Internet User Behavior Model Discovery Process

Dragoş Marcel VESPAN, dragos.vespan@ie.ase.ro

The Academy of Economic Studies has more than 45000 students and about 5000 computers with Internet access which are connected to AES network. Students can access internet on these computers through a proxy server which stores information about the way the Internet is accessed. In this paper, we describe the process of discovering internet user behavior models by analyzing proxy server raw data and we emphasize the importance of such models for the e-learning environment.

Keywords: *Internet, User Behavior, e-Learning.*

Introduction

Students access the Internet from inside AES by using about 3000 computers at the same time, all this access being recorded by the proxy server cache.ase.ro in squid log files. A proxy server acts like an intermediary caching level between the client browser and the web server. Proxy caching can be used to minimize a requested webpage loading time and also to reduce network traffic load both at client side and at server side. The performance of a proxy server depends on its ability to correctly predict future page requests. Proxy logs can reveal HTTP requests of more users to more web servers. This can be used as a data source for identifying browsing behavior of anonymous user group using the same proxy server.

Squid is a caching proxy server for the web clients which has very high performance and supports HTTP, gopher and FTP protocols. The log data recorded by squid server contains information about the IP of workstation which made the request, the URL requested, date and time, HTTP response, size and duration of the request.

The process of discovering internet user behavior models

Figure 1 presents the process of discovering internet user behavior models from raw log data. The input is represented by the raw log data, the didactic activity schedule, the location of the workstations and visited website files. The output is represented by the models of internet user behavior discovered from the log data.

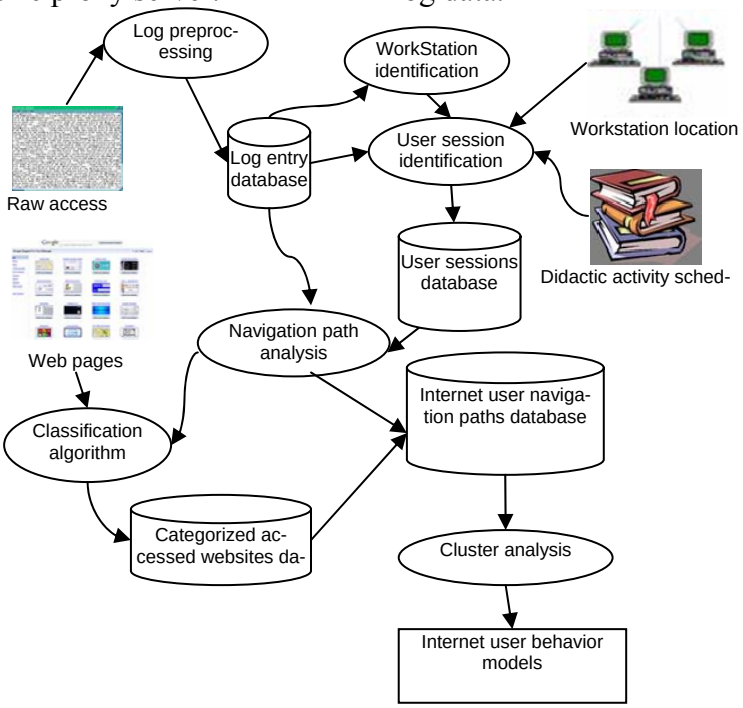

Fig. 1. Tasks of discovering internet user behavior models process

Server log filtering techniques which remove unimportant items are very important for any web log analysis. The discovered associations or reported statistics are useful only if the data represented by the log server offers a precise image of the user access on websites. HTTP protocol requests a separate connection for each file requested from the web server. This way, a user request to see a certain page often results in more log entries because images and scripts are loaded in addition to HTML files. In most cases, only the log entry of the HTML file request is relevant and should be kept for the user session database. This happens because, usually, a user does not explicitly requests all the images on a web page, but these are automatically loaded due to HTML tags. Because the main purpose of web analysis is to get an image of internet user behavior, it makes no sense to keep file requests that the user didn't explicitly make. Removal of these elements can be done by verifying URL's suffix. For instance, all log entries with file extension suffix like JPEG, GIF, JPG, PNG can be removed. In addition, the common scripts like „count.cgi" can also be removed.

For weblogs which represent longer time periods, it is very alike that a certain workstation requests access to a website more times. The purpose of session identification is to divide page accesses of each workstation in individual sessions. For the workstations identified to be located in the seminar classes, this is done accordingly to the didactic activity schedule, considering that the user session ends the same time as the didactic activity. For the other workstations, this is done by using inactivity time period. If this period is greater than a threshold (most log analyzers use a 30 minutes threshold) it may be assumed that another user session has started.

In order to identify user access models, the log entries must be processed separately for each user. Instead of processing the log file according to date and time, first, the log data is sorted by workstations first. This allows analyzing traffic from a single workstation at a certain time and gathering general statistics after processing each entry. The entries in the log database will be sorted by the workstation ip and by time stamp and then will be processed in order to get user sessions.

Semantically speaking, a user session can be defined as the set URLs accessed by a certain user (on the same workstation) with a certain purpose. We do not try to guess the purpose of the user but consider some heuristic aspects, like the length of the reference which is based on the assumption that the time spent by a user to examine an object is correlated to the interest of the user for the content of that object. On this basis, a model of a user session is obtained by differentiating between navigational objects (e.g. which contain only links the user is interested in) and content objects (e.g. which contain the information the user is looking for). The distinction between navigational and content accesses is related to the time distance between a request and the following one. If between two accesses A and B there is a time difference greater than a threshold, than A can be considered to be a content URL; otherwise a navigational URL.

Instead of using raw logs for internet user behavior models discovery, each log is converted into integer tuples and key words. Each user will be associated with an integer unique identifier and time stamps will be converted to integers (seconds passed from the beginning of the epoch). Each URL will be represented by a key word which will represent the accessed domain.

Tuples representations are built by a single parsing of the filtered logs. The association of users with unique identifiers is done by using a hash table for each user. The representation of URLs by keywords will be done by using an association table between keywords and the site of the accessed URL. For instance, the presence of words „job" or „career"/"cariera" inside the url or the accessed website name could reveal a job seeking oriented internet user behavior, words like „seminar", „ase", "biblioteca"/"library", „.edu" could reveal a study oriented behavior and words like „fun", „haz", „entertain" could describe an entertainment oriented be-

havior

In order to identify rules in internet behavior user it is very important to determine the domain of the visited websites. According to Sven Mezer zu Eissen and Benno Stein, the most frequent information searches on WWW are on didactic material, shopping and products information, consultancy (forums), entertainment (music/games/movies/comic materials/news), downloads, health and programming.

The requested page type of content is often revealed by the URL field of server logs but, sometimes, the access to the information on the addressed page is necessary in order to determine its content type. In order to run content exploration algorithms, the information must be converted into a quantifiable format. For this, the vector-space model is used, where files are divided in word vectors. Images and multimedia can be substituted by keywords or text descriptors. The content of a static web page can be easily preprocessed by analyzing the HTML code and reformatting information by running additional algorithms.

The internet user navigation paths database is built on the basis of identified navigation path in which the visited urls are replaced by the category of the website the urls point to. Cluster analysis allows grouping navigation paths with similar characteristics. Clustering algorithms aim to dividing the set of objects into clusters where the objects in each cluster are similar to all the objects in the same cluster (but non-similar to objects from other classes). Objects that do not fit in any of the detected clusters are considered to belong to a special cluster formed by exceptions.

Cluster analysis represents the last step in web usage mining. The reason behind this analysis is that of identifying disjunctive internet user behavior models.

The analysis of the squid log obtained from the cache.ase.ro proxy server revealed, according also to Sven Mezer zu Eissen and Benno Stein theory, six basic models of internet user behavior:

- users interested in didactic material (study and research) where the visited sites are mainly sites of universities
- users interested in daily life where visited sites are mainly on-line newspapers, news sites and sites with information about how to spend the spare time
- users interested in shopping and products/services information, which mainly visit on-line auctions, on-line shops and services sites
- users interested in on-line communication, which visit mainly mail websites, on-line communities, forums and other entertainment sites
- users interested in finding a job, where the mainly visited sites are those which post job offers
- users which use on-line chat applications.

Conclusions

The continuous development of on-line education tools and methods requires that the education management focuses on the on-line behavior of students. The behavior models obtained through the described process represent a starting point for analyzing what kind of information is interesting for the students browsing on the internet and how the educational management can meet this need of information.

References

[1] Jan Kerkhofs, Koen Vanhoof, Danny Pannemans – Web Usage Mining on Proxy Servers: A Case Study, 2001

[2] Paulo Batista, Mario J. Silva – Mining Web Access Logs of an On-line Newspaper, 2002

[3] Risto Vaarandi – A Data Clustering Algorithm for Mining Patterns From Event Logs, 2003

[4] Soumen Chakrabati – Mining the Web. Discovering Knowledge from Hypertext Data, 2003

[5] Ian H. Witten, Eibe Frank – Data Mining Practical Machine Learning Tools and Techiques, Second Edition, 2005, Morgan Kaufmann Publishers