Automated Discovery of Patient-Specific Clinician Information Needs Using Clinical Information System Log Files

Elizabeth S. Chen, MPhil and James J. Cimino, MD
Department of Biomedical Informatics, Columbia University, New York, NY, USA

Knowledge about users and their information needs can contribute to better user interface design and organization of information in clinical information systems. This can lead to quicker access to desired information, which may facilitate the decision-making process. Qualitative methods such as interviews, observations and surveys have been commonly used to gain an understanding of clinician information needs. We introduce clinical information system (CIS) log analysis as a method for identifying patient-specific information needs and CIS log mining as an automated technique for discovering such needs in CIS log files. We have applied this method to WebCIS (Web-based Clinical Information System) log files to discover patterns of usage. The results can be used to guide design and development of relevant clinical information systems. This paper discusses the motivation behind the development of this method, describes CIS log analysis and mining, presents preliminary results and summarizes how the results can be applied.

INTRODUCTION

The availability of clinical information to the clinician at the point of care is essential to the health care process. Inability to locate needed information can be costly in terms of time and quality of care. Clinical information systems have been developed to assist clinicians with their decisions; however, these systems need to ensure that they provide the information in optimal ways. In order to develop a useful and effective system, the users and uses of such systems must be understood. Without a thorough understanding of the potential users and their information needs, developers may produce a system that does not serve its purpose and perhaps inhibits the decision-making process, leading to medical errors

Information Needs

A variety of information is used by clinicians during the decision-making process. The following types of information needs have been defined: patient data, population statistics, medical knowledge, logistic information and social influence. The patient record, medical textbooks, on-line resources and colleagues are among the sources that have been identified for satisfying such needs[1,2]. A number of qualitative methods have been employed for assessing clinician information needs such as surveys, interviews, observations and self-reporting.

Log File Analysis

Log files are files that contain a list of actions that have occurred in a system. Analysis of these files can not only tell who, what, when and where but how information in the system was sought and used. Log file analysis is a quantitative method that is used to monitor usage of systems and gain an understanding of users in many domains. Results of such an analysis can be used in a number of ways such as making system improvements to make the user's experience more rewarding and satisfactory or determining if suspicious activity is occurring so that necessary actions can be taken. Because different types of log files exist, variations of log file analysis have emerged including Web, transaction, and audit log analysis.

The Web has become a popular medium for disseminating information. Web log analysis can help developers gain insight into searchers, determine what types of information to include in their sites and determine how to organize the information. In the health care setting, Web log analysis has been used to evaluate usage of Digital Health Science Libraries (DHSLs)[3,4], medical education Web sites[5,6], and on-line databases of medical images[7].

Analysis of transaction logs has commonly been done to evaluate information retrieval systems and catalog systems such as OPAC (On-line Public Access Catalogue)[8]. Numerous studies have reported using transaction log analysis to measure success of their systems.

To ensure confidentiality of patient information while allowing such data to be readily available to clinicians, security management of clinical systems is necessary. One security management strategy is maintenance of audit logs or audit trails for the various clinical systems at health care institutions[9,10,11]. These logs can be used to monitor accesses, determine if there have been security breaches and identify if authenticated users are using the system inappropriately. Additionally, these audit logs can be used to monitor basic operational aspects of applications.

Mining Techniques

Data mining and Web mining are both concerned with discovering meaningful patterns in data. Aspects of these techniques can be applied to analyze log files. Data mining is often considered part of a larger process called knowledge discovery in databases (KDD) and is concerned with the exploration and analysis of large quantities of data[12]. Data mining and KDD are aimed at developing methodologies and tools that can automate the data analysis process and create useful information and knowledge from data to help in decision-making. Web mining utilizes Web server logs to understand and better serve the needs of Web-based application users[13,14]. Web usage mining is one category of Web mining, which focuses on analysis of Web logs to understand user behavior and web structure so that improvements can be made.

Many clinical information systems maintain as standard practice some type of log file or usage log. At New York Presbyterian Hospital (NYPH), WebCIS (Web-based Clinical Information System) enables clinicians to browse the content of patients' medical records[15]. The usage log of WebCIS conveys how users are interacting with data in the patient's record. Analysis of these usage logs can convey what patient-specific information users are looking at, in what order users look at this information and if the order is dependent on user characteristics, patient characteristics and/or type of result. From this, we can identify patient-specific information needs (need for particular patient data); this discovered knowledge can be valuable for making appropriate and effective system improvements.

METHODS

We define CIS log analysis as the process of analyzing clinical information system log files to uncover usage patterns and rules. What is discovered will give us insight into what data users are interested in and how they access them. We call the technique for this discovery CIS log mining.

CIS log mining emerges from data mining and Web usage mining. This new mining technique consists of four phases: data collection, preprocessing, pattern discovery and pattern analysis.

We use Perl and UNIX commands to perform the various CIS log mining techniques. Associative arrays or hashes are used as the main data structure for storing the different data formats and transformations.

Data Collection

We have chosen to analyze WebCIS log files. Logs are generated that record the actions of all users of this system in chronological order. Each line in the logs consists of seven fields: timestamp, application name, username or user ID, client machine name or IP address, 7-digit medical record number (MRN), data type (with subtypes if applicable) and action (with modifiers if applicable). Figure 1 contains some lines that can be found in a WebCIS usage log. Each day's log goes through preprocessing and pattern discovery. Pattern analysis can then summarize the results over a given period of time.

Preprocessing

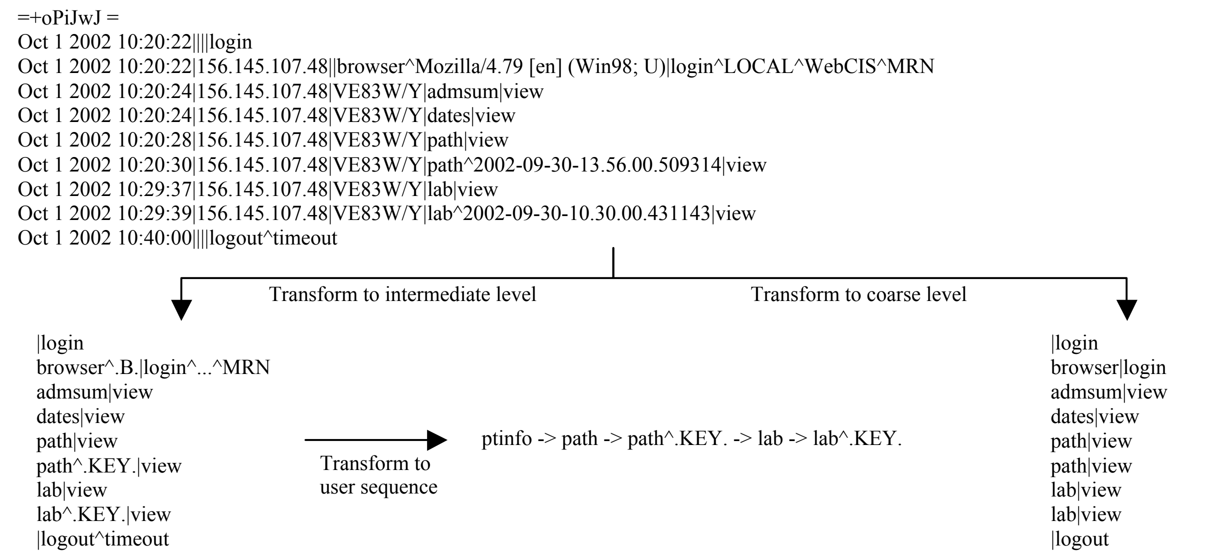
Preprocessing tasks include de-identification, data cleaning, enrichment and transformation. It is also in this phase that missing data are identified.

In order to maintain the privacy and confidentiality of the users and patients, any identifying information must be removed. We de-identify the data by encrypting all usernames and MRNs using the MD5 hash function.

Data cleaning is concerned with removing as many pollutants as possible. This may involve de-duplication and filtering out unnecessary data. For our logs, we found that the timestamp field contained extraneous information and that the application field was not necessary. We also discovered lines not adhering to the 7-field format, which needed to be expanded to follow this format.

```
Oct  1 00:08:59 webcis3-i syslog: |WebCIS|user1|156.145.130.17|mrn1|admsum|view
Oct  1 00:09:18 webcis3-i syslog: |WebCIS|user1|156.145.130.17|mrn1|clinnote.^2002-09-30-04.51.20.456331|view
Oct  1 00:19:47 webcis3-i syslog: |WebCIS^LDAP4|user2||||login
Oct  1 00:20:19 webcis3-i syslog: |WebCIS|user2|156.145.43.33|mrn2|lab|view
Oct  1 00:21:43 webcis3-i syslog: |WebCIS|user1|156.145.130.17|mrn3|rad^2002-09-04-11.43.00.000599|view
Oct  1 00:21:55 webcis3-i syslog: |WebCIS|user2|156.145.43.33|mrn2|lab^2002-09-30-20.05.00.740657|view
Oct  1 00:26:03 webcis3-i syslog: |WebCIS|user1|156.145.130.17|||logout
```

Figure 1: WebCIS usage log lines: Each line has 7 fields that are delimited by "|". The fields are: timestamp, application name, username, MRN, IP address, data type and action. Data types may have subtypes, which are delimited by "^" and actions may have modifiers that are also delimited by "^". Subtypes such as "2001-01-10-00.00.00.257515" are keys that are used to retrieve results from the clinical data repository and are patient-specific. These are lines for two WebCIS users that demonstrate how one viewed an admission summary, a specific clinical note, a specific radiology result and logged out, and the other logged in and viewed a list of laboratory results and then a specific one.



```
=+oPiJwJ =
Oct 1 2002 10:20:22||||login
Oct 1 2002 10:20:22|156.145.107.48||browser^Mozilla/4.79 [en] (Win98; U)|login^LOCAL^WebCIS^MRN
Oct 1 2002 10:20:24|156.145.107.48|VE83W/Y|admsum|view
Oct 1 2002 10:20:24|156.145.107.48|VE83W/Y|dates|view
Oct 1 2002 10:20:28|156.145.107.48|VE83W/Y|path|view
Oct 1 2002 10:20:30|156.145.107.48|VE83W/Y|path^2002-09-30-13.56.00.509314|view
Oct 1 2002 10:29:37|156.145.107.48|VE83W/Y|lab|view
Oct 1 2002 10:29:39|156.145.107.48|VE83W/Y|lab^2002-09-30-10.30.00.431143|view
Oct 1 2002 10:40:00||||logout^timeout
```

Transform to intermediate level
Transform to coarse level
Transform to user sequence

```
|login
browser^.B.|login^...^MRN
admsum|view
dates|view
path|view
path^.KEY.|view
lab|view
lab^.KEY.|view
|logout^timeout
```

```
ptinfo -> path -> path^.KEY. -> lab -> lab^.KEY.
```

```
|login
browser|login
admsum|view
dates|view
path|view
path|view
lab|view
lab|view
|logout
```

Figure 2: Preprocessing phase of CIS log mining: Original log file is de-identified, cleaned and transformed into user sessions. Each user session starts with the encrypted username and is followed by the lines associated with the session, which contain timestamp, IP address, encrypted MRN, data type, and action. The example session indicates the user logged into the system, selected a patient, got basic information about the patient (automatically displayed for each patient), viewed a list of pathology results and then a specific one, viewed a list of lab results and then a specific one, and logged out. The data type and action fields in the user session are transformed to the intermediate and coarse levels of granularity. ".KEY." replaces all references to patient-specific department results (i.e. 2002-09-30-13.56.00.509314). User sequences are then generated from the intermediate level session. For user sequences, we remove references to logging in and logging out and create the data type "ptinfo" to encompass "admsum" and "dates".

Data enrichment involves adding extra information that is not found in the log file. Because one of our goals is to determine if there is a correlation between user or patient characteristics and usage patterns, we need to obtain these characteristics. We use the NYPH LDAP (Lightweight Directory Access Protocol) server to retrieve user characteristics such as position and department for each username found in the WebCIS logs. For patient characteristics such as age, gender and diagnosis, we can use our CDR (Clinical Data Repository) for each MRN in the log. Additionally, some of the information found in the logs are in coded form, specifically they are MED (Medical Entities Dictionary) codes. The MED can be used to translate these codes into more meaningful MED names[16].

Because a number of pattern discovery techniques can be used that may take data in different formats, the log file needs to be transformed in several different ways. Most of the techniques will be interested in viewing the log as user sessions. We define a user session as the time from when a user logs in until the time that user logs out. A user may login and logout multiple times in a day but these are considered separate sessions. Figure 2 contains a single user session, which depicts how the log is de-identified, cleaned and transformed.

The data contained in the data type and action fields are at a fine level of granularity. This makes it difficult to make generalizations since there are so many possible data type-subtype and action-modifier combinations. We therefore want to have two additional transformations of the log: one that contains only data types and actions and one that contains data types with generalized subtypes and actions with generalized modifiers. These will be referred to as the coarse and intermediate levels of granularity respectively. We are also interested in user sequences at the intermediate level, which depict how the user moves from one piece of information to another. For these sequences, we are only interested in the data type. Figure 2 shows how user sessions are transformed to the intermediate and coarse levels of granularity as well as to user sequences.

Pattern Discovery

The four pattern discovery techniques we have chosen to perform are descriptive statistics, path analysis, association rule generation and sequential pattern discovery.

Descriptive statistics provide a basic overview of users. Statistics can be gathered that describe features of the logs such as size in terms of number of lines. Statistics can also be gathered for each field that can describe where the application is being used based on IP addresses, the patient records being accessed, the number and types of users, the frequency of data types and subtypes at the different levels of granularity, and the frequency of actions and modifiers at each level of granularity.

We use *path analysis* to identify the frequently visited pages in WebCIS. Graphs represent the logical layout of the system, where the pages are nodes and the directed edges are links between pages. We specify data types as nodes and the edges indicate how the user moves from one piece of information to another in a user session. We use *dot*, a tool for making hierarchical layouts of directed graphs to generate the graph output[17].

We use *association rule generation* to relate data types that are most often referenced together in a user session, disregarding the order. The Apriori algorithm, which uses two measures called support and confidence, identifies items that commonly occur together to produce frequent n-item sets[12].

Finally, we use *sequential pattern discovery* to generate rules that take into account the order or sequence of data types in a user session. We use a variation of the Apriori algorithm to generate frequent n-sequences.

Pattern Analysis

Each of the pattern discovery techniques provides a different view of the logs. We aggregate and summarize the results from each technique so that patterns can be analyzed and visualized.

The descriptive statistics can give us an idea of who our users are, what data they commonly look at and how they typically interact with that data. We combine association rules and sequential patterns to identify those that are the most frequent over time. By knowing which data users commonly look at together in a session, we can make those data available to them in this fashion in relevant systems.

RESULTS

We have applied our CIS log mining techniques to one year's worth of WebCIS log files. The results are presented in Table 1.

DISCUSSION AND CONCLUSION

Through CIS log analysis, we can gain insight into clinical information system users and their usage patterns. WebCIS log files are rich resources for providing information on thousands of users. The CIS log mining techniques we have developed identify which patient data they are commonly interested in and how they access these data. In addition, we can use qualitative methods to support and enhance the findings of CIS log analysis by informing us of why users are behaving in the manner that they are and what users found missing.

Results from all of our pattern discovery techniques indicate that WebCIS users commonly view laboratory and radiology results in a session. The frequent association rules and sequential patterns we have obtained only give us a general idea of the access patterns of users. We need to refer back to the user sessions and the Clinical Data Repository to provide meaning to the results. For example, we might discover that abdominal ultrasonography (USG) results are commonly viewed after liver function test (LFT) results. With this knowledge, we

Descriptive statistics	
Size of logs (lines)	58,789,808
Users	7,072
User sessions	3,431,419
IP addresses (on-campus)	7,318
Patient records (unique)	408,156
Data types (intermediate level)	58,447,470
lab^.KEY.	15%
lab	5%
rad^.KEY.	3%
rad	2%
Actions (intermediate level)	58,447,470
view	80%
login	6%
Path analysis (See Figure 3)	

Association rules (% of user sessions)	
lab,lab^.KEY.	36%
rad,rad^.KEY.	20%
dsum^.KEY.,lab^.KEY.,phar	6%
lab,lab^.KEY.,rad,rad^.KEY.	10%
lab,path,path^.KEY.,rad,rad^.KEY.	3%
card,card^.KEY.,lab,lab^.KEY.,rad,rad^.KEY.	2%
Sequential patterns (% of user sessions)	
ptinfo -> lab	57%
lab^.KEY. -> rad -> rad^.KEY.	5%
rad^.KEY. -> lab -> lab^.KEY. -> lab^.KEY.	2%
lab^.KEY. -> ptinfo -> lab -> lab^.KEY. -> lab^.KEY.	6%
ptinfo -> rad -> rad^.KEY. -> rad^.KEY. -> rad^.KEY. -> rad^.KEY.	1%
ptinfo -> lab -> lab^.KEY. -> lab^.KEY. -> lab^.KEY. -> lab^.KEY. ->lab^.KEY.	13%

Table 1: Results from analysis of one year of WebCIS logs: Daily WebCIS logs from January 2002 to December 2002 underwent preprocessing, pattern discovery and pattern analysis. The numbers and percentages presented are for a year of WebCIS logs. The more frequent data types and actions at the intermediate level of granularity, association rules (2 to 6-item sets), and sequential patterns (2 to 7-sequences) are displayed.

can add a “shortcut” from LFT to USG results in patient record systems[18]. Additionally, we will use another pattern discovery technique, *classification* to look for correlations between user characteristics, patient characteristics and usage patterns. For the user sessions retrieved for the frequent association rules and sequential patterns, we will retrieve the user and patient characteristics (obtained during data enrichment) associated with them and determine whether or not there is a correlation. When such correlations are found, classes can be developed. By identifying classes of users or patients, we can customize systems to present data depending on the user or patient characteristics.

Although we focused on analyzing patient record system log files for discovering patient-specific information needs, CIS log mining can be applied to other clinical information system logs and used to discover other types of information needs. With the knowledge about system users and their information needs obtained from CIS log analysis, developers can more appropriately and effectively design and develop clinical information systems. Providing users with quicker access to relevant data when they are needed can facilitate the task at hand and improve the health care process.

Acknowledgments

This work has been supported in part by National Library of Medicine Training Grant NO1-LM07079. The authors thank Soumitra Sengupta and Richard Gallagher for their assistance in obtaining the WebCIS log files.

References

1. Covell DG, Uman GC, Manning PR. Information needs in office practice: are they being met? Ann Intern Med. 1985 Oct;103(4):596-9.
2. Gorman PN. Information needs of physicians. J Amer Soc Inform Sci. 1995;46(10):729-736.
3. D'Alessandro DM, Kreiter CD. Improving usage of pediatric information on the Internet: The Virtual Children's Hospital. Pediatrics. 1999 Nov;104(5):e55.
4. D’Alessandro MP, D’Alessandro DM, Galvin JR, Erkonen WE. Evaluating overall usage of a digital health sciences library. Bull Med Libr Assoc. 1998 Oct;86(4):602-9.
5. Nieder GL, Nagy F. Analysis of medical students' use of web-based resources for a gross anatomy and embryology course. Clin Anat. 2002 Nov;15(6):409-18.
6. Dev P, Rindfleisch TC, Kush SJ, Stringer JR. An analysis of technology usage for streaming digital video in support of a preclinical curriculum. Proc AMIA Symp. 2000;:180-4.
7. Ribaric S, Todorovski L, Dimec J, Lunder T. Presentation of dermatological images on the Internet. Comput Methods Programs Biomed. 2001 May;65(2):111-21.
8. Atlas MC, Little KR, and Purcell MP. Flip charts at the OPAC: using transaction log analysis to judge their effectiveness at six libraries of the University of Louisville. Reference and User Services Quarterly. 1997;37(1):63-69.
9. Barrows RC and Clayton PD. Privacy, confidentiality, and electronic medical records. J Am Med Inform Assoc. 1996 Mar-Apr;3(2):139-48. Review. 1996;3(2):139-148.

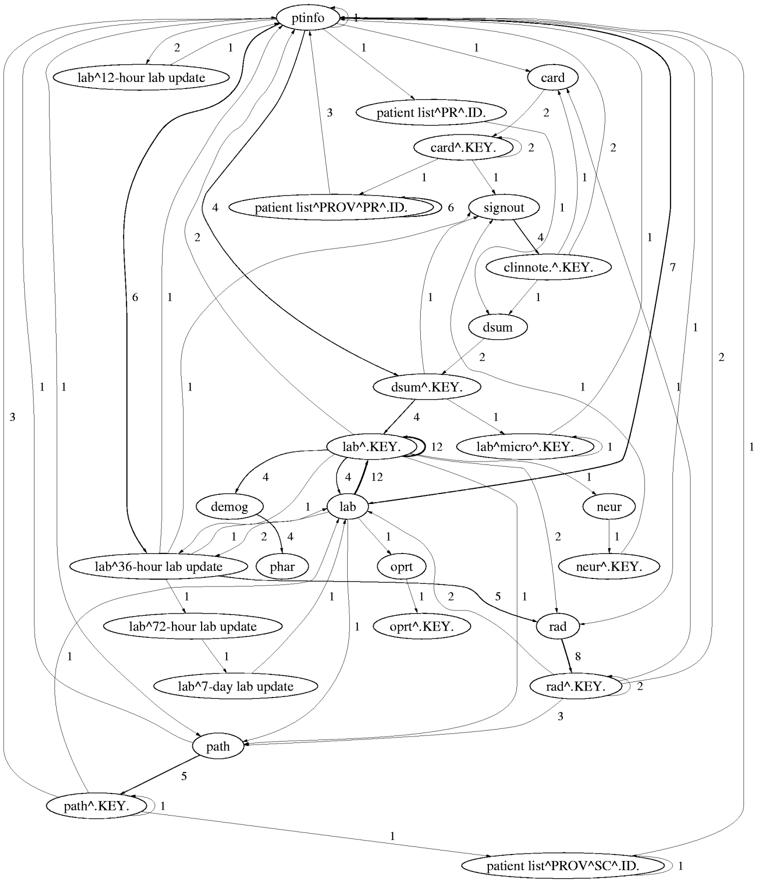


Figure 3: Graph for fifteen WebCIS user sessions generated from path analysis: Nodes are data types. Edges show how user moves from one piece of information to another. Edge thickness and label indicates frequency of the edge.

10. Gallagher RJ, Sengupta S, Hripcsak G, Barrows RC, Clayton PD. An audit server for monitoring usage of clinical information systems. Proc AMIA Symp. 1998;:1002.
11. Asaro PV, Ries JE. Data mining in medical record access logs. Proc AMIA Symp. 2001;:855.
12. Fayyad U, Piatetsky-Shapiro G, Smyth P, Uthurusamy R, editors. Advances in knowledge discovery and data mining. Menlo Park, CA: AAAI Press/MIT Press; 1996.
13. Cooley R, Mobasher B, Srivastava J. Web mining: information and pattern discovery on the world wide web. Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), November 1997.
14. Srivastava J, Cooley R, Deshpande M, Tan P. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations. 2000;1:2:12-23.
15. Hripcsak G, Cimino JJ, Sengupta S. WebCIS: large scale deployment of a Web-based clinical information system. Proc AMIA Symp. 1999;:804-8.
16. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. J Am Med Inform Assoc. 1994 Jan-Feb;1(1):35-50.
17. AT&T. Graphviz – open source graph drawing software. Available at: http://www.research.att.com/sw/tools/graphviz/. 2002.
18. Chen ES, Hripcsak G, Patel VL, Sengupta S, Gallagher RJ, Cimino JJ. Automated identification of shortcuts to patient data for a wireless handheld clinical information system. Proc AMIA Symp. 2003. (in press)