Proceedings of the 6th WSEAS Int. Conf. on EVOLUTIONARY COMPUTING, Lisbon, Portugal, June 16-18, 2005 (pp76-79)

Data Mining Technique for Collaborative Server Activity Analysis

JELENA MAMČENKO, REGINA KULVIETIENĖ
Information Technology Department
Vilnius Gediminas Technical University
Saulėtekio av. 11, Vilnius
LITHUANIA
http://gama.vtu.lt

Abstract: - Most people, companies and organizations put information on the Web because they want it to be seen by the world. Their goal is to have visitors come to the site, feel comfortable and stay awhile. The goal of Vilnius Technical University Information Technology department is constructive and convenient web site for study purposes, gathering useful information, finding free workplaces in labour exchange, electronic library and others.
To make know better web site visitors needs we have analysed access log file using modern Data Mining technique. This article treats data mining possibility for web site analysis.

Key-Words: - Data Mining, Clustering, Intelligent Miner for Data, Intelligent Analysis, Demographic Clustering, Access log

1 Introduction

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

Nowadays, the use of Data mining is widespread in any industry. In common, everywhere where 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 unstructed 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. Analyzed server *gama*, LDAP directory is included, is a part of Distance Education Information System at Vilnius Gediminas Technical University (Fig. 1). Together with it there are *kappa, teta, irma a*nd *beta* collaborative servers.

2 Log file analysis using Data Mining technique

Server log files are records of web server activity. They provide details about file request to a web server response t those request. In the access file which is the main log file, each line describes the source of 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 describes 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:

2004.11.29 11:35:42 CN=Jelena Mamcenko/O=Vtu/C=LT 193.219.147.87 GET /mail/jm.nsf

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

For data analysis we've used IBM Intelligent Miner for Data software which supports not only mining but statistics functions as well [8]. For data analysis was selected demographic clustering function, which searches characteristics that most frequently occur in common, and groups the related records accordingly. The results of the clustering function contain the number of detected clusters and the characteristics that make up each cluster. Demographic clustering provides 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.

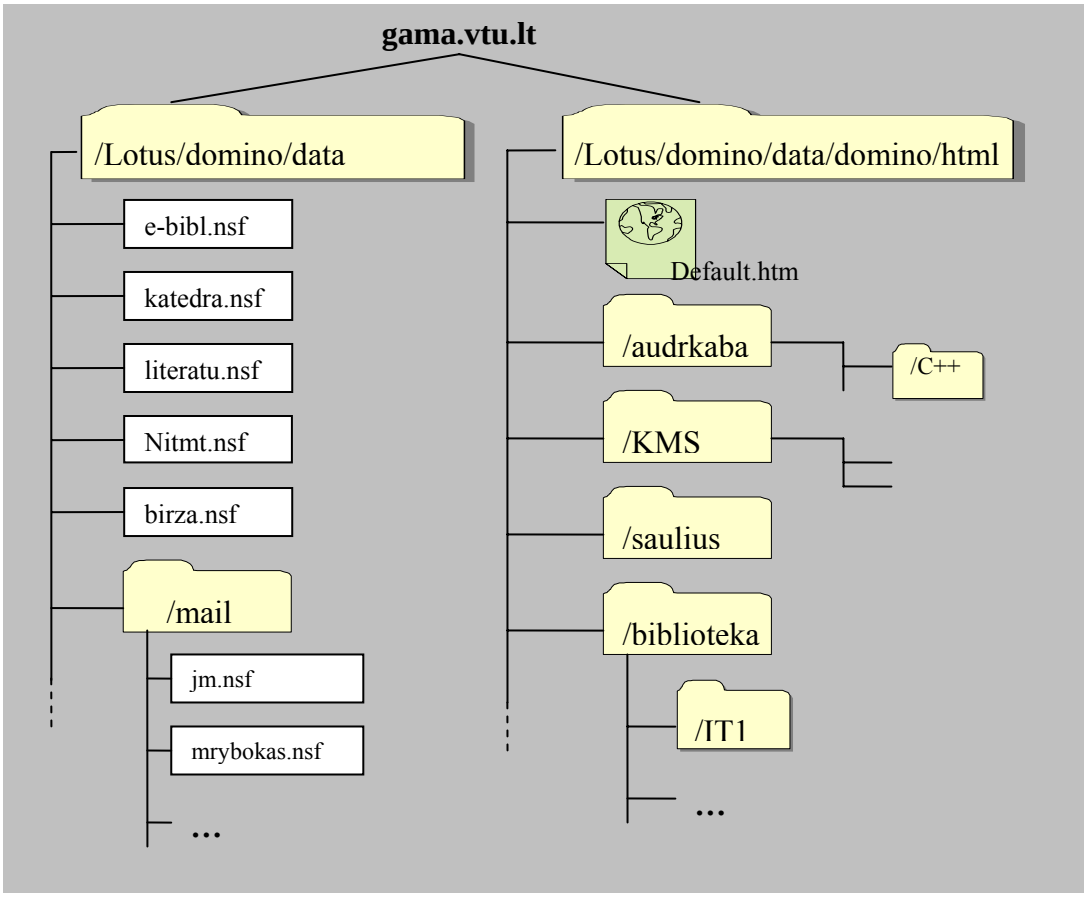

Fig 1. Web site structure

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 database file exported to flat file format with historical data range from 2004 November 29 11:09:08 to 2005 April 12 16;36:23 with additional fields shown in table 1.

However, some problems with data transformation were occurred. As we know, Intelligent Miner for data requires database or flat file data format. But it can't take any 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.

2.1 The mining run tasks

There are five phases of data mining tasks:

- Defining the data. We specify a data object that points to a flat file (Table 2).
- Building the model. Define a demographic clustering settings object. This model contains information that describes the clusters identified during the mining run.
- *Applying the model.* Define a demographic 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 combine several mining tasks into one.
- *Analyzing the results.* Define a bivariate statistics function. This function analyzes the data object and produces an output object, flat file and result object.

Table 1. Fields which are used

Date	The format is year, month, day (yyyy:mm:dd)
Time	Hour 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, PUT, SEARCH or PUT
Request	Is the path and file retrieved

Table 2. Domlog.nsf

Date	Time	Authenticated User	IP Address	Method	Request
2004.11.29	11:09:18	-	193.219.180.226	GET	/mobilus/nokia.htm
2004.11.29	11:35:42	CN=Mindaugas Rybokas/O=Vtu/C=LT	193.219.147.87	POST	mail/mrybokas.nsf/vwdrafts.gif!OpenImageResource
…	…	….	…	…	…
2005.04.01	10:51:16	-	193.219.146.95	POST	/orc/up.htm
2005.04.01	10:51:41	-	65.54.188.85	GET	/robots.txt

2.2 Generated results

The result generated by mining function (IBM Intelligent Miner for Data) are shown in table 3.

This table shows nine rows, each represented one of the nine clusters identified by mining run. The numbers down the left side represents 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.

Table 3. Textual interpretation

Name	Size	Characteristics
[6] 5	25,88%	Date is predominantly 2005.02.08, User is predominantly -, [Protocol] happens to be predominantly HTTP/1.1 , [Method] happens to be predominantly GET, IP Address is predominantly 193.219.145.99, File is predominantly / and Time is predominantly 18:06:21.
[7] 6	15,52%	[Protocol] happens to be predominantly HTTP/1.1 , Date is predominantly 2005.01.13, User is predominantly -, [Method] happens to be predominantly GET, IP Address is predominantly 65.54.188.98, File is predominantly /birza.nsf/header!OpenPage and Time is predominantly 17:15:24.
[3] 2	12,04%	Date is predominantly 2004.12.04, [Protocol] happens to be predominantly HTTP/1.1, User is predominantly -, [Method] happens to be predominantly GET, IP Address is predominantly 62.80.224.226, Time is predominantly 20:17:52 and File is predominantly /birza.nsf/navigation!OpenPage.
[8] 7	10,35%	[Protocol] happens to be predominantly HTTP/1.1 , Date is predominantly 2004.12.21, User is predominantly -, [Method] happens to be predominantly GET, IP Address is predominantly 193.219.145.200, File is predominantly /default_files/image001.jpg and Time is predominantly 13:43:06.
[9] 8	9,94%	[Protocol] happens to be predominantly HTTP/1.0 , Date is predominantly 2004.12.15,

		User is predominantly -, IP Address is predominantly 193.219.146.95, [Method] happens to be predominantly GET, Time is predominantly 09:57:57 and File is predominantly /domjava/view.properties.
[5] 4	8,62%	User is predominantly CN=Rytis Lastauskas/O=Vtu/C=LT, [Method] happens to be predominantly GET, [Protocol] happens to be predominantly HTTP/1.1 , Date is predominantly 2005.02.02, IP Address is predominantly 212.59.0.201, File is predominantly /domjava/view.properties and Time is predominantly 13:43:11.
[2] 1	8,56%	User is predominantly CN=Mindaugas Rybokas/O=Vtu/C=LT, [Protocol] happens to be predominantly HTTP/1.1 , [Method] happens to be predominantly GET, Date is predominantly 2004.12.15, IP Address is predominantly 84.32.61.190, File is predominantly /icons/ecblank.gif and Time is predominantly 10:16:38.
[4] 3	5,41%	[Protocol] happens to be predominantly HTTP/1.0 , Date is predominantly 2005.02.18, User is predominantly -, [Method] happens to be predominantly GET, IP Address is predominantly 65.54.188.100, File is predominantly /audrkaba/stud/index.html and Time is predominantly 17:01:15.
[1] 0	3,68%	[Protocol] happens to be predominantly HTTP/1.1 , Date is predominantly 2005.03.01, [Method] happens to be predominantly GET, User is predominantly -, IP Address is predominantly 193.219.146.90, File is predominantly /MRybokas and Time is predominantly 20:14:45.

3 Conclusions

There are a lot different log files analysing tools such as Accure Software, Sane Solutions, WebTrends and etc. 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 for your sharply formulated questions whereas Data mining gives 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 powerful tool for effective analysis and tasks optimization.

With the combination of Domino and the intelligent analysis tools, we have several options for analyzing 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 the 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 access log flat file, indicating date, time, user, IP address, method and requests by different users. We have analyzed 520 636 records (about 137 Mb) of access log file using demographic clustering method and it took 3:08:38 hours. The results are presented.

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, 41-46.

[4]. Gavin Meggs, Internet Usage Analysis, *Handbook of Data Mining and Knowledge Discovery,* University press. Oxford 2002, 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,* 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, 189-196.