ISSN 0146-4116, Automatic Control and Computer Sciences, 2007, Vol. 41, No. 4, pp. 232–235. © Allerton Press, Inc., 2007.
Original Russian Text © R.M. Alyguliev, F.F. Yusifov, 2007, published in Avtomatika i Vychislitel'naya Tekhnika, 2007, No. 4, pp. 73–77.

# Development of a Model for Studying Web User Activity by an Analysis of Web Traffic

**R. M. Alyguliev and F. F. Yusifov**

*Institute of Information Technologies, National Academy of Sciences of Azerbaijan,*
*ul. F. Agaev 9, Baku, Az-1141 Azerbaijan*
*e-mail: farhadyusifov@gmail.com*

Received January 12, 2006; in final form, February 7, 2007

**Abstract**—The article considers methods of intelligent data analysis (data mining) used in problems involved in the analysis of Web traffic, and also considers the application of the method of cluster analysis and a newly developed model for the study of Web user activity.

**DOI:** 10.3103/S0146411607040074

**Key words:** Web traffic, log file, cluster analysis

## 1. INTRODUCTION

An intelligent method of analysis of Web traffic is considered to be one possible approach for determining the level of utilization of information resources and for knowledge extraction [1–3]. It may be characterized briefly in the following way. At the present time enormous information arrays have been accumulated by different Web servers. The problem is that it is difficult to determine and analyze the useful information. The absence of any possibility of obtaining the most necessary and most complete information according to a particular criterion makes most of the accumulated resources quite useless. Since the study of a particular problem requires very great labor costs for the direct discovery and analysis of useful information on a particular topic, many solutions are adopted on the basis of an incomplete representation of the problem [3, 4].

The log files present on a server have always been the basic source of information about the traffic passing through a Web site and about user behavior. The overwhelming majority of tools for monitoring and analyzing Web sides is based on the use of these files [1, 3]. Such parameters as the volume of traffic, addresses of site visitors, methods of exiting from the site, and the response to the content of the site are evaluated on the basis of data on calls to the given site. However, the need to obtain more extensive information about Web sites than just the basic information, moreover, information that is also more detailed and more reliable, is becoming increasingly more urgent.

Through the use of methods of intelligent data analysis, or "data mining," it is possible to reduce the number of man-hours expended in extraction of data, represented in the log file of Web servers. The application of intelligent methods makes it possible to find and analyze useful information, while the use of methods of cluster analysis makes it possible to study Web users.

## 2. FORMALIZATION OF PROBLEM

With the increase in the popularity of the World Wide Web, Web sites are playing an important role in communicating knowledge and information to users. The discovery and processing of information is an important problem for determining an effective marketing strategy and optimal utilization of a Web server [2, 5]. Most tools that have now been created tor the analysis of Web traffic supports the collection of statistical information without extraction of useful knowledge for Web managers. The analysis of useful information is being spurred on due to the large and growing volume of Web traffic.

It is possible to identify the user through the use of information that has accumulated in a log file, such as the date and time of a user visit, the IP address of the user computer, the name of the user browser, the URL of the user-requested page, the user referer, etc. [5].

By observing the traffic on a server, we may determine whether the visitors encountered any errors, from which reference they encountered an error, etc. If a site visitor was redirected to us by a computer search,

we can see precisely which computer and which key words were used. Such information may prove to be very opportune when planning an advertising campaign or in the development of new content [4–6]. However, it is extremely tiresome to sit for hours by the monitor and analyze information by means of instructions. The use of professional analyzers of log files is, of course, a better alternative, one that, moreover, assures a sufficiently detailed analysis of Web traffic.

New tools for extraction of data utilize such complex methods as data mining [7, 8]. The methods of data mining represent a new scientific trend, one that has appeared and has developed on the base of advances in applied statistics, methods of artificial intelligence, database theory, and a number of other fields.

Most of the program tools that have recently appeared for the analysis of traffic make it possible to filter and obtain statistical information about users. Such toolkits help determine the number of calls to different files and servers and the addresses of individual users. Moreover, such systems are designed for a small or limited flow of data and rarely make it possible to perform an analysis of the relationship between calls to files and the underlying logic of the location of these files.

In the present study the method of clustering will be used to analyze Web traffic and study user activity. For this purpose, statistical data on the daily activity of each user over the course of a week are processed. Using these data, the following mathematical model may be constructed:

$$\text{user}_i(w_{i1}, w_{i2}, w_{i3}, w_{i4}, w_{i5}, w_{i6}, w_{i7}), \quad i = \overline{1, n}, \tag{1}$$

where $\text{user}_i$ is the $i$-th user and $w$ the weight of the activity of the $i$-th user on each day of the week.

Using the method of clustering, it becomes possible to partition users into clusters in terms of daily activity, to analyze user activity, and to arrive at effective decisions.

## 2. METHODS OF CLUSTER ANALYSIS

Clustering of Web users is the basic domain of application of methods of cluster analysis in Web usage mining [9, 10]. The user is allocated to one of several categories, after which information inferred for this user is appropriately varied. Yet another domain of application that is traditional for clustering is decision support in data analysis.

Methods of cluster analysis are used to study and analyze the user data of Web servers. Of particular interest in the present study is the selection of the subject of clustering. The rejection of the traditional approach is due to the fact that difficulties arise when making a choice of metric; a second factor is that the volume of user data is excessively large. In clustering it is always necessary to solve two problems, first, that of choosing a metric and, second, choosing an algorithm. The choice of a matric is the basic problem in clustering of user data. For a number of reasons, classical Euclidean metrics often prove to be inefficient.

Existing approaches cluster Web users and are fundamental in the analysis of user sessions. They ignore the dynamic nature of user data. In the present article we will concentrate on the discovery of new knowledge that is basic in the development of new methods (clustering of Web users).

Introduction of the concept of resemblance of objects derived from observable variables is the basic problem in creating a classification. Objects that possess similar characteristics must be placed in the same cluster. The concept of a metric is introduced to arrive at a quantitative assessment of similarity. The degree of similarity or difference between objects is determined by the metric distance between them. Different measures of distance between objects are used.

The choice of a measure of distance and of weights for classified objects is a very important stage of cluster analysis. This is because the composition and number of clusters that are created as well as the degree of similarity of objects depend on the particular measure and the particular weight that are chosen.

There are many different methods of clustering. The most commonly used methods are different modifications of the $K$-means method [8, 11]. In the present study we will use the $K$-means method. On the first stage the method will be used to identify groups of users in terms of daily activity. On the second stage the daily activity over the course of a week will be determined with respect to each of the obtained groups.

## 3. CLUSTERING OF WEB USERS

One of the most interesting methods of analysis of log files is that of clustering of Web users [10, 12, 13]. Clustering of Web users involves partitioning the set of users into clusters using Web access logs so that the users that fall within the same cluster are considerably more similar to each other than to any of the users of the other clusters. Clusters may assist in the analysis of Web content and provide a guide on operational transformation of the content of a Web site. It is also worth noting that implementing clustering of Web users

may provide the Web administrator with information about the behavior of users, their range of interests, and information about frequently queried Web resources.

Hierarchical methods of cluster analysis are not suitable where there is a large volume of data. In such cases nonhierarchical methods based on partitioning are used; these include iterative methods of fragmenting the initial population. In the course of partitioning, new clusters are formed until a halting rule is satisfied. The *K*-means algorithm constructs *k* clusters situated at the greatest possible distances from each other. The basic type of problem which this algorithm may be used to solve concerns the existence of hypotheses concerning the number of clusters, moreover, these clusters must be as different as possible. The choice of the number *k* may be based on results of preceding studies, theoretical considerations, or intuition.

Using the *K*-means clustering method, sets of users may be partitioned in terms of daily activity into clusters thus:

$$\text{User} = \bigcup_{i=1}^{n} \text{user}_i, \quad C_k \cap C_p = \varnothing,$$

$$C_1 \Rightarrow (\text{user}_1, \text{user}_2, \ldots, \text{user}_{n_1}).$$

Using this technique, the activity of users may be analyzed and, using the number of users in each cluster, the daily activity may be determined, as shown below:

$$\frac{\sum_{i=1}^{n_k} w_{ij}}{n_k} = W_j^k, \quad j = \overline{1,7}, \quad k = \overline{1,q},$$

where $W_j^k$ is the day of the week and *k* the number of clusters.

An implementation of the algorithm by the *K*-means method may be demonstrated in the following form:

(1) Initial distribution of objects (users) into clusters.

Select points *k*; these are to be considered the "centers" of the clusters.

As a result, each object is assigned to a definite cluster.

(2) Iterative process.

Compute the centers of the clusters, which are then and subsequently considered the coordinate centers of each cluster. The objects are again redistributed.

The process of computing the centers and redistributing objects continues until one of the following two conditions is satisfied:

• cluster centers are stabilized, i.e., all observations belong to the cluster to which they had been assigned prior to the current iteration;

• the number of clusters is equal to the maximal value.

After the results of a cluster analysis by the *K*-means method have been obtained, it is necessary to check that the clustering is correct (i.e., to estimate the extent to which the clusters differ from each other). For this purpose, the mean value is computed for each cluster. For a good clustering, strongly different means for all the measurements or for at least most measurements must be obtained.

The following are among the advantages of the *K*-means algorithm [11]:

• simplicity and speed of use;

• intelligibility and transparency of algorithm.

Among the drawbacks of the *K*-means algorithm, we may note the following:

• the algorithm is excessively sensitive to outliers, which may distort the mean. One possible solution is to use a modification of the algorithm, called the *K*-median algorithm.

• the algorithm may function slowly on large databases. One possible solution is to use data samples.

It should be noted that besides the use of data mining for the analysis of Web server data, there exist other, traditional analytic techniques oriented towards the graphic representation of data of server operation. These techniques provide a representation of the quantitative characteristics, e.g., number of visitors or number of calls to a particular document, and may be considered the simplest techniques for the representation of reports. But these techniques do not supply information about possible laws, including hidden laws.

## 4. CONCLUSION

There is much attention being devoted today to the creation of methods and computational techniques for the analysis of log files, monitoring of network activity, and on tracking viruses and hacker programs. In the present study a model was developed for the analysis of the activity of Web users through the use of statistical data of a Web server.

The widespread use of the World Wide Web continues to increase, and, therefore, there is a need for the development of a technique and toolkits for the detection and processing of information. The basic software problem for the analysis of Web traffic is that of extracting useful information from server log files. For this purpose, it is necessary to create professional Web analyzers. It should be noted that the state of Web analyzers at present does not completely satisfy the above requirements. The development of models for studying the activity of Web users may assist in the creation of professional analyzers.

## REFERENCES

1. Bartolini, G., Web Usage Mining and Discovery of Association Rules from http Servers Logs, www.prato.linux.it/~gbartolini/en/view-a/2/pdf/wum.pdf (2001).
2. Baglioni, M.,Ferrara, U., Romeil, A., Ruggieri, S., and Turini, F., Preprocessing and Mining Web Log Data for Web Personalization, www.di/unipi.it/~ruggieri/Papers/aiia2003.pdf, 2003.
3. Wanga, X., Abrahamb,A., and Smitha, K.A., Intelligent Web Traffic Mining and Analysis, *J. Network Comp. Appl.*, 2004, vol. 28, p. 147–165.
4. Rabin, D., Study Log Journals, *Seti i Sist. Svyazi*, no. 1 (121); http://ccc.ru/magazine/depot/05_01/read.html?htm, 2005.
5. Ivancsy, R. and Vajk, I., Different Aspects of Web Log Mining, in *6th Intern. Symp. Hungarian Researchers on Computational Intelligence,* Budapest, Nov., 2005.
6. Ivancsy, R. and Vajk, I., Frequent Pattern Mining in Web Log Data, www.bmf.hu/journal/Ivancsy_Vajk_5.pdf, 2006.
7. Chakrabarti, S., Data Mining for Hypertext: A Tutorial Survey, *SIGKDD: SIGKDD Explorations: Newsletter Special Interest Group (SIG) on Knowledge Discovery and Data Mining*, ACM, vol. 1, no. 2, pp. 1–11, 2000.
8. Dyuk,V.A. and Samoilenko, A.V., *Data Mining, Uchebnyi Kurs* (Data Mining. Textbook), St. Petersburg: Piter, 2001.
9. Kosala, R. and Blockeel, H., Web Mining Research: A Survey, *SIGKKD Explorations*, vol. 2(1), July, 2000.
10. Fu, Y., Sandhu, K., and Shih, M.Y., Clustering of Web Users Based on Access Patterns, in *Proc. ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining (KDD(99) Workshop on Web Mining)*, San Diego (U.S.), vol. 5, pp. 560–567, 1999.
11. *Metody klusternogo analiza* (Methods of Cluster Analysis), http://www.intuit.ru/department/database/datamining, 2006.
12. Morzy, T., Wojciechowski, M., and Zakrzewicz, M., Web Users Clustering, in *Proc. 15th Intern. Symp. Comp. Inform. Sci.*, Istanbul (Turkey), pp. 374–382, 2000.
13. Xie, Y. and Phoca, V.V., Web User Clustering from Access Log Using Belief Function, *Proc. 1st Intern.Conf. Knowledge Capture (K-CAP 01)*, ACM Press, 2001, pp. 202–208.