Multi-Dimensional Visualization of Web Access Logs

Ágoston Winkler, Sándor Juhász, Renáta Iváncsy

Department of Automation and Applied Informatics, Budapest University of Technology and Economics
{winkler.agoston, juhasz.sandor, ivancsy.renata}@aut.bme.hu

Abstract: Web content providers need to obtain precise information about the visits on their site in order to improve their services. They often contract with independent third party auditing companies. These companies collect an enormous amount of multi-dimensional data that is quite difficult to visualize. This paper describes four visualization techniques that can be employed for visualizing four dimension relationships expressively and shows their application with real web log data. It mentions some other methods that can be applied for more dimensions as well.

Keywords: multi-dimensional data visualization, histogram matrix, area block chart, parallel coordinates, web access logs, web auditing, data mining

1 Introduction

The rapid penetration of the Internet has caused a significant increase in the number of webpage downloads. Web content providers need to know precisely when and which pages are visited the most frequently in order to discover the interest of their visitors, to improve the contents of their pages and to determine the hardware solutions required for operating the site more effectively. Valuable visitor statistics can be used for finding advertisers, too. Content providers often contract with independent, reliable third party companies [1, 2, 3, 4] for auditing the activity of their site. These companies must collect a huge amount of data continuously.

Web access logs are interesting for marketing, sociological, and computer science researchers as well [5]. However, it is quite difficult to visualize such an enormous amount of data in such a way that it can be understood easily by the users [6]. It is not trivial either, how to show multi-dimensional relationships using two-dimensional media (e.g. paper or computer display).

Interesting results can also be achieved by using data mining and computational intelligence algorithms [7, 8]. As an example, one of the current researches aims at identifying the real persons behind the so-called Internet users that can be detected and tracked by using different cookies and user IDs [9]. The presentation of these results needs some special solutions as well.

This paper deals with the problem of visualizing multidimensional data, especially focusing on data obtained from web access logs. The data that was used in the researches is obtained from real world web access log files that were collected by a large Hungarian opinion and marketing research institute.

The organization of the paper is as follows. Section 2 presents the creation and the structure of the web access logs that was used during the researches, as well as the methods of obtaining the essential data from them, required for the visualization. Section 3 summarizes the main types of multi-dimensional visualization techniques, and describes four methods in details. It demonstrates how the proposed methods can be used for visualizing the web access logs as well. Finally, in the conclusion some more visualization methods are suggested that are worth examining in the future.

2 Web Access Logs

Web access logs are the files in which the webpage downloads are recorded by a server. This server can be the same as the one that contains the accessed page (most web server programs automatically write a log of the incoming requests) but it can be an external server as well, typically the server of an auditing company. In the latter case it is not trivial how to notify the auditor's server about the activity without adding a too large overhead.

A widely used technique that solves this problem is that a small, in most cases invisible picture (e.g. one single white pixel) that origins from the auditor's server is inserted into the page code. This means that it is not the original page accesses, but the picture downloads that are recorded in the auditor's log file. However, this method works properly because the URL of the original page can be found in the picture request (as the so-called HTTP referrer) so this important piece of information will be recorded as well [10].

As mentioned earlier, the data examined in this paper was obtained from a large Hungarian opinion and marketing research institute [1] that audits 450 web sites with more than 40 million page accesses every day. The server-side log files recording these downloads have a daily size of approximately 16-20 gigabytes.

Table 1 shows the data structure of the log entries recorded in the log that was examined during this research. 3rd party cookies are placed by the auditing

company, they are used to identify real persons by connecting their activity on different sites using 1st party cookies (placed by the browsed sites) and user IDs (used by certain sites with a user registration system) [9].

It can be seen well that there are some dimensions (e.g. time, URL, browser etc.) that are worth visualizing even without applying special data mining or computational intelligence algorithms to the data.

Name	Length	Description
3rd party cookie	22 bytes	Special identifiers used by future research in order to identify real persons
1st party cookie	26 bytes	
Medián user ID	40 bytes	
IP address	4 bytes	IP V4 address of the client
Timestamp	4 bytes	Number of seconds since 1 January 1970
Thematic code	16 bytes	A special code used by future research in order to group pages based on their topics.
URL	100 bytes	The URL of the accessed page
Browser	180 bytes	Information about the browser and some optional plug-ins installed on the client's computer

Table 1
Structure of the web access log entries

This paper focuses on the visualization of the following four dimensions: the day type (working day, Saturday, Sunday), the two-hour interval which contains the time of the visit, the site, and the number of page downloads. The values for working days were calculated as the arithmetic means of the values of Monday, Tuesday, Wednesday, Thursday, and Friday. Table 2 shows the 10 most frequently visited sites that have been examined.

Name	Description	Weekly downloads
myvip.com	Community network portal	54 312 464
ctnetwork.hu	Web advertising agency	39 152 474
citromail.hu	Free mail service	15 096 426
love.hu	Lonely hearts service	14 261 655
origo.hu	General news portal and search engine	12 741 269
freemail.hu	Free mail service	12 290 429
startlap.com	Thematic link collection	10 301 617
index.hu	General news portal	8 716 198
p**a.hu	Erotic site	7 867 608
kep.tar.hu	Photo sharing service	5 003 462

Table 2
The 10 most frequently visited sites audited by Medián Webaudit with the number of visitors between 2 and 8 September 2006

In order to create the charts, a table with four columns had to be derived from the original log files with the appropriate data fields. As log files can contain a maximum of 5 million entries, 7-9 log files are created each day. Each file has a size of approximately 2 gigabytes. In order to obtain the required table, a log file analyzer has been written that was run separately for each data file. By using an asynchronous buffered file reader and a hash table implemented in .NET 2.0 framework, a quite promising processing time has been reached. It took 140 seconds to process one file by using a PC running Microsoft Windows XP 5.1 SP2 with an Intel Mobile Pentium Dothan 2MB/FSB400 1.7 GHz processor, 512 MB RAM, and a 40 GB hard disk with 8 MB cache.

The next step was to concatenate the results obtained from the log files of a whole week into one table. After this transformation visualization could have been started.

3 Visualization Techniques

Multi-dimensional visualization techniques are widely used in different areas of science [11]. Some of them aim at showing all the data points from different aspects. As an example, scatter plot matrices visualize all dimensions versus each other. Other methods use some kind of aggregation to provide the viewer with a more abstract view, such as the multi-dimensional versions of histograms and pie charts. If the purpose is to find hidden relationships, it may be enough to find data points that are close to each other, exact dependences are not important at this level. Some special visualization techniques like Radviz and Gridviz [12] are based on this principle. In this case even the number of dimensions can be decreased: this technique is called dimension reduction, an example for it is the so-called Sammon plots [13].

This section presents four relatively simple, but quite powerful techniques based on aggregation: histogram matrices, area block charts, parallel coordinates, and circular parallel coordinates. Then it shows the results that can be obtained by applying these methods to the examined data.

3.1 Histogram Matrices

Histograms [11] are very expressive tools for presenting two-dimensional relationships. The simplest way to extend this ability to 4 dimensions is placing the histograms into a two-dimensional matrix. Figure 1 shows such a histogram matrix.

In the example, the columns of the matrix represent the day types and the rows represent different sites. The histograms inside the matrix show the number of the

page downloads in the specified two-hour intervals. In addition, a line shows the weekly average values for the given site. To improve expressiveness, histograms of different day types can be shown in different colors.

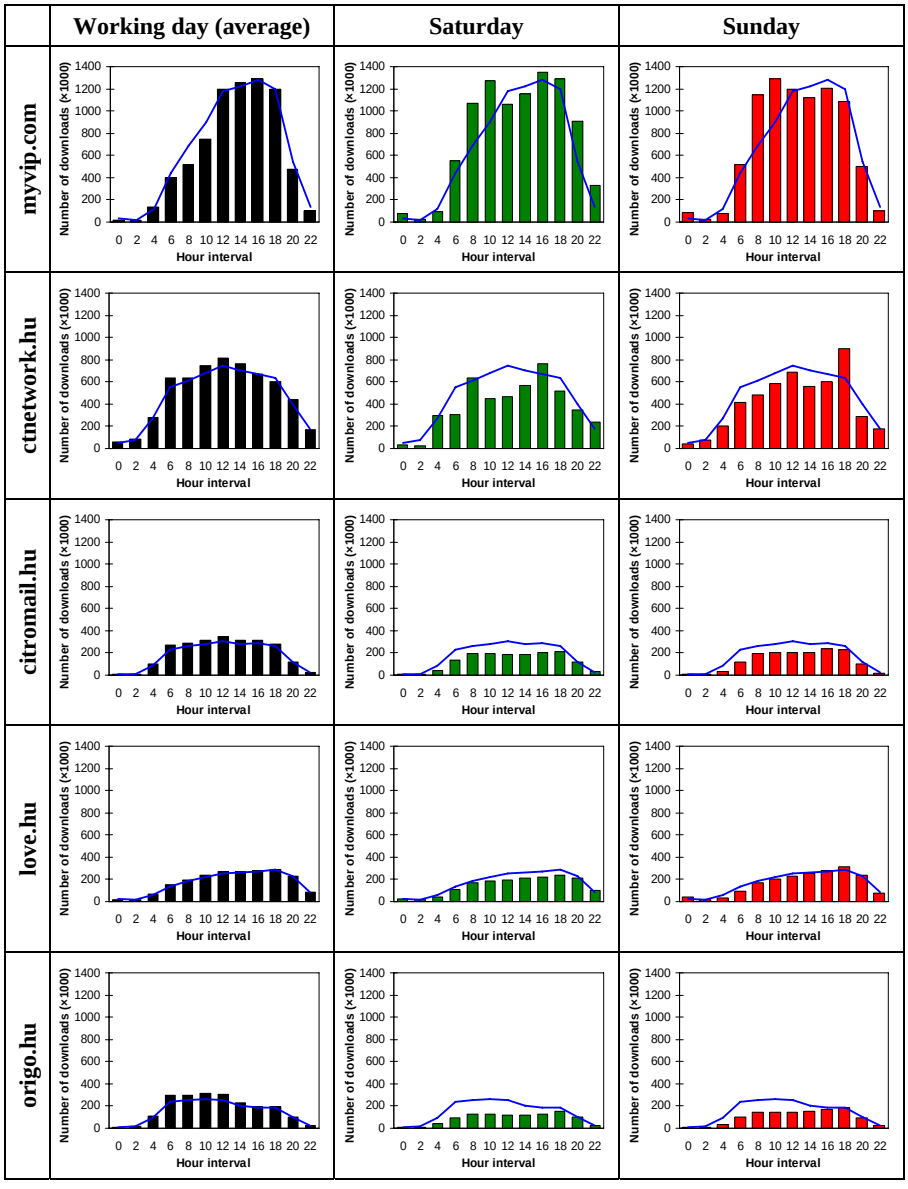

Figure 1
Histogram matrix showing the number of page downloads for two-hour time intervals, on different sites and day types (top 5 sites)

A great advantage of histogram matrices is that precise values are easy to read. However these matrices can be quite large and as they are actually 'simple' histograms arranged side by side, nontrivial relationships are relatively difficult to find as similar histograms are not necessarily near each other.

On the other hand, many interesting relationships can be seen in the matrix as well. As an example, the examined community network is more frequently visited at the weekend than during working days. However, the advertising agency and the news portal actually show the opposite trend.

3.2 Area Block Charts

4 dimension data can also be visualized using one single chart. Figure 2 shows a so-called area block chart that utilizes that a three-dimensional coordinate system can be presented in two dimensions so that one can imagine the relationships in the original three-dimensional space.

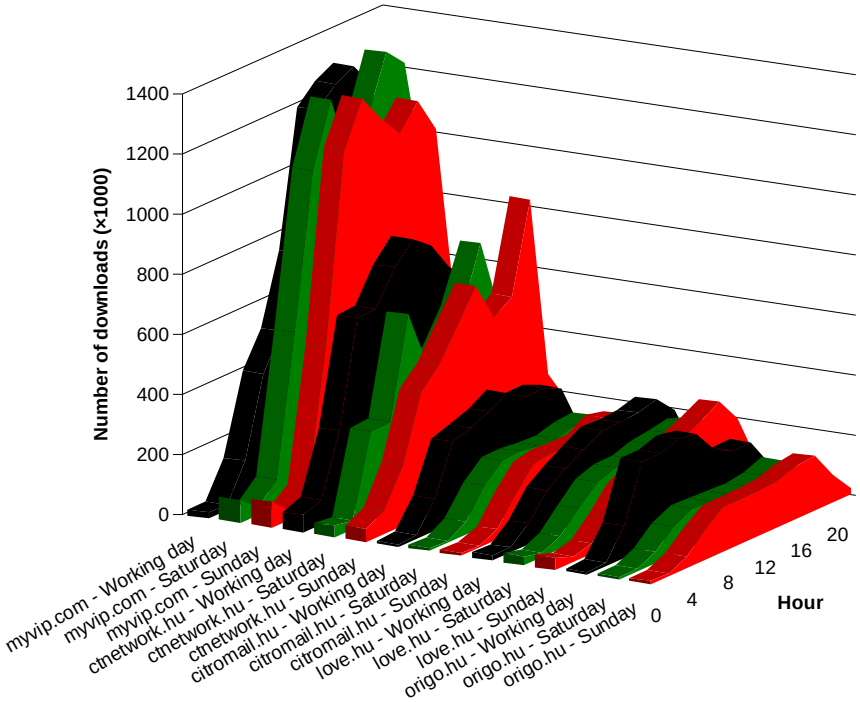

Figure 2
Area block chart showing the number of page downloads for two-hour time intervals, on different sites and day types (top 5 sites)

In the example, three axes represent the site, the hour and the number of page downloads. As the fourth dimension, different day types are shown as separate blocks but may be distinguished by using different colors or labels (the same way as in the histogram matrix).

The advantage of this method is that it provides an overall picture, all the aggregated data can be seen in one chart making easier to find special relationships. However, similar data sequences are not necessarily near each other by using this technique either. Furthermore, it is more difficult to read the precise values at certain positions of the chart: histogram matrices are better for the latter purpose. Another disadvantage is that relatively few data sequences (it this example, sites) can be presented.

3.3 Parallel Coordinates and Circular Parallel Coordinates

Webpage downloads can be presented by using the technique called parallel coordinates [11, 14] that works with lines.

On Figure 3 hour intervals are separated by vertical lines, these show the relative frequency of downloads (on a certain day type and web site). The values of the same site and day type are connected by lines. Day types can be color-coded in the same way as in the previous cases.

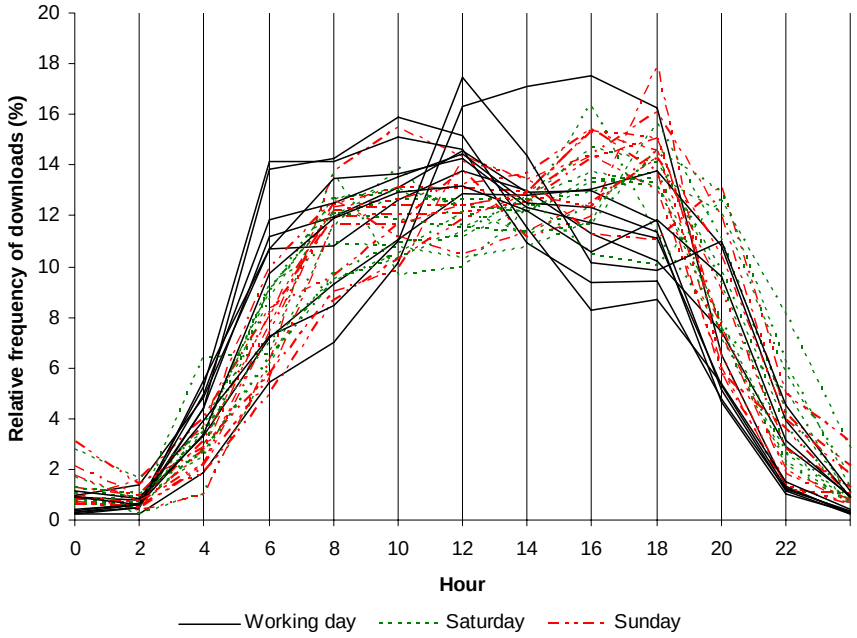

Figure 3
Parallel coordinates showing the relative frequency of page downloads for two-hour time intervals, on different sites and day types (top 10 sites)

This technique could be used for displaying precise values as well, but it is not really intended for this purpose, rather to show trends. If more value sequences are shown using parallel coordinates, like in this example, the different sequences cannot be distinguished. A few sequences can be labeled but that would show the same relationships as an area block chart which is much more spectacular. Thus, if the aim is to show precise values, the usage of area block charts or histogram matrices is rather suggested. However, much more data sequences can be presented by using this method.

A modified version of parallel coordinates is called circular parallel coordinates [11]. The only difference is that values are displayed on no vertical lines, but the radii of a circle. Figure 4 shows how the examined data looks like by using this technique.

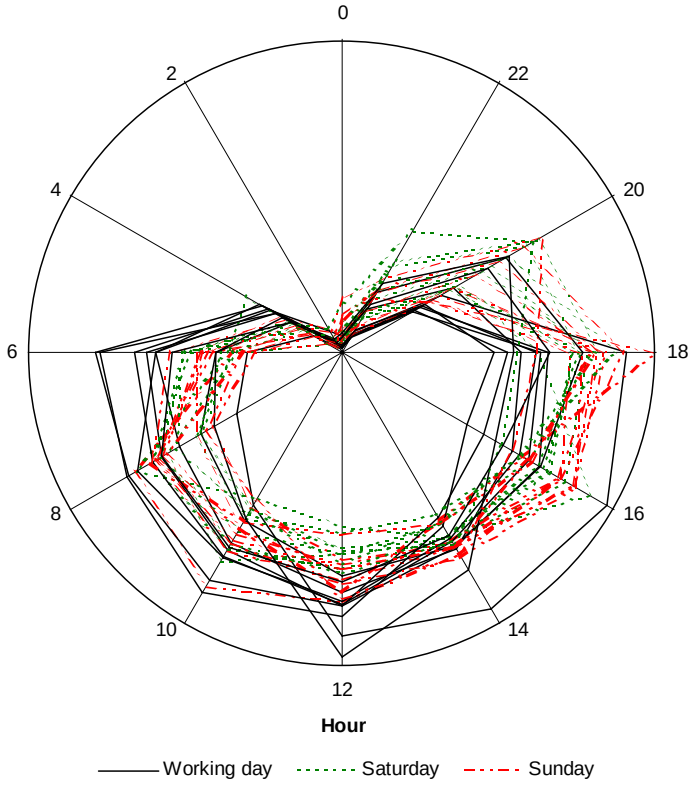

Figure 4
Circular parallel coordinates showing the relative frequency of page downloads for two-hour time intervals, on different sites and day types (top 10 sites)

Figure 3 and 4 show well that night Internet usage is not significant, furthermore, most of the pages are accessed in the afternoon. Different web sites show similar trends especially at the weekend, differences are greater on working days.

Conclusion

This paper summarized the main types of multi-dimensional visualization techniques and proposed four of them for visualizing the relationships that can be found in the logs: histogram matrices, area block charts, parallel coordinates and circular parallel coordinates. It presented their application with some data obtained from real web access log files. It showed that the four-dimensional data of the log files can be visualized expressively by using these methods.

However there are some other techniques that can be employed successfully for such purposes even in more than four dimensions. This possibility will be really important when not only page downloads but the activity of real users will have to be visualized. In this case various kinds of demographic data will be available for the users, multiplying the opportunities of multi-dimensional visualization and the number of dimensions as well. The examination of these techniques is the subject of future research.

Acknowledgement

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

References

[1] Medián Webaudit, http://www.webaudit.hu/

[2] Auditonline, http://www.auditonline.hu/

[3] Gemius/Ipsos Audience, http://www.szondaipsos.hu/hu/ipsos/gemius

[4] Coremetrics First Party Data Collection, http://www.coremetrics.com/technology/first_party.html

[5] Ed H. Chi: Improving Web Usability Through Visualization, in IEEE Internet Computing, Volume 6, Issue 2, Piscataway, 2002, pp. 64-71

[6] Harry Hochheiser, Ben Shneiderman: Coordinating Overviews and Detail Views of WWW Log Data, Workshop on New Paradigms in Information Visualization and Manipulation (NPIVM 2000), ACM Press, 2000

[7] Jeffrey Heer, Ed H. Chi: Identification of Web User Traffic Composition Using Multi-Modal Clustering and Information Scent, in Proceedings of Workshop on Web Mining, SIAM Conference Data Mining, SIAM Press, Philadelphia, April 2001, pp. 51-58

[8] Amir H. Youssefi, David J. Duke, Mohammed J. Zaki, Ephraim P. Glinert: Visual Web Mining, 13th International World Wide Web Conference (poster proceedings), New York, May 2004

[9] Csaba Legány, Attila Babos, Sándor Juhász: Cookie-Chain-based Discovery of Relation between Internet Users and Real Persons, 5th

International Conference on Information System Development (ISD 2006), 2006

[10] Jack Powers: Counting Clicks: Auditing Your Web Site Activity, Fall 96 Internet World, New York, December 1996

[11] Patrick Hoffman, Georges Grinstein: Visualizations for High Dimensional Data Mining - Table Visualizations, http://home.comcast.net/~patrick. hoffman/viz/MIV-datamining.pdf

[12] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, Eugene Stanley: DNA Visual and Analytic Data Mining, in IEEE Visualization '97, Phoenix, 1997, pp. 437-441

[13] John W. Sammon Jr.: A nonlinear mapping for data structure analysis, in IEEE Transactions on Computers, C-18 (5), May 1969, pp. 401-409

[14] Alfred Inselberg: The Plane with Parallel Coordinates, in Special Issue on Computational Geometry, The Visual Computer, Volume 1, 1985, pp. 69-91