International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:10, 2007

Analysis of Web User Identification Methods

Renáta Iváncsy, and Sándor Juhász

***Abstract*—**Web usage mining has become a popular research area, as a huge amount of data is available online. These data can be used for several purposes, such as web personalization, web structure enhancement, web navigation prediction etc. However, the raw log files are not directly usable; they have to be preprocessed in order to transform them into a suitable format for different data mining tasks. One of the key issues in the preprocessing phase is to identify web users. Identifying users based on web log files is not a straightforward problem, thus various methods have been developed. There are several difficulties that have to be overcome, such as client side caching, changing and shared IP addresses and so on. This paper presents three different methods for identifying web users. Two of them are the most commonly used methods in web log mining systems, whereas the third on is our novel approach that uses a complex cookie-based method to identify web users. Furthermore we also take steps towards identifying the individuals behind the impersonal web users. To demonstrate the efficiency of the new method we developed an implementation called Web Activity Tracking (WAT) system that aims at a more precise distinction of web users based on log data. We present some statistical analysis created by the WAT on real data about the behavior of the Hungarian web users and a comprehensive analysis and comparison of the three methods

***Keywords*—**Data preparation, Tracking individuals, Web user identification, Web usage mining

I. Introduction

WEB usage mining is a heavily researched area in the field of data mining. The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that are collected on the servers of the web sites. From these data various types of information can be derived, such as information about frequently visited pages or page sets, user navigational patters, behavior of different user groups etc. For this reason data mining techniques are adapted to the data originated from the internet as well, and new difficulties are raised that have to be solved in an efficient way.

The information obtained form web log data can be used for various purposes. For example information about association rules obtained from web log data [1]-[4] can help detecting pages which are visited together even if they are not directly connected, thus revealing associations between groups of users sharing a specific interest [5]. This information can be used when restructuring Web sites by adding links between those pages that are visited together. Another data mining technique adaptable to the web environment is sequence mining [6] used for discovering the web pages that are accessed immediately one after another. Using this knowledge the trends of user activities can be determined and predictions to the next visited pages can be given.

A knowledge discovery process executed on a web log database is similar to that of a traditional database regarding the main steps of the process. However, there is one key step in web log mining that does not exist in the traditional process. This step is the user identification that filters and labels unique users of the log data. Because of the nature of the internet and the features of the technology that is used for browsing the Web, the users cannot be identified in a trivial way. Without using an explicit authentication method, they cannot be fully distinguished, thus some heuristics have to be used for separating the activities of the unique web users.

In certain cases not only the web users have to be distinguished but also the individuals behind the users should be detected. This is a more challenging task that needs some form of authentication as well. Identifying the activities of the unique individuals on the web is not a straightforward problem even after the separation of web users. The same individual can appear in the log file as two different web users, when using for example different computers and two individuals can appear in the log as one single web user when they share the same computer and browse without logging in and out. For this reason a novel method has to be introduced for data collection that store more information allowing us to obtains the information needed for discovering the individuals' behavior.

This paper focuses on the preprocessing phase of a complete web log mining system. The main scope covers the problem of identifying web users, and furthermore, identifying real individuals behind them. We present a system called WAT (Web Activity Tracking system) that aims at solving the problem of identifying individuals' web behavior based on log files. A novel approach is presented as well, that uses a complex cookie-based data collection method for gaining enough information for tracking the behavior of web users and individuals as well. For verifying the usability of our approach the novel method is compared to two of the most commonly used web user identification methods that are based on IP addresses and cookies respectively.

The organization of the paper is as follows. Section 2 introduces the process of web usage mining, and explains the

Manuscript received August 28, 2007. This work was supported by the Mobile Innovation Center, Hungary and accomplished with active cooperation of Median Public Opinion and Market Research Institute. Their help is kindly acknowledged.

R. Iváncsy and S. Juhász are both with Department of Automation and Applied Informatics at Budapest University of Technology and Economics, Hungary (e-mail: renata.ivancsy@aut.bme.hu, sandor.juhasz@aut.bme.hu, fax: +36-1-4633478).

most important tasks regarding the scope of this paper. In Section 3 some user identification approaches are introduced. Section 4 describes the complex cookie-based data collection method, which is the basis of the presented WAT system. Section 5 introduces the WAT system and its operation. Comprehensive measurement results can be found in Section 6. Conclusion and further research directions are presented in Section 7.

II. The Process of Web Usage Mining

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals [5] or to improve the Web structure and Web server performance [6]. In this case, the mined data are the log files stored on the servers of the different web content providers.

In general the Web log mining process consists of 8 steps as follows [7]:

1) Data collection. This is done mostly by the web servers; however there exist methods, where client side data are collected as well.
2) Data cleaning. As in all knowledge discovery processes, in web usage mining can also be happen that such data is recorded in the log file that is not useful for the further process, or even misleading or faulty. These records have to be corrected or removed.
3) User identification. In this step the unique users are distinguished, and as a result, the different users are identified. This can be done in various ways like using IP addresses, cookies, direct authentication and so on. Because the focus of this paper is put on the analysis of the different user identification methods, this step will be discussed later in detail.
4) Session identification. A session is understood as a sequence of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. There are Web server logs that do not contain enough information to reconstruct the user sessions, in this case (for example time-oriented or structure-oriented) heuristics can be used as described in [8].
5) Feature selection. In this step only those fields are selected, that are relevant for further processing.
6) Data transformation, where the data is transformed in such a way that the data mining task can use it. For example strings are converted into integers, or date fields are truncated etc.
7) Executing the data mining task. This can be for example frequent itemset mining, sequence mining, graph mining, clustering and so on.
8) Result understanding and visualization.

As it can be seen, the main steps of a web usage mining process are very similar to the steps of a traditional knowledge discovery process. The only main difference is the appearance of the user and session identification steps. However, because of the presence of this step the whole process including the former steps as well have to be adapted. Because the system presented in this paper deals with the first three steps of the whole process, we focus exclusively to these steps in the subsequent part of the paper.

The data collection step seems to be a trivial task in data mining; however, it needs a lot of attention regarding the object of the mining task. In several cases the common log format (CLF, [16]) provides enough information for the further steps, but in certain cases some extentions of the log format suit better mining process. Furthermore, in certain cases a special data collection method is needed for obtaining more information.

In most web log mining systems data cleaning means removing multimedia objects from the log data [9, 10]. Because to one single page request multiple log entries can belong containing the different pictures, movies and other multimedia objects; these are removed from the log data. In most cases these entries are not useful for further processing, for example when searching for frequent visited pages or navigational paths. Another cleaning task is to remove the log entries of web crawlers [10].

The WAT system presented in this paper is not a full web log mining system. Its aim is to better identify web users and individuals behind the users. In this manner it realizes the first three steps of a web log mining process. The results provided by our system can be used for further processing by any data mining algorithm.

III. Web User Identification – Related Work

Web user identification is one of the most challenging steps in the process of web usage mining. In case of simple market basket analysis, the customer is identified exactly by its customer ID. However, in case of web users, it is not trivial which page downloads belong to which user. The same individual can use multiple computers, and more persons can use the same computer as well. Furthermore, proxy servers can hide relevant information about unique users as multiple computers appear on the internet using the same IP address through the proxy server.

For user identification an important issue is how exactly the users have to be distinguished. It depends mainly on the task for the mining process is executed. In certain cases the users are identified only with their IP addresses [11, 12]. This can provide an acceptable result for short time periods (minutes or hours) or when the expected results from the data mining task do not need more precisely information about the unique web users. For example in case of selecting frequently visited pages for server side caching, or preloading the next page of common navigational paths, it is irrelevant, whether a page is visited by two different individuals or by one individual twice. The key point is that the page is visited twice. However, in case of an advertisement, it is important, whether two unique individual has seen the page or not.

In other cases some heuristics are used for better identification of the users. In [13] the different methods are grouped into two classes, the one is the class of the proactive methods and the other is that of the reactive methods. Proactive strategies aim at differentiating the users before or during the page request while reactive strategies attempt to associate individuals with the log entries after the log is written. Proactive strategies can be simple user authentication with forms, using cookies or using dynamic web pages that are associated with the browser invoking them. Reactive strategies work with the recorded log files only, and the different users will be distinguished by their navigational patterns, download timing sequence or some other heuristics based on some assumption regarding their behavior. For example in [14] web users are distinguished based on their navigational patterns using clustering methods.

IV. Cookie-Based User Identification

Our research focuses on the problem of differentiating the unique individuals behind the web users as precisely as feasible. As the simple or even the extended log format of the web servers does not provide enough information, a more complex approach is needed for recording more information about the different users. This is done by using cookies. Our research partner, the Medián Public Opinion and Market Research Institute (Medián for short) has developed a system that logs the page visits of more than 400 web sites of the Hungarian internet market, and this collected information provides the raw data for our efforts. The process of recording page visits is as follows.

A. Cookie Handling

For tracking the users' behavior the log files are extended with cookies and some other fields as well. Cookies are the most common way of client side data storing, as the web sites that provide the pages (and write the log file) send a data packet (cookie) to the client's browser at the first visit, then this data is sent back to the server each the user navigates to one of the pages of the same site. Downloading a unique identifier in the data packet for each client will allow recording additional information on the server side. Medián aims at tracking the users' activity globally for the more than 400 web sites. For this reason the logging is done on a central server. In this way not only the user's behavior on a single web site can be observed but on several sites that are part of the investigation. Thus, comparative measurements for the different web sites can be achieved as well. This is done by using third party cookies (denoted with C3).

Technically it is realized by embedding a small picture in all the pages of all web sites that are part of the investigation. This small picture (usually a 1*1 pixel GIF in background color) is a web reference to the logging server, and it is downloaded each time a page request is coming to the content provider. This reference places a cookie on the client's browser by the first request. As this C3 belongs the the central server (domain), it will be unique for the browser, and will remain the same for all web sites that are involved in the investigation. It means that the browser uses the same cookie regardless to the investigated site. Afterwards, when the given browser downloads the same page or other page that is involved in the investigation, the cookie is sent back, and the server writes the appropriate record to the log.

The main problem with this type of information collection is that security software and also the users themselves often delete third party cookies. [16]. It distorts the results because in this case new cookie will be assigned to the client, and it will be considered as a new user by the system. To overcome this difficulty the web sites that are involved in the measurement use first party cookies (denoted with C1) as well. In this case the central server places a small script in the page, and this script generates the first party cookie in the name of the investigated web site. The security software do not delete first party cookies; because they are considered as an integral part of the site (some user specific information is stored in C1 by the site like personal settings, persistent information or baskets of an on-line store etc.).

This method ensures that the user's browser contains a unique C1 for each web site, and one single C3 for all the web sites. When the client's browser resolves the reference embedded in the page requested and turns to the server it appears on the logging serves with the C3 belonging to the user along with the C1 belonging to the pair of user and the web page downloaded by the user. In this way the different third party cookies that are deleted and regenerated for the same user can be connected by the first party cookie belonging to the given user. In this way the user behavior can be tracked more precisely.

For tracking the individuals behind the web users a kind of authentication is needed. Some sites requires authentication like web mail systems, web stores, forums and so on. Medián concluded a contract with some content providers that requires user authentication, that in case a user authenticates himself on one of these sites a user id (MID – Medián ID) is included in the log file besides the C1 and C3 ids. In this way, the web users tracked by the first and third party cookies can be joining to individuals.

B. An Example for Cookie Handling

The cookie-handling in the Median logging system is presented in Fig. 1 and Fig. 2. In the figures the first party cookies are denoted with a_n, and the third party cookies are denoted with b_m, where n and m are unique indexes for distinction. The Medián web log file is created by the central server. Two client computers (Client 1 and Client 2) download pages of four different content providers. One of them (Content provider 3) needs also a user authentication. Client 2 uses two types of web browsers (Firefox and Internet Explorer), but not necessarily in the same time.

Fig. 1 shows the way how C1 and C3 are distributed. It depicts also the cookies that are stored by the browsers. In Fig. 2 the steps of logging a page download is depicted. The temporal sequence of the steps depicted with numbers in circles.

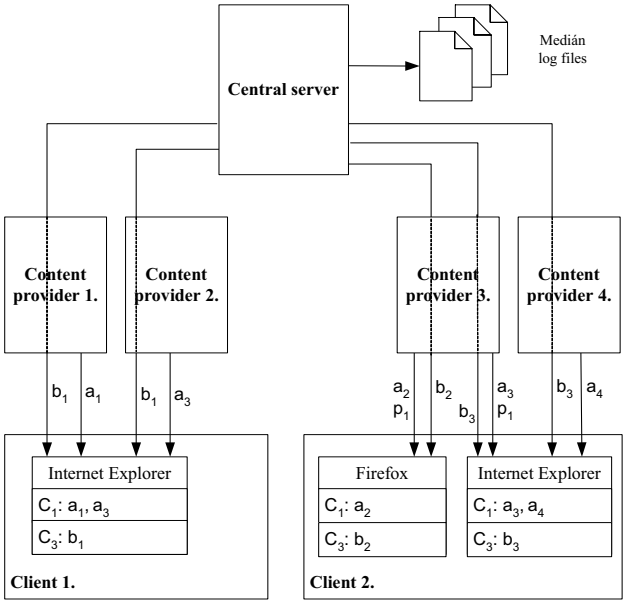

Fig. 1 Distributing C1 and C3 cookies

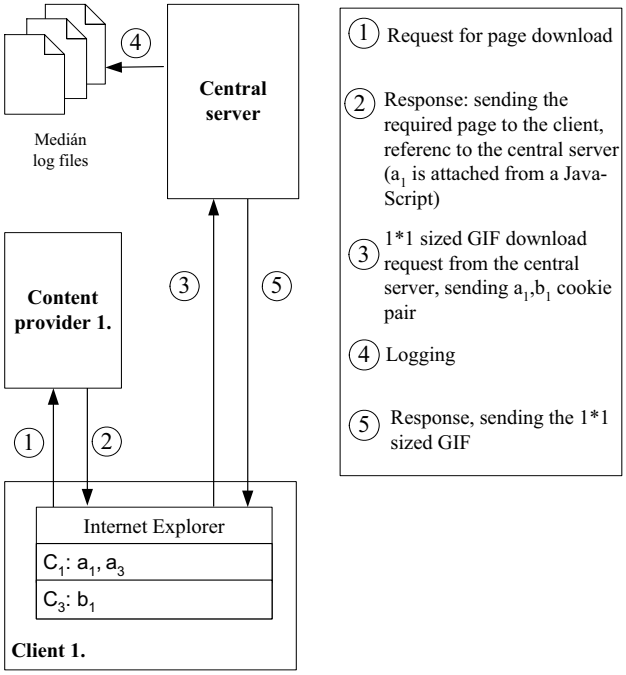

Fig. 2 Forwarding C1 and C3 cookies to the central server

V. Web Activity Tracking System

Preprocessing is a complex step of the complete knowledge discovery process that can take up to 80% of the whole process' time [17], thus developing methods for efficient preprocessing is an essential task from the point of view of performance. The WAT (Web Activity Tracking) system is a preprocessing system that takes a raw web log file as an input, and labels the entries of the log file with the identifiers of real individuals' who created them.

A. Overview of the System

The log recorded by Medián consists of the following fields: (i) C3 cookie ID, (ii) C1 cookie ID, (iii) MID, (iv) IP address, (v) timestamp, (vi) Ucode, that is a code distributed by the Medián to the subparts of each site, (vii) URL, and (viii) browser type.

Assigning the web users (identified by the C3 IDs) to real individuals consists of the following steps:

1) Those records belonging to the same web user must be selected:
   a. As the first step those records have to be joined that contain the same C3 ID, because they certainly (by definition) belong to the same web user. In this way the so called C3-chains are generated.
   b. Two different C3-chains should be joined, when they contain at least one common C1. In this way those C3-chains that seemed to be different based on the first step are joined. Thus the problem of deleted and recreated C3 IDs can be overcome. The same C1 indicates, that the two C3-chain belongs to the same web user. The chains, that are joined based on the C1 IDs are called Cookie-chains (CC). A CC is a record sequence belonging to one web user.
2) The cookie-chains can be grouped together that contain the same MID. In this way so called Cookie-networks (CN) can be produced. The different CNs can be the basis for further processing when identifying individuals.
3) Identifying the individuals based on the Cookie-networks. For this step several rules have to be introduced that are based on real life observations. In this step such kind of rules have to be taken into consideration that overcome the problem of having two individuals with the same MID, or having more than one MID belonging to the same individual. Forming these types of rules is the object of further research.

The WAT system realizes the above steps; however, before these steps can be performed a compression and cleaning step is completed. Fig. 3 shows the overview of the WAT system.

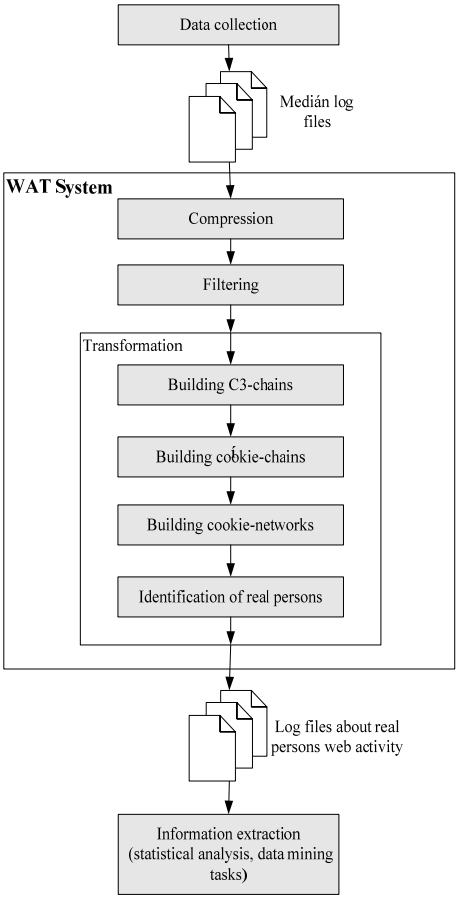

Fig. 3 Overview of WAT system

One of the key issues is how to handle the large size of log files. Each page that belongs to a site which is part of the investigation of the Medián generates a log entry in the log file on the central server of the Medián. Each log file contains up to 5 million records, and it is reached a new file is opened for further recording. Independently of the current size of the last file, at the beginning of each day a new log file is opened. In average 25-30 log files are generated each day. As the size of one record is 392 byte, and a log file contains 5 million records, thus the size of each log file is approximately 2Gb. This means that 50-60 Gb data is collected each day. Processing this huge amount of data is not a trivial task.

Because of the limited memory capacity and the huge number of records all the log cannot be loaded at the same time. For efficient processing of the huge amount of data the concept of working units were introduced. Working unit are processed separately, only the records belonging to one working unit should be kept in the main memory at the same time. However, the results have to be created based on all the data collected so far, thus the results generated on working unit bases must be merged. In this manner each step of the whole process is split into two separate phases, the first one is the so called builder phase that is applying the algorithm on the working unit. The second phase is the merging step combining the results generated in the previous phase with the global results generated so far. The processing of a working units of the same size can be considered to be constant. The merging phase is implemented in a way, that it requires to sorted files to be combined to a new resulting sorted result file, which can also be done linear time. In this way each step can be preformed in a time that is linearly proportional to number of records collected by the servers.

B. Data Compression and Filtering

Data compression is performed by using two methods. On one hand the fields of the log records are compressed by coding (converting character data to integer data, truncating, rule based transformation, code tables). On the other hand the records are grouped into smaller files which can enhance further processing, as the size of the data handled by the following steps can be decreased (not the whole record, but only the relevant fields are read from the disc). The two types of compression can be done together, because the split records can be written directly in the compressed form to the new log file.

The original Medián log (MLog) is split into four separate log files. These are the following:

- CLog (Cookie Log): that contains the C1, C3 and MID fields of the MLog and the timestamp.
- DLog (Detailed Log): that contains the IP address and the UCode field.
- ULog (URL Log): that contains the URL field.
- BLog (Browser Log): that contains the browser information.

The compressor component also generates a so called TempRecordInfo file that speed up the filtering component by extracting new information in working unit sized chunks. It contains the C1, C3 identifiers, the type of the browser, and a hash code created based on the browser type and on the domain part of the URL field (this is called domain hash).

After the compression the filtering step follows. In this case there is no need for removing multimedia objects or web robots from the log file as this type of data is filtered out directly during the logging phase. In our case filtering means removing those cookie identifiers that were distributed more than once. Because of some technical mistakes and the inefficiency of random number generation in the early phase of the data collection period the same first and also third party cookies could have been distributed to different web users. These identifiers not unique, therefore misleading, thus they have to be deleted from the records. The filter component searches for duplicate cookies and creates a so called ban list for them. The identifier collision is detected by checking the browser type and URL domain that should be the same at each appearance of the same identifiers. Two different ban lists exist, one for the duplicated first party cookies and another for the duplicated third party cookies. The ban lists are the input of the subsequent components and they handle these cookie ids as missing information.

C. User Identification

The user identification step contains the following substeps, as mentioned above: (i) generating C3-chains, (ii) generating cookie-chains, (iii) generating cookie-networks and (iv) determine individuals based on the cookie-networks and some business rules.

i. Generating C3-chains

The C3-Chain Processor component of the WAT system performs two main steps. Firstly it builds C3-chains based on the given working unit. Afterwards it merges the actually built C3-chains with the global C3-chains built so far. Fig. 4 shows the overview of the C3-chain Processor.

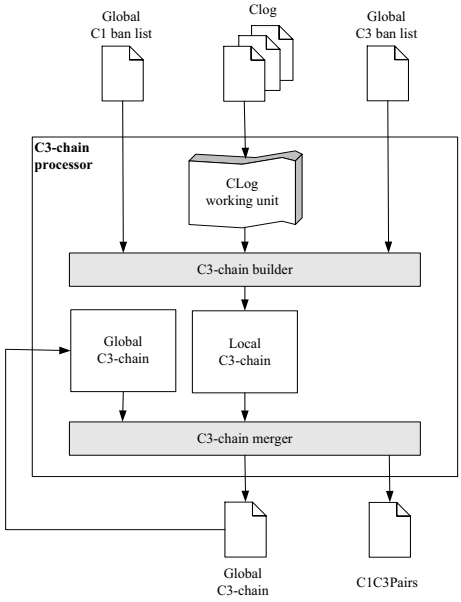

Fig. 4 Overview of C3-chain Processor

When building C3-chains the records sharing the same C3 have to be joined. The input of the C3-processor component is the Clog file and the C1 and C3 ban lists.

The data structure that is built for a C3-chain contains the information about the C1 and MIDs that occur together with the given C3 identifier. These are stored in the form of lists in the structure. For further processing it is needed to store the first and the last occurrence timestamp of the given C3, because using this information will allow later the detection of the time overlaps of certain C3-chains (a more complex method for detecting C1 and C3 collisions). A C3-chain structure contains the following information:

- C3 identifier
- C1List: the C1 identifiers that occur in the chain
- MIDList: the MIDs that occur in the chain
- FirstTime: the timestamp of the first MLog record that belongs to the chain
- LastTime: the timestamp of the last MLog record that belongs to the chain

ii. Generating Cookie-chains

The task of the Cookie-chain processor component is to join those C3-chains that have the same C1 identifiers in their C1List. The structure generated in this way is called Cookie-chain. A Cookie-chain structure contains the following information:

- A unique CC identifier
- C3List: the list of the C3 ids contained by the cookie-chain
- C1List: the list of the C1 ids contained by the cookie-chain

After generating the cookie-chains for the current working unit they are merged into the global list of previously found cookie-chains. The overview of the CC-chain processor is depicted in Fig. 5.

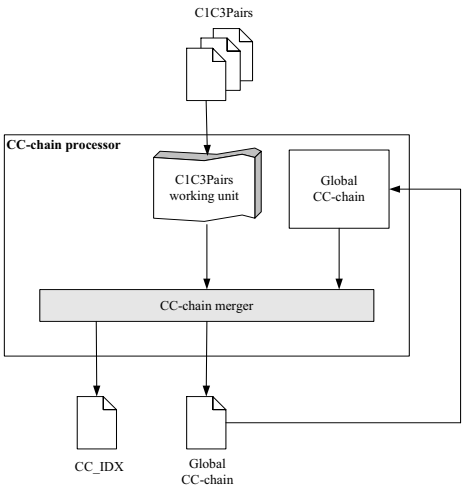

Fig. 5 Overview of the CC-chain processor

At the moment, the WAT system works with the CC-chains when discovering the activity of web users. The step for identifying individuals is the objective of future research.

VI. Measurement Results

In order to demonstrate the usefulness of the WAT system, and that of the cookie-based data collection method, some measurements were carried out. The measurements were executed on a database obtained from the Median containing real data about the navigational behavior of the web users of Hungarian web sites. The data were collected using the technique described in Section 4.

First of all the performance of WAT system was investigated. As mentioned in Section 5 the system has a preprocessing phase that compresses and filters the raw data for further processing. Afterwards it builds C3-chains from which cookie-chains are created by joining the C3-chains having any C1 identifier in common. Afterwards the system collects some statistical results as well. On Figure 6 the execution time for the different phases of WAT system can be tracked while statistical results were created on a data set

collected during one month (approximately 1500 Gb in size). It can be seen well, that the most time consuming step is the compression of the raw data that takes 69% of the total time, while creating statistics from the preprocessed data takes only 1%. The whole process takes 101.5 hours that is approximately 4 days.

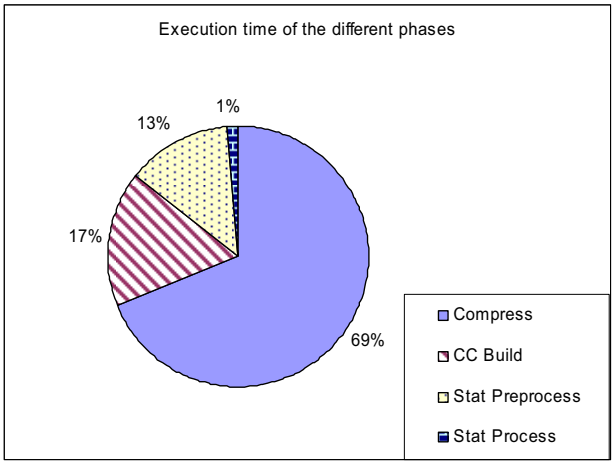

Fig. 6 Execution time of the different phases of the WAT system

In order to compare the accuracy of the different user identification techniques, we counted the number of the C3 ids, CC-chains and IP addresses for different pages of one of the most popular sites of Hungary, namely the index.hu. This statistics is depicted in Fig. 7. In this figure it can be observed, that the number of IP addresses is significantly lower than that of the C3 and CC-chains. It means, that using the cookie-based method more precisely distinction of the web users can be achieved. The insignificant difference between the number of the C3 ids and that of the CC-chains can be explained by the fact, that the statistics were created out of data collected during one day, and during this time period few cookies were deleted on the client side.

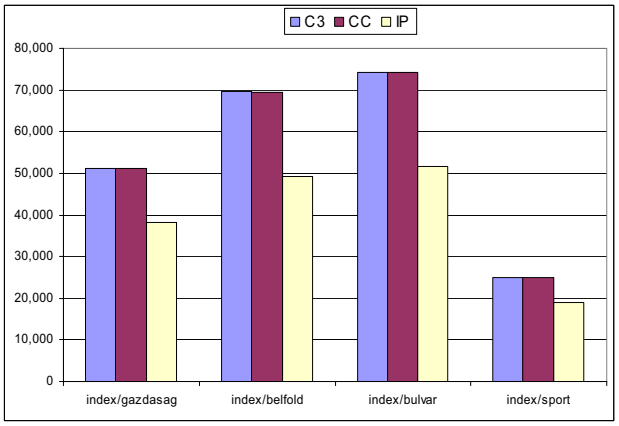

Fig. 7 Comparing measuring methods on 8th January, 2007

An interesting result can be seen on Fig. 8. It depicts the hourly visitor distribution of some important sites of Hungary with different topic like economy, sport, politics etc. We can find out which topic in which hour is mostly read by the users. For example the topic sport is popular in the early hours and in the afternoon, while yellow press is mostly read in the midday hours.

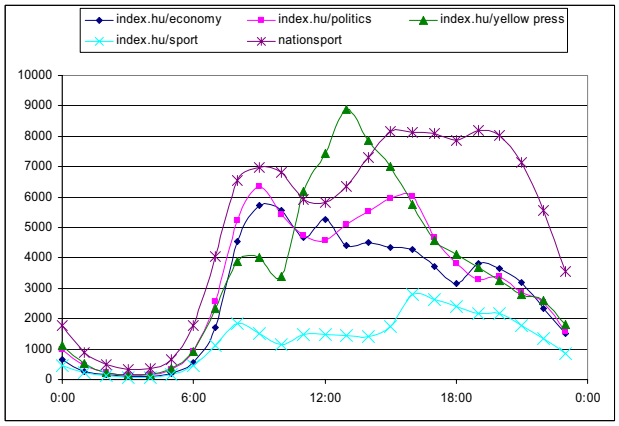

Fig. 8 Hourly visitor distribution in the second week of January 2007, from Monday to Friday

On Fig. 9 the weekly visitor distribution of some important Hungarian web portals is depicted. It can be seen that the number of visits is low in the weekend, on Monday is a bit higher, and on the other weekdays its value is nearly constant.

The different content providers on this figure can be divided into three main groups:

- mail providers (freemail, citromail)
- communal portals (iwiw, myvip) and
- news portals (index, origo, nemzetisport, hirado).

We can draw the conclusion that the most popular portals are the communal portals, especially the iwiw, which is the greatest communal portal in Hungary. Also the mail providers are frequently visited portals, while the new portals are the least visited out of the portals presented in the chart. However, these sites are the most visited out of the portals involved in the investigation.

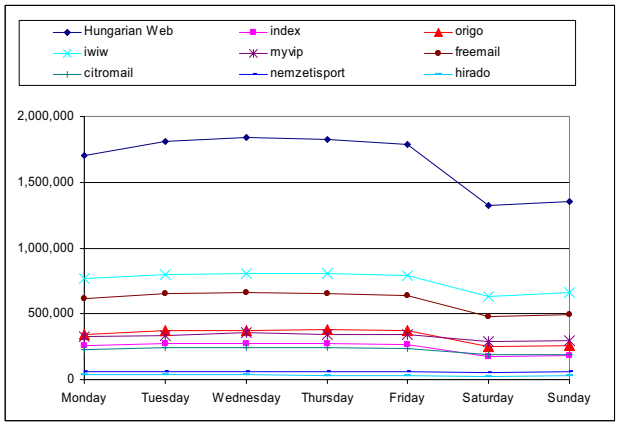

Fig. 9 Weekly visitor distribution from 1st to 28th January, 2007

VII. Conclusion

The main aspect of this paper was the analysis of different user identification methods of web log mining. We introduced a novel method that exploits the benefit of using first and third party cookies in parallel for collecting data about the user's activity. Furthermore the data collection was realized through a central server that collects visitors' data of more than 400 Hungarian content providers. This makes possible to compare the web activity of the different users through more than one single site. A novel system was presented that identifies the web users using the novel complex cookie-based method. Furthermore, by introducing an identifier, it can differentiate real individuals behind the web users. However, this step of the system is the object of future research. The system can also be used for creating different statistics about the Hungarian web usage and about the different web user identifying methods. We showed some interesting measurement results about the popularity of the different content provider, and about the visit behavior of the web users.

Acknowledgment

This work was supported by the Mobile Innovation Center, Hungary and accomplished with active cooperation of Median Public Opinion and Market Research Institute. Their help is kindly acknowledged.

References

[1] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment," in Sixteenth International Conference on Distributed Computing Systems, 1996, pp. 385–392.

[2] J. Punin, M. Krishnamoorthy, and M. Zaki, "Web usage mining: Languages and algorithms," in Studies in Classification, Data Analysis, and Knowledge Organization. Springer-Verlag, 2001.

[3] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper," 2002

[4] O. R. Zaiane, M. Xin, and J. Han, "Discovering web access patterns and trends by applying olap and data mining technology on web logs," in ADL '98: Proceedings of the Advances in Digital Libraries Conference. Washington, DC, USA: IEEE Computer Society, 1998, pp. 1–19.

[5] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM* Trans. *Inter. Tech.*, vol. 3, no. 1, pp. 1–27, 2003.

[6] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*. London, UK: Springer-Verlag, 2000, pp. 396–407.

[7] Z. Pabarskaite and A. Raudys, A process of knowledge discovery from web log data: Systematization and critical review, Journal of Intelligent Informatin Systems, Vol. 28. No. 1. 2007. pp. 79-104.

[8] J. Zhang and A. A. Ghorbani, "The reconstruction of user sessions from a server log using improved timeoriented heuristics." in CNSR. IEEE Computer Society, 2004, pp. 315–322.

[9] Robert Cooley and Bamshad Mobasher and Jaideep Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns, Knowledge and Information Systems, Vol. 1. No. 1. 1999, pp. 5-32

[10] M. Spiliopoulou and C. Pohle and L. Faulstich, Improving the Effectiveness of a Web Site with Web Usage Mining, WEBKDD '99: Revised Papers from the International Workshop on Web Usage Analysis and User Profiling, 2000. pp. 142-162.

[11] M. Gery, H. Haddad: "Evaluation of Web Usage Mining Approaches for User's Next Request Prediction", *Fifth International Workshop on Web Information and Data Management (WIDM'03),* 2003. pp. 74-81.

[12] O. Nasraoui, H. Frigui, A. Joshi, and R. Krishnapuram, Mining Web Access Logs Using Relational Competitive Fuzzy Clustering, Eight International Fuzzy Systems Association World Congress - IFSA 99, 1999

[13] M. Spiliopoulou and B. Mobasher and B. Berendt and M. Nakagawa, A Framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis, INFORMS Journal on Computing, 15, 2003.

[14] T. Morzy, M. Wojciechowski, and M. Zakrzewicz. Web users clustering. International Symposium on Computer and Information Sciences 2000.

[15] Brandt Dainow, „3rd Party Cookies Are Dead,", Web Analytics Associations, http://www.webanalyticsassociation.org/en/art/?2

[16] W3C, Common Log Format, http://www.w3.org/Daemon/User/Config/Logging.html

[17] Ansari, S., Kohavi, R., Mason, L., & Zheng, Z. (2001). Integrating e-commerce and data mining: Architecture and challenges. Data mining. San Jose, CA: IEEE Computer Society.