Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008

# INVESTIGATION OF INTERNET SYSTEM USER BEHAVIOUR USING CLUSTER ANALYSIS

**DARIUSZ KRÓL, MICHAŁ ŚCIGAJŁO, BOGDAN TRAWIŃSKI**

Wrocław University of Technology, Institute of Applied Informatics, Wyb. Wyspiańskiego 27, 50-370 Wrocław, Poland
E-MAIL: dariusz.krol@pwr.wroc.pl, bogdan.trawinski@pwr.wroc.pl

**Abstract:**

**The method of the investigation of information web system users' activity using a clustering method is presented in the paper. On the basis of a web server log, anonymous sessions are determined in the form of a 65 dimensional vector, where dimensions represent individual web system pages. Each dimension comprises the value of a measure of user interest in a page during a given session. This value is calculated as a ratio of time user spent visiting a given page to the total time of a session. Then the whole set of sessions is clustered using HCM (Hard C-Means) algorithm. The resulting clusters are assumed as the user activity patterns and among them clusters dominated by a page are selected as those where the user interest value exceeds a given threshold value e.g. 50 per cent. The sessions of named users, registered in the system, are determined using an application log of user activity. The frequencies of named user sessions, comprised by individual clusters, are calculated for a given period of time e.g. one month. The user activity can be assessed by analyzing frequencies obtained. For example, the user behavior can be regarded as deviated from normal pattern when the frequency of a session in a cluster dominated by a page is below a determined threshold value e.g. 10 per cent. The method was evaluated using data from a cadastral web system exploited in an extranet.**

**Keywords:**

**Web system; clustering; user activity; server log; HCM algorithm**

## 1. Introduction

Researchers are seeking ways to know more about similarities and patterns in users' Web behavior to make the Web more friendly. For better user understanding, cluster analysis is used [4]. The user activities are grouping on the basis of the similarity metrics.

A number of clustering approaches have been proposed in the literature. For example, clustering techniques have been used to group queries by semantic [12] and to group users by activity patterns [7]. There are three main approaches. First technique uses page URLs to construct a hierarchy which is then used to categorize the pages. Second utilizes the time spent on a page and Longest Common Subsequences [5] to cluster the user sessions. Third proposes the combination of user profile and standard clustering algorithms. Most relevant to our work is the last one.

A few methods for automatically determining the number of sessions have been proposed [3]. Studies of users' behavior based on log analysis usually begin with an assumption that a time threshold can be set to separate an individual user's sessions [9]. Other approaches to session identification take into account query similarity by changes in terms. However, users often take a break from a work task.

## 2. The real estate cadastre system

Cadastre systems are mission critical systems designed for the registration of parcels, buildings and apartments as well as their owners and users and they are comprised by the governmental information resources. Those systems have complex data structures and sophisticated procedures of data processing. They can be constructed in client-server architecture for LAN as well as in Web technology to be used in intranets and extranets.

The cadastral system, which users were investigated is an internet information system designed for the retrieval of cadastral data and is complementary to the main system in which cadastre database is updated. The system is deployed in intranets and extranets in local governments. Using the internet information system mainly rests on formulating queries, browsing the list of retrieved objects, choosing the objects to reports and generating reports in PDF format (see Figure 1).

978-1-4244-2096-4/08/$25.00 ©2008 IEEE

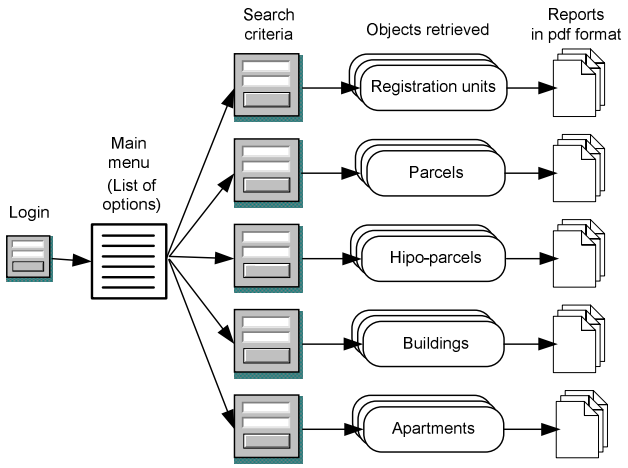


Figure 1. Generalized schema of the internet cadastre information system

The users of the system are the members of an organization e.g. a local government or a corporation and fulfill their everyday duties frequently use information obtained from the system. For some days they focus on specific topics and after completing one task they move to another one and change their topics of interest. The access to the system is limited. Each user should be registered in the system and the rights should be assigned to the data from a given territory. The users of the system are the workers of local governments who utilize data to prepare administrative decisions, to inform real estate owners and to prepare reports for management boards of local governments.

## 3. The method of user activity investigation

The method presented in the paper is based on data gathered in a web log and an application log of an internet cadastral system in one of bigger cadastral information centre in Poland. Standard Microsoft's ISS log covering the period of whole two years 2004 and 2005 was analyzed. Data comprised IP user's address, date and time of request, and name of resource requested. Web server log was main source of data to study behavior of cadastral system users and was used to determine users' anonymous sessions. In Figure 2 the number of requests of the most popular pages is shown. Some pages were introduced to the system in 2005; therefore they have usage frequency bars only for that year.

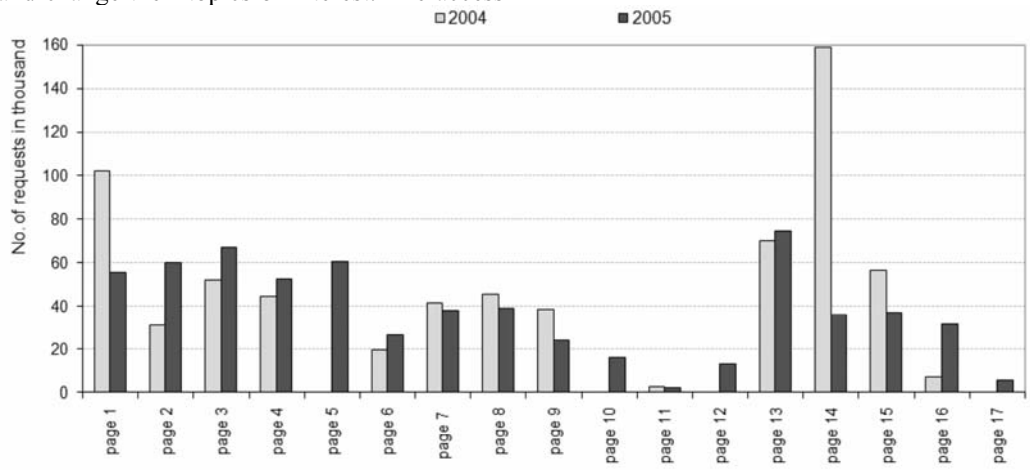


Figure 2. Generalized schema of the internet cadastre information system

On the basis of a web server log, anonymous sessions are determined in the form of a 65 dimensional vector, where dimensions represent individual web system pages. Each dimension comprises the value of a measure of user interest in a page during a given session. This value is calculated as a ratio of time user spent visiting a given page to the total time of a session. In the case of anonymous sessions a user is represented by an IP address. The number of dimensions of data space analyzed equals to the number of web pages visited by the named users when accessing

the cadastral internet system.

The anonymous session was defined as a series of requests called from one IP address when a gap between two subsequent requests was no longer than given time. In our investigation we experimentally determined this time equal to 30 minutes. Creating an anonymous session we did not trace the sequence of pages requested but only summed time a user visited individual pages. So the sequence of pages accessed was neglected in our approach. It enabled us to determine the measure of user interest in a page during a given session, expressed by the following formula (1):

$$SPI_i = \frac{\sum_{j=1}^{V_i} T(i,j)}{\sum_{k=1}^{P}\sum_{l=1}^{V_k} T(k,l)} \quad (1)$$

where:

$SPI_i$ - user interest in a page during a single session,

$V_i$ – number of visits of i-th page during a single session,

$T(i,j)$ – time of j-th visit of i-th page during a single session,

$P$ – number of all pages in a system,

$V_k$ – number of visits of k-th page during a single session,

$T(k,l)$ – time of l-th visit of a k-th page during a single session.

Then the anonymous session is represented by a 64-dimensional vector where dimensions correspond to individual pages and comprise values of the measure of user interest in a page during the session. Above 24 thousand anonymous sessions were detected in the web server log. The whole set of sessions was clustered using a HCM algorithm [3] and the best value of Dunn's index was obtained for 20 clusters. The resulting clusters were assumed as the user activity patterns and among them clusters dominated by a page were indicated as those where the user interest value exceeded a given threshold value e.g. 50 per cent.

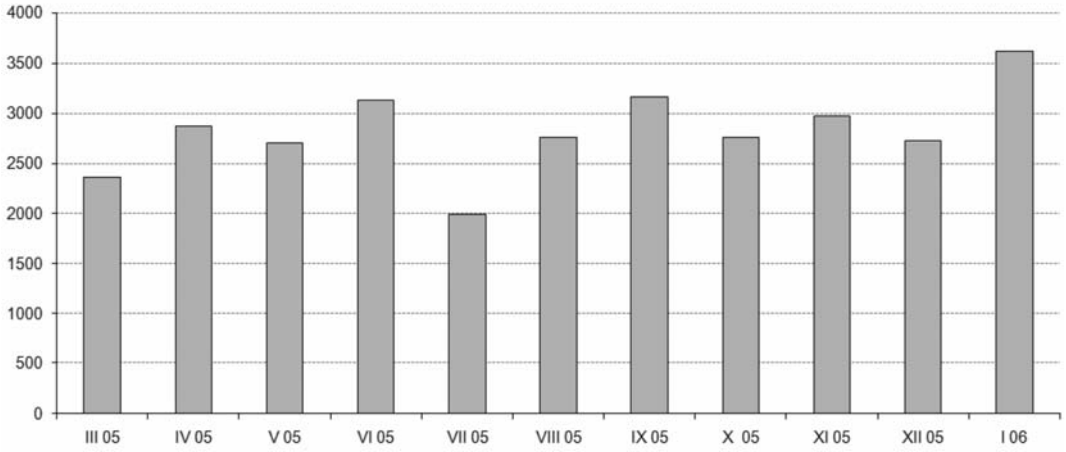


Figure 3. Monthly users' logging frequency from March 2005 to January 2006

In order to extend the monitoring possibilities of users' behavior an application log was introduced to the cadastral system. Data gathered in this log were registered in six relationally bound tables and comprised among others user identifiers, date and time of logging, date and time of accomplishing a search operation, number and localization of objects retrieved or contained in reports. In Figure 3 monthly frequency of users' logging is presented. Application log was developed in January 2005 so that data presented cover the period from March 2005 to January 2006.

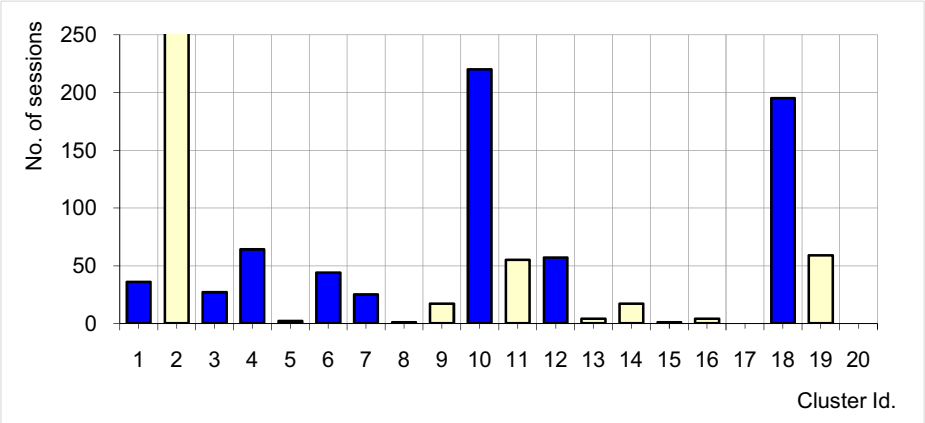


Figure 4. Number of named user sessions from November 2005 to January 2006 classified into 20 clusters

In Figure 4 the number of named user sessions from November 2005 to January 2006 falling into individual session clusters is presented. Dark color bars indicate clusters dominated by a page whereas light color ones the clusters not dominated. About half of clusters are dominated by a page. These dominating pages are:

- list of reports and the report managing – clusters 6, 8 and 20;
- generating the standard extracts from land, building or apartment registers – cluster 10;
- generating the report of parcel owners – cluster 18;
- main pages, error pages – clusters 1 and 7;
- generating lists of parcels – clusters 3 and 12;
- browsing registration units – cluster 4.

The analysis of the history of user session classification into clusters enables to detect whether the user, who was acting in a determined way, has not devoted his time to the job. The deviation from normal, repeatable work should be checked by a supervisor. For example, the user behavior can be regarded as deviated from normal pattern when the frequency of a session in a cluster dominated by a page is below a determined threshold value e.g. 10 per cent. In Figure 5a the sessions of the user 127 classified into clusters are presented. This user has a uniform characteristic of his work. However the cluster 1, dominated by a page, can be neglected because the page it comprises only system start pages which are not important for user activity analysis. Majority of time the user spent working with pages qualified to the clusters 19 and 10. Both behavior patterns are connected with the page for generating standard extracts from registers. The cluster 10 is dominated by a page but the percentage of user sessions is greater than a determined threshold 10. The other clusters are not dominated by any page.

In turn Figure 5b shows data for the user 91, who is engaged mainly in generating the extracts from system registers. Three sessions assigned to the cluster 12 should be checked. Probably the reports were generated which were not normally used by the user. Maybe he only changed the characteristics of his work, but it is also possible he retrieved data which were not required by his everyday duties.

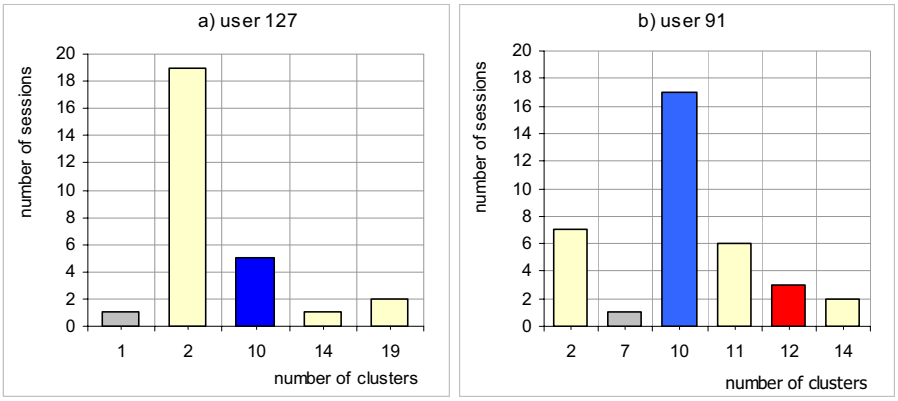


Figure 5. Number of sessions of the user 127 and 91 classified into clusters

## 4. Conclusions and future works

As the WWW grows every day, the task of understanding the users' activity still remains important. In this paper, we have shown that the novel method based on anonymous activity patterns combined with sessions of registered users can identify possible anomalies. Such information was not available using common web log analyzer. It is found that clustering the Web pages is the first step; we need then organizing them into the page hierarchy.

Our project is still ongoing. Though it is unlikely that a perfect solution can be found, it is planned to implement an intelligent system which enable the detection of anomalies automatically. Also, the threshold values should be determined on the base of further analysis.

## Acknowledgements

This paper is supported by the Institute of Applied Informatics of the Wroclaw University of Technology, and Computer Association for Information BOGART Ltd.

## References

[1] A. Abraham, "An overview of fuzzy modeling for control", Journal of Information & Knowledge Management, vol. 2, 2003, pp. 375-390.

[2] M. Eirinaki, M. Vazirgiannis, "Web mining for web personalization", ACM Transactions on Internet Technology, vol. 3, 2003, pp. 1-27.

[3] Y. Fu, K. Sandhu, M.Y. Shih, "Clustering of Web users based on access patterns", In Proceedings of the 1999 KDD Workshop on Web Mining, San Diego, CA, 1999.

[4] A. Jain, M. Murty, P. Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31(3), 1999, pp. 264-323.

[5] D. Król, G. Kukla, "Distributed Code and Data Propagation Algorithm for Longest Common Subsequence Problem Solving", In Proceedings of KES-AMSTA, Wroclaw, Poland, 2007.

[6] C. Kruegel, C. Vigna, "Anomaly detection of web based attacks", In Proceedings of the 10th ACM Conference on Computer and Communications Security, 2003.

[7] G. Murray, J. Lin, A. Chowdhury, "Identification of User Sessions with Hierarchical Agglomerative Clustering", In Proceedings of ASIST, Austin, Texas, 2006.

[8] A. Smirnov, et al., "Ontology-Based Users and Requests Clustering in Customer Service Management System", LNCS 3505, 2005, pp. 231-246.

[9] J. Srivastava, R. Cooley, M. Deshpande, P.T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, vol. 1, 2000, pp. 1-12.

[10] M. Scigajlo, "Investigation of internet system users' behavior using data mining methods", M.Sc. Thesis (in Polish), Wroclaw University of Technology, 2006.

[11] B. Trawiński, M. Wróbel, "User activity investigation of the Web CRM system based on the log analysis", Third International Atlantic Web Intelligence Conference AWIC, Lect. Notes Artif. Intell., LNAI 3528, 2005, pp. 427-432.

[12] J.R. Wen, J.Y. Nie, H.J. Zhang, "Clustering User Queries of a Search Engine", 10th World Wide Web Conference, May 1-5, Hong Kong, 2001.