Case Study: E-Commerce Clickstream Visualization

Jeffrey Brainerd Barry Becker
Blue Martini Software
[brainerd, becker]@bluemartini.com

Abstract

We have developed an interactive, scalable visualization tool for analyzing the behavior of users of a web site. Our system not only shows site topology and traffic flow, but by segmenting site traffic data based on user attributes, including demographic data and purchase history, we can present a more complete picture of web site usage. This can lead to a more focussed analysis that allows direct comparison between user segments, and ultimately a deeper understanding of how users interact with a site. The tool is designed for real world use, and we present a usage study of the tool by analyzing the data of a failed "dot-com".

1. Introduction

As the internet matures as a medium for doing business, companies are trying to move beyond the basics of business transactions into a deeper level of understanding of what is occurring at their web site. Primary to this is the understanding of how their customers are interacting with the site, which includes not only navigation patterns, but other customer information such as demographic data and buying habits. In traditional ("brick and mortar") retail, it is common for managers or other specialists to analyze customers' behavior in a physical store by following them around, noting the path through the store, watching how their eyes filter through the merchandise and seeing their facial expressions [13]. A key component here is not only seeing the navigation patterns of the customer, but combining that with other information about that customer, including demographic and purchase information. It is this deeper level of understanding of the customer and how the customer is behaving in the store that allows the retail manager to make more informed business decisions. In the virtual marketplace, there is no direct, physical interaction with customers, so we must come up with new ways of understanding customers' behavior in a web site.

Our goal is to aid in the process of understanding the interactions between users of a web site, and the web site itself. We present an easy to understand visualization tool that combines navigation patterns with relevant customer attributes to provide a powerful view of customers' experiences inside a web site. Direct comparisons between customer segments, for instance between different age groups, are possible. We are taking a step forward in understanding who the users of a web site are and how they are behaving.

There have been several attempts to visualize the structure of web sites [3,9]. Frecon and Smith [6] added a notion of history to the topological representation. Others have done work to visualize the geographic connectivity of the web [4,10] rather than web site connectivity, but the display issues are similar. WebQuilt [2] allows one to graphically compare an optimal path for a task as specified by a web site designer with alternatives taken by users. Chi et al [16] present an algorithm to infer the information need of users from well-traveled paths. Some interesting ideas on navigation through large graphs can be seen in Munzner's Hyperbolic Viewer [9] and Abello & Korn's Massive Graph Visualizer [1]. There has also been some interesting work on the effective use of color to distinguish partitions within a graph [4,7,11]. Our work adds dynamic segmentation of site traffic and other forms of interaction.

2. Data Collection and Preparation

The collection and preparation of the data used by the Clickstream Visualizer (ClickViz) is an essential step in the process of analyzing web usage behavior. The data used in developing this tool, including clickstream data, purchase history data, and demographic data, is collected live from a web site. Additional demographic data can be merged in from third party vendors. Logging support and unlimited data object attribution tie together clicks from the web site with customer information and purchase history. An offline data warehouse is built, and the data is transformed using filter and aggregation operations. A series of first-order Markov models [5,12] are built on a server, segmented by any number of pre-selected attributes and their corresponding values. Each model is simply a matrix of transition probabilities between two states (e.g. web pages), and is visualized directly.

Maintaining both scalability and tool interactivity has been a constant issue, especially in using real clickstream data, with millions of clicks. By allowing users to preselect attributes of interest, we set up a foundation that allows the user a good deal of interactivity within the tool while keeping the data size manageable on the client. The data returned to the client is proportional to the number of templates (or pages) squared times the number of segments. It is independent of the number of clicks.

3. Using ClickViz

The main ClickViz window is shown in Figure 1. The viewer shows a single directed graph and a control panel for the user interface controls. Each node in the directed graph can be thought of as representing a web page in the web site, and is labeled with the name of the web page. In our system, each node is actually a jsp page [16], which can be compiled and run to deliver a custom web page. This keeps the number of nodes of even large web sites low enough for effective graph visualization. Each node is also given an icon representation based on its group. This can greatly

Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)
1522-4048/01 $17.00 © 2001 IEEE

aid in the comprehension of the graph, especially with large numbers of nodes. Visually grouping nodes can often help to organize a complicated graph structure. One might be able to easily detect a pattern through the checkout pages by identifying the checkout icons on certain nodes, for example, whereas that pattern would be much more difficult to pick out using just the node labels as visual identifiers. Groups are assigned based on the path name of the web page in the file system. Typically a large web site will group related pages, such as "Search" and "Checkout" pages. Explicit "Enter" and "Exit" nodes are automatically created during the server side processing of the data. Thus every path into the web site comes from the "Enter" node, and every path out of the web site terminates in the "Exit" node. This makes isolating and viewing the top entrance and exit pages, a common analysis scenario, very easy. Each edge in the graph represents a set of transitions from one web page to another. Edge width is proportional to the number of transitions in the set.

Frequently there is so much data in total that the resulting graph becomes incomprehensible. Filtering mechanisms are therefore very important in improving the usability of the tool. Edges can be filtered out by weight, allowing a user to easily see the most heavily traveled paths. Edges can also be filtered based on the current node selection. Nodes are selected in the viewer by direct mouse selection or by using the integrated table control, which allows selection of groups of similar nodes through sorting. In addition, users can choose to show only edges into or out of selected nodes. A typical usage scenario that is improved by this type of filtering is selecting a group of nodes, say checkout nodes, and filtering to see only edges out of those nodes. Thus a user can detect whether or not web site users are flowing through the checkout process as intended. Typical questions that can be answered in this manner include: Are users exiting at the last checkout page? Are users jumping off to other parts of the site?

Users can further increase the effectiveness of the display by changing the layout of the graph. The default layout is hierarchical, which lays out the nodes top to bottom such that nodes with higher out-degree are positioned high, and nodes with higher in-degree are positioned lower. This has the effect of creating a graph that in general mirrors the actual flow of site users through a web site. Another very useful layout is circular, which lays out groups of related nodes together in circles, so that similar web pages, such as the search-related pages, are in close proximity. This layout is particularly effective in accentuating the relationships within and between groups of nodes (see Figure 2). We use third party software from Tom Sawyer Software to aid in automatically generating the complex graph layouts [14].

One of the primary strengths of ClickViz is the ability to visually differentiate clickstreams based on customer segments. This is accomplished using edge color. Every value of an attribute is assigned a different color, which determines the color of each edge that corresponds to that attribute value. Users choose from a preselected list of attributes, then can further choose which of those values will be shown in the graph. This allows both the analysis of a single segment as well as the comparison of different segments in a single graph.

4. Analysis of a "Dot-com"

In this section we describe how we used the ClickViz tool to understand the data collected from Gazelle.com, a now defunct online retailer of leg care products. This data was used as part of the KDD Cup 2000 competition [15]. We discuss several scenarios, including identifying differences in navigation behavior based on gender differences, problems in the checkout process, and analysis of an online newsletter.

The data consists of approximately 4,000 customers who produced 900,000 requests (clicks), 400,000 sessions, and 2,000 orders on 1,000 distinct products over a two month period between 1/29/00 and 3/29/00. Demographic data, such as gender, was collected via an online survey form that was filled out by site users during a registration process, and supplemented with data from Acxiom, a third-party data supplier.

4.1 Gender Differences

In this section we analyze the similarities and differences in the user navigation patterns of men and women. One possible business use of this exercise is to enhance real-time personalization of the site based on the user's gender. Figure 3 shows how gender influences navigation. Males are shown in Figure 3(a) and females in 3(b). Notice in both frames the general flow from top to bottom of the graph. We can immediately identify certain sections of the site such as the enter and exit pages (top and bottom), the search pages and product pages (middle) and the checkout pages (near the bottom). From this framework we can not only confirm that there are differences in the overall navigation patterns between men and women, but we can also generate an overview narrative of what those differences are. Men, in general, tend to navigate straight through the site: entering, getting what they need, and then exiting the site via the checkout process. Women on the other hand, take a less direct, but still identifiable route through the site. There is much more of a browsing process that includes use of the search pages, the product detail pages, and other supporting pages that may have general information related to the primary interest of the site (leg care). The preceding discussion demonstrates the strength of this tool in visualizing navigation patterns at a high level, as the first, "overview" step of the analysis process. We have framed a potential avenue for investigation. We can then dig down, selecting various pages to get more detailed information backing up the trends that we saw at a high level.

4.2 Checkout Process Analysis

In almost all cases, the checkout process is a very important component of the web site from a navigational and operational point of view. A web site user who enters the checkout process has indicated a desire to purchase something, and the checkout process is designed to expedite that purchase. It is very bad if a potential customer abandons the site from within the checkout process, and this may in fact point to a confusing process, or a broken link. In our analysis, we are therefore primarily interested in identifying any behavior that deviates from a straight-through, predictable navigation pattern through the checkout pages. Figure

4 shows a section of graph that highlights the checkout process. We easily identify the straight-through pattern that most purchasers take, which is expected. The non-purchasers have much less directed paths. There are self-loops on some of the checkout pages and edges from intermediate steps in the checkout process to exit, indicating broken links or simply a confusing checkout process that needs to be investigated further.

4.3 Analysis of the "Gazette"

Finally we looked at some of the "supporting" web pages in this site. In this site there was an online newsletter, a series of articles that provided information relating to leg care issues. We wanted to see if analyzing the navigation patterns of these pages could help us understand if these pages were helpful in driving sales on the site.

We observed that non-purchasers were following a pattern of navigating from one newsletter to another in succession; but instead of returning to the main site for a possible purchase, they were instead exiting the site. Thus we could see that not only were the newsletter pages not driving sales, but they may have been detrimental to sales, as users who started reading the newsletters often simply left from there.

5. Conclusion

We have implemented a tool for performing clickstream visualization. Our design goals centered around making this tool truly usable in the real world, with real data. To this end we have tried to achieve a balance between interactivity, to enhance the user's experience in the manipulation of the graph; and scalability, to handle the huge amounts of data that are typical in clickstream data. Scalability is achieved using server-side processing to sample and aggregate the data into transition matrices, so that the amount of data on the client side is independent of the number of clicks or sessions. The primary focus of interactivity in the tool is on filtering and layout manipulations to improve comprehensibility of the graph. One of the primary strengths of ClickViz is the ability to take advantage of support for clickstream logging and customer attribution. Customer attributes, including demographic data and purchase history, can be directly associated with particular clicks and sessions, allowing ClickViz to uniquely segment the edges in the graph. The result is a visual separation of the navigation patterns of customer segments, allowing direct comparisons of those segments.

The ultimate goal of such a tool is to provide insight into customer interactions with the web site. This can be a key element in improving site design and site personalization, alerting for broken links and performing other diagnostic functions. The analysis performed with this tool differs from other forms of site usage analysis such as reports, which may list top enter and exit pages and most common paths, because these may miss much of the insight that comes from visualizing the interconnections of the navigation patterns.

References

[1] J. Abello, J. Korn, Visualizing Massive Multi-Digraphs. In *Proc. IEEE Information Visualization 2000*, pp 39-47, October 2000.

[2] E. Chi, P. Pirolli, J. Pitkow, The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage and Usability of a Web Site. In *Proceedings of the Human Factors in Computing Systems, CHI 2000*, pp 161-168, April 2000.

[3] K. Andrews, M. Pichler, P. Wolf, Towards Rich Landscapes for Visualizing Structured Web Spaces. In *Proc. IEEE Information Visualization '96*, pp 62-63, October 1996.

[4] K.C. Cox, S.G. Eick, 3D Displays of Internet Traffic. In *Proc. IEEE Information Visualization '95*, pp 129-131, October 1995.

[5] M. Deshpande, G. Karypis, Selective Markov Models for Predicting Web-Page Accesses, *University of Minnesota Technical Report 00-056*, October 2000.

[6] E. Frecon, G. Smith, WebPath – A Three Dimensional Web History. In *Proc. IEEE Information Visualization '98*, pp 3-10, October 1998.

[7] I. Herman, M.S. Marshall, G. Melancon, Density Functions for Visual Attributes and Effective Partitioning in Graph Visualization. In *Proc. IEEE Information Visualization 2000*, pp 49-56, October 2000.

[8] J.L. Korn, A.W. Appel, Traversal-based Visualization of Data Structures. In *Proc. IEEE Information Visualization '98*, pp 11-18, October 1998.

[9] T. Munzner, H3: Laying Out Large Directed Graphs in 3D Hyperbolic Space. In *Proc. IEEE Information Visualization '97*, pp 2-10, October 1997.

[10] T. Munzner, E. Hoffman, K. Claffy, B. Fenner, Visualizing the Global Topology of the MBone. In *Proc. IEEE Information Visualization '96*, pp 85-92, October 1996.

[11] T.J. Overbye, J.D. Weber, New Methods for the Visualization of Electric Power System Information. In *Proc. IEEE Information Visualization 2000*, pp 131-136, October 2000.

[12] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw Hill, 1991.

[13] P. Underhill, *Why We Buy: The Science of Shopping*. Simon & Schuster, 1999.

[14] www.tomsawyer.com

[15] A competition within KDD-2000 (The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), Boston, MA, August 2000. The data is available for download from www.bluemartini.com/kddcup2000.

[16] K. Avedal and others, *Professional JSP*. WROX, 2000.

[17] J. Hong, J.Landay, WebQuilt: A Framework for Capturing and Visualizing the Web Experience. In *Proc. of the 10th International WWW Conference*, Hong Kong, May

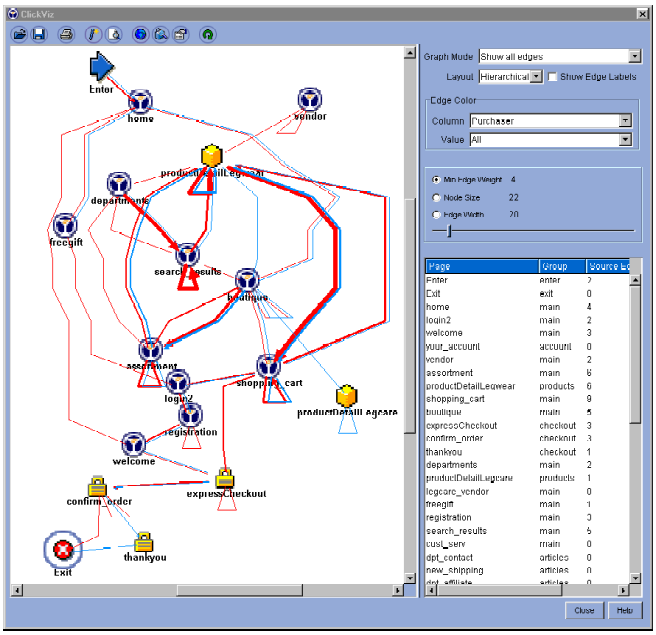

Figure 1: Main ClickViz window showing hierarchical layout

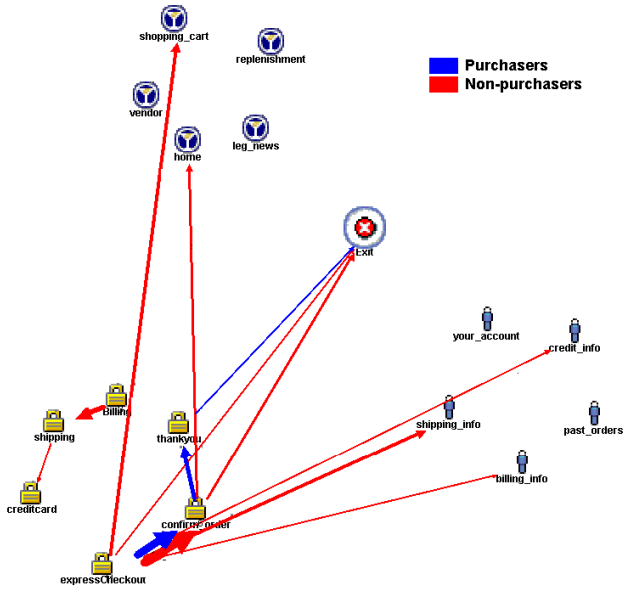

Figure 2. Circular layout. All the checkout pages are grouped (lower left). Red edges that emanate from the checkout pages to other parts of the site represent non-purchasers who are abandoning the checkout process.

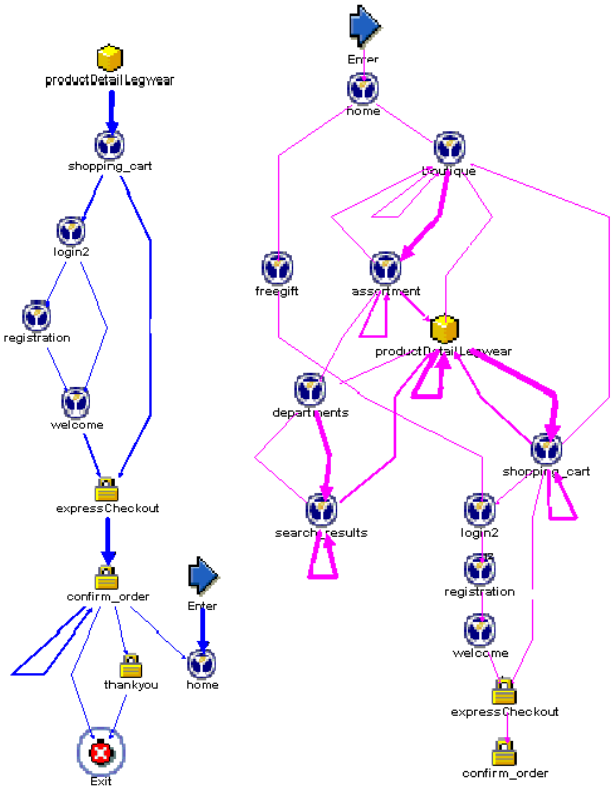

Figure 3. Gender Differences: Males tend to navigate in specific, direct patterns, whereas women's navigation patterns include much more browsing, utilizing much more of the site.

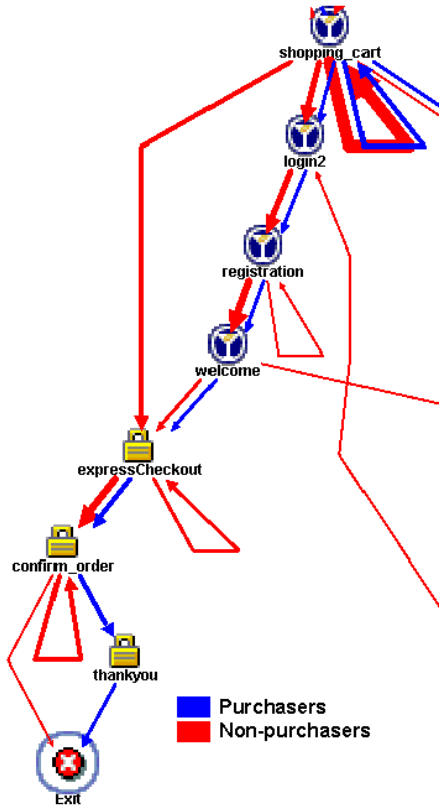

Figure 4. Checkout process. Purchasers take a direct route through the checkout process, whereas non-purchasers show a more haphazard route, including self-edges and early abandonment, possibly indicating a confusing checkout process.