2001

Visualization of usage patterns for web personalization

Victor Perotti

Adam Burke

VISUALIZATION OF USAGE PATTERNS FOR WEB PERSONALIZATION

Victor Perotti, Assistant Professor of Management Information Systems
Adam Burke, Graduate Research Assistant
Rochester Institute of Technology College of Business
107 Lomb Memorial Drive
Rochester, New York 14623
Phone: 716-476-7753
Fax: 716-475-7055
Email: vic@mail.rit.edu

INTRODUCTION

The rise of the Internet has created new and unique demands for information management. Perhaps none of these demands is more important than the need to understand users and their usage patterns. By leveraging an understanding of usage patterns, personalization and customization of web content have become commonplace features on many eBusiness web sites. Consumers, however, enjoy their right to privacy and are increasingly wary when their personal information is requested. A less invasive approach is to look at the aggregate behavior of all users, and to try to identify trends therein. Once these trends are identified, a user can be classified as a member of a particular group, and customized web content can be delivered.

The process of deciding what content is relevant to a user based on the aggregate behavior of earlier users is a topic of active research today. Most of the work in this area has focused on how web site structure can be modified on the fly so that it is more relevant to each user. Several new algorithms and approaches have been designed to do exactly this. In parallel to this research, several studies have examined how web site usage can best be communicated to human experts. Visualization is an obvious tool to help this process, since it can make clear patterns that are difficult to discover in the extensive amount of data generated through Internet browsing.

The present paper will discuss both on these research streams, and then explain how we build on them to create a new visualization of web sites. Our new visualization, the *web usage plot*, will make use of a scatter plot to show groupings of web pages based on their usage.

WEB USAGE MINING

Recently, several researchers have begun to implement clustering algorithms to capture this aggregate user information on the web. In this research, the primary source of information about web usage patterns is the web access log continuously created by the web server. Perkowitz & Etzioni (1998) introduce the "PageGather" algorithm to perform cluster mining of these web access logs. Their approach involves the generation of "index pages", which provide organized sets of links to web pages that are related, but are not currently connected. Mobasher, Cooley and Srivastava.(1999) present a usage clustering approach that serves as a recommendation engine, suggesting certain related pages to web page visitors as they visit.

Common approaches to Web Usage Mining

The web usage clustering approaches from Mobasher, et al. and Perkowitz & Etzioni provide a powerful set of tools for performing web usage mining. The first steps for these two groups is remarkably similar, and can be summarized as a three-part process. Figure 1 is an illustration of this process. First, the web log data is cleaned by eliminating the web log items that result from images, and other non-structural web elements. Second, the cleaned data is divided into visitor sessions. A session is simply a group of web pages that are visited during one "trip" to a web site. Each session will be from one client, and the duration of the session is limited so that if a significant amount of time passes between page requests, a new session is begun. In the third and final step, session information is summarized with a co-occurrence matrix, which records how many times two pages are visited together in the same session. The co-occurrence matrix is a real representation of the usage patterns on the web site. For example, the largest number in the co-occurrence matrix indicates the two pages at the web site that were visited together the most frequently. From this common beginning, the algorithms of the two research groups differ, however the creation of the co-occurrence matrix is a crucial first step for our visualization of web usage patterns.

VISUALIZATION

An alternative approach to the two automated procedures above is the use of human expertise to make intelligent choices about web site updates. Certainly, human insight, goals or relevant experience can be valuable in understanding how best to match a web site's structure to its usage. Unfortunately, the pace of change and the sheer volume of traffic can create a significant obstacle for people trying to get a sense of how a particular organization of web content is functioning. As an example, the huge and complex web logs created by web server software provide little value without some additional tools to help understand them.

Visualization represents a natural tool to make clear the complex set of usage patterns for a site. Already, there are several companies (e.g. SiteStats.com or WebTrends.com) that provide visualization of a web site's usage. At this point, however, their tools are fairly simple. Figure 2 shows a bar graph from sitestats.com depicting the total number of hits for a site for the early part of January 2001. In addition to bar graphs like Figure 1, line charts and pie charts summarize the hits by day, week or month, or show which pages are most popular. Similar charts can also help to show which geographic regions are sending the most visitors. These tools provide a first step, but do not enable the user to identify relationships between pages.

Hochheiser and Shneiderman (1999) have recently proposed the use of interactive starfield visualizations to explore and discover new relationships in web access log data. Starfield visualizations are powerful tools since the designer can use point size, placement and color to communicate different dimensions. As an example, Figure 3 shows a visualization comparing the Referrer to the URL being requested. Although the starfield visualization offers new prospects for understanding a web site, it does not directly help with the problem of understanding the usage patterns across multiple web pages.

A NEW VISUALIZATION OF WEB USAGE PATTERNS

Our present research seeks to build on the web usage mining and visualization approaches to create a new visualization of web usage patterns. To do that, we follow the first steps of web usage mining (shown in Figure 1) to create a co-occurrence matrix. Once the co-occurrence matrix is formed, the statistical technique of multidimensional scaling will be used to provide a graphical representation of the web usage.

Multidimensional Scaling

Multidimensional Scaling (MDS) is a multivariate data analysis tool frequently used to create visual representations of complex data sets. MDS assumes that the elements being described have an underlying spatial structure in two or more dimensions, and it seeks to find that structure through statistical analysis. A common example of MDS is the reconstruction of spatial configurations of cities: Given a set of cities, and a table of the distance between each pair of these cities, the MDS algorithm can create a map of the cities that correctly shows their locations relative to each other.

To apply MDS to the problem of web usage analysis, the co-occurrence matrix can easily be converted to such set of "distances". This is done by subtracting each value in the matrix from a fixed number greater than the maximum number in the matrix. This way, the pairs of web pages that are visited in the same session the most frequently are mapped to have a small distance, while those that are seldom visited together have a large distance.

Web usage plots

The output of the MDS process is a set of coordinates for each web page. These coordinates can be visualized using many popularly available packages. We call the resulting visualization a *web usage plot.* Figure 4 shows one such visualization for RIT's College of Business (COB) web site (http://www.cob.rit.edu). Each point in the visualization represents one web page at the COB site. The groups of points in the MDS output represent clusters of the web pages that are visited together the most frequently.

One of the first things that can be read from this visualization is the **primary usage pattern** for the web site. For example, the large cluster in the lower right-hand quadrant reflect all of the web pages dedicated to the graduate programs offered in the College of Business. Because they are close together, we know that when a user goes to one of these pages, they are likely to visit the other pages during the same session. Similarly, the cluster at the top, center includes the web pages dedicated to the undergraduate program information. Pages showing faculty biographical information are found in the cluster in the lower left quadrant of the chart. These three clusters represent the three most consistent patterns of user visits: graduate program information, undergraduate program information and faculty information.

While these clusters are heavily impacted by the web site structure, the visual groupings are not constrained to follow the link structure on the site. Indeed, one of the great advantages of this type of analysis is the indication of "**missed opportunities**." For example, the group of points in the lower right-hand quadrant that lead from the vertical axis to the graduate program cluster are representations of our alumni pages. Because they are close to the graduate pages, it is clear that users are often visiting the alumni pages and the graduate information pages in the same session. However, there is NO link from the alumni page to the graduate programs page. Instead, visitors have to go back to the site's Home page to get from the alumni page to the graduate programs page. This missed opportunity is made obvious by the plot.

DISCUSSION AND FUTURE WORK

The research in progress presented here describes a new visualization of web usage patterns called *web usage plots*. This simple technique can yield powerful results, since a human agent can see clearly from the web usage plot how visitors are using the site.

Of course, there is much opportunity to build on this simple tool. As it stands now, the visualization process described in this paper requires the coordinated use of a variety of tools, including custom-written Java code, a multidimensional scaling package and a visualization package. Unifying these algorithms into a single application would create an interactive tool that could enable a better understanding of web usage. For example, by allowing the selection of subsets of the web access log, a unified tool could show how the visitors from New York, the visitors from the North America, or all visitors are using the web site up to the present moment.

Overlaying the web site structure onto this visualization would make the web site structure judgements described above much easier. A user could directly determine the mapping between the visual points and the web pages without referring to a separate document. Also, it would be possible to overlay our scatter plot with visual images of the statistical clusters generated by Mobasher, Cooley and Srivastava (1999) or Perkowitz and Etzioni (1998). Doing this would enable a comparison of the automated statistical clustering algorithms and the visual clusters that human experts consider important.

BIBLIOGRAPHY

Hochheiser, H. and Shneiderman, B. (1999) *Understanding Patterns of User Visits to Web Sites: Interactive.* Technical Report: CS-TR-3989 University of Maryland Institute for Advanced Computer Studies. Department of Computer Science, University of Maryland.

Mobasher, B. Cooley, R. Srivastava, J. (1999) *Automatic Personalization Through Web Usage Mining*, Technical Report TR99-010, Department of Computer Science, Depaul University, 1999.

Mobasher, B. Cooley, R. Srivastava, J. (2000) *Automatic Personalization Based on Web Usage Mining*, Communications of the ACM, 43-8.

Perkowitz, M. and Etzioni, O. (1998). *Adaptive Web Sites: Automatically Synthesizing Web pages*, Presented at the American Association of Artificial Intelligence conference.

Perkowitz, M. and Etzioni, O. (1997). *Adaptive Sites: Automatically Learning from User Access Patterns*,
Presented at the World Wide Web 6th conference.

Perkowitz, M. and Etzioni, O. (1997). *Adaptive Web Sites: An AI Challenge*,
Presented at International Joint Conferences on Artificial Intelligence.

RIT College of Business web site: http://www.cob.rit.edu

Sitestats web site: http://www.sitestats.com

Webtrends web site: http://www.webtrends.com

FIGURES

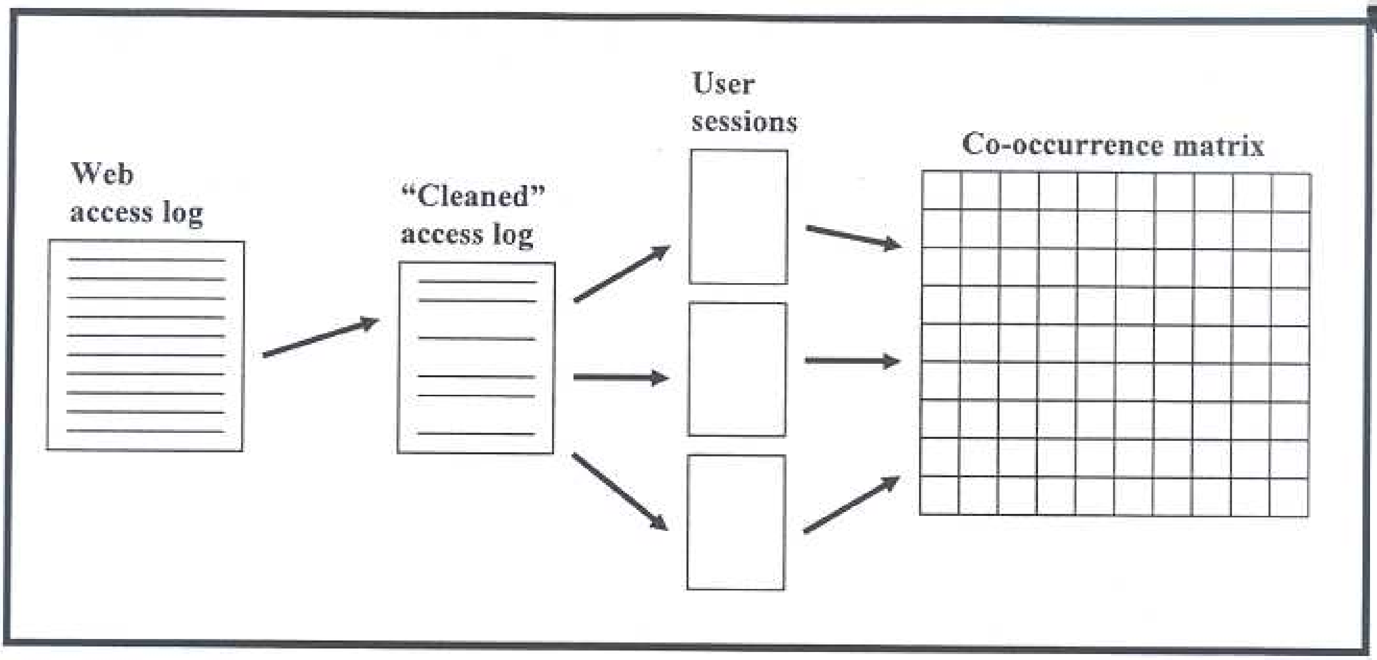

Figure 1: Process used to develop the co-occurrence matrix.

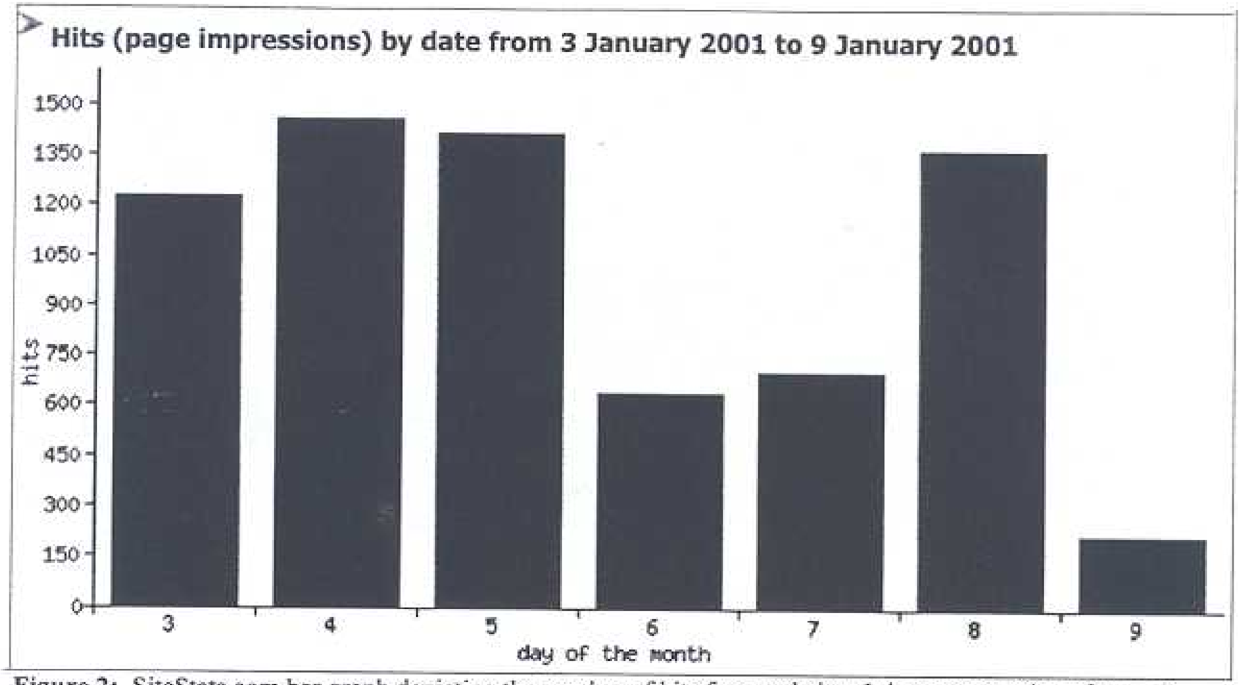

Figure 2: SiteStats.com bar graph depicting the number of hits for a website. It is representative of one of any number of simplistic visualizations commercially available today for websites.

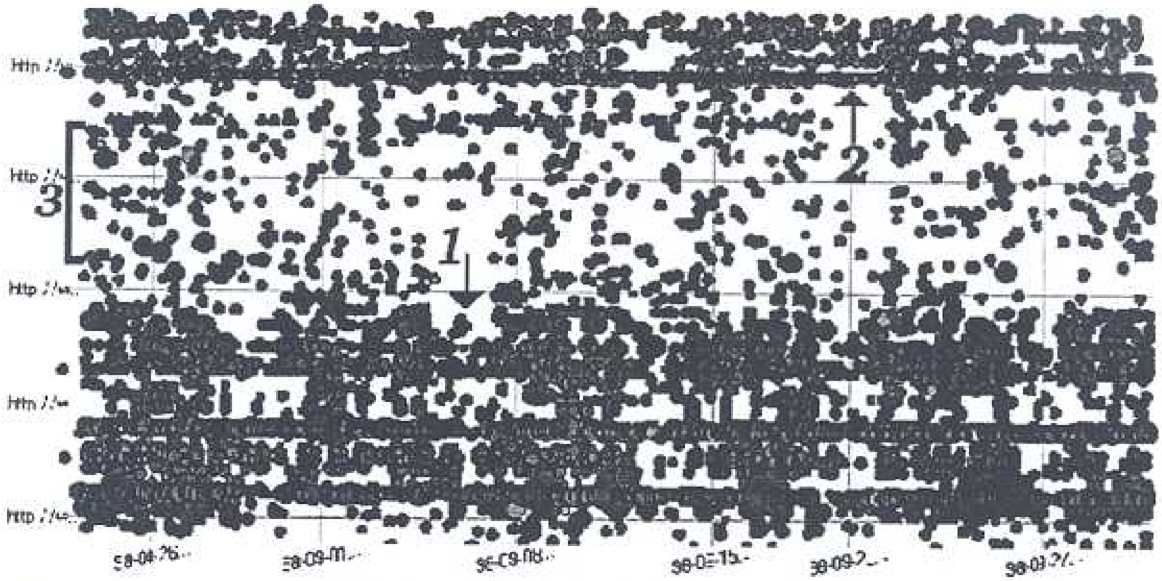

Figure 3: An example of a starfield visualization of web page usage trends, comparing Referrer to URL.

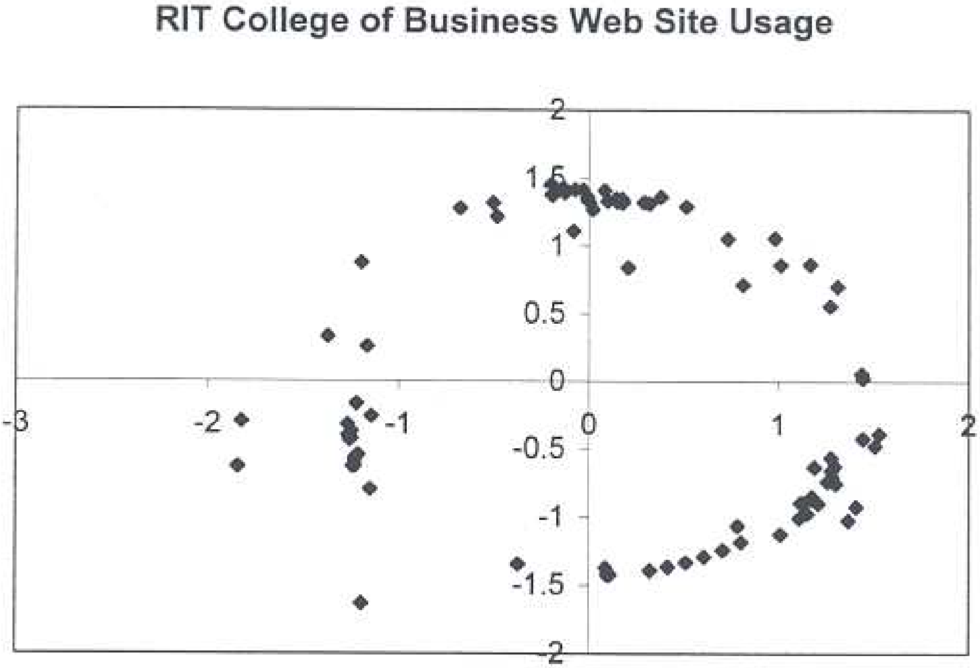

Figure 4: A web usage plot showing the patterns for the Rochester Institute of Technology College of Business web site. Closely spaced points indicate web pages that were visited together frequently.