Proceedings of the 4th WSEAS/IASME Int. Conf. on System Science and Simulation in Engineering, Tenerife, Spain, December 16-18, 2005 (pp217-220)

FROM LOG FILES TO VALUABLE INFORMATION USING DATA MINING TECHNIQUES

JELENA MAMČENKO, REGINA KULVIETIENĖ
Information Technology Department
Vilnius Gediminas Technical University
Saulėtekio av. 11, Vilnius
LITHUANIA
jelena@gama.vtu.lt, regina_kulvietiene@gama.vtu.lt

Abstract. Most people, companies and organizations put information on the Web because they want to be seen by the world. Their goal is to have visitors come to the site, feel comfortable and stay awhile. The aim of this work is to conduct an in-depth analysis of the data kept in the log files of the server that hosts the web pages of the Vilnius Technical University's Information Technology department. It is a constructive and convenient web site for study purposes, gathering useful information, finding free jobs in a labour exchange, electronic library and others.
To better know the needs of web site visitors we have analysed its access log file using a modern Data Mining technique. This article treats data mining potential for web site analysis.

Keywords: world wide web, log files, data mining, statistics, clustering.

1. Introduction

Educational institutes are confronted with the problem what to do with huge amount of data collected in log files and how to use and interpret them. It would be very useful if patterns could be derived from the traces that users leave when they navigate around a university's web page. For instance, a university may be interested in questions such as: how many people look at the courses of the university, which courses are most popular, what is the country of origin of people who are interested in a specific course, etc. These are all questions that are relevant to a university, and answers to which lie implicit in the data kept in log files on the server that hosts the respective web pages.

The new Data mining technology development and tools to process data into useful information was stimulated by a huge growth of data in database [1, 2, 3].

Nowadays, the use of Data mining is widespread in any industry. In common, everywhere there are immense volumes of data. An increasing amount of information is being stored in electronic form such as log files [4].

In fact, the Internet environment, especially World Wide Web has very unstructured data format from Data mining point of view [3].

Lotus Domino is a server that provides an ideal communications infrastructure by tightly integrating the robust functionality of enterprise-ready, client/server messaging and groupware with the open standards and global reach of the World Wide Web. Domino enables individuals and organizations to communicate with colleagues, collaborate in teams, and coordinate business processes within and beyond their organizational boundaries to achieve a competitive edge. Domino supports a variety of clients and devices, including Web browsers, Lotus Notes clients, and various mail and mobile clients.

The analysed server *gama*, an LDAP directory included, is a part of the Distance Education Information System at Vilnius Gediminas Technical University. Together with it there are *kappa, teta, irma a*nd *beta* collaborative servers.

2 Log file analysis using Data Mining

Server log files are records of web server activity. They provide details about file request from a web server and its response to the request. In the access file that is the main log file, each line describes the source of a request, the file requested, the date and time of the request, the content type and length of the transferred file, and other data such as errors and the identity of referring pages [5, 6].

We began with gathering data in the form of access logs that describe users' behaviour [4, 7]. Our goal is to group them together based upon the similarity of their activity.

A log file in the common log format contains a separate line for each request that comes in to the

server. Here's a sample access log entry in the domlog.nsf database:

Date: 2005.09.30 14:54:37
User Address: 82.135.222.5
Authenticated User: CN=Roma Siugzdaite/O=Vtu/C=LT
Status: 200
Content Length: 268
Content Type: image/gif
Request: GET /mail/rsiugzda.nsf/MailFS%20-%20Button%20Web!OpenImageResource HTTP/1.0
Browser Used: Mozilla/4.0 (Windows XP 5.1) Java/1.5.0_04
Error:
Referring URL:
Server Address: gama.vtu.lt
Elapse Time (ms): 0
Translated URI: d:/Lotus/Domino/Data/mail/rsiugzda.nsf
Cookie: LtpaToken=AAECAzQzM0QyNzc0NDMzREQwMzRDTj1Sb21hIFNpdWd6ZGFpdGUvTz1WdHUvQz1MVAo/BkoIxNQiMCfSX8hYL3uS/aSt

This entry uses the following syntax, where each access criteria is separated with a space.

For data analysis we've used the IBM Intelligent Miner for Data software that supports not only mining but statistics functions as well [8].

Clustering analysis allows one to group together clients or data items that have similar characteristics. Clustering of client information or data items on Web transaction logs can facilitate the development and execution of future marketing strategies, both online and off-line, such as automated return mail to clients falling within a certain cluster, or dynamically changing a particular site for a client, on a return visit, based on past classification of that client.

The results of the clustering function contain the number of detected clusters and the characteristics that make up each cluster. Clustering provides a fast and natural clustering of very large database. It automatically determines the number of clusters to be generated. Similarities between records are determined by comparing their field values.

The clusters are then defined so that Condorcet's criterion is maximized. Condorcet's criterion is the sum of all record similarities of pairs in the same cluster minus the sum of all similarities of pairs in different clusters.

So, we have a database file exported to a flat file format with the historical data range from 2004 November 29 11:09:08 to 2005 April 12 16:36:23 with additional fields shown in Table 1.

2.1 The mining tasks

There are five phases of data mining tasks:

- Defining the data. We specify a data object that points to a flat file.
- Building the model. Define a clustering settings object. This model contains information that describes the clusters identified during the mining run.
- *Applying the model.* Define a clustering settings object. It runs in application mode using *building model* results and produces an output data in flat file. This output file identifies the subgroup associated with a user record.
- *Automating the process.* To automate the process we created a sequence object containing the *build model* settings object and *apply model* settings object. A sequence is an object containing several other objects in a specific sequential order. This allows one to combine several mining tasks into one.
- *Analyzing the results.* Define a bivariate statistics function. This function analyses the data object and produces an output object, a flat file and a result object.

Table 1. Fields that are used

Date	The format is year, month, day (yyyy:mm:dd)
Time	Time in 24-hour clock, minute, second (hh:mm:ss)
Authenticated User	Using local authentication and registration, the user's log name will appear, if no value is presented, a "-" is substituted.
IP address	IP address
Method	Method is: GET, POST, OPTIONS, PROPFIND, CONNECT, HEAD, SEARCH or PUT
Request	Is the path and file retrieved

2.2 The results generated

The result generated by the mining function (IBM Intelligent Miner for Data) is shown in Figure 1.
This table shows nine rows, each representing one of the nine clusters identified by the mining run. The numbers down the left side represent the cluster size as a percentage for instance, the top cluster represents 26% of the data, the next 16% and so on. The numbers in brackets in Name column identify the cluster ID.

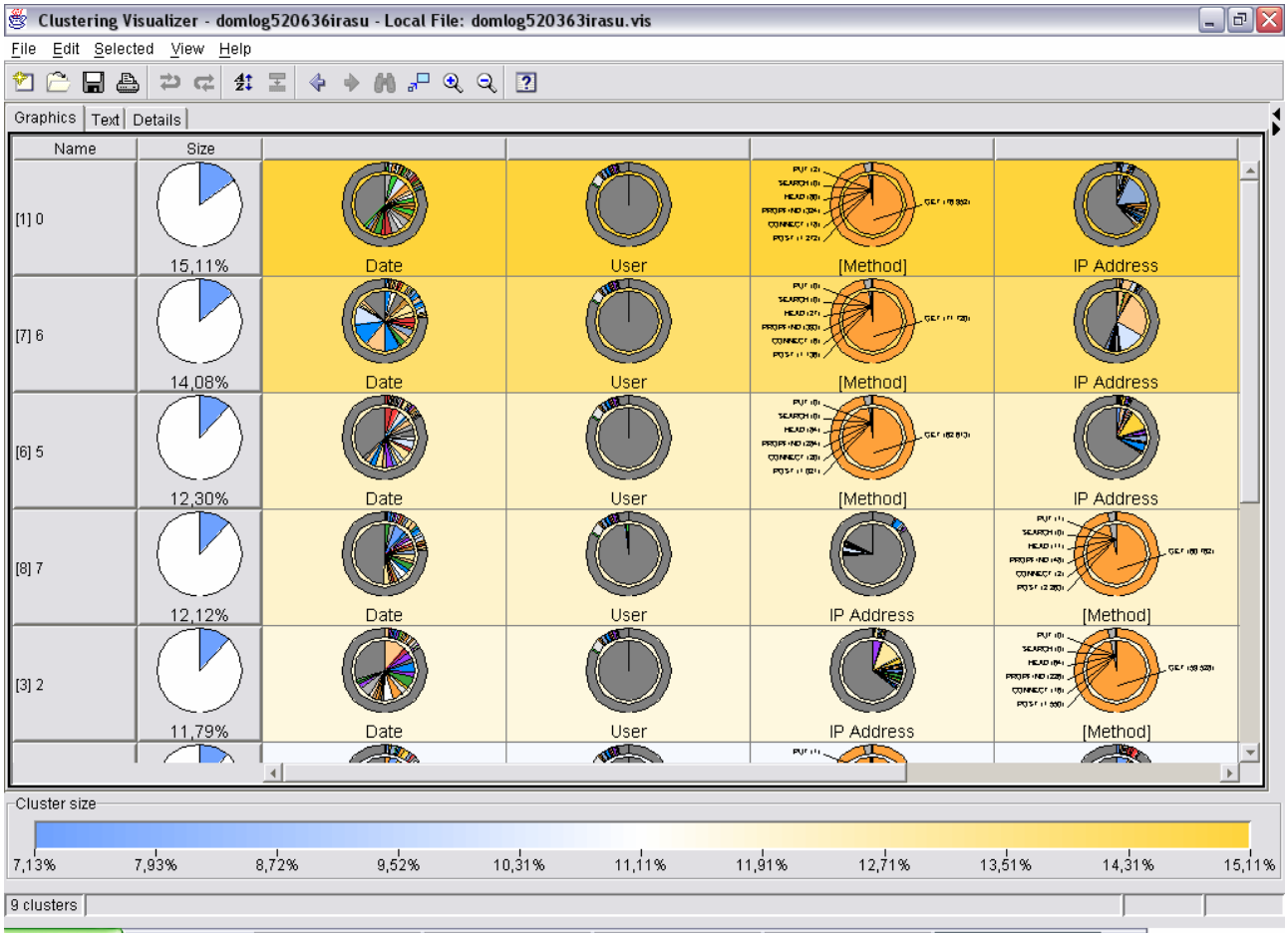

Fig. 1. Graphical view

Figure 1 shows nine rows, each representing one of the nine clusters identified by the mining run. Within each cluster, the pie chart diagrams represent active and supplementary fields used in the cluster. In this case, fields that had the greatest influence on forming the cluster are displayed on the left, while fields with the least influence are displayed on the right. Each pie chart produced by the Intelligent Miner shows two distributions. The outside ring shows the distribution for the entire sample. The distribution for the associated cluster is shown by the inside ring. The top row is the cluster with the largest number of users (15%). Supplementary fields are indicated with square brackets around the field names. The biggest cluster and information in it mean, that 2004.12.13 at 17:20:53 from 65.54.188.98 IP address was predominantly access to the file /audrkaba/pics/honey.jpg (authentication is unnecessary). The second one (14% of all users) - 2005.03.01 at 20:08:14 from 193.219.146.90 IP address (authentication is unnecessary as well) and file was predominantly /icons/ecblank.gif and etc.

After full cluster analysis we found out that the most popular and necessary part of gama.vtu.lt (42% users) is /icons/ecblank.gif and predominantly the user is unauthenticated, but almost 11% of all users are using the main page of Information technology website, without knowing definite address for necessary information.

3 Problems with log files format and analysis tools

However, some problems with data transformation have occurred. As we know, Intelligent Miner for data requires a database or flat file data format. But it can't take just flat file. First of all, data in that file should be aligned by columns and every record's end must have enough spaces. It means that any

shorter last column record must have the same number of spaces as the longest last column record has symbols.

Second, most of the existing Web analysis tools provide mechanisms for reporting user activity in servers and various forms of data filtering. Using such tools, for example, it is possible to determine the number of accesses to the server and the individual files within the organization's Web space, the times or time intervals of visits, and domain names and the URLs of users of the Web server. However, in general, these tools are designed to handle low to moderate traffic servers, and furthermore, they usually provide little or no analysis of data relationships between the accessed files and directories within the Web space.

Another note related to this work is that it's necessary to clean a server log file to eliminate irrelevant items as it is important for any type of Web log analysis, not just data mining. The discovered associations or reported statistics are only useful if the data represented in the server log give an accurate picture of the user accesses of the Web site. Elimination of irrelevant items can be reasonably accomplished by checking the suffix of the URL name. For instance, all log entries with filename suffixes GIF, JPEG, JPG, and map can be removed [9].

4 Conclusions

There is a lot of different log files analysing tools such as Accure Software, Sane Solutions, WebTrends and etc. Their prices differ and we can find programs for or much more sophisticated programs at very high prices. The choice depends on organizational needs. But most of them are based on statistical methods. There are some limitations of using such methods. First of all, statistics summarizes information and answers precisely for formulated questions whereas Data mining provides a rich picture from mining domain.

Data mining technology discovered previously unknown and potentially useful information from raw data while statistics can't discover such patterns. It gives big potential and makes data mining technology a powerful tool for effective analysis and tasks optimisation.

With the combination of Domino and the intelligent analysis tools, we have several options for analysing what users like about Information Technology Department's site and what we need to improve. The key to understand every site's logs is to understand the structure of that site. Then, without problems, we can make sure that intelligent analysis tools give the results that we expect.

In this work we tried to show one of the possible Data mining applications. For this reason from the gama.vtu.lt server was taken an access log flat file, indicating date, time, user, IP address, method and requests by different users. We analysed 520 636 records (about 137 Mb) of anaccess log file using demographic clustering method and it took 3:08:38 hours. The results have been presented.

In addition to that, it would be very interesting to know the specific key words that were used by the students in the various search engines so that they can be used to manage the site's metadata more effectively. Finally, the layout of the site can be improved considerably by placing new banners and links to the most popular sites.

References:

1. Weiss, S.H. and Indurkhya, N, *Predictive Data Mining: A practical Guide*, Morgan Kaufmann Publishers, San Francisco, CA, 1998.
2. Piatetsky-Shapiro, G. and Frawley, W.J., *Knowledge Discovery in Database*, AAAI/MIT Press, 1991.
3. Sang Jun Lee and Keng Slau, A review of Data Mining Techniques, *Industrial Management & Data Systems* 101/1, 2001, pp. 41-46.
4. Gavin Meggs, Internet Usage Analysis, *Handbook of Data Mining and Knowledge Discovery,* University press. Oxford 2002, pp. 920-927.
5. Dorothy Bailey, Why analyse logs? *Is available on site http://slis-two.lis.fsu.edu/~log/.*
6. Glenn Fleishman, Web Log Analysis: Who's Doing What, When? *Web Developer® magazine*, Vol. 2 No. 2 May/June 1996 ©, 1996.
7. Bauer, K. Who goes there? 2000, January, *Online Magazine, 24,* pp. 25-31.
8. Karen A.Forcht and Kevin Cochran, Using data mining and data warehousing techniques, MCB University Press. *Industrial Management & Data Systems* 99/5, 1999, pp. 189-196.
9. R.Cooley, B.Mobasher, J.Srivastava. Web Mining: Information and Pattern Discovery on the World Wide Web.