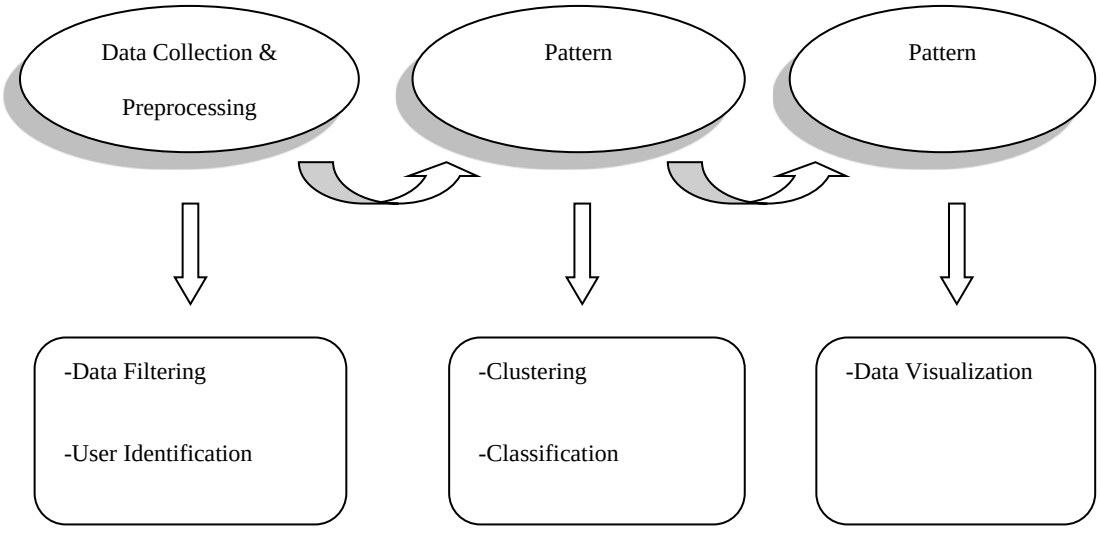


Figure 1. Web Mining Process.

Web Usage Mining

"Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interactions with Web resources on one or more Web sites. The goal is to capture, model, and analyze the behavioral patterns and profiles of users interacting with a Web site" (Liu, 2007).

The purpose of Web usage mining is to gather useful information about navigation patterns. This information can be exploited to help improve the user's satisfaction. Information obtained by mining Weblogs can have several applications (Facca & Lanzi, 2005; Srivastava, Cooley, Deshpande, & Tan, 2000):

1. Personalization
2. Improving navigation through pre-fetching and caching
3. Improving Web design
4. E-commerce

In the health area, it can be seen that user modeling techniques, such as personalization and Web design, offer opportunities for health providers to improve patients' satisfaction.

Personalization and Web design

"Web Personalization is simply defined as the task of making Web-based information systems adaptive to the needs and interests of individual users" (Pierrakos, Paliouras, Papatheodorou, & Spyropoulos, 2003). The object of personalization is to provide a particular information package dynamically. The most common application in personalization is a recommendation system (Facca & Lanzi, 2005). Typically, this type of system compares a user's profile to some reference characteristics, and then tries to

predict the preference that a user may have for an item he or she had not yet considered. Dynamic recommendation would be a very attractive development from a user's perspective, and the technique used to achieve this goal is discussed in many research articles (Aghabozorgi & Wah, 2009; Eirinaki & Vazirgiannis, 2003; Mobasher, Cooley, & Srivastava, 2000; Pierrakos et al., 2003) .

With the expansion of information volume, the complexity of health web design is increasing correspondingly. It is easy to understand that a well-designed website is a critical factor that contributes to gaining the satisfaction and loyalty of commercial customers. In the health field, it is assumed that the impact is also profitable. As the Web has become the leading source of health information, making the website efficient and convenient for patients is very important. The object is to provide different users the most appropriate information in the shortest time from an adaptive website.

Past Studies

Past studies employed query analysis and Web usage mining as their methods to discover the users' information needs and searching behaviors.

For query analysis, Shuyler and Knight (Shuyler & Knight, 2003) analyzed what people are searching for when they use a health-education website offering orthopedics and sports-medicine topics. They examined the queries users submitted to the Ask a Question function, allowing the users to describe what they were looking for in more detail than that provided by the small keyword search box. They performed content analysis to discover the overall trends observed in the raw data. The study suggested the five most common reasons for users to visit the website are to seek information about a

condition, a treatment, or a symptom, or to ask advice about symptoms or treatments. Therefore, they suggested the site managers organize their health-information websites according to these topics. In order to examine the characteristics and topics of medical and health information queries, (Spink et al., 2004) collected about ten thousand query terms on Excite.com and AlltheWeb.com and then classified them according to topics. Their findings showed that only a small percentage of web queries are medical or health related, and the top five categories of medical or health queries were general health, weight issues, reproductive health and puberty, pregnancy/obstetrics, and human relationships. (Scherer, Zitterbart, Mildenstein, & Himmel, 2010) analyzed the content of a Web-based expert forum for migraine and headache information (www.lifeline.de). They examined more than eight hundred queries over four years and found out that users of this Internet forum usually had questions about symptoms and their interpretation, as well as drugs and therapies.

For Web usage mining, Chen and Cimino (Chen & Cimino, 2003) analyzed a Web-based clinical information systems (New York Presbyterian Hospital) logs to discover patients' pattern of usage informing design and development of future clinical systems. In the first stage, one year of system log files from a Web-based clinical information system were collected. Data preprocessing included de-identifying usernames and medical record numbers, removing duplicated and unnecessary data, formatting the log files, and converting medical code into names more understandable to humans. User sessions were defined by log-in and log-out time. In the second and third stage, descriptive statistics were used to describe features of the logs; path analysis was used to identify the frequently visited pages. Association rule generation and sequential

pattern discovery were also employed to discover the frequency of use and the search patterns of the users. The result of data analysis indicated that users commonly view laboratory and radiology results in one session. Hence, the researcher suggested adding "shortcuts" in these Web pages to provide patients a quicker access to the information.

Graham and Keselman (Graham, Tse, & Keselman, 2006) researched the navigation patterns on a consumer health website (ClincialTrials.gov). One of their findings showed that many of the users like to use the *Back* button after viewing one page. Therefore, they suggested including more descriptive text or a site index all through the Web documents to encourage the users to explore lower level pages. Hence, users could reach the deeper site hierarchy and also reduce information-seeking episodes. Rozic-Hristovsk and Hristovski (Rozic-Hristovsk, Hristovski, & Todorovski, 2002) investigated the usage of the central medical library of their university by exploring weblog files. They found that the request amount of the website was increasing rapidly. The three main interests of the users are database, electronic journals, and site-search engines. Hence, they decided to increase the availability and stability of the database and also reconstruct more intuitive reference pages to fulfill the needs of the increasing number of visitors.

The studies reviewed demonstrated that applications of content analysis and Web usage mining are quite popular in recent research. In the e-commerce domain, some of the applications are very practical. However, in the health domain, applications are relatively immature. Most of the current efforts have focused on extracting users' usage patterns to better understand the users' navigational behavior, so that dynamic customization and decisions concerning site redesign or modification can then be made to provide better service to consumers.

Deficiencies in Past Literature

Although there is much research into "personalized websites," most of the research is focused on the common Web users instead of specific users. As Rozic-Hristovsk and Hristovski (Rozic-Hristovsk et al., 2002) stated in the limitation discussion in their study, "The analysis adequately reveals overall usage patterns but can only provide an estimation of individual user characteristics." Although it is hard to predict every single user's preference, it may be useful to divide users into groups based on their information needs and/or other characteristics, such as age, education, disease, and social role.

In past studies, there has been some research examining the difference in preferences of different types of groups, but most of the criteria were focused on the patient demographic information like age, gender and race. In the health area, patients and doctors are two distinct user groups, and they usually have different motives when searching the Internet. However, there is a lack in the literature of detailed studies and comparisons of the information needs for these two groups, especially in the same searching environment. Therefore, in this study, the different preferences of these two groups are going to be examined by comparing their Web queries and navigation patterns. For a more significant result, the raw data of patients and doctors came from the same website.

Website Research by User Groups

In past studies, researchers investigated health information seeking behavior from either a patient's or a physician's perspective. A study (Morita et al., 2007) of cancer patients' information needs showed that participating patients mostly want basic information such as general information about their disease and its symptoms, rather than every detailed stage of their illness and treatments. Conversely, another study (Gonzalez-Gonzalez et al., 2007) has shown that primary physicians would like to spend more time gathering information about the diagnosis and treatments. It is also assumed that there are different preferences among Web users when they seek information. For example, HON.ch ("Health On the Net Foundation,"), a European not-for-profit organization guiding both lay users and medical professionals to reliable digital sources of health information, provides different search options for patients, medical professionals, women, men, seniors, children and so on. For the same search term, searching results are different for different user groups. For example, if "heart failure" is searched, for patient groups the organization provides some consumer health links to websites like MedlinePlus, WebMD and Family Doctor. For health professionals, on the other hand, the results are more focused on professional peer reviewed articles from online journals and medical resources, like articles from *eMedicine*, which is an online clinical medical knowledgebase maintained by WebMD. It's clear that depending on their social roles, users have different information preferences.

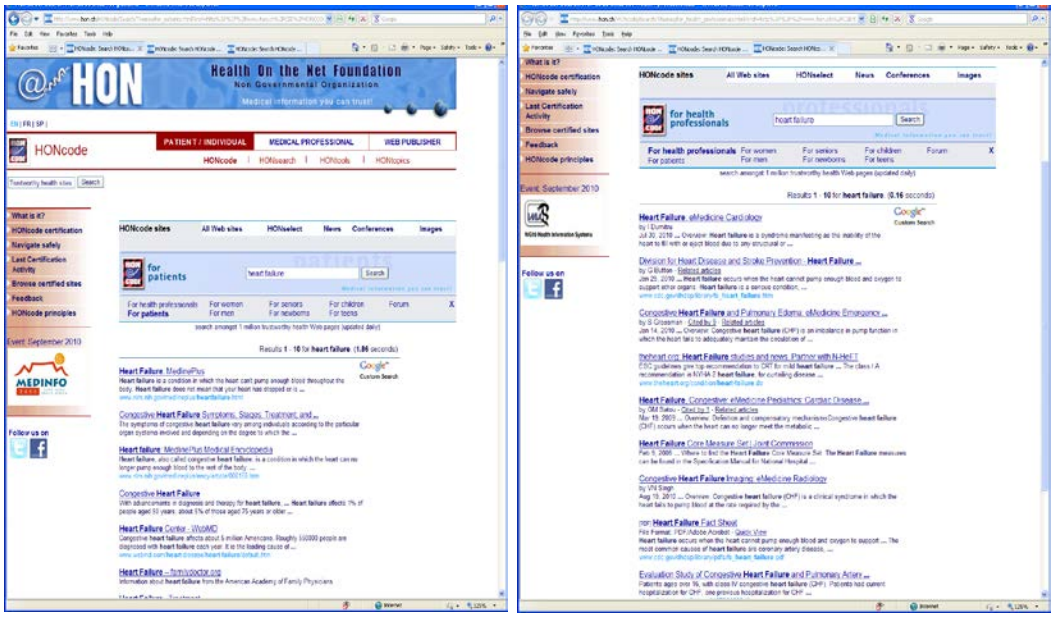


Figure 2. Different Search Results from HON.ch

Research Questions

The prerequisite to satisfy the users is to know what information the different types of users are looking for and how they search for it on the Web. Suggestions for website organization and or redesign will be culled from the results of this research.

The proposed study will answer the following research questions:

1. What are the topics of concern to users logged in as (1) physicians and (2) patients?
2. What needs to be altered on the website to match user information needs and search behavior?

CHAPTER THREE: METHODOLOGY

Summary of Methodology

For convenience, an existing health provider's website will be employed for this research. The website logs of Clarian Health ("Clarian Health,") were collected and used for analysis. Clarian Health, now renamed as IU Heath, was first formed in 1997. It is a private, nonprofit organization that owns more than 20 hospitals and health centers throughout Indiana. There are two main reasons to choose this website for the pilot study: First, the Clarian Health organization is large enough to have a sufficient number of users to provide sufficient and diverse data; second, the homepage is organized around two types of users, providers and patients, which will facilitate user group classification. The raw data is the daily access log for a five-month period in 2007.

Web mining technology was involved in the entire process, especially query analysis and Web usage mining. In addition, some tools and programming were employed to achieve certain goals. The detailed introduction of each tool will be given later. The main programming language is Perl, since this language provides powerful text processing facilities which are necessary to process the log files.

Data Collection

The study was based on five months' of daily access weblogs collected in 2007. The usage of this website is sufficient, and the website has already built up the navigation bars for patients, physicians and the visitors, which greatly facilitates the user classification. However, if one user searches for a term in the site search engine, no matter which group (s) he selects, the results are exactly the same. Figure 3 is a

screenshot of this website in 2009. The difference in the homepages between 2009 and 2007 will be discussed later.

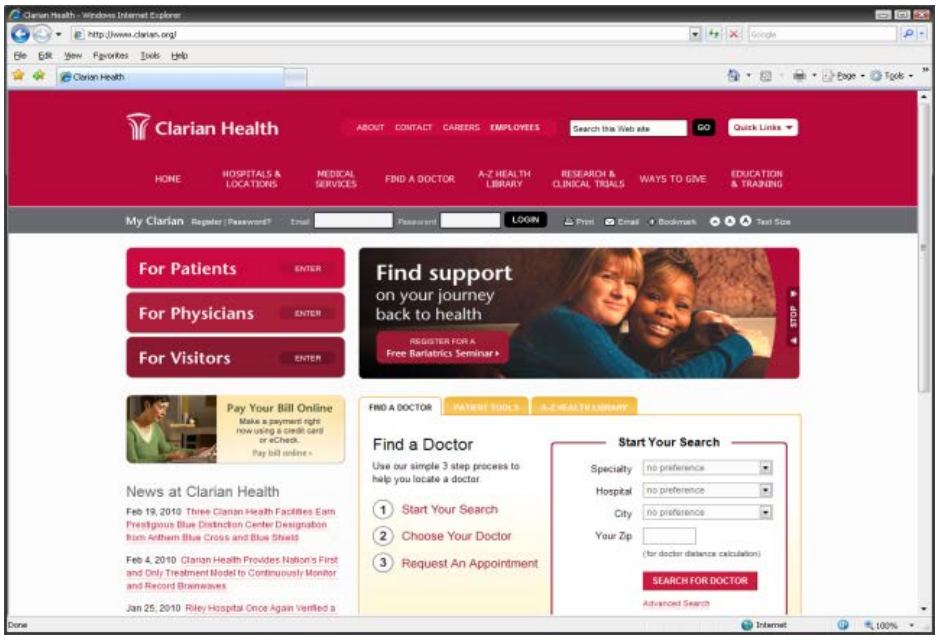


Figure 3. Screenshot of Clarian's website in 2009.

Data Preparation

Generally, preparing Web server log files for mining requires the following steps(Pohle, 2003):

1. Conversion of the log files into the suitable format

2. Removal of irrelevant requests and duplicate requests

3. Removal of robot requests

4. Definition of sessions

5. Application of specific data preparation according to project needs

In this research, all the steps above were involved. The last specific preparation was to identify the user groups and extract the query terms by programming. Raw log data were processed to reduce the noise according to the general process discussed above. First, the log file format was identified and the template was set up for programming. Below is an example of the log file and the regulated format we set up.

Example: 192.168.11.12 - - [28/Feb/2010:01:44:01 -0500] "GET http://www.google.com/search?q=anterior+spinal+fusion&hl=en&start=40&sa=N HTTP/1.1" 200 10603 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)" "-"

Format: IP/time stamp/method/path/protocol/status/sc_bytes/referrer/agent/cookie

Second, identification information of the users, Web spiders, and irrelevant and duplicate records were removed. Web spiders, also called Web robots or Web crawlers, are programs that automatically collect relevant content from Web pages, so the search queries generated by these spiders do not represent the actual information needs of the real users; these data needed to be removed before any analysis could be done.

And then user sessions were defined using cookies and a 30-minute time constraint. This time constraint is recommended by the tool used, and it is also employed in other research (Graham et al., 2006). The user groups of patients, doctors and visitors are separated by URL. If the user has clicked any one of the buttons shown on the front page as “patient,” “physician,” or “visitor,” the URL will clearly show it. For example, if one user clicks on “patient,” the URL will contain “/portal/patients/”. So since we know this pattern, the user’s role can be easily identified.

After the above cleaning process, the log files are ready for pattern analysis.

In order to further preparing for query analysis, the last step is to extract the search terms. From the example, it can be seen that the log file is semi-structured data. Although the format is regularly ordered, when users are searching from a search engine, the search terms or sentences they use are free text. These are the query terms that are going to be examined. In the above log file example, the free text is the "anterior+spinal+fusion." Figure 4 describes the whole process of data preparation.

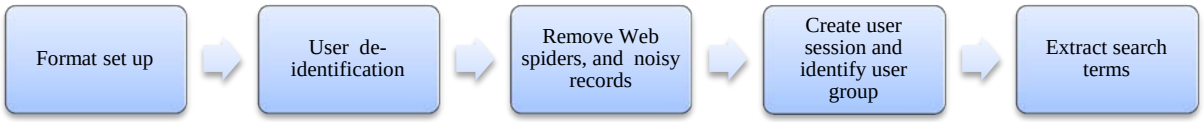


Figure 4. Query Data Preparation Process.

Table 2 provides a summary of the data obtained for analysis.

Table 2

Summary of Data Obtained for Analysis

Data for query analysis	**Data for pattern analysis**
Term with session number, for both patient and doctor groups	Cleaned Logs with session number , for both patient and doctor groups
Eg. 541137: 148532 anterior 541137: 148532 spinal	Eg. 541137:148533|66.231.189.55-- [31/Mar/2007:00:04:00-

541137: 148532 fusion	0400]"GET/portal/patients/registrationjsessionid=Q MQSANUR3TK3BLAQA5MSFEQ?paf_dm=full& paf_gm=content HTTP/1.0"200 192527"- ""Gigabot/2.0" "-"

Data Analysis

Descriptive statistics are used to generally describe the visit volume, search engine usage rate, and distribution. Except for the common data analysis and visualization tool Microsoft Excel, there are three other electronic tools used in the data analysis.

The first one is called Web Utilization Miner (WUM), which is a tool that aims to discover navigation patterns over the aggregated view of the web log (Spiliopoulou & Faulstich, 1998), to realize the pattern analysis. This tool is focused on the user pattern discovery by following the process shown in Figure 5.

Figure 5. WUM Process.

The second tool is WUM-prep, a Perl-based tool supporting data preparation for mining Web server log files. WUM-prep is also used as a primary tool to handle the data preparation part of the first tool. This tool is used mainly for data cleaning.

The last tool is called RapidMiner, which is a free tool to provide data mining and machine learning procedures involving many algorithms. This tool is employed to generate and compare classifiers. The detailed process will be introduced in the next section for better understanding.

CHAPTER FOUR: RESULTS & DISCUSSION

Log File Descriptive Statistics

Log File Volume

Figure 6 describes the volume statistics of the log files. Originally, there was a total of 11 million log records, but after the cleaning process, there were only 6.38 million (58%) left to be used for processing. So for this particular website, the Web spider covered more than half of the visit records, and the data generated by the spider need to be removed from the original data set so that the remaining data can be analyzed.

During the five months' time period of the study, April and May received the most visits. This may be for seasonal reasons or because of events that took place during those two months. In the future, the health provider could explore the data for these months and recall for whether there were certain activities in April and May that could account for the increased number of visits. That would make it possible for us to know if

there are any special system requirements needed to accommodate the increased visits during the spring or during certain events.

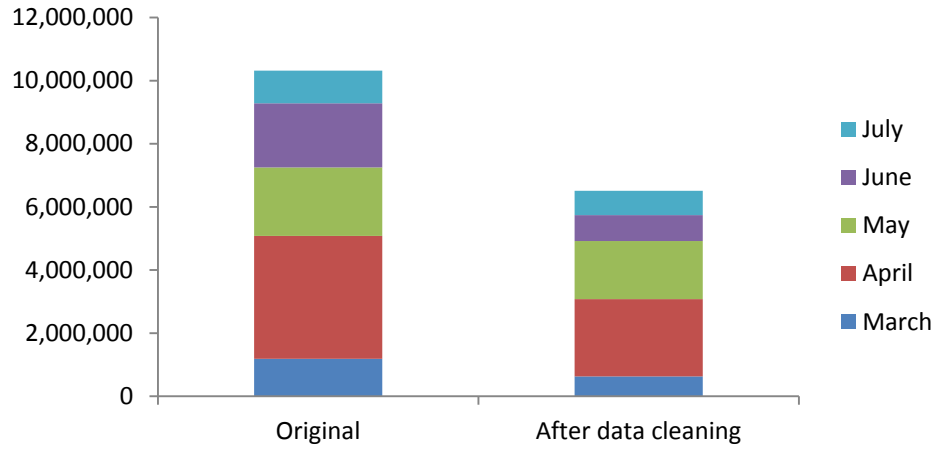


Figure 6. Log File Volume.

User Session

Figure 7 demonstrates the user sessions when users access the site through the homepage. Users logged as patients have around 200,000 user sessions, which is almost 10 times more than doctors. No user logged in as a visitor during the five-month period. Seventy-three percent of the users did not log in as any of the user groups when they were surfing the website. Although this website includes the log-in button, the majority of the users still didn't log in, so they might not have the special services provided based on user groups.

As we used the website from 2009 as a reference point, it was surprising to see that no users logged in as visitors during all five months. After reviewing the archives of the site it was found that the logs dated from 2007 and the choice *visitor* was not available at that time. Therefore, since no visitor portal was built, the data of the visitor

section should be zero. Although the 2007 homepage did not include a visitor button, it still had portals for patients and doctors.

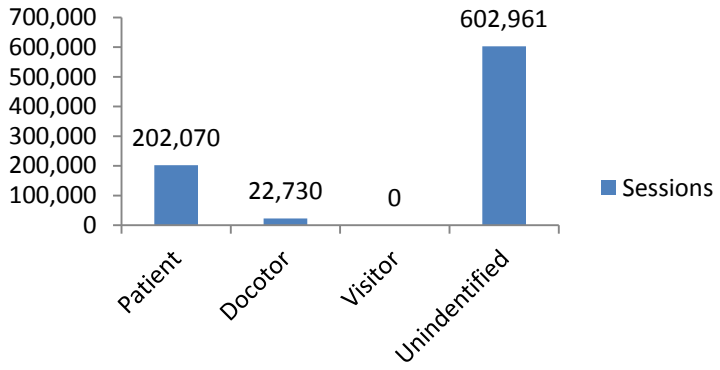


Figure 7. User Sessions.

Search Engine Distribution

Query requests were examined from the four most popular search engines: Google, Yahoo, MSN (which is now Bing), and Clarian's site search. The log file containing general search engine information indicates that the users searched terms in a general search engine and then were directed by the search result to Clarian's website. Table 3 gives the number of sessions for each search engine for each of the available portals

Table 3

Number of Sessions by Search Engine

	Google	**Yahoo**	**MSN**	**Clarian**
Patient	36486	7465	2659	39426
Doctor	2697	624	291	5292

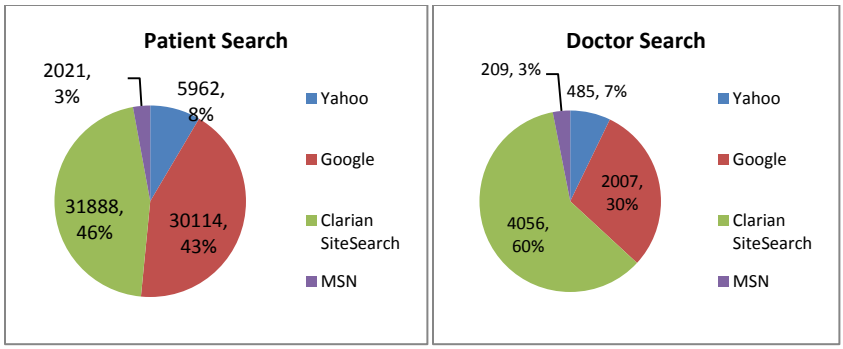


Figure 8. Search Engine Distribution.

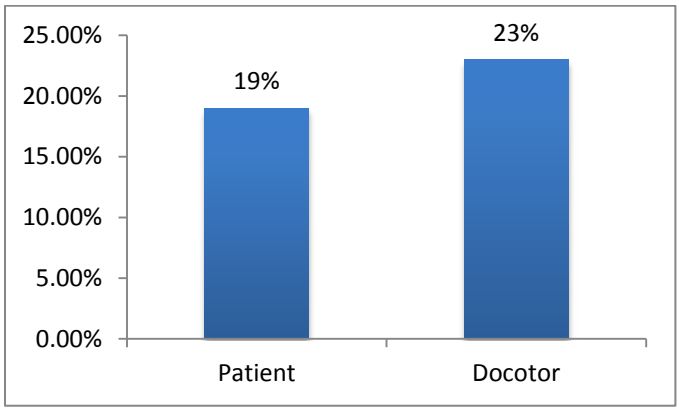


Figure 9. Site Search (Clarian Search) Usage Rate.

Figure 8 gives the search engine distribution chart for both the patient and doctor groups. The result shows that more than half of the search behavior was from general search and that google.com is the most popular search engine for users logged in as both patients or doctors. It can be seen that the intranet is well used through the Web and we recommend that this website could consider increasing the server's support ability and optimize the website to Google.

For the doctor group, 60% of the search behavior came from its intranet search engine (Clarian search), while for patients, this number is 40%. Also seen from Figure 9,

nearly 23% of the doctor sessions and 19% of patient sessions included searching in Clarian's site. The data show that doctors relied on the site search engines more than patients and that they searched more than patients. These data may indicate that when doctors were browsing this website, they usually were focused on one particular topic such as "breast cancer" rather than overall general ideas of "cancer information." It may be that because of time constraints related to clinical practice, doctors need to find the appropriate information quickly and are less tolerant of long lists of search results. Therefore it can be assumed that they searched Clarian's internet by preference, expecting that the site most likely has the information requested.

Query Terms Analysis

Single Term Analysis

The top 200 search terms from Clarian's website search engine were examined for patient and doctor groups; the first 50 terms are shown in Table 4.

For the patient group, it is surprising to see that many of the top terms are related to employment and education information, like "job," "employment," "class," and "program." In other words, users logged in as patients cared about jobs and training rather than health information. This finding may suggest that even if they logged in as patients, they are not the real patients, but rather job seekers, a category not provided on the site. Since job seekers had nowhere to directly access the information needed on the homepage, they clicked on the generic "patients" to start their search. As a consequence, it can be suggested that the homepage interface is not intuitive for these users.

Compared to the patient group, doctors used more medical terms to search, like "pulse," "cancer," "pathology," and "pain." Also, there are some words like "Dr." and names as "John" and "David." Looking through the original data revealed that the doctors like to search using the "Dr. + name" combination for more information about other doctors, like phone numbers or specialties. This may be another reason why the users logged as doctor when they browsed. Other words, like "careweb" and "cerner," that are shown in the result are some tools for Clarian's doctors to look for medical records or for resorting the knowledgebase. In summary, doctors were more likely to use this website as a handy tool to search auxiliary information, such as detailed doctor information, patient medical records, lab or surgery data, and to access the knowledgebase.

Table 4

Top 50 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
transplant	258	cerner	26
pulse	258	group	24
birth	238	md	23
surgery	234	director	22
pharmacy	228	medicine	22
lab	217	outlook	22
record	206	pediatric	22
education	204	iu	21
map	203	neurology	21
clinic	201	patient	21
number	194	women	20

radiology	185	service	20
phone	184	oncology	20
therapy	183	staff	20
nurse	181	directory	20
baby	175	test	18
application	171	order	18
volunteer	168	hospital	18
information	168	employee	18
riley	168	radiology	17
nursery	165	family	17
group	164	department	16
bill	161	david	16
child	153	employment	16
north	151	west	16
cancer	151	library	16
cpr	147	record	15
directory	146	program	15
student	146	scott	14
life	142	web	14

Association Discovery

When users are searching information, they sometimes use several words or a short sentence as key words. Also seen from the above, “human” and “resource” both appeared as high frequency search terms. Thus, it may be predicted that “human resource” is the actual search query. In order to investigate more of the search terms, the phrases which the users searched were examined in this step. Phrases are defined as two or more terms used together. Identifying these phrases would give us a more detailed look at the users’ information-seeking terms. Another reason for finding these associations among terms is to give dynamic search suggestions based on the associations. When there is a high confidence in certain term associations, the Web builder can give dynamic search suggestions based on these findings. In this way, users may be suggested to use a more precise search query that would more likely result in higher precision and recall.

For search topic discovery, the result is similar to the findings above for single terms. Table 5 describes the top 20 combinations with the number of times they were used.

Table 5

Top 20 Phases for Patient and Doctor Group

Patient	**Doctor**
HUMAN RESOURCES,322	MEDICAL GROUP,16
MEDICAL RECORDS,136	METHODIST MEDICAL,16
METHODIST MEDICAL,107	WOMEN'S HEALTH,14
CHILD LIFE,93	ORDER SETS,12

CLARIAN WEST,90	CLARIAN WEST,12
CLARIAN NORTH,90	FAMILY PRACTICE,11
OCCUPATIONAL HEALTH,81	MEDICAL RECORDS,10
PHONE NUMBER,86	MIKE DENTON,10
METHODIST HOSPITAL,81	MULTIPLE SCLEROSIS,9
STUDENT NURSE,76	CLARIAN NORTH,9
PATIENT INFORMATION,73	PULSE PAGE,9
MEDICAL GROUP,74	METHODIST HOSPITAL,9
CLARIAN HEALTH,71	SPEECH PATHOLOGY,8
PHYSICAL THERAPY,72	COLEMAN CENTER,8
EMPLOYMENT OPPORTUNITIES,56	PATHOLOGY LABORATORY,8
DAN EVANS,50	IU MEDICAL,8
CARE CENTER,50	INFECTIOUS DISEASE,7
METHODIST GROUP,50	MEDICAL LIBRARY,7
EMPLOYEE HEALTH,47	IU GROUP,7
METHODIST HEALTH,49	METHODIST GROUP,7

In addition to the methods of analysis discussed above, market basket analysis (Agrawal, Imieli\, \#324, ski, & Swami, 1993) was employed to find some associations of the popular search terms. Market basket analysis (Liu, 2007) is a modeling technique based on the theory that if one person buys a certain group of items, (s)he is more or less likely to buy a certain other group of items. It provides an insight into customer behavior based on observations their buying habits. In query analysis, the individual search terms

can be seen as items, and thus the whole search term, a phrase, could be predicted based on the existing observations. If Web builders know which terms users are most likely to search together, they could provide dynamic site search suggestions to users. Therefore, the users could get an idea of what popular words other people are likely to search and thus get a more precise keyword. For example, “people mover,” which is a transportation vehicle between Clarian hospitals, is a popular high-frequency phrase. Yet if a user doesn’t know the name “people mover,” and instead searches for “move,” that user will probably not get the desired result. In this case, if the Web could dynamically suggest “people mover” as a key word, the user may get a quick result with good precision.

Table 6 lists some of the associations found with a confidence rate of 50% or better. The association finding process is done using Microsoft Excel with the data mining plug-in ("Data Mining Add-ins," 2011). The Microsoft SQL Server Data Mining Add-in for Microsoft Office provides a tool to derive patterns and trends that exist in complex data and visualize those patterns in charts ("Data Mining Add-ins," 2011). The included market basket analysis function is employed in this section to analyze search transactions quickly and identify search combinations. The strength of association between terms is calculated using a statistical measure called the Confidence rate (Liu, 2007). This measure represents the percentage of searches which contain term1 and also contain term 2. It can be seen as an estimate of the conditional probability, Pr(Term2 | Term1). The higher the rate, the higher the reliability to predict term 2 from term1. The confidence rate is computed by the following equation (Liu, 2007):

$$\text{Confidence} = \frac{(\text{Term1} \cup \text{Term2}).\text{count}}{Term1.count} * 100\%$$

Table 6

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%			

Topic Classification

General Topics

As the top search terms and phrases of both user groups are examined, the next step is to attempt a classification of the topics. This classification makes it easier to see which topics the patient and doctor groups have in common and into which categories the interests of each individual group fall. General topic classification of both groups is done manually, without using a tool, based on the search terms. The results are listed in Table 7.

The patient group has four categories: employment, medical record, general information, and education. The doctor group has three categories: doctor's detail information, supplemental information, and general hospital information. The results show that two topics, "general information of hospitals" and "medical record," are common to both groups. Because these topics are important to both groups, the organization should pay attention to this point so that the website builder can include improvements in the amount and accessibility of information.

Table 7

General Topics for Patient and Doctor

Patient	**Doctor**
Employment information	Doctor's detail information
Medical record	Supplemental information, like *medical record*, lab result, radiology result
General information of hospitals, like address, telephone, services	General information of hospitals
Education information, like intern, program， CPR class, CAN class	

Health Topics

As for health-related information, the top two concerns for patients are children's health and occupational health. Other frequently searched health topics are physical therapy, liver transplant, kidney transplant, urgent care, weight loss, poison control and

sports medicine. It was unexpected to see that occupational health appeared as the second most popular health topic in patient groups, which is unusual because occupational health is generally not a popular topic. The high number of searches indicates that people are not easily finding the information they need, so it may be concluded that there is a lack of information about occupational health in this website. Actually, among all the health topics, health providers usually pay less attention to occupational health in primary healthcare, so their websites have less information about it. However, patient groups are interested in this topic. Since they find it hard to get the information easily, patients searched "Occupational health" for more details. The Clarian professionals should pay attention to this point.

For doctors, the top two topics of concern are cancer and women's health. Other topics searched are multiple sclerosis, speech pathology, pulse page and infectious disease. There is a possible reason for the top search being cancer. That is because the IU Simon Cancer Center, which is a hospital belonging to the Clarian Health organization, is the only National Cancer Institute (NCI)-designated cancer center with such distinction in Indiana that provides patient care. Therefore, doctors may prefer this website to look for cancer information.

Another point worth stressing is that, from all the hospitals organized by Clarian Health, Methodist, Clarian West, and Clarian North are the top three hospitals both patients and doctors searched. This is another finding the Clarian professionals should pay attention to so that the website builder can check to see if the accessibility of these sites is sufficient.

In summary, from the comparison it can be seen that patients and doctors searched for different health topics. In addition, patients searched more about general health problems, and they preferred to use consumer terms. On the other hand, doctors searched with more professional terminology, and they cared about more special health problems.

Table 8

Health Topics for Patient and Doctor

Patient	**Doctor**
Children' health *Occupational health*	*Cancer* *Women's health*
physical therapy, liver transplant, kidney transplant, urgent care, weight loss, poison control, sports medicine	family practice, multiple sclerosis, speech pathology, pulse page , infectious disease Breast cancer, severe, acute

User Group Classification

Solutions for Classification

As seen previously, 73% of the users did not log in as any group when browsing, so no tailored service was available to them. Two solutions are suggested to solve this problem.

The first solution is to build a pop-up hint page once the users access the entry page. The pop-up will remind them to log in with the group information or “force” them to log in. The advantages of this solution are it is easy to do and it is complete. The solution does not require complex technology, yet it can classify all the Web users. However, the pop-up may be annoying to some users and thus make them lose interest in this website. Another solution is to build an automatic classifier based on the data mining classification technology. The classifier can automatically identify the user role based upon search terms that users input. This solution may be more favorable among users, but it can only auto-classify part of the users for whom it searched. With either of these two solutions, a majority of the users can still get a benefit even they don’t log in.

For the second solution, two popular classifiers, naïve Bayesian and Support Vector Machine (SVM) were tested and compared (Sholom Weiss, 2004). The naïve Bayesian classifier is a simple probabilistic classifier based on applying Bayes' theorem with independence assumptions. It assumes the presence of a particular feature based on the statistics of the presence of other related features, and gives classification. The SVM classifier is a binary-liner classifier. A linear classifier makes a classification decision based on the value of a linear combination of the characteristics. In this study, these two classifiers were tested based on 600 patient queries and 600 doctor queries randomly selected from the five months of data. The reason to select the same number of query terms is to reduce the negative performance effect of the in-balanced sample. This process is done by RapidMiner (Miner, 2011), which is a free Java-based software that provides data mining and machine learning procedures, including data loading,

transformation , preprocessing, modeling and visualization. For this research, the main function of this software that was used is classifier generation and performance testing.

For each classifier, general text mining process (Sholom Weiss, 2004) was followed to do the tokenization (breaking a stream of text in meaningful words or symbols) and remove stop words (words that do not contain important meanings, such as *the*, *after*, or *a)*. Each classifier was generated and tested by the cross-validation function. Cross-validation indicates that the software will use sufficient data to generate a classifier and then use the rest of the data to generate the testing. Figure 10 shows the whole process.

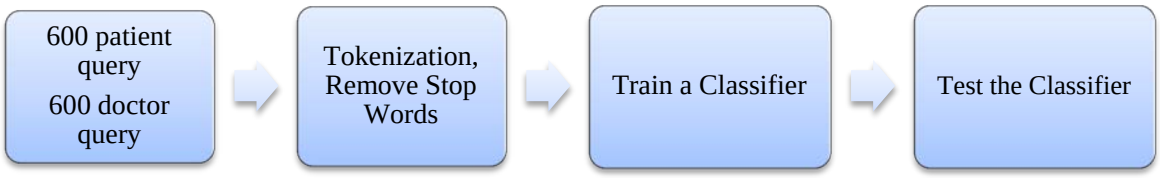


Figure 10. User Group Classification Process

As a result, the SVM classifier has a better F-score than naïve Bayesian. With this classifier, it is possible to categorize the users. So when people search information, they can receive suggestions or be directed according to their user roles, no matter whether they are logged in or not.

Table 9

Performance of Bayes Classifier

		Patient **(F = 71.13%)**	**Doctor** **(F = 44.06%)**	
Predict	Patient	True Positive (TP) = 563	False Positive (FP) = 420	Precision= TP / (TP +FP) =57.27%
	Doctor	False Negative (FN) = 37	True Negative (TN) = 180	Precision = TN / (FN + TN) = 82.95%
		Recall: = TP / (TP + FN) =93.83**%**	Recall: = TN / (FP + TN) = 30.00%	

Table 10

Performance of SVM Classifier

		Patient **(F = 78.25%)**	**Doctor** **(F = 80.19%)**	
Predict	Patient	True Positive (TP) = 447	False Positive (FP) = 96	Precision =82.32%
	Doctor	False Negative (FN) = 153	True Negative (TN) = 504	Precision = 76.71%
		Recall: = 74.50%	Recall: = 84.00%	

From the comparison in Figure 11, it can be seen that although the patient recall rate and doctor precision rate of the Bayesian classifier are a little bit higher than that of the SVM classifier, the overall performance, the F-Score, of SVM is much better than the Bayes classifier, especially for doctor prediction. This is because the SVM classifier greatly reduces the FP number in the experiment. Therefore, SVM should be chosen as the primary classifier to do classification for this website.

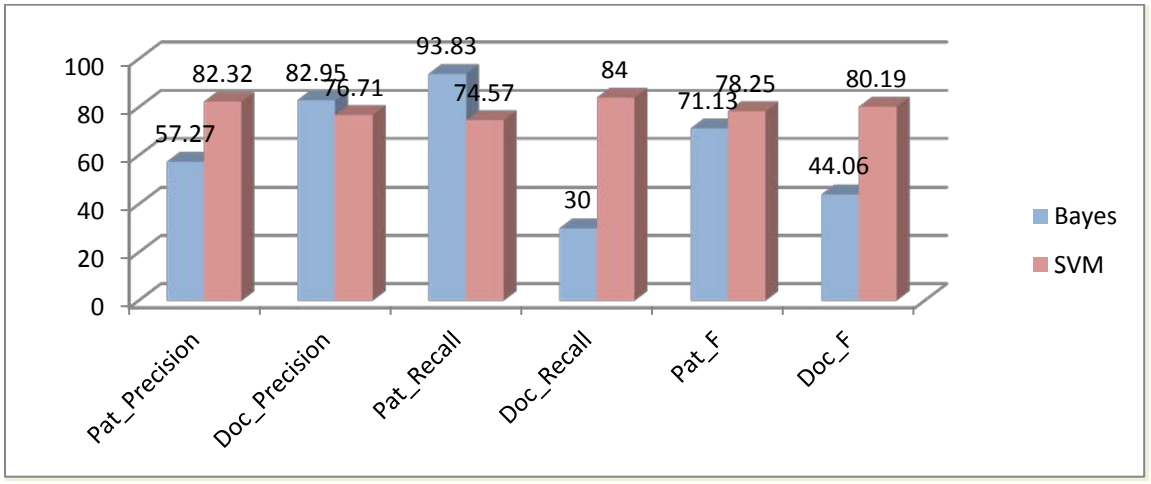


Figure 11. Comparison of SVM and Bayes Classifier.

Test of SVM Classifier for Unidentified Users

For future application of the SVM classifier, a deeper investment was done to test the classifier with the unidentified users. The study was done mainly based on the user's IP address.

In the first step, all the IP addresses were collected from identified patient and doctor groups. Then, sessions with the same IP but without identifying information were collected. After extracting the query terms from these sessions, the SVM classifier was applied to test the performance. Figure 12 shows the whole process.

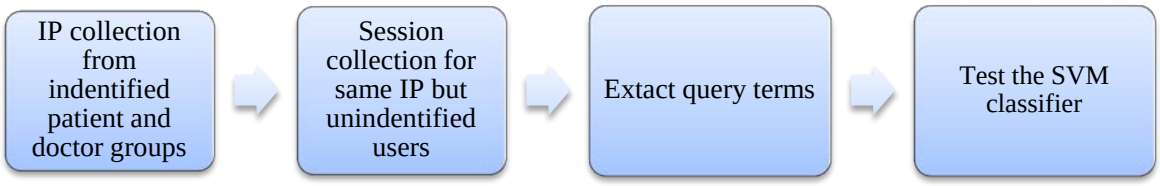


Figure 12. Process to apply SVM Classifier to Unidentified users.

Similar to identified-user study, the SVM classifier was also tested based on 600 patient queries and 600 doctor queries randomly selected from the five months of data with RapidMiner software. Table 11 shows the performance of the SVM classifier to unidentified users.

Figure 13 shows the performance of the SVM classifier for the unidentified users compared to the identified ones. It can be clearly seen that the F-score of unidentified users is slightly lower than that of the identified ones. However, the performance is still better than the Bayesian classifier and acceptable.

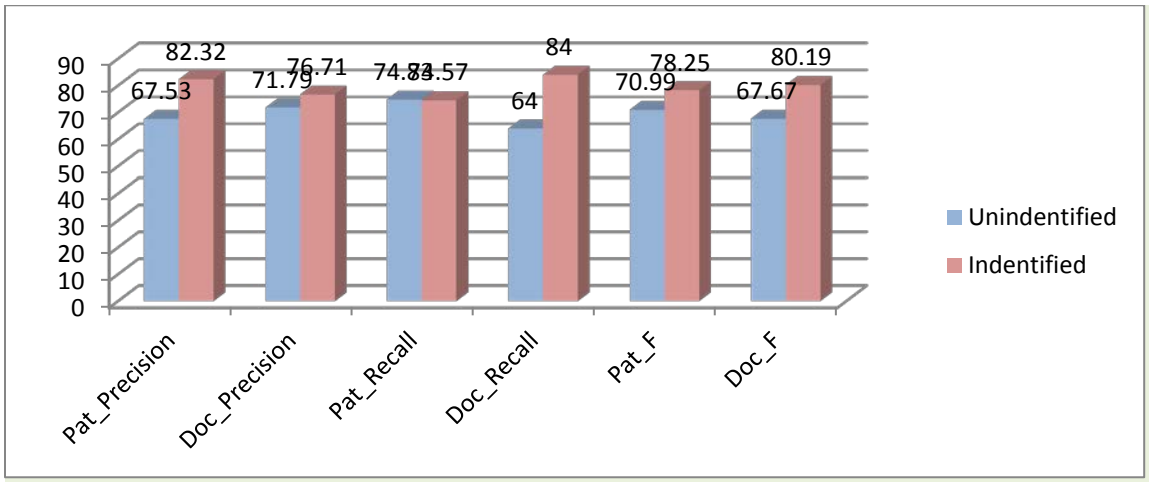


Figure 13. Performance Comparison of Unidentified and Identified users for SVM.

From the number in Table 11, it can be seen that the increased false positive number mainly contributes to the decrease of the F-score. One thing that needs to be stressed is that the precondition made for the test was the assumption that the users with the same IP address belonged to the same user groups. However, this assumption is not precise because different users may have the same IP address. For example, the people from the same company may share a same IP address, and there could be hundreds of employees in one company who could visit the website as either a patient, a doctor or a visitor. Therefore, the variation of the users from the same IP address will lead to an increase in false positives and thus decrease the performance.

Table 11

Performance of SVM Classifier to Unidentified users

		Patient **(F = 70.99%)**	**Doctor** **(F = 67.67%)**	
Predict	Patient	True Positive (TP) = 449	False Positive (FP) = 216	Precision =67.53%
	Doctor	False Negative (FN) = 151	True Negative (TN) = 384	Precision = 71.79%
		Recall: = 74.83%	Recall: = 64.00%	

User Pattern Analysis

The last analysis was to discover the navigation pattern. Discovering users' navigation patterns is the fundamental approach for generating recommendations. Knowing how the patients or doctors locate a page is very important to optimize the contents and structure for them. As the amount of data is huge and would consume a lot of time to generate the result, only one-month of log data was used to analyze the pattern.

Figures 14 and 15 visualize the click streams for both users. But because the website had already been changed, discovered links could not be reproduced to see the exact pages. However, some trends can still be seen clearly by comparing these two figures. The number at the end of each path indicates how many users get back to this page.

The pattern of patients is relatively longer and denser than that of doctors. Patients are more aimless and tolerant than doctors; they often have a longer pathway and are more likely to return to the page again and again. In contrast, the doctors' pattern is cleaner and shorter. They surf this website more intentionally, always going directly to the final page in the shortest way and then leaving without going back.

In sum, this pattern provides some evidence for the prediction made above: Doctors are more likely to go to target topics directly. So to them, the shorter pathway seems to be more efficient. The patients, however, spend more time than doctors to browse the website, and they receive more general information than results about one particular topic. Compared to the doctors, they seem to have more tolerance for longer

lists and are more receptive to generic and comprehensive information rather than specific information.

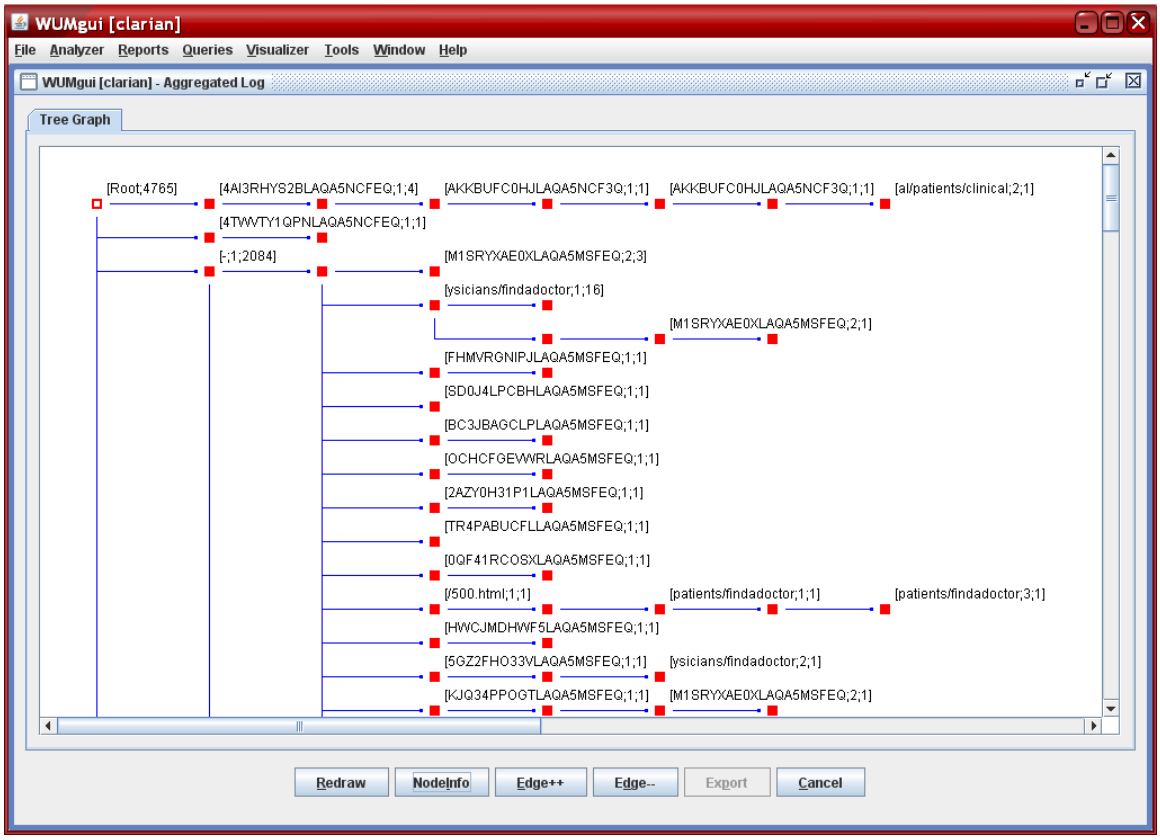


Figure 14. Pattern for Doctors.

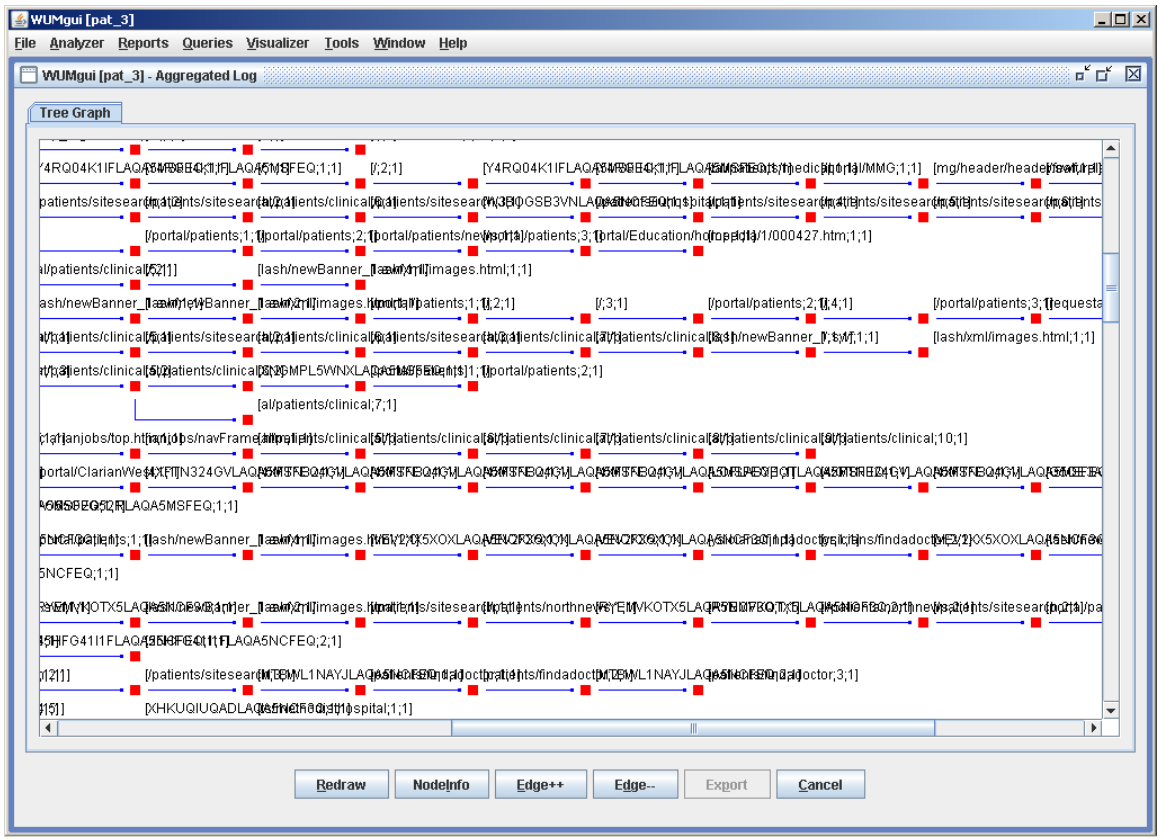


Figure 15. Pattern for Patients.

CHAPTER FIVE: CONCLUSION

Strengths and Limitations

The results of the study provide evidence in the form of quantitative data with which to compare and contrast the searching behavior of patients and doctors on the same website. Findings are meaningful not only for the pilot website but also for constructing other health websites. Results show that patients have more tolerance when browsing the website; they get back to the previous page more often. Conversely, doctors usually go to the final page directly and then leave the page without returning to it. Knowing this could not only help the Web builder to restructure this website, but also provide a fundamental clue to other website and Web application developers to establish long-term user profiles for either the patient or doctor user group. The findings could make up for the lack of detailed studies and comparison of the information needs of patient and doctor groups in the literature.

Other important findings indicate that patients are not searching health information as much as expected, since a relatively small proportion of their searching on Clarian's website is medical or health related. The most popular topic of patients' searches is employment- related issues, which is a good hint for the website builders to consider reorganizing the user category. As for the health issue, the result reminds health providers to check whether occupational health information is deficient among general health issues on their websites. And using the same method, other website builders could be able to check the comprehensiveness and accessibility for other health subjects.

Although it is shown that patients and doctors have different preferences for health topics and terms, they both include “medical record” as an important search term. This shows that “medical record” is a common health topic in demand by both patients and doctors. The necessity to build an easy-to-access medical record portal for health websites is conspicuous.

Besides the overall findings for health website builders in general, the results suggest some detailed suggestions for reconstructing Clarian’s particular website. The following recommendations aim for a more user-friendly interface for different users.

- For the homepage, build a log portal for employment seekers, like “employer” or “future employee.”

- Differentiate the entry pages for different user groups. For the patient group, build friendly links to training, education programs and general information. For the doctor group, build intuitive links to doctor contact directory, knowledgebase, and auxiliary medical data access.

- As a majority of the users did not log in as either a patient or doctor, it is suggested to build a direction service for this website, like a pop-up page to lead the users, or to implement the SVM classifier to auto classify the users.

- For search engines, increase the server support to Google and provide dynamic searching suggestions in the site search engine to facilitate the search criteria.

Because this study represents only the users’ seeking pattern from one website, the results can only be used as an estimate for other health websites. The user separation

is based on the log-in information, so the user groups of patients or doctors may not be the real patients and doctors. As pointed out previously, the users logged in as patients may be some employment seekers who are just looking at this website for jobs.

Future Study

Future studies may be more focused on the navigation pattern of the different groups, such as what path is used to find the same topic, and are there any wasted steps in the process to get to the final page.

Other small topics could involve general search engine and volume discovery. General search engine discovery is to see which term is the top direction from each general search engine, so more specific detailed information can be better provided. Volume discovery is to examine the search terms in the highest volume months, April and May, so would be possible to know whether the users have special needs related to the season or to events.

REFERENCE

Aghabozorgi, S. R., & Wah, T. Y. (2009). *Recommender systems: Incremental clustering on web log data*. Paper presented at the Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human, Seoul, Korea.

Agrawal, R., Imieli, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases*. Paper presented at the Proceedings of the 1993 ACM SIGMOD international conference on Management of data, Washington, D.C., United States.

Ayantunde, A. A., Welch, N. T., & Parsons, S. L. (2007). A survey of patient satisfaction and use of the Internet for health information. *Int J Clin Pract, 61*(3), 458-462. doi: IJCP1094 [pii]10.1111/j.1742-1241.2006.01094.x

Chang, P., Hou, I. C., Hsu, C. L., & Lai, H. F. (2006). Are Google or Yahoo a good portal for getting quality healthcare web information? *AMIA Annu Symp Proc*, 878. doi: 86564 [pii]

Chen, E. S., & Cimino, J. J. (2003). Automated discovery of patient-specific clinician information needs using clinical information system log files. *AMIA Annu Symp Proc*, 145-149. doi: D030002693 [pii]

Chiu, Y. C. (2011). Probing, impelling, but not offending doctors: The role of the internet as an information source for patients' interactions with doctors. *Qual Health Res*. doi: 1049732311417455 [pii]10.1177/1049732311417455.

Clarian Health. from www.clarian.org

Cooley, R., B. Mobasher, & Srivastava, J. (1999). Data preparation for mining World Wide Web browsing patterns. *Journal of Knowledge and Information Systems*.

Data Mining Add-ins. (2011), from http://office.microsoft.com/en-us/excel-help/data-mining-add-ins-HA010342915.aspx

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for web personalization. *ACM Trans. Internet Technol., 3*(1), 1-27. doi: http://doi.acm.org/10.1145/643477.643478

Eysenbach, G. (2003). The impact of the Internet on cancer outcomes. *CA Cancer J Clin, 53*(6), 356-371.

Eysenbach, G., & Köhler, C. (2002). 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.

Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng., 53*(3), 225-241. doi: http://dx.doi.org/10.1016/j.datak.2004.08.001

Fernandez-Luque, L., Karlsen, R., & Bonander, J. (2011). Review of extracting information from the social web for health personalization. [Research Support, non-U.S. Gov't review]. *J Med Internet Res*, *13*(1), e15. doi: 10.2196/jmir.1432

Fox, S. (2005). Health Information Online. *PEW Internet & American life project*. Retrieved from http://www.pewinternet.org/pdfs/PIP_Healthtopics_May05.pdf

Fox, S., & Fallows, D. (July 16, 2003). Internet health resources. *Internet & American life project* Retrieved August 30, 2003, 2003, from http://www.pewinternet.org/

Gonzalez-Gonzalez, A. I., Dawes, M., Sanchez-Mateos, J., Riesgo-Fuertes, R., Escortell-Mayor, E., Sanz-Cuesta, T., & Hernandez-Fernandez, T. (2007). Information needs and information-seeking behavior of primary care physicians. *Ann Fam Med, 5*(4), 345-352. doi: 5/4/345 [pii]10.1370/afm.681

Graham, L., Tse, T., & Keselman, A. (2006). Exploring user navigation during online health information seeking. *AMIA Annu Symp Proc*, 299-303. doi: 86156 [pii]

Greenberg, L., D'Andrea, G., & Lorence, D. (2004). Setting the public agenda for online health search: a white paper and action agenda. *J Med Internet Res, 6*(2), e18. doi: v6e18 [pii]10.2196/jmir.6.2.e18 [doi]

Hallingbye, T., & Serafini, M. (2011). Assessment of the Quality of Postherpetic Neuralgia Treatment Information on the Internet. *J Pain*. doi: S1526-5900(11)00628-6 [pii] 10.1016/j.jpain.2011.05.005.

Health On the Net Foundation. from http://www.hon.ch/

Hoeber, O. (2008). *Web information retrieval support systems: The future of web search*. Paper presented at the Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 03.

Kosala, R., & Blockeel, H. (2000). Web mining research: a survey. *SIGKDD Explor. Newsl., 2*(1), 1-15. doi: http://doi.acm.org/10.1145/360402.360406

Lambert, S. D., & Loiselle, C. G. (2007). Health information seeking behavior. *Qual Health Res, 17*(8), 1006-1019. doi: 17/8/1006 [pii] 10.1177/1049732307305199

Lawrentschuk, N., Abouassaly, R., Hackett, N., Groll, R., & Fleshner, N. E. (2009). Health information quality on the internet in urological oncology: a multilingual longitudinal evaluation. *Urology, 74*(5), 1058-1063. doi: S0090-4295(09)00874-7 [pii]. 10.1016/j.urology.2009.05.091

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. New York, NY: Springer.

Rapid Miner, R. (2011), from http://en.wikipedia.org/wiki/RapidMiner

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on Web usage mining. *Commun. ACM, 43*(8), 142-151. doi: http://doi.acm.org/10.1145/345124.345169

Morahan-Martin, J. M. (2004). How internet users find, evaluate, and use online health information: A cross-cultural review. *Cyberpsychol Behav, 7*(5), 497-510.

Morita, T., Narimatsu, H., Matsumura, T., Kodama, Y., Hori, A., Kishi, Y., . . . Kami, M. (2007). A study of cancer information for cancer patients on the internet. *Int J Clin Oncol, 12*(6), 440-447. doi: 10.1007/s10147-007-0707-5 [doi]

Pierrakos, D., Paliouras, G., Papatheodorou, C., & Spyropoulos, C. D. (2003). Web usage mining as a tool for personalization: A survey. *User Modeling and User-Adapted Interaction, 13*(4), 311-372. doi: http://dx.doi.org/10.1023/A:1026238916441

Pohle, C. (2003). WUMprep-logfile preparation for data mining with WUM.

Rice, R. E. (2006). Influences, usage, and outcomes of Internet health information searching: Multivariate results from the Pew surveys. *International Journal of Medical Informatics, 75*(1), 8-28.

Risk, A., & Dzenowagis, J. (2001). Review of internet health information quality initiatives. *J Med Internet Res, 3*(4), E28. doi: 10.2196/jmir.3.4.e28 [doi]

Rozic-Hristovsk, A., Hristovski, D., & Todorovski, L. (2002). Users' information-seeking behavior on a medical library Website. *J Med Libr Assoc, 90*(2), 210-217.

Scherer, M., Zitterbart, S., Mildenstein, K., & Himmel, W. (2010). [What questions do headache patients pose in the internet? Content analysis of an internet expert forum]. *Gesundheitswesen, 72*(5), e28-32. doi: 10.1055/s-0029-1234128

Sholom Weiss, N. I., Tong Zhang, Fred Damerau. (2004). *Text mining: Predictive methods for analyzing unstructured information.* New York, NY: Springer.

Shuyler, K. S., & Knight, K. M. (2003). What are patients seeking when they turn to the Internet? Qualitative content analysis of questions asked by visitors to an orthopaedics Web site. [Evaluation Studies Research Support, Non-U.S. Gov't]. *J Med Internet Res, 5*(4), e24. doi: 10.2196/jmir.5.4.e24

Spiliopoulou, M., & Faulstich, L. C. (1998). WUM: A Web Utilization Miner

Spink, A., Yang, Y., Jansen, J., Nykanen, P., Lorence, D. P., Ozmutlu, S., & Ozmutlu, H. C. (2004). A study of medical and health queries to web search engines. *Health Info Libr J, 21*(1), 44-51. doi: 10.1111/j.1471-1842.2004.00481.x HIR481 [pii]

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl., 1*(2), 12-23. doi: http://doi.acm.org/10.1145/846183.846188

Stavri, P. Z. (2001). Personal health information-seeking: a qualitative review of the literature. *Stud Health Technol Inform, 84*(Pt 2), 1484-1488.

Tangri, V., & Chande, N. (2011). Quality of Internet-based information on gastrointestinal diseases. *Can J Gastroenterol, 25*(2), 93-96.

Trotter, M. I., & Morgan, D. W. (2008). 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. doi: 14/3/175 [pii] 10.1177/1081180X08092828

APPENDICES

Appendix A: Data Preparation Scripts

The following script was used to separate patient and doctor groups by URL.

```
#!/usr/bin/perl

$LINE_STEP = 100; #Define after processing how many lines the program should give progress report
```

```
$line_count = 0; #Count current processing line number

open (INFILE, "log2007.nobots.log.clean.sess")|| die ("cannot open  input file");

#Read whole data file line by line, and mark type of sessions
while ($line = <INFILE>){
        chop ($line);      #Remove \n as last character
        @log = split(/[(|)("*")]/, $line);#change the log files into an array
        @session = split(/\:/, $log[0]); #retrieve session ID
                #following if check if the URL contains any login information
        if ($log[2]=~m!/portal/patients!) {
    $patientLOG[$session[1]] = 1;
        }
        if ($log[2]=~m!/portal/physician!) {
    $physicianLOG[$session[1]] = 1;
        }
        if ($log[2]=~m!/portal/visitor!) {
    $visitorLOG[$session[1]] = 1;
        }
        $line_count++;  # Update line number count
        if (0 == $line_count % $LINE_STEP) {
                print "-$line_count- \n";
        }
} #end while
print "\n";
close (INFILE);

print "==================Finish the first read.================\n";
open (LOG, "log2007.nobots.log.clean.sess")|| die ("cannot open  input file");
```

```
open(PATIENT, ">patient_final.txt") || die ("cannot open output file");
open(DOCTOR, ">doctor_final.txt") || die ("cannot open output file");

open(VISITOR, ">visitor_final.txt") || die ("cannot open output file");

$line_count = 0; #Count current processing line number
while($line = <LOG>){
        @log = split(/\|/, $line);  #change the log files into an array to pick up the first number
        @session = split(/\:/, $log[0]); #retrieve session ID

   if ($patientLOG[$session[1]]) {
            print PATIENT ($line);
   }
   if ($physicianLOG[$session[1]]) {
            print DOCTOR ($line);
   }
   if ($visitorLOG[$session[1]]) {
            print VISITOR ($line);
   }
        $line_count++;  # Update line number count
        if (0 == $line_count % $LINE_STEP) {
                print "-$line_count- \n";
        }
}
print "\n Finished second read.\n";

close (LOG);
close (PATIENT);
close (DOCTOR);
```

```
close (VISITOR);
```

The following script was used to extract the query terms from the four search engines.

```
#!/usr/bin/perl

open (LOG, "doctor_final.txt")|| die ("cannot open  input file");
open(GOOGLE, ">google_d_fianl.txt") || die ("cannot open output file");
open(MSN, ">msn_d_final.txt") || die ("cannot open output file");
open(YAHOO, ">yahoo_d_final.txt") || die ("cannot open output file");
open(CLARIAN, ">clarian_d_final.txt") || die ("cannot open output file");
$google_query=0;#count how many queries were from google
$msn_query=0;#count how many queries were from msn
$yahoo_query=0;#count how many queries were from yahoo
$clarian_query=0;#count how many queries were from clarian

while ($line=<LOG>){
        chop ($line);#remove the \n
        @log = split(/[(|)("*")]/,$line);#change the log files into an array
        #the following codes finds the queries in google, yahoo, MSN and clarian's search engine
        if ($log[4]=~ /google(\S+?)q=([^&]+)/) {   #match contains google and q=,ends with &
    $google_search=$2;
     print GOOGLE ("$log[0] $google_search \n");#google $log[0]$log[1]
     $google_query++;
        } #end if google
        if ($log[4]=~ /yahoo(\S+?)p=([^&]+)/) {   #match contains yahoo and p=,ends with &
    $yahoo_search=$2;
     print YAHOO ("$log[0] $yahoo_search \n");#yahoo $log[0]
     $yahoo_query++;
```

```
                } #end if yahoo
        if ($log[4]=~ /msn(\S+?)q=([^&]+)/) {    #match contains msn and q=,ends with &
    $msn_search=$2;
     print MSN ("$log[0] $msn_search \n");#msn $log[0]
     $msn_query++;
        } #end if msn
                if ($log[2]=~ /sitesearch(\S+?)query=([^ ]+)/) {    #match contains sitesearch and
query=,ends with &

    $clarian_search=$2;
          print CLARIAN ("$log[0] $clarian_search \n");#clarian $log[0]$log[1]
          $clarian_query++;

        } #end if clarian
        #the following codes print the array by index
}#end while
print "Google query is $google_query in total.\n";
print "Yahoo query is $yahoo_query in total.\n";
print "MSN query is $msn_query in total.\n";
print "Clarian query is $clarian_query in total.\n";
close (LOG);
close (GOOGLE);
close (MSN);
close (YAHOO);
close (CLARIAN);
```

The following script was used to remove duplicated query terms and some URL codes.

```
#!/user/bin/perl
```

```
open (INFILE, "clarian_p_final.txt")|| die ("cannot open  input file");
open (OUTFILE, ">clarian_P_ST_clean_final.txt") || die ("cannot open output file");


$count_line = 0;
while($line = <INFILE>){ #read each row of the table
        if ($line =~ /([\S\s]+?)&/){
                $line = "$1\n";
        }
        $line =~ s/\+/ /g;
        $line =~ s/\%20/ /g;
        $line =~ s/\%2E/ /g;
        $line =~ s/\%2C/,/g;
        $line =~ s/\%27/'/g;
        $line =~ s/\%22/"/g;
        push(@line, "$line");
        $count_line++;
}
print OUTFILE "$line[0]";
for($i=0;$i<=$count_line;$i++){
        if($line[$i] ne $line[$i+1]){
                print OUTFILE "$line[$i+1]";
        }
}
print $count_line;
close (INFILE);
close (OUTFILE);
```

Appendix B: Data Analysis Scripts

The following script was used to count how many unique sessions of each group.

```
#!/usr/bin/perl
open (INFILE, "log2007total.nobots.clean.log.sess")|| die ("cannot open  input file");
open (OUTFILE, ">countsession.txt")|| die ("cannot open  input file");
while ($line=<INFILE>){
        #@log = split(/[(|)("*")]/, $line);#change the log files into an array for session
        @log = split(/[(|)("*")]/, $line);#change the log files into an array for session
        push(@session, "$log[0]")
}
foreach $session (@session){
        if ( ! grep( /$session/, @uniqse ) ){
                push( @uniqse, $session );
        }
}
$count = @uniqse;
print $count;
print OUTFILE "there is the $count session.\n";
print OUTFILE "@uniqse\n";
close INFILE;
close OUTFILE;
```

The following script was used to discover the phrases user used to search.

```
using namespace std;
```

```
using namespace stdext;

const string NEGLIGIBLE_WORD_FILE("NegligibleWord.txt");

const int MAX_LINE_LENGTH = 1000;

// Make a string's all letters upper case

string uppercase_all(string source)

{

  std::transform(source.begin(), source.end(), source.begin(), ::toupper);

  return source;

}

int main(int argc, char* argv[])

{

  //check command line argument number

  if (argc != 2)

  {

    cout<<"Usage: weicomp3 [filename]"<<endl;

    exit(0);

  }

  hash_set<string> negligible;  //store negligible words
```

```
string token;

/// ---------------- Read negligible words----------------

ifstream fneg(NEGLIGIBLE_WORD_FILE.c_str());

        if(!fneg.is_open())

{

 //if cannot open file

 cout<<"Cannot open file: "<<NEGLIGIBLE_WORD_FILE<<endl;

 exit(0);

        }

while(!fneg.eof())

{

 fneg>>token;  //

 fneg.ignore(50000, '\n');   //skip rest of line



 negligible.insert(uppercase_all(token));

}

fneg.close();

/// ----- Read input file for enumerating combinations of key words----
```

```
//open file
ifstream fin(argv[1]);
        if(!fin.is_open())
{
  //if cannot open file
  cout<<"Cannot open file: "<<argv[1]<<endl;
  exit(0);
        }
hash_map<string, int> combined_keywords; //store combined keywords
string line;    //store each line of data file
//read each line and extract combination of key words
getline(fin, line);
while(fin.good())
{
  istringstream session(line);


  if (line.find("http") ==
  int sessionID = 0;
```

```
int seq = 0;

char colon;

session>>sessionID>>colon>>seq;

vector<string> session_words;   //store key words of current session

//read each key word

session>>token;   //read in one word

while(!session.eof())

{

  token = uppercase_all(token);

  if (negligible.find(token) == negligible.end())

  { //if the word is not in the list of negligible words

    session_words.push_back(token);

  }

  session>>token;   //read next word

}//while()

//enumerate all comination of current session key words

for (vector<string>::const_iterator i = session_words.begin();

  i != session_words.end(); ++i)
```

```
 {
  for(vector<string>::const_iterator j = i + 1;
   j != session_words.end(); ++j)
  {
   string temp = *i;
   temp.append(" ");
   temp.append(*j);
   if (combined_keywords.find(temp) == combined_keywords.end())
     combined_keywords[temp] = 1;
   else
     combined_keywords[temp]++;
  }//for(j)
 }//for(i)
 getline(fin, line); //read next line
}//while
fin.close();
multimap<int, string> combination_sorting;
for (hash_map<string, int>::const_iterator i = combined_keywords.begin();
```

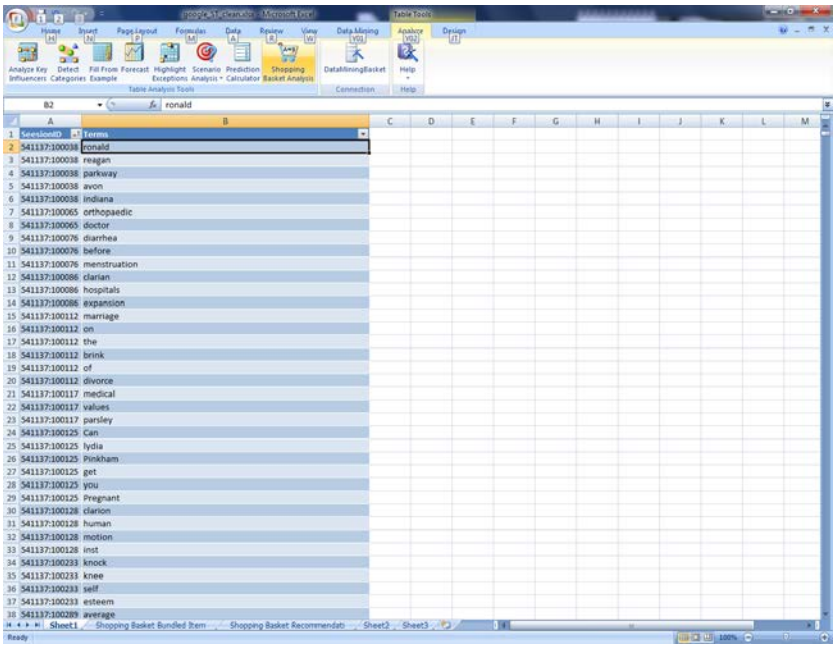
```
  i != combined_keywords.end(); ++i)
 {
  combination_sorting.insert(pair<int,string>(i->second,i->first));
 }//for(i)

 cout<<"=========Sorting result:============"<<endl;
 for (multimap<int, string>::const_iterator i = combination_sorting.begin();
  i != combination_sorting.end(); ++i)
 {
  cout<<i->second<<","<<i->first<<endl;
 }//for(i)
}
```

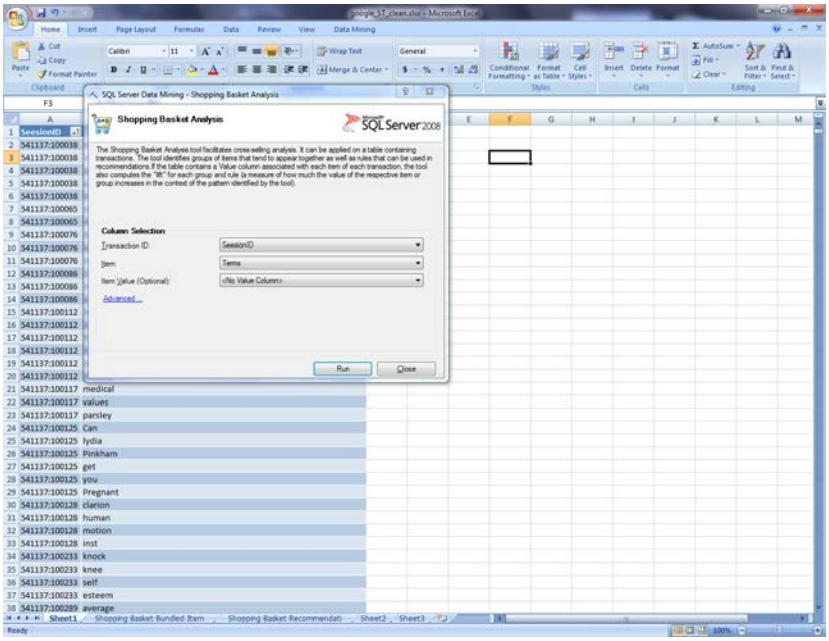
Appendix C: Use Excel to process "Market Basket Analysis"

This function is provided by Microsoft Excel with a data mining add-in

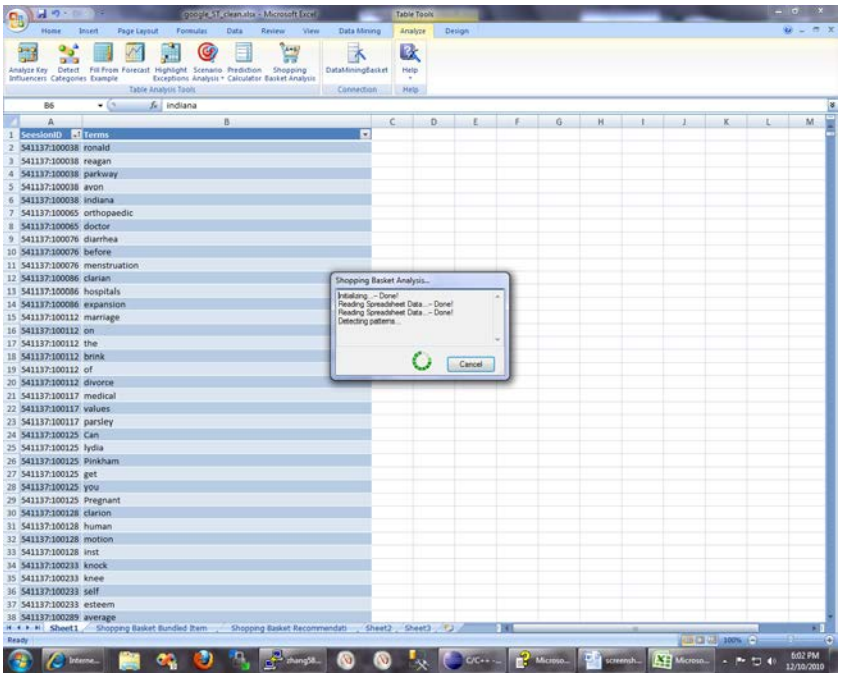
1. Transfer the data into required format: One Session ID one Term



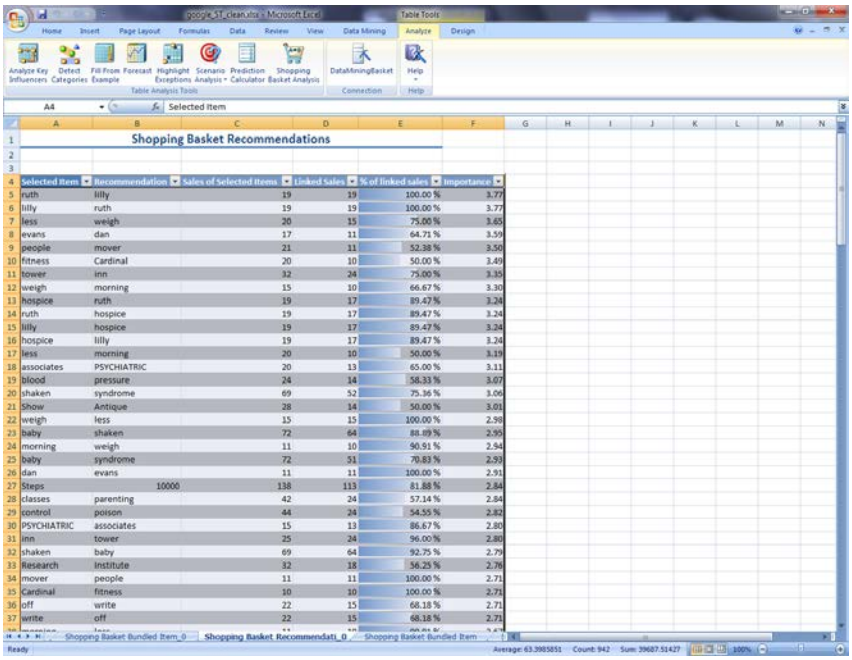
2. Provide the corresponding column information



3. Run



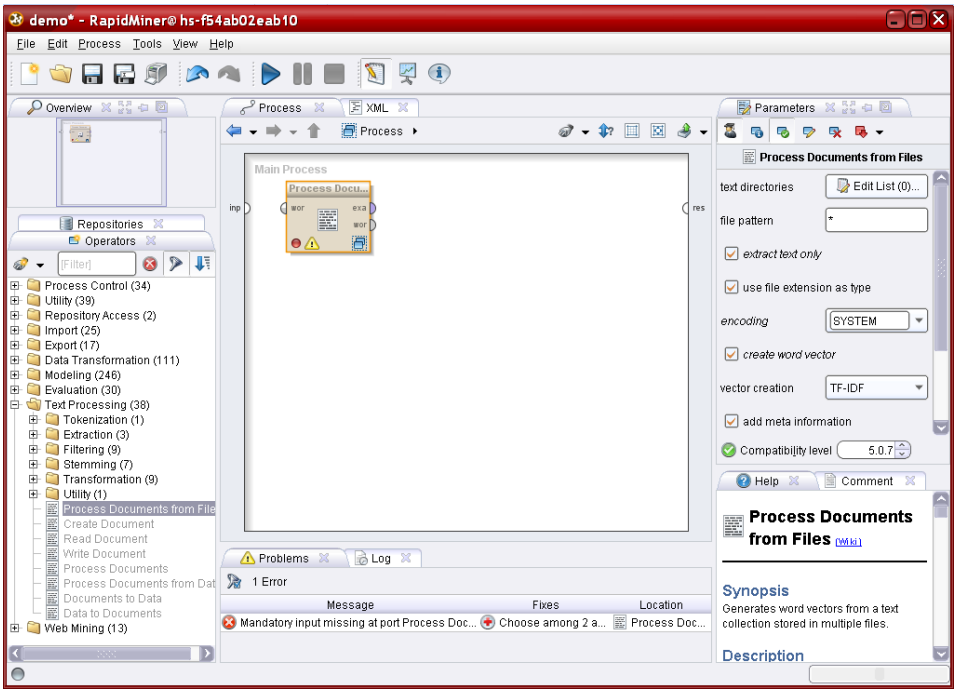
4. Get the result

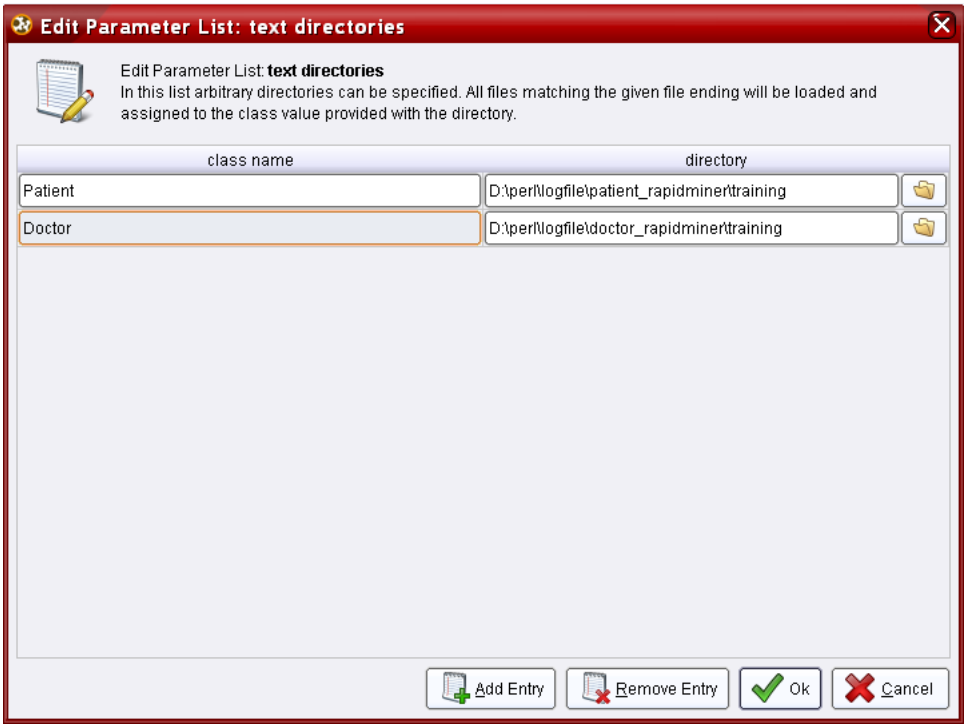


Appendix D: Use RapidMiner to Train & Test a Classifier

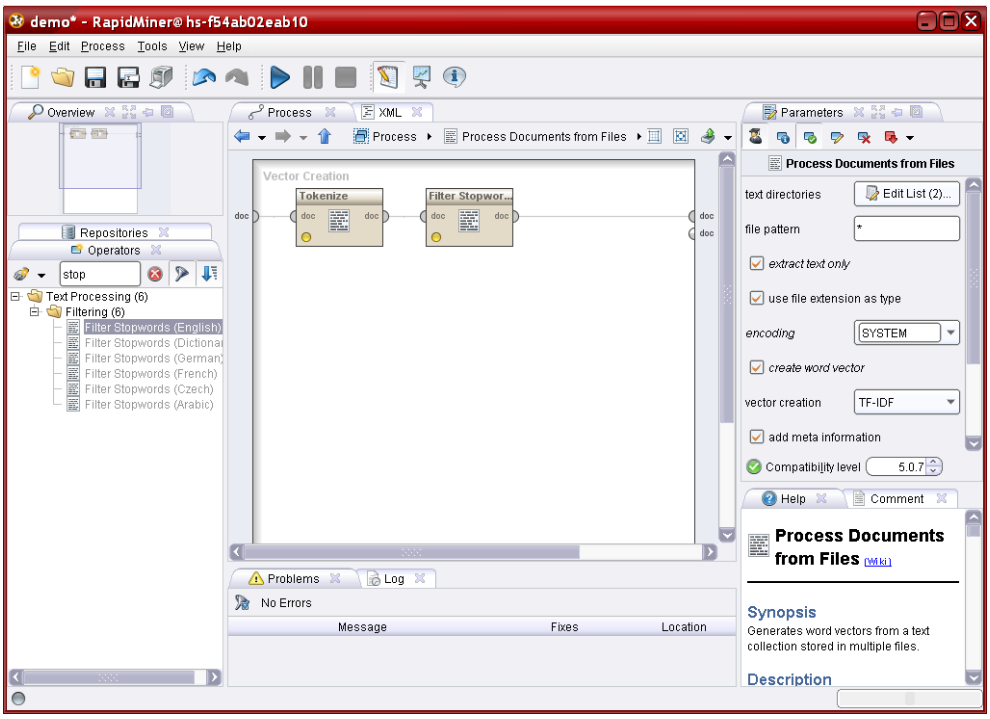
This function is provided by RapidMiner5.0 with cross validation

1. Process the labeled documents, choose the source data, and vector set to be TF-IDF

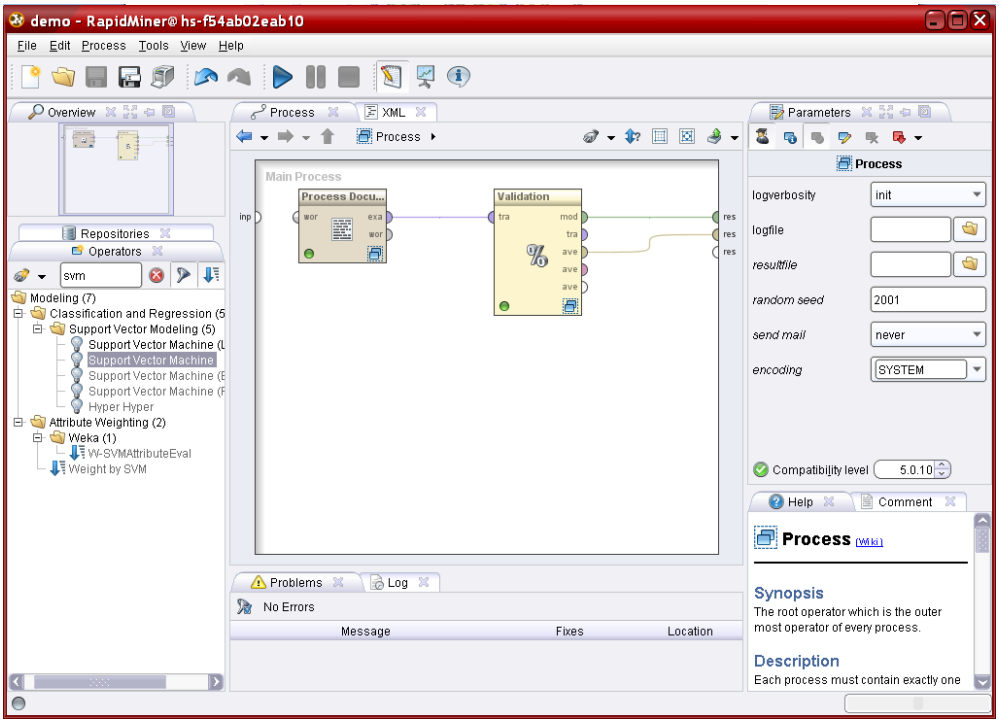




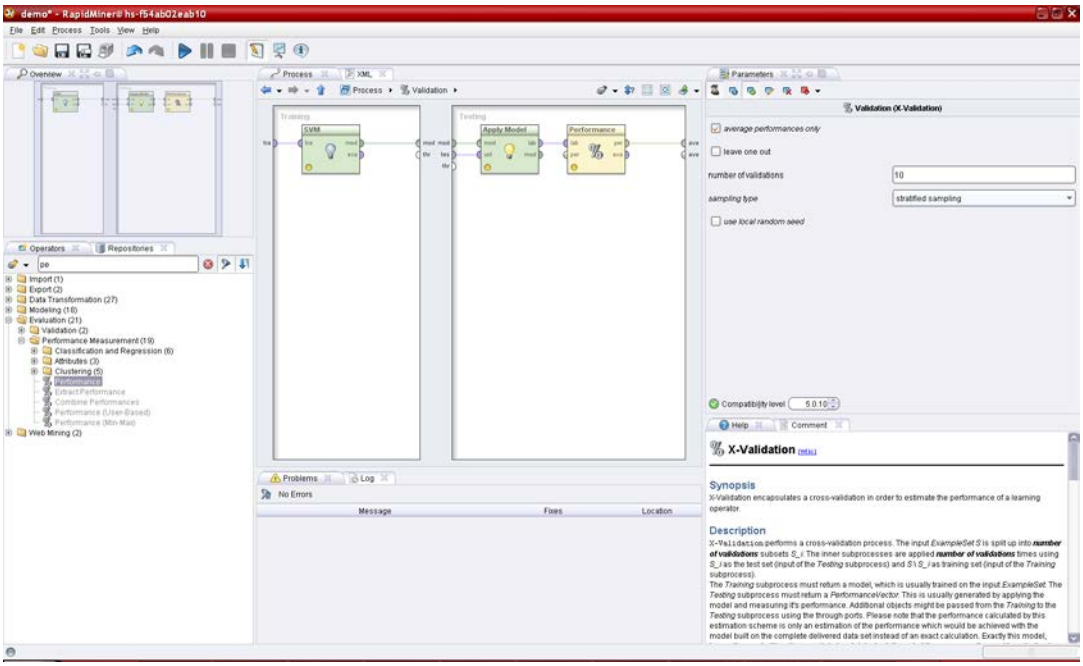
2. Add tokenization and filter stop words models



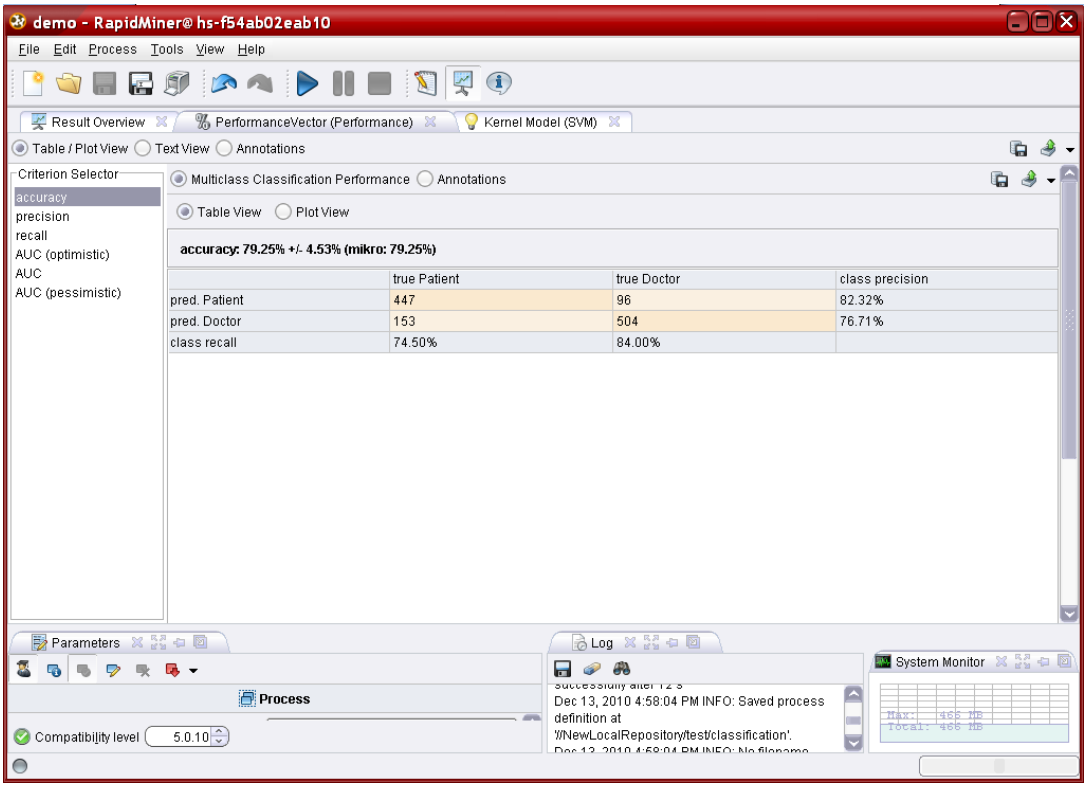
3. Add validation model



4. Add testing model



5. Run and get the results



Wei Kong

EDUCATION:

Indiana University, School of Informatics, Indianapolis, IN **Jan 2009 – Present**

Master of Science, Health Informatics GPA: 3.9/4.0

Thesis: Web mining for a health provider's website

CERTIFICATES:

- Web Markup & Style Coding, UITS, IUPUI
- Web Site Development Fundamentals, UITS, IUPUI

Sichuan University, China **Bachelor of Engineering** **July 2003**

Class Rank: 5th out of 56

GPA: 3.2/4 Major GPA: 3.6/4

Major: Polymer Science and Engineering

Thesis: PC resin application in beer storage industry

WORKING EXPERIENCE:

University UITS center, Biomedical Group **Nov. 2010 – Mar. 2011**

Biomedical Intern

- Software design and development (Perl, SQL)

Cientive Group Inc. Indianapolis, Indiana **May 2010 – August 2010**

EMR(Electronic Medical Record) system Analyst (Internship)

- OpenMRS analysis and data model design
- User interface development

General Electric, Advanced Material. **July 2003 – July 2006**

Environmental Health & Safety (EHS) Engineer

- Ensure compliance with requirements of Health framework audit program
- Electronic tools development

PROJECTS EXPERIENCE:

Course projects:

- Clinical Information System (CIS) Data Sources and Information Utilization
- My Clinic System (MySQL, HTML, XML, PHP)
- E-Vet Health (MySQL, XHTML, PHP, Unix)

Work projects:

- EMR development
- PPE Management
- Medical Surveillance system

HONORS AND AWARDS:

2009~2010 Informatics School Master Fellowship, IUPUI

2004~2005 GE outstanding staff awards

2000~2003 First level Student Scholarship, Sichuan University

Publications:

Wei Kong, Josette Jones, "Exploring Health Website Users by Web Mining", SMART Interfaces for Consumer Health Applications II, 2011

Patents:

Qinjian Yin, Hong Yang , Wei Kong "Mini Polymerizer of Polycondensation Reaction" China utility model patent Application No.CN200420060483.6 Publication No.CN2756292 Publication Date: 02/08/2006

Qiang Fu, Qinjian Yin, Hong Yang , Wei Kong "Gas Transmission Rate Tester" China utility model patent Application No.CN200420061664.0 Publication No.CN2750316 Publication Date: 01/04/2006

Qinjian Yin, Hong Yang, Wei Kong "Mini Polymerizer of Polycondensation Reaction" China invention patent Application No.CN200410040329.7 Publication No.CN1597730 Publication Date: 03/23/2005