The Internet Visit Rate, Its Monitoring and Analysis

JIRI KOHOUT, ANTONIN SLABY
Department of Informatics and Quantitative Methods, Faculty of Informatics and Management
University of Hradec Kralove
Rokitanskeho 62 500 03 Hradec Kralove
CZECH REPUBLIC
http://www.uhk.cz

Abstract: - Currently the Internet visit rate is a topic of increasing importance. It is a tool of monitoring success of marketing campaigns and comparing the popularity rate to other competitors. This article aims at analysing the process of measuring the frequency distribution of online customer's behaviour, highlighting the risks in interpretation of the results and presenting methods used in the data analysis process. Part One outlines a way of collecting data for the analysis by recording them to the web log. Then another method working on the active content principle is presented. The comparison of both methods follows, supported by the overview of advantages and disadvantages. Consequently, the gained results are summarized and result in the list of essential preconditions necessary for objective and complex analysis of visit rate data.

Key-Words: Internet visit rate, Monitoring

1 Introduction

Most of Internet communication connections is based on the server-client relation. The web server with contains stored information replies to many client requirements by sending demanded data, usually web sites. Having written the URL or clicked on the link, the DNS server translates the domain name to IP address and finds the server where the demanded web page is available. The server receives the demand and having found the file on its discs, it sends the demanded web page to the client.

Naturally, the whole process is much more complicated, especially with dynamic web pages (.NET, PHP etc.), where the page is processed with the use of client data form having been sent. The demand on any web page, either successfully replied or not, is recorded into so called web server log. . This information is essential for understanding the first method. In addition to it, most web pages present other links to images, media, cascade styles etc. If a web page contains 100 links, the basic one plus each linked file are counted and recorded, i.e. total 101 items.

2 Internet and Intranet

The Internet and intranet environments differ from the point of architecture, number of connected devices and, as for the visit rate, by the length of Internet path between the server and client. In the Internet the server and the client may reside on the opposite sides of the Earth. The client's demands and results are sent through numerous devices. (That is why the secured HTTPS protocol is used as otherwise the data might be displayed and read by administrators of these devices.) The main problem in monitoring and measuring of visit rate is in caching the web pages when once required web page is saved on a device which is closer to the client than the server which the web page is demanded from. Another caching may be present and run on the client-side if the browser is set not to upload the demanded web site again but to use the local copy saved on the disc. This process of saving copies misrepresents (distorts) objectivity of the measured results in the way explained below.

Cache on the client-side causes that only the first demand for the page from the computer is recorded, not the following ones. If the client requires the same address in a few-day period, the requirement may reach the server only in the first time, even if the web page is required by another person from the same computer.

Caches on the Internet provider-side cause the same misrepresenting but in much larger extent. Theoretically, thousands of clients may demand the same web site in one-hour time but only the first requirement reaches the server, the other clients receive the saved copy of the file of the provider only.

Originally, the Internet was not designed so as to record number of web page requirements. It prioritised the load decrease by cutting the data flow between Internet nodes and fast client service instead. This attitude is presented in the first definition of the HTTP protocol which states that every participant of the communication must have a cache memory so that the path and way of replying the demand is shorter [1], following HTTP protocol versions contain whole chapters on this problem, e.g. [2].

It is obvious from the above presented that the method of web server log analysis is not very exact one and it suits more to be used in the intranet of big corporations to analyse employee's web site visits, because caching on the whole path may be forbidden in this environment.

3 The Log Web Server Structure

Let's have a brief look at the web server log structure. It is usually organised as a text file where each demand for a file is presented on a single line. Demanded file may be a static or dynamic web page, image or any other type of file. It is possible to set the recording (output) directly in the databases. which may analysed and elaborated by special programmes, e.g Microsoft Reporting Services. The scope of recorded information and its format can be set by server administrator, e.g. recordings of some files (for example images) which are not relevant to the measurements can be switched off [6]. The process usually starts with the W3C Extended Log File Format which contains information about client demands, which include.

- Date and time when the demand was satisfied (solved)
- Client IP address
- User name if the client is to prove authenticity to the server
- The port where the communication run (was established)
- Methods used for transporting demands (GET, POST)
- The path to the demanded file
- Parameters of the demands for file
- Code of return of HTTP (200 – OK, 404 – not found, 500 – internal web server error, …). More information about codes can be found in [2]
- The type of client's browser
- Client's cookie (the unique client's identifier). It can but be deactivated by the client and the information is not sent to the server in this case.)
- URL of the previously visited web page by the client
- The file size of the reply, in bytes (i.e. number of bytes in the reply).

4 Server Log Abilities and Restrictions

We have presented the log structure. Now let's focus on what can be learnt from it and what is expected to happen during its processing. It is mainly:

- Monitoring of the visited web pages and so monitoring of user's activities (browsing) on the server. It is interesting in connection with data mining methods, that can be used for elaborating it. The problem is discussed below.
- Obtaing the sum all of web pages and unique visitors (visitors units) per time unit. It is used for the comparison of success rate in competing Internet presentations of some business subjects.
- List of demands resulting in errors (the web page does not exist or there occurred a script error in processing) and proportion to the satisfied demands. This is useful information for server administrators and application programmers.
- Information about which www addresses are preferred by server visitors. It is important feedback of running advertising campaigns and finding out the most frequently used addresses which the new visitors access the server from.
- What robots and of what parameters browse the server content for indexing it then in the internet search engines. It is important for content optimization of Internet search engines (so called SEO) which enables to display server links at top positions.
- Counting the data volume which is handed on clients. This information is useful for monitoring the total server load factor per time unit or average size of data handed on one client.
- Server load factor distribution during the period of the day, week, year. This is useful for marketing, server administrator, defining administrator's optimal working hours as he deals with users´ orders and inquiries.
- The visit rate statistics of single (individual) web pages. Useful for displaying important advertising campaigns on frequently visited web sites.
- Counting the users of various kinds of Internet browsers. Important for programmers to optimize applications for the particular browser. Currently but there are minimal differences in displaying the web sites on various browsers.

All the above mentioned statistics are very interesting. Unfortunately the results can be used on the technical level only. Because of the use of cache between the server and client, not all demands are recorded on the log. So it provides valuable results only to server administrators and programmers who are interested in data volume and accessibility of the service, and script errors – all the information can be found in log.

When using log for marketing purposes we should be very careful – either only part of data is recorded and/or

we are not able to identify the unique visitor. This is caused by the fact that the HTTP is a stateless protocol, which means that the client sends a demand to the server which having handled it finishes the connection with the client. The following demand is not related to the previous ones and so for the purpose of the visit rate analysis the unique visitor must be distinguished from the others by some identifier. The client IP address is not very suitable because the same one can be shared by hundreds of users within a company intranet. The user name or cookie suit more but most users are not under the server authentification or they switch the cookie off.

Recently, selected common characteristics have been used to identify the unique visitor. They are not contained (included) in the web server log but can be easily added. The unique chain of characters called "Session ID" is generated during the first client connection to the server, and then used in each action. The identifier is lost if all browser windows are closed by the client or no demands are sent to the server for a period predefined on the server. The "Session ID" is recorded in the log in case of customizing web links so that in addition to the standard parameters there is displayed "Session ID" also, and consequently the client identifier is part of URL requirements.

Another parameter, the web site expiry date and time, can be added to prevent the use of cache between the client and server. The client demand in this option reaches the server every time when required. However, the demand for server is supposed to be required each time when client wants to display the page, which is not always true.

5 Collecting data by active content

The principle of collecting data in the way of measuring by active content is the most often used substitute (complement) of the web server log method. It completely differs from previous methods and has both advantages and disadvantages in comparison to the previous method. A small piece of code (JavaScript, CGI script, …) is included into each web page represented by a transparent 1x1 pixel which cannot be seen (detected) by human eyes.

The code is required when the web site is loaded. It generates URL of the image which differs each time. Either the cache is switched on or off, the local cache, the providers cache do not register the image. The requirement is sent to the server which works (records data about client) in a similar way as does it web server by making record into web server log.

The server administrator has to switch the Log himself. If using the active content principle but, even some external company may generate the code, include it on each web site and in such a way measure the visit rate by use a special software. This method does not record some characteristics as in the web server log but it contains other ones, e.g. type of operating system, the version of the Internet browser, monitor resolution, support of other technologies in the Internet browser (Flash, Java, …).

Of course, other methods exist but their use brings some restrictions and disadvantages in comparison to those presented above. These methods include measurements on the client-side, measurements based on the Internet provider data, analysis of the packet transport by catching packets etc.

6 Comparison of the two methods

Both main methods have advantages and disadvantages, so it is optimal to use their combination and have both types of data processed by the same company. Let's pay attention to the comparison of them so as to learn what it may result in [3].

Advantages of the web server log measurements:

- The volume of data flowing from server to clients can be monitored.
- Browsers not supported JavaScript and similar technologies used in measurements by active content are also included into consideration.
- Accesses of the Internet browser robots is recorded and thus the web site can be optimized according to their requirements.
- Demands on all files, not only Internet web pages, are logged.
- Unsuccessful attempts resulting in errors are also recorded.
- It is a server log, so the data can be analysed back until the results of previous analysis is reached.
- Once collected data may be analysed by various tools. It is not necessary to rely on an external company methodology.
- The analysed data are not transported to other subjects, they are processed on the same server where the log file arose.

Disadvantages of the web server log measurements:

- Because of caching the visit rate analysis carried out by this method may not be exact enough.
- It is difficult to identify the unique visitor.
- No information on visitor's computer settings is presented in the recordings.
- Recordings are saved on the local server, if it crashes, all data are lost, including log files.

Advantages of measurements by active content:

- It solves the problem of caching web pages and records access of all currently online visitors.
- Detailed information on user's computer settings is provided.
- The topics of collecting and evaluating data are not necessary to be understood, you only enter the code and receive the results processed by an external company.

Disadvantages of measurements by active content:

- Some user accesses are not recorded, i.e. of those who have text type of Internet browser, displaying of images switched off, technologies used for measurements by active content forbidden, etc. (consequently access of indexing robots is not recorded to)
- It is necessary to include the code for measurements to each web page as a web sites without a code cannot be recalled.
- Possible problems in case of changing the visit rate service provider may occur.

7 Preparation for the data analysis

In the part above some basic methods of measuring the visit rate have been introduced. Having collected the data, we will clear them to receive valuable information. We will deal with processing the web server log because if the method of measuring by active content is applied, the web site provider has the data processed by an external company and there is no need to prepare anything. However, in the text the recommended way of gaining data by active content by one's own means is briefly described. This type of data is also under following clearing procedures, with several exceptions.
To extract information from the collected data, the following procedures must be kept [4]:

- Record data and join the server logs into one item in case of the web application is distributed on more servers.
- Remove unimportant information, especially demands on other types of files but web pages.
- Filter out the accesses of Internet search robots, employees, services checking the server state, automatically opened windows and other recordings which were not caused by any direct web page visitor's demand.
- In any real visit rate analysis remove recordings containing other HTTP return code than 200 (i.e. especially the codes generated in error demand or when forwarded to another web page).
- Add the data with information about customers, if available (see more in the following text)
- Calculation of comprehensive metrics if the data are collected from more sources.
- Presenting correct results to the right target groups of recipients.

In the process of data filtering the most complicated activity is to distinguish the Internet search robots from real users. There are not only good robots which generate relatively small number of demands but also those aiming at the DOS attack (i.e. service failure because of server overload), or browsing web site to get e-mail addresses for sending spams. These are the most frequent ways to filter them out:

- Save (place) the "robots.txt" file into the www presentation root directory before measuring the visit rate (This file contains exceptions for The Internet search robots which read it before they start to search pages).
- Remove recordings where the type of browser contains the word "bot".
- Download and analyse the current list of names of robots and robots' IP addresses from the Internet – if either the name of browser type or client's IP address agree, filter the line out.
- The most complicated method is to identify the robot according to its behaviour, e.g. it demands lots of web pages in a short period, especially if the intervals between demands is constant ones and the first demand aims at "robots.txt" file.

8 Effective measurements

What should be done to measure the visit rate effectively? To set the monitoring into log file of the web server of maximum number of characteristics because in the future we may be interested in other type of information than these days. To set the "caching forbidden" option for peak efficiency so as the maximum number of demands could be recorded on the web server. It is important to focus on the "robots.txt" file to minimize deviations caused by the Internet search robots.

Because of disadvantages of this method we also use the accompanying measurements by active content. If the financial situation is good, pay attention to the measuring code being part of each web site and follow the visit rate by several external companies which will be processing web server logs at the same time.

Has the maximum been done to collect complex information on the visit rates? The answer is both yes and no. If providers are only interested in visit rate statistics, the results are authentic (adequate). But if the

company pays enormous sums of money on advertising campaigns, does million-crown businesses, has thousands of customers and needs to learn who of them are perspective and who are not – in this situation the visit rate is not the suitable tool to apply.

A specialised software should be used to connect (i.e. relate, link or match) all types of data, i.e. the visit rate measured by active content and information about the users gained from the CRM (customer relationship management) systems.

9 Active content measuring special firm script

If the company uses a common measuring of the visit rate and hands web server log to an external processor, it relates the results to those gained from measurements by active content which results in reliable information. Information of this type cannot be used for processing in or by other systems as it is an aggregate quantity which has lost information about single users.

If the information about the visit rate of regular customers is required, it is recommended to have the custom-made programme to measure the active content created. The client-side contains a code similar to the ready-made one but the volume and structure of the data sent to the server by reading the measuring code differ. E.g. when visiting the server an unique user name parameter is sent there. The server adds information from the CRM system. In most cases it only records the user name and links it to statistic data which are integrated with the CRM information at the time of elaborating information about access.

10 Active content measuring by an external company

Creating of one's own solution of active content measuring might be very expensive, so why not to choose from an external company offer. It is still aimed at measurements and analysis, and the unprocessed data are not provided to the customer. When the companies understand there is a market gap, and rich web site providers require aggregate/non-aggregate data (i.e measurements and/related to CRM information) to receive the reliable results of visit rate, they will quickly react to the situation.

The unique identifier of each visitor available from the CRM system has to be added to the recordings, e.g. the user name is connected to the URL address and is functioning per every visit.

11 ntegration with CRM – Yes, or No?

The visit rate analysis related to user data will contribute more than common analysis offered by most companies. If you hesitate whether to use services of a special company for complex statistics of the CRM system data, define your answers to the following questions first:

- What do you expect to receive from the analysis? What sum of money are you going to spend? What profitability do you suppose to receive by gaining new knowledge about web site visitors?
- Do you prefer a non-recurring analysis or a special solution for ad-hoc connection to the CRM systems and visit rate recordings, where you can dynamically ask questions and receive answers without external company interventions?
- Are your data consistent enough in the CRM systems? First think about a solution optimizing data flows within intranet, then concentrate on the detailed visit rate analysis. If the data about customers are not complex or current, the analysis might be counter-productive.

12 Conclusion

This contribution is only a brief introduction to the problem of monitoring of internet visit rate. In journal publication will be given more information and some samples of solution

References:

[1]: Hypertext Transfer Protocol - HTTP/1.0. (October 2007). Internet access: URL: http://www.w3.org/Protocols/rfc1945/rfc1945

[2]: Hypertext Transfer Protocol - HTTP/1.1. (October, 2007). Internet access: URL: www.w3.org/Protocols/rfc2616/rfc2616

[3]: Collecting Web Data: A Look at Web Analytics Methodology (October, 2007). Internet access: URL:http://www.roirevolution.com/blog/2006/05/how_web_analytics_are_collected.html

[4]: IAB Ad Measurement Study (October, 2007). Internet access:
URL: http://www.iab.net/standards/pwc_report.pdf

[5]: Metriky a metodologie pro Internet reklamu (October 2007). Internet access: URL: http://www.park.cz/metriky_a_metodologie_pro_internet_reklamu

[6]: PETERSON, E. T. Web Site Measurement Hacks. O'Reilly, 2005, ISBN 0–596–00988–7