International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:7, 2007

Tracking Activity of Real Individuals in Web Logs

Sándor Juhász, and Renáta Iváncsy

***Abstract*—**This paper describes an enhanced cookie-based method for counting the visitors of web sites by using a web log processing system that aims to cope with the ambitious goal of creating countrywide statistics about the browsing practices of real human individuals. The focus is put on describing a new more efficient way of detecting human beings behind web users by placing different identifiers on the client computers. We briefly introduce our processing system designed to handle the massive amount of data records continuously gathered from the most important content providers of the Hungary. We conclude by showing statistics of different time spans comparing the efficiency of multiple visitor counting methods to the one presented here, and some interesting charts about content providers and web usage based on real data recorded in 2007 will also be presented.

***Keywords*—**Cookie based identification, real data, user activity tracking, web auditing, web log processing

I. Introduction

WEB log files recorded on the server side providing a valuable source of information about the users' behavior can be exploited for several purposes. The most important areas cover navigation improvement (information placement, server side caching), restructuring websites, intelligent adverts, efficiency monitoring of web sites, detecting user types and creating customized content for them [1]. As web servers record all the events within single file in the order they happen, in most of the above mentioned areas the separation of the activities (log entries) belonging to different users plays a key role, and in many cases further subdivision follows to break up the activity of individual users into separate sessions.

Because of the wide scope of above domains no universal solution exists, which could cover all types of data extraction, but each domain claims for different tools and methods instead. In this paper the attention is focused on measuring the popularity of web pages from a very specific point of view. Content providers in the commercial sector attribute primary importance to the number of visitors of their pages, as in the competitive environment this is a major metric to rank them: that is their professional prestige and the direct advertisement value of their web pages is measured in the number of people they reach. In this domain the exactitude and trustworthiness of the counting method represents a direct financial value, this is why this task is usually propagated to trusted third party web auditing companies, who are more likely to provide comparable and authentic measurements, than that originating from the content providers themselves.

Our work was done in cooperation with Median Public Opinion and Market Research Institute [2] that is, next to its other profiles, one of the biggest and most trusted web auditing companies in Hungary. (Hungary is a small county in Central-Europe with 10 millions of inhabitants.) Median has a contract with all the important content providers of the country, which allows them to collect exact and meaningful statistics about the national browsing habits. The different methods and processing steps described in this paper were executed against this enormous amount of real web log data, reviewing the user activity on the server of several hundred content providers. The log files contain the activity of 2.2 millions of users in average each day, and 2.9 million a week. As official surveys report a 35% penetration of Internet in Hungary (that is 3.5 millions of people out of the whole population of 10 million use the Internet at least once in a month), we can declare that the log file includes records from all frequent internet users of the country. This is a massive amount of data as 140-150 millions of entries are recoded on a daily basis, occupying a storage space of 50-60 GB (i.e. 22 TB of data in a year).

The monolithic common log file gathered from the audited web sites treated in the same way allows not a fair comparison of visitor statistics, but also enables to add specials marks (user and topic identifiers, some kinds of personal attributes) to the log file entries. These additional marks permit following the activity of the same user on different sites, thus conclusions about countrywide social issues (Internet penetration, web browsing habits) both in general and regarding specific groups (by age, sex, location of the users) can be drawn. Of course these statistics must be based on real-life individuals who are usually hard to discover behind their faceless and not necessarily unique computers and browser agents.

The rest of the paper is organized as follows. Section 2 describes the structure of web logs and introduces the most common methods for detecting and identifying users belonging to the web log entries. Section 3 outline the duality and the difference between web users and real individuals, and presents a completely new method for finding the web log

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.

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

entries belonging to the same person. Section 4 briefly introduces our log processing system and includes some interesting statistical results about the Hungarian internet users and main Hungarian content providers. We conclude by summarizing the work completed till this point, and we detail what further improvements are currently in sight.

II. Counting Visitors in web logs

As mentioned above web processing may serve several goals. While some of them (like web server performance measurement or optimization) consider the log entries as independent entities, in most of the applications (personalization or tuning of web services, presentation of promotional contents) it is an essential prerequisite to identify and separate log entries belonging to the same user, although the definition of the user might vary depending on the final goal.

Web log data is usually generated by web servers and written to special log files. Whatever user identification method is chosen it must be based on the information stored in the log entries. The entries contain all information about visitors' activities [3], but various servers may generate their logs in different formats (Apache, Netscape Flexible, Lotus Domino, Microsoft proxy, ISS standard/extended etc.). Nevertheless, the most common one is so called common (or combined) log format (CLF, [4]) containing the following fields in each entry: IP address, HTTP authentication data, date, time, GTM zone, request method, page name, HTTP version, status of the page retrieving process and number of bytes transferred.

The IP address stored in the common log format generally proves to be insufficient to identify the web users, thus often different additional pieces of information are added to the records. Extended log format [5] is the most well known special type of the common log format storing some additional information like the referrer URL (the page the user was coming from), the user agent (browser type and client side operating system) and HTTP cookies. These extensions prove to be extremely valuable for more precise user identification.

Due to some technical reasons (IP addresses might be shared or change over the time) and to constraints of the HTTP specification (no user identification is required) the requests contained by the log files cannot be uniquely assigned to the individual who has performed them. Spiliopoulou [6] distinguished two main strategies ("proactive" and "reactive") on how sessions of different individuals are detected. Proactive strategies aim at an unambiguous association of each request with an individual before or during the individual's interaction with the Web site (user authentication, the activation of cookies that are installed on a user's machine). Reactive strategies attempt to associate requests to individuals after the interaction with the Web site, based on the existing, incomplete records, thus sessions are created afterwards, based on the information stored in the web logs. As the whole processing builds on the server side logs, the presence of proxies and caches may interfere with the results as some requests may not get to server at all.

The same difficulties apply when counting visitors, which is the log processing task in the focus of our interest. The number of visits during a given time period can be determined by counting log entries, IP addresses, cookies belonging to a given site, or the estimation can be based on session reconstruction or direct authentication. These methods described bellow can be enhanced by heuristics and can also be combined with each other.

As web logs contain both the URL and the timestamps belonging to the request, the simplest way is to *count the number of log entries* recorded with the given URL during a specific time period. The advantage of counting "clicks" is the speed and simplicity, the main disadvantage is the unpredictable distortion it causes in the measurements, as one user can click multiple times on different links of the audited site (or page) depending on the amount and the type of the page content. When having some information on the exact number of visitor from an external source (e.g. from surveys or the site itself needs authentication of users) a distortion coefficient can be attributed the raw results, but this kind of calibration must be completed for each audited site and must be regularly repeated, as it depends on the changing contents, user behavior and even on the time of the day or week, as it will be shown in Section 5.

The majority of authors dealing with web usage mining suggest *identifying the web users by their IP address* [7][8][9]. IP based identification is also simple, and more exact, but underestimates the number of visitors as it cannot separate several web users sharing the same IP address behind a NAT server or a firewall (extremely common scenario for users at home and at their working places as well). Same individuals changing their IP address (users of mobile computers, IP addresses leased for a fixed period of time from the pool of an internet provider) also distort the statistics, as they are counted more than one time.

Deploying cookies to the client agent is the most straightforward among the proactive methods. The biggest advantage is that cookies uniquely identify the web users while being completely insensitive to IP address changes. As web robots generally do not handle cookies this method also acts as a prefiltering step against undesired data. Cookies are believed in all published works available for the authors to be a ultimate solution for web user identification but are rejected nearly instantly because of the high level of cooperation they require form both the web user and the web provider side as cookies raise a number of issues [1][10]. Firstly, there are those related to the privacy among users. Secondly, cookies are not stored into web logs, as not all web site developers and web server administrators understand the importance of collecting and storing complete web log information. In Section 3 we show that the cooperation level required on the client side can be eased by slightly raising it on the provider side. We will also demonstrate that even successfully setting cookies still does not provide sufficient information for

identifying the individuals behind the browser agents.

Several methods exist for reactive *reconstruction of sessions* using heuristics. Although sessionizing is not primarily used for user identification, it can be helpful when trying to separate users who share a common IP address, or human users who share the same computer (see Section 3 for details). Two basic approaches exist: the time oriented and navigation oriented methods [10]. The first one tries to group log entries with the same IP address sufficient close to each other, not exceeding a maximum of time of inactivity between them. In [11] a 9.3 minute was measured as a mean time of inactivity within a site, by adding 1.5 standard deviations the maximal session length was limited in 25.5 minutes. Looking at the fact that human web user rarely type URLs directly, but they usually click on links found on the pages, the navigation oriented approach considers the availability of pages from each other. Not only direct links are allowed, but allowing some invisible backward steps (going back to a previously visited pages is usually served form the local cache of the client) a log entry that is available from any pages that already belongs to the session will be added [12]. A session reconstruction only separates sessions belonging to different web user, but they do not address the question what sessions belong to the same individual over a longer time period (a day, a week or a month), so this method cannot be used directly for counting visitors on a larger time scale.

The simplest way of user identification is *requiring an obligatory authentication* on a login page. HTTP login has direct footprint in fields in the log file rfcname and logname (see Common Log Format [4]). Higher level forms of authentication can add extra fields in order to identify uniquely a visitor. This method has a limited use, as humans usually try to avoid web sites where registration information is required [1], unless it is inevitable and part of the functionality (special purpose sites such as on-line banking, community sites, or e-mail providers). However no evidence of such cases was found in the literature.

The next section introduces a new combined method that makes an extensive use of cookies, session reconstruction and direct user authentication in order to find the real individuals behind web users. The method builds on the willing cooperation of the content providers by creating a situation of common interest and requires no extra effort from the web visitors at all.

III. Detecting Individuals behind the Log Entries

Today the most common and straightforward way to identify visitors is by using cookie information. When a user visits the web site for the first time, no corresponding cookie exists yet, thus no information is sent from the browser to the web server. This causes the web server to recognize that it is a new user and passes unique cookie information to the browser together with the requested content. The web browser saves the just received cookie in a special directory. Next time when the visitor accesses the same web site, the browser sends the already existing cookie to the web server together with the request. The web server recognizes the cookie and only sends back the requested page. The cookie information can be saved into the log (extended format) together with the other information about the served request. Cookie based identification is cited as the best known user identification method in the literature [1][10] where the cooperativeness of the browsing persons (whether they refuse cookies from privacy, technical or any other reason) represents the highest risk.

The experience with real-life web auditing (visitor counting) system shows that cookie-based identification suffers from two major problems [2]. The first one is the durability of cookies (how often they disappear), whereas the second one is related to the duality and the difference between web users and real individuals. In the following sections we will show a new method that significantly eliminates both these difficulties.

A. Increasing the Durability of Cookies

Cookies always belong to the URL domain from which they originate. If a server serving domain *a.b.c* once set a cookie, the browser will send it back to the server every time it sends a request to that same domain *a.b.c* (e.g. asking for page *a.b.c/subdir/main.html*). This implies that in order receive the cookies a third party web auditing company has to ask the content providers to set a reference to his domain from their pages (in practice this is done by inserting an invisible picture of size 1 by 1 into the content pages that is downloaded form the auditor's domain). The advantage of this method is that the same web user passing through any of the audited domains (*a.b.c*, *example.org*, etc.) will keep the same identifier, as they are always in contract with auditor's domain where the picture is coming from.

The web browsers have to separate storage place for cookies. They distinguish 1st and 3rd party cookies (referred later as C1 and C3 cookies respectively) depending on whether they were set directly from the browsed domain (C1), or by a third party content (banners, promotions and the above mentioned picture serving as measurement point). While C1 cookies are usually considered as essential part of the functionality of a web page, C3 cookies are more likely to be blocked or removed regularly by browser agents, users, anti virus and anti-spyware software and other privacy tools [13][14]. After a cookie is deleted, the browser agent will have no information to send along with the request thus it will appear as a new user to the server, and a new C1 or C3 identifier will be generated and sent back to the client.

As C1 and C3 cookies are handled separately, they are not likely to be deleted at the same time. In order to increase the life time of the cookie based identification our method combines the use of C1 and C3 cookies. All audited sites place a C1 cookie on the client's browser as well, independently of the C3 cookies set by the auditor. Instead of the simple reference to the measurement point (the picture in the auditors' domain) a small Java Script is inserted into the

audited pages, which allows sending up the C1 cookie from the audited domain to the central server, that logs both the C1 and C3 cookies belonging to the given page request. Should the C1 or C3 cookie disappear from the client's agent, the other cookie type maintains a continuity of tracking the same user. The log entries belonging to the same web user kept together by the interleaving C1 and C3 cookies identifiers is called a ***cookie chain***. These chains contain several continuous ranges with different C3 identifiers, where inside the range of entries the same C3 value is present. In a cookie chain several C1 might appear, one for each audited site on which the web user turned up. Any C1 common for two C3 ranges causes the entries to be considered part of the same range. Cookie chains can have a time span of months or even years.

Of course this method provides no protection against deleting all the cookies together (regular clean-up for privacy or maintenance reasons, changing or reinstalling browser agents, reinstalling operating system).

B. Web Uses and Real Individuals

After extending the temporal durability of cookie-based method, we address the issue of the dual nature of web users and real individuals. Except for direct authentication of the human user all the identification method suffer from the problem of seeing the user's browser agent only instead of the human being behind. A web user visible through its agent is an abstract person, whose activity can be logged and tracked at servers. The extent of their possible identification depends on the technical parameters. As cookies are stored by the browser agents in the storage space of the local user who is current logged in, the web user visible at the server is nothing else then the triple of an operating system, a user just logged into it, and a browser agent. The relationship between these "web users" and real individuals can vary. More individuals using the same browser agent (frequent case in schools, families or community web access points) appear as one web user in the server logs, whereas the same person having multiple access to the Internet (working place, home, community web access points) can be overrepresented by the multiple web users he appears to be.

Accurately finding the relation between web users and real individuals is prerequisite of most social statistical analysis, as they principally aim at handling and describing human behavior. The primary and single faithful source of identification is requiring the users to authenticate themselves. To spare the users the extra efforts of filling in registration forms and logging in every time (and avoid the massive emerge of useless forged or fake logins) in our case the already present information is reused.

Median Institute encouraged the content providers, where authentication was already in place and part of their everyday functionality (community sites, forums, e-mail providers, public advertising pages, etc.), to cooperate by attaching a unique user identifier (MID) to every log record produced by an authenticated user. For privacy reasons the MIDs are randomly generated numbers only usable for statistical processing. As they are only meaningless numbers, there is no way of finding out which individual is behind a specific MID, but the provider attaches some demographical information such as the sex, the age or area code to it. The cooperation (the transfer of this data) is in the direct interest of content providers, as this allows the auditor to create more valuable visiting statistics with demographical data for them.

The presence of MIDs in the log file not only allows impersonating cookie chains by attribution of demographic data, but the common MIDs can connect completely independent cookie chains of the global data as shown in (Fig. 1). For example if an individual with one computer at home, and one at work uses an e-mail service from both places with his own login name, his unique MID will appear in the cookie chains generated by both computers. The set of cookie chains connected by MID are called cookie networks. The cookie networks are supposed to cover and combine all the activities of one (occasionally more) real individual.

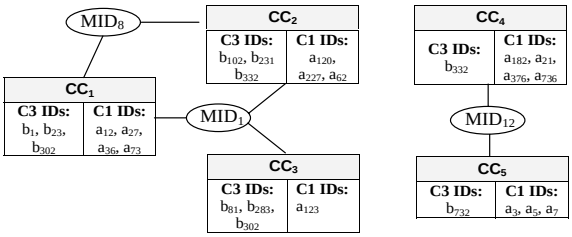

Fig. 1 Coockie chains connected by MIDs

C. Further Extensions

MIDs are also helpful, when multiple individuals share the same computer without logging in and out (families, community sites). In this case the presence of contradicting demographical data (e.g. male-female, of different birth year) might give a starting point to the sessionizing algorithms.

The starting sessions can be formed either by following the page availability structure or simply grouping the log entries that form a sequence where the difference between two consecutive timestamps is lower than a specified limit. In the next steps these session are split at the points where a new MID with contradicting demographical data appears. This method allows detecting the change of real individuals sitting before a single computer even if they use the same browser without logging out. The resulting partial cookie chains (sub-chains) can be connected through their common MIDs similarly to the previously presented full cookie networks.

IV. Web Activity Tracking System

As traditional data mining and statistical methods are not suitable for web log mining as it exists, and neither is web log data suitable for data mining algorithms, therefore data, algorithms and applications require specific preparations. [1] Our main goal was to create a system that is able to do the preprocessing steps on the collected log data in order to allow data mining algorithms to process it directly. The primary task

of Web Activity Tracking (WAT) System developed by us is to assign an additional field to each of the log entries containing the identifier of the individual the entry was created by. This extension enables creating e.g. age or sex based visit statistics, and gives a significant freedom to data mining algorithms that can assign demographical data to the browsing behavior. The WAT System is charged with more traditional transformation steps as well (data compression, data cleaning, filtering erroneous records) that are completed before the cookie chain and cookie network building steps. The whole processing procedure is organized in pipeline, and uses sophisticated transformation algorithms with a restricted time limit linearly increasing with number of records to process.

The detailed description of the algorithm is behind the scope of this paper, thus we only present some interesting results produced by WAT in its current state. Although cookie network handling and session processing is not fully functional yet, cookie chain building is fully in place, allowing the comparison of the results produced by different visitor counting methods (IP address, traditional cookie based and cookie chain based user identification).

Fig. 2 compares three different approaches for counting visitors appearing during a single day. All the three methods are applied on the log data of the 8th of January 2007 for three different topics of a big Hungarian news site called "Index". The methods count the different IP addresses, (third party) cookies and cookie chains that appear in the log during the day, respectively. The number of cookie chains is slightly lower than the number of cookies, that means, that the more complicated method brings little change only, which was suggest that a very few cookies (about 0.1%) get deleted during a single day. The number of IP addresses seen is indeed very different from the other two methods. The ratio between the cookies and IP addresses is between 1.31 and 1.44, depending on the topic. This suggests that IP address counting introduces a significant distortion that cannot be considered to be a constant even inside a single site.

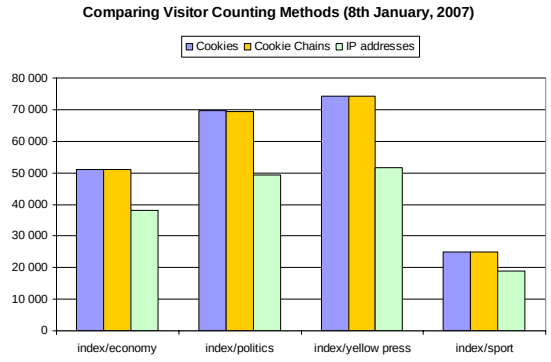

Fig. 2 Comparing visitor counting methods on daily data

The second series of measurements represented in Fig. 3 performs a similar comparison, where the scope of counting is extended to cover a whole month (January, 2007). Next to the previously mentioned three methods two new methods are also introduced. The new ones are extended versions of the cookie and cookie chains counting methods, where the number of the elements is weighted by their lifetime. (E.g. if a cookie chains spans from the 1st till the 10th of January it has a 10/31 weight in the statistics.)

Weighting helps to eliminate the effect of counting multiple times the users who change/loose their cookies during the sampling period, while on the other hand it attributes a much lower weight (1/31) to user paying a short visit only to the site than the expected 1.00. Although not all the effects are desirable of such a weighting, the external measurements (questionnaires and other traditional public opinion research methods) seem to justify the validity of this approach.

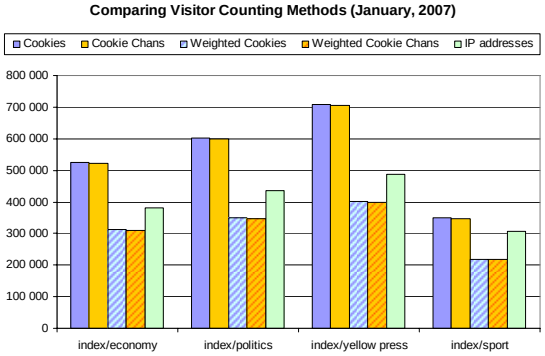

Fig. 3 Comparing visitor counting methods on monthly data

WAT system also allows comparing the number of visitors of different sites. Fig. 4 shows the daily visitor statistics of different important news, community and e-mail provider sites during the whole month on January. The periodicity of the weekly traffic changes is clearly visible, as the user activity drops during the weekends. A line that monitors the internet usage of the country visible to our system was also included in the figure.

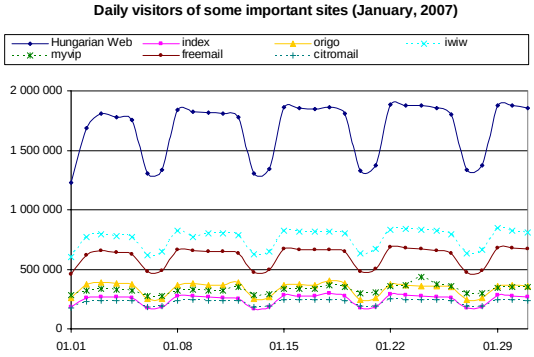

Fig. 4 Comparing visitor counting methods on monthly data

By increasing the resolution we can draw interesting conclusions regarding the browsing habits at the working places as well. Fig. 5 applies an hour by hour resolution to show the ratio between the cookie and IP address based

measurements. The chart visualizes well the beginning and the ending of working hours with the significant increase and decrease of the above ration. As working places usually have private networks that are connected with special equipments like NATs, proxies, firewalls to Internet. These equipments enable the computers behind to share a common IP address when seen from the outside world. The figure not only gives an overview of the working hours, but looking at the different topics allows drawing conclusions about the site that are typically visited (or not visited) during working hours. For example the higher baseline and the lowers peaks of the sport site in the figure hints on balanced popularity of sport between the home and the working place accesses. According to the same statistics politics and yellow press seem to be more the pleasure of the users at browsing from their working places.

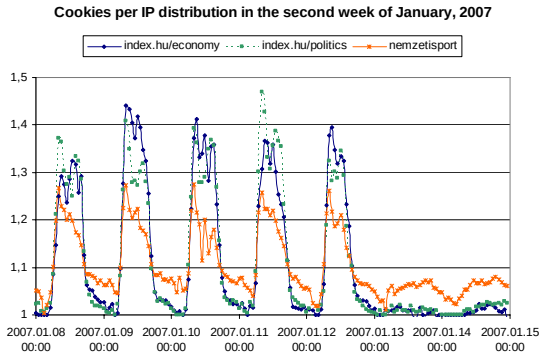

Fig. 5 Detecting working hours by web log analysis

V. Conclusion and Future Work

In this paper we presented a new method of user identification that extends the traditional cookie based approach. The method produces high quality, representative results for a single site and for the whole country as well, due to the wide support of the Hungarian content providers. Our experiments with real data shows, that 99.9% of the Hungarian users accept cookies, so the doubts about cookie-based method because of the user side security restrictions seem not to be founded. According to our experience more than 1,3 million out of 2,8 millions of user in the scope of our system do not delete their cookie regularly, while the ones who do unfortunately remove first and third party cookies at the same time. That is why merging third party cookie chains by first party cookies was not particularly efficient (the number of merged chains remained below 1% of all the cookies detected). The distortion effect of the non connectable chains can be significantly decreased by the simple heuristics of weighting the cookie chains by their life times.

It is also important to mention, that the next step of this method will be extended with the use of the quasi-personal identifiers (MIDs), which are derived for real (directly typed) user authentications, that are present in 75% of all the records, allowing to efficiently connect the cookie chains otherwise seeming to have no relation. According to the first experimental results not presented here, the introduction of MIDs is much more efficient in connecting independent chains than the use of first party cookie identifiers, thus the effect of the weighting heuristics will have a much more moderate importance during the next phase.

The preliminary results presented in this paper show the processing potential of the system. WAT System developed by us is capable of processing the log data collected from a whole country, and the labeling of the log entries with the identifier of the browsing individual. After this preprocessing step not only different providers, topics and timescales can be compared, but demographical data can also be assigned to describe the browsing behavior. As the system is not completely finished yet, the result presented here do not reflect this full capacity of the system. Connecting cookie chains by using MIDs and the introduction session handling methods require additional work in the near future.

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] Z. Pabarskaite, A. Raudys, “A process of knowledge discovery from web log data: Systematization and critical review.” *Journal of Intelligent Information Systems* 28(1): pp. 79-104, 2007.
[2] *Median Public Opinion and Market Research Institute*. http://www.median.hu/ and http://www.webaudit.hu/
[3] R. Kosala, H. Blockeel, “Web mining research: A survey.” *ACM SIGKDD Explorations*, 1, pp. 1–15, 2000.
[4] W3C, Common Log Format, http://www.w3.org/Daemon/User/Config/Logging.html
[5] G. Fleishman, “Web log analysis, who’s doing what, when?” *Web Developer*. 1996.
[6] M. Spiliopoulou, “Managing interesting rules in sequence mining.” *3rd European Conference on Principles and Practice of Knowledge Discovery in Databases* PKDD’99. Prague, Czech Republic: Springer-Verlag. 1999.
[7] H. Ishikawa, M. Ohta, Sh. Yokoyama, J. Nakayama, K. Katayama, “On The Effectiveness of Web Usage Mining for Page Recommendation and Restructuring,” *Lecture Notes In Computer Science*; Vol. 2593, pp: 253-267.
[8] S. Baron, M. Spiliopoulou, “Monitoring the Evolution of Web Usage Patterns, Web Mining: From Web to Semantic Web,” *First European Web Mining Forum*, (EMWF 2003), Cavtat-Dubrovnik, Croatia, September, pp. 181-200, 2003.
[9] M. Spiliopoulou, C. Pohle, L. C. Faulstich, “Improving the Effectiveness of a Web Site with Web Usage Mining, International *Workshop on Web Usage Analysis and User Profiling*,” WEBKDD, pp. 142-162. 2000.
[10] M. Spiliopoulou, B. Mobasher, B. Berendt, M. Nakagawa, “A Framework for the Evaluation of session reconstruction heuristics in Web-usage analysis.” *INFORMS Journal on Computing* 15: pp. 171-190, 2003.
[11] L. D. Catledge, J. E. Pitkow, “Characterizing browsing strategies in the world-wide web.” *Computer Networks and ISDN Systems*, 6, 10–65, 1995.
[12] R. Cooley, P. Tan, J. Srivastava, “Discovery of interesting usage patterns from Web data.” B. Masand, M. Spiliopoulou, eds. *Advances in Web Usage Analysis and User Profiling*. LNAI 1836, Springer, Berlin, Germany. 163-182, 2000.
[13] Brandt Dainow, “3rd Party Cookies Are Dead,” *Web Analytics Associations*, 2005. http://www.webanalyticsassociation.org/en/art/?2
[14] “WebTrends Advises Sites to Move to First-Party Cookies Based on Four-Fold Increase in Third-Party Cookie Rejection Rates,” *WebTrends*, 2005. http://www.webtrends.com/CookieRejection.