

Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising

Juhnyoung Lee and Mark Podlaseck

IBM T. J. Watson Research Center
P. O. Box 704
Yorktown Heights, NY 10598
{jyl, podlasec}@us.ibm.com

Abstract

Clickstreams are visitors' path through a Web site. Analysis of clickstreams shows how a Web site is navigated and used by its visitors. Clickstream data of online stores contains information useful for understanding the effectiveness of marketing and merchandising efforts, such as how customers find the store, what products they see, and what products they purchase. In this paper, we present an interactive visualization system that provides users with greater abilities to interpret and explore clickstream data of online stores. This system visualizes the effectiveness of Web merchandising from two different points of view by using two different visualization techniques: visualization of sessions by using parallel coordinates, and visualization of product performance by using scatterplot graphs. Furthermore, this system provides facilities for zooming, filtering, color coding, dynamic querying and data sampling. It also provides summary information along with visualizations, and by maintaining connection between visualizations and source database, it dynamically updates the summary information. To demonstrate how the presented visualization system provides capabilities for examining online store clickstreams, we present a series of parallel coordinates and scatterplot visualizations which display clickstream data from an operating online retail store. A framework for understanding Web merchandising is briefly explained. A set of metrics referred to as *micro-conversion rates*, which are defined for Web merchandising analysis in our previous work, is also explained and used for the visualizations of online store effectiveness.

1. Introduction

Clickstream is a generic term to describe visitors' path through one or more Web sites. A series of Web pages requested by a visitor in a single visit is referred to as a *session*. Clickstream data in a Web site is a collection of sessions in the site. Clickstream data can be derived from raw page requests (referred to as *hits*) and their associated information (such as timestamp, IP address, URL, status, number of transferred bytes, referrer, user agent, and, sometimes, cookie data) recorded in Web server log files. Analysis of clickstreams shows how a Web site is navigated and used by its visitors.

In an e-commerce environment, clickstreams in online stores provide information essential to understanding the effectiveness of marketing and merchandising efforts, such as how customers find the store, what products they see, and what products they buy. (While not all this information may be available from Web server log files, it can be extracted from associated data sources such as commerce server database and tied together with HTTP request data.) Analyzing such information

embedded in clickstream data is critical to improve the effectiveness of Web marketing and merchandising in online stores. Interest in interpreting Web usage data in Web server log files has spawned an active market for Web log analysis tools that analyze, summarize, and visualize Web usage patterns. While useful to some extent, most of existing tools have the following shortcomings: (1) the summaries they provide obscure useful detail information, (2) the static displays such as histograms and pie charts restrict users to passive interpretation, and (3) the weak (or lack of) connection between purchase data and navigation data limits the ability to understand the site's effectiveness in terms of return on investment.

In this paper, we present an interactive e-commerce visualization system that can be used to provide users with greater abilities to interpret and explore clickstream data of online stores in the Web. In order to help users actively explore and interpret data of interest, this system provides facilities for zooming, filtering, color coding, dynamic querying, and data sampling. This system visualizes the effectiveness of Web merchandising from two different points of view by using two different visualization techniques: visualization of sessions by using parallel coordinates, and visualization of product performance by using scatterplot graphs. Finally, it augments visualizations of parallel coordinates and scatterplot graphs with summary information. It also dynamically updates the summary information by maintaining connection between visualizations and source database.

A *scatterplot* graph is a well-known general-purpose visualization tool useful for finding patterns in multidimensional data. *Parallel coordinates* are a visualization method developed by Inselberg for displaying multivariate data sets to identify the relationship among the variables. A parallel coordinate system comprises a series of parallel lines that are placed equidistantly and perpendicular to the x-axis of a Cartesian coordinate system. Each parallel axis is assigned a specific dependent variable and dependent variable values are plotted along the respective axis. The independent variable is represented by polygonal lines which connect the corresponding dependent variable values relating to the independent variable.

To demonstrate how the presented visualization system provides capabilities for examining online store clickstreams that exceed those of traditional Web log analysis tools, we present a series of visualizations which displays clickstream data from an operating online retailer. The visualizations reflect the a framework developed for understanding online merchandising efforts introduced in our previous work. Especially, a set of metrics referred to as *micro-conversion rates*, which are defined for online merchandising analysis, is used for the visualizations. The results show that the interactive parallel coordinate system is useful in validating various hypotheses about Web merchandising as well as finding interesting patterns in clickstreams that are not identified previously.

The rest of this paper is structured as follows: Section 2 explains a framework and metrics for understanding Web merchandising. In Section 3, we discuss the types of data required for visual analysis of Web merchandising, and briefly describe how the data can be collected and integrated in an online store. Sections 4 and 5 discuss how parallel coordinates and scatterplot graphs, respectively, can be used to visualize online store clickstream data, and help understanding the effectiveness of merchandising tactics. Section 6 describes the design of an interactive e-commerce visualization system and its facilities. Section 7 summarizes an empirical study of analyzing clickstreams from an operating online store by using the presented visualization system. In Section 8, related work is evaluated and summarized. Finally, in Section 9, conclusions are drawn and further work is outlined.

2. Analysis of Web Merchandising

Web merchants generally analyze their sites' effectiveness from two perspectives: marketing and merchandising. Marketing on the Web is broadly defined as the activities used to acquire customers to online stores and retain them. Techniques for online marketing include the use of banner ads and e-mail campaigns. Examples of marketing-related business questions include the followings: Which banner ads generate the most traffic and sales? Which portal sites are pulling in the most qualified traffic? Metrics that are used for answering these questions include banner ad *clickthrough rate* (the percentage of viewers who click on a banner ad), *conversion rate* (the percentage of visitors who purchase from the store), and banner ad *return on investment* (the amount of revenue and profit generated by visitors referred by a banner ad). The area of reporting and analyzing Web marketing is relatively well-understood, while useful metrics and analysis tools for Web merchandising lag behind. In this paper, we focus on the analysis of Web merchandising.

Merchandising consists of the activities involved in acquiring particular products and making them available at the places, times, and prices and in the quantity to enable a retailer to reach its goals [2]. In general, there are four areas for Web merchandising analysis: product assortment, merchandising cues, shopping metaphor, and Web design features [9]. The first analysis area, *product assortment*, deals with whether the products in an online store appeal to the visitors. If the product assortment is not optimal, the merchants may adjust, for example, brands, quality, selection, inventory or price of the products they carry. Examples of business questions related to product assortment include the following: What are the top sellers for a specific period of time, e.g., this week? What is the conversion rate for a particular department? In what frequencies and quantities are products purchased? What characterizes the products that end up being abandoned?

Merchandising cues are techniques for presenting and/or grouping products to motivate purchase in online stores. Examples of merchandising cues are cross-sells, up-sells, promotions and recommendations. Merchandising cues are associated with hyperlinks on Web pages. For example, a cross-sell is a hyperlink which refers visitors to a Web page marketing an item complementary in function to the item marketed on the current page. Online merchants need to understand the effectiveness of the merchandising cues in their stores in terms of traffic and sales driven by them. Examples of business questions related to merchandising cues include the followings: How much did cross-sells and up-sells contribute to gross revenue? What are the best performing cross-sell pairs? And worst? What is the overall conversion rate for cross-sells? How much do promotions contribute to gross revenue? At which levels in the site hierarchy are the best promotions located?

Shopping metaphors in an online store are the means that shoppers use to find products of interest. Examples include browsing through the product catalog hierarchy, various forms of searching, and configuration for "build-to-order" type products. The effectiveness of different shopping metaphors in the store is a concern for online merchants. Like merchandising cues in online stores, shopping metaphors are associated with hyperlinks on Web pages. This allows one to categorize and group together hyperlinks in an online store by their types of merchandising cue and shopping metaphor. Examples of business questions related to shopping metaphors in online stores include the followings: What generates the most sales value, e.g., search or browsing? How much does search contribute to gross revenue? What is the conversion rate for search?

The effectiveness of *Web design features* presents another area of analysis for Web merchandising. The design features of hyperlinks include media type (e.g., image or text), font (if text), size, color, and

location. Examples of business questions related to Web design features include the followings: What are the features of links customers most frequently click? What are the features of links customers most frequently buy from? What are the parts of Web pages customers most frequently buy from? Do products sell better in the upper left corner?

Just as Web marketing uses banner ads and/or referral sites to attract customers from external sites to an online store, online merchandising uses hyperlinks and image links within the store to lead customers to click to Web pages selling products. Web merchants employ a variety of tactics for merchandising by using hyperlinks. From this perspective, the problem of tracking and measuring the effectiveness of different merchandising tactics in an online store can be partitioned into three sub-problems:

- 1) classifying hyperlinks by their merchandising purposes,
- 2) tracking and measuring traffic on hyperlinks and analyzing their effectiveness (e.g., profit), and
- 3) attributing the profit of hyperlinks to their merchandising cue type, shopping metaphor type, and design features.

Having identified the areas of Web merchandising analysis, we now introduce a set of metrics, referred to as *micro-conversion rates*, which can be used for measuring the effectiveness of efforts in these merchandising areas. The metrics are based on the conversion rate which is used for measuring online marketing performance. Traditionally, the conversion rate of an online store indicates the percentage of visitors who purchase from the store. While this measure is useful for evaluating the overall effectiveness of the store, it does not help understand the possible factors within the store that may affect the sales performance. The notion of a micro-conversion rate extends this traditional measure by considering the four general shopping steps in online stores, which are:

- 1) *product impression*: the view of hyperlink to a Web page presenting a product.
- 2) *clickthrough*: the click on the hyperlink and view the Web page of the product.
- 3) *basket placement*: the placement of the item in the shopping basket.
- 4) *purchase*: the purchase of the item - completion of a transaction.

Basic micro-conversion rates are computed for each adjacent pair of these measures, resulting in the first three rates in the following list. In addition, the aggregation of the first three is also interesting. By looking at this look-to-buy rate, online merchants can tell if a product is overexposed or underexposed and take action to change the presentation of the product:

- 1) *look-to-click rate*: how many product impressions are converted to click-throughs.
- 2) *click-to-basket rate*: how many click-throughs are converted to basket placement.
- 3) *basket-to-buy rate*: how many basket placements are converted to purchases.
- 4) *look-to-buy rate*: what percentage of product impressions are eventually converted to purchases.

Note that the first of these, look-to-click rate, is similar to the clickthrough rate used for measuring the amount of traffic on banner ads. Also note that the micro-conversion rates relate the traffic-related measure to sales which happen later in the shopping process. By precisely tracking the shopping steps with these metrics, it is possible to spot exactly where the store loses how many customers. The micro-conversion rates extend the traditional measure by considering the merchandising purposes associated with hyperlinks viewed in the first shopping step, i.e., product

impression. In this way, the micro-conversion rate is related to tactics of merchandising, and can be used for evaluating the effectiveness of different merchandising aspects of the store.

3. Data Requirements

In this section, we briefly describe several data requirements for the analysis of Web merchandising explained in the previous section. While some source data is readily available from most online store sites, others are not and need to be collected with some special tools. Also, the collected data has to be integrated to show micro-conversions over shopping steps and to provide insight into the merchandising effectiveness of online stores.

First, the visualization of merchandising effectiveness based on micro-conversions requires the combination of the site traffic data and sales data. In most online stores, the two types of data are typically stored in separate storage systems in different structures: the traffic data in Web server logs in a file format, and the sales data in the database of the associated commerce server. The commerce server database also contains information about customers and products (including product taxonomy) that may also be useful and interesting to visualize with micro-conversions. It is important to combine data from the two different sources with a common key and to construct an integrated database system or a data mart system for business visualizations.

Second, showing a complete set of micro-conversions requires product impression data. Capturing product impressions involves tracking the content of served Web pages, which is challenging because more and more Web pages are dynamically generated. (A simple example of a dynamically created Web page is a search result page commonly found in online stores.) Currently, the standard Web server logging mechanism does not capture the content of Web pages. One possible method is to enhance the Web server logging as a way to dynamically parse the content of served Web pages and extract useful data such as product impressions and information on hyperlink types. This ability of dynamically scanning Web pages as they are served is critical for tracking Web usage, because more and more Web pages are dynamically created from databases and contain personalized and adaptive content.

Finally, it is important to classify and identify hyperlinks by their merchandising purposes, so that later to attribute the profit generated from the hyperlinks to their merchandising cue type, shopping metaphor type, and/or design features. For this purpose, Web pages and hyperlinks in an online store need to be tagged with semantic labels describing their merchandising features. Semantic labels of a hyperlink may include, for example, a product label, a cross-sell or promotion label, and a tag indicating where the product is being displayed. Such semantic labels for hyperlinks in a site may be explicitly provided in a form of meta-data during the site creation. If this is not the case, semantic labels need to be inferred from various sources such as the file name and/or path portion of URLs, types, values and orders of parameters in URLs of dynamic Web pages, and the location of a hyperlink in the page.

4. Parallel Coordinates for Online Store Clickstreams

Clickstream data of an online store can be visualized by displaying the progression of sessions in terms of micro-conversions among shopping steps described earlier. This approach provides visualizations that help users identify where the store loses how many customers, understand and compare the shopping behavior of different groups of customers, and understand the effectiveness

of different merchandising tactics. In this section, we introduce a set of parallel coordinates visualizing the micro-conversions of sessions in an online store.

The system of parallel coordinates was developed for displaying multivariate data sets to identify the relationship among the variables in the set [8]. A parallel coordinate system comprises a series of parallel lines that are placed equidistantly and perpendicular to the x-axis. Each parallel axis is assigned a specific dependent variable and dependent variable values are plotted along the respective axis. The independent variable is represented by polygonal lines which connect the corresponding dependent variable values relating to the independent variable and illustrate a relationship between an independent variable and the dependent variables appearing on each axis. For the parallel coordinate visualizations presented in this paper, we used the Diamond system developed at IBM T. J. Watson Research Center for visualizing multidimensional data [14, 15]. Some examples of parallel coordinates such as a visualization of families of atomic weights can be found in [14, 15].

Figure 1 illustrates how we use parallel coordinates to display clickstream data of an online store. In this figure, each polygonal line represents a single session and its progression in the online store. The first parallel axis represents a session categorizer, in this figure, the referrer of each session. In all the clickstream data visualizations presented in this paper, we use the first parallel axis for presenting a session attribute that can be used to categorize sessions. In Figure 1, the data points in the categorizer axis are the site names of referrers, and there are 33 different referrers shown.

The next three parallel axes represent shopping steps in the online store, i.e., clickthrough, basket placement, and purchase. (The product impression data was not available for the data set used for the visualizations presented in this paper.) For data points in the shopping step axes, timestamps (i.e., start times in second granularity) of sessions are used. An advantage of using timestamps for data points in shopping step axes is that because they are unique to sessions, no two sessions share the same data points in these axes. Using data points unique to individual sessions prevent the problem of overlaying lines between two parallel axes, which is serious in parallel coordinate visualizations, because it sometimes obscures the accuracy of visualizations. One disadvantage of using timestamps for data points is that they do not carry any sense of volume. Namely, the existence of a data point in a shopping step axis does not tell how many products were viewed, placed in basket, or purchased. Rather, it merely says that one or more products were viewed, placed in basket or purchased. The last parallel axis represents the total dollar amount of the completed transaction in the corresponding session.

Clickstream of a session is visualized by a polygonal line that connects its value in each parallel line in the system. In this figure, each polygonal line displays where the customer came from, if the person saw one or more product information pages in the store, if the person inserted one or more product items into shopping basket, if the person purchased one or more items from the store, and if so, how much the purchase value was in dollar amount?

In Figure 1, it is important to notice that some polygonal lines stop before they reach the last shopping step, i.e., purchase. In preparing the data set for the visualizations in this paper, we did not give a session a data value for the next shopping step and on, if the session did not convert to the next step. Hence, polygonal lines stops at the last step the session reached, which indicates the point the session left the store. Figure 1 clearly shows that the numbers of lines connecting two adjacent parallel axes decrease, as polygonal lines go from left to right up to the purchase line. Dropouts of polygonal lines visualize where the store loses its customers. Also, the micro-conversion rates can be

computed directly from the visualization by using the numbers between two adjacent axes. That is, the click-to-basket rate is the ratio of the number of lines connecting the clickthrough and basket axes to the number of lines connecting the referrer and clickthrough axes. The basket-to-buy rate is the ratio of the number of lines connecting the basket and buy axes to the number of lines connecting the clickthrough and basket axes. In Figure 1, the click-to-basket rate is about 19% (210/1094), the basket-to-buy rate about 43% (90/210), and the click-to-buy rate about 8% (90/1094). Polygonal lines which reached the purchase and value axes (i.e., ones representing sessions which made purchase) are selected and colored blue in the visualization. With coloring, the referrers and micro-conversions of the selected group are easily identified.

5. Starfield Visualization of Online Store Clickstreams

While using parallel coordinates to visualize clickstream data from the view point of customer visits, i.e., sessions, we use scatterplot graphs to display the same data from a different point of view, i.e., products. The information about product-oriented view of micro-conversions will be particularly useful for understanding the product assortment aspect of the store. The product visualization, a scatterplot graph, shown in Figure 2 augments the interactive starfield model [1], a general-purpose analysis tool useful for finding patterns in multidimensional data. With the associated tree controls given in Figure 3, users can filter on hierarchical dimensions found in the E-commerce domain such as product taxonomy (pictured here) and site architecture. Selection of one or more branches of the tree causes the products under that branch to be pictured in the graph area. The color key associated with a particular branch in the tree can be inherited from a parent (the default) or overridden with a color unique to that child.

Each rectangle or *glyph* in the graph space represents a distinct product (a certain brand and type of T-shirt, for example, but not its colors or sizes). The color of each glyph in this example corresponds to the product's department, as indicated by the color key on the tree control described above. The area occupied by a glyph describes a product's relative significance: width is an indicator of the product's price, and height, its relative (profit) margin.

The x-axis and the y-axis of the starfield graph can represent any two of the micro-conversion rate metrics. In the example in Figure 2, the user configured the axes to analyze product exposure relative to customer interest ("Are the right products being promoted? How do I optimize the exposure of all my products to maximize my revenue stream?"). The x-axis thus represents raw impressions, that is, the number of times a hyperlink to a product was served. (Product hyperlinks can occur anywhere on the commerce site: the home page, category pages, search result pages, as well as other product pages.) The y-axis represents the percentage of impressions that resulted in a click-through (that is, of the number of customers that saw a hyperlink to this product, the percentage that clicked on the link).

The scatterplot graph makes evident the heavy over-promotion of a product represented by the small glyph in the lower right quadrant. While it has had more impressions than almost any other product, its click-through is almost the lowest. To make matters worse, it is a low-priced, low-margin product. Its exposure could be reduced by moving its promotion to a less-trafficked page, or eliminating it entirely.

On the other hand, the large glyphs in the upper left quadrant represent products that are under-exposed. Although links to these products have few impressions, a relatively high percentage

of customers are clicking on them. This level of interest might be maintained if the number of impressions of these product links were incremented. If that is not the case, it is possible that the products are niche products, appealing only to the small group of customers that are specifically looking for them (left-handed joy-sticks, for example). Therefore, depending on the nature of the products in the upper-left corner, one might chose to display or to promote them more, and then monitor the results carefully.

Reconfiguring the graph space allows one to explore other questions. For example, one might reassign the x-axis to represent click-throughs and the y-axis to represent the percentage of click-throughs that resulted in a product being placed in the shopping basket. Products with a high click-through rate, but low basket-placement rate would occupy the upper left quadrant of the graph. These are products in which customers were interested enough to click on, but not interested enough to consider buying. Causes to explore here include the quality of the information on the product detail page, surprise pricing, or misleading product links.

6. An Interactive Visualization System for Web Merchandising Analysis

Earlier in this paper, it was pointed out that the static displays of existing Web server log analysis software packages restrict users to passive interpretation of Web usage data. In order to help users actively explore and interpret data of interest, we design a tightly-coupled interface of an interactive system that provides visualizations of parallel coordinates and scatterplot graphs with facilities for color-coding, filtering, zooming, data sampling, dynamic querying, and summary data augmentation. In this section, we will briefly describe the use of each facility.

Filtering facility allows users to remove some polygonal lines from visualizations by using certain criteria and see only the lines of interest. Criteria for filtering of sessions include their category types (e.g., referrer, host name, timestamp, or length), and then their category type values (e.g., AOL, Yahoo, or Excite@Home for the referrer category). The proposed visualization system provides an interface for users to select one or more filtering criteria. Figure 4 illustrates a screenshot of the proposed interactive parallel coordinate system where sessions are filtered and categorized by their referrers, and two groups of sessions are displayed and color-coded (sessions from referrer A colored yellow and sessions from bookmark colored in blue). Zooming can be performed on any parallel axes in the visualization by using arrows attached to both sides of axes.

The visualizations of parallel coordinates and scatterplot graphs are augmented with a summary of the micro-conversions and average order values of the selected session groups presented in a table at the bottom. The summary of micro-conversion rates is also visualized in a bar chart. The idea is that presenting parallel coordinate/scatterplot visualizations together with summary information help users understand the data better, because the two different types of information complement with each other: visualizations give users insight into the relationship among multiple variables and their patterns, while summary information delivers specifics. Note that the visualization system maintains a connection to the source database, and supports dynamic updates of the summary information in the table and bar chart, as users select, deselect, filter, and sample data in the parallel coordinate or scatterplot window.

Data sampling is an important issue in many data visualizations. The issue becomes even more serious for visualizing Web usage data which tends to grow fast and huge. Especially when visualizations are used to identify changes in Web usage patterns over time, data sets of huge size

(e.g., hundred millions of sessions) need to be processed and displayed. Unfortunately, such expanded visualizations might exceed the capabilities of most visualization tools, e.g., the performance of the parallel coordinates/scatterplot graphs degrades noticeably on data sets containing more than a thousand data points even with zooming and filtering. Data sampling provides a plausible solution to this scalability problem of Web usage data, if samples accurately represent the mother data set, and can be obtained in flexible ways, i.e., the size of samples should be controllable, and various parameters need to be incorporated into the sampling process. The proposed visualization system provides a data sampling facility that allows users to adjust the size of data sets for visualization by using various criteria such as session attributes. Data sampling is also useful to balance the size of two or more groups of data points when these groups are visualized together in a window and compared.

Finally, the visualization system provides fly-over information for individual polygonal lines. The information boxes are triggered as a user moves the mouse over polygonal lines in the visualization and display useful information about the session represented by the selected polygonal line, e.g., session id, timestamp, referrer, host name, length, etc. The interactive visualization system presented in this section is currently under implementation as part of the E-Commerce Intelligence project at IBM T. J. Watson Research Center. We envision to provide the interactive visualization system as a Web application and allow users to run the system from their browsers.

7. An Empirical Study

In order to understand the applicability and usefulness of the proposed interactive visualization system, we have performed an empirical study with clickstream data from an online retailer, creating a series of visualizations of parallel coordinates and scatterplot graphs. For a five day period during May 1999, the data set consists of over 35,000 sessions and 7,500 basket placements and 3,500 completed transactions. In an attempt to generate meaningful clickstream data, basic Web usage data, i.e., raw hits recorded in Web server log files was processed and cleansed in a way similar to one described in [6]. Then the generated session data was integrated with data of basket placements and transactions extracted from the commerce server. Because the online store actively uses cookies for identifying customers and their visits, and records them in both Web server logs and commerce server database, the processes of sessionization and data integration were done in an accurate and reliable fashion. Note that product impression data was not available for these visualizations. Below we will examine a subset of created visualizations one by one. Figures 4, 5, 6 and 7 visualize micro-conversions of sessions by using parallel coordinates, and Figures 8, 9, 10 and 11 visualize micro-conversions of products sold in the store by using scatterplot graphs.

Various attributes of sessions that can be used to categorize them were identified. The session categorizers include referrers, host names, timestamps (not just of session start time, but also of every commerce-related activity in the session, e.g., basket placements and purchase), the length of time spent in the store, the types and numbers of shopping metaphors used, the types and numbers of merchandising cues used, and the categories and numbers of products viewed and purchased. Note that design features of hyperlinks and customer profile information were not available for this analysis. Visualizations presented in this section are for a demonstration purpose, only a small subset of the visualizations that could be generated from the available data. The visualizations in this section were created for the following four session categorizers: referrers, ISPs, session length, and the number of shopping metaphors used.

Figure 4 provides a visualization of sessions categorized by their referrers, i.e., a group of sessions from a well-known portal site which was labeled as A, and another group of sessions which came to the store through bookmarks which the customers had in their browsers. Note that the number of sessions that had clickthroughs in the first two visualizations are arranged to be roughly the same for a balanced visual comparison. We used the data sampling facility for balancing the data set sizes.

It is commonly believed that the visitors who come to the store through bookmarks are repeating customers who shop at the store frequently, and probably, know what they want to buy from the store. Visualization in Figure 4 confirms this speculation by showing a relatively high number of click-to-basket conversions. The summary table in the figure also confirms the high micro-conversion rates: the click-to-basket rate of the bookmark group is almost two times higher than that of the baseline, while the click-to-basket rate of the referrer A group is a few points lower than that of the baseline. The baseline data was computed for the entire data set which include over 35,000 sessions, 7,500 basket placements, and 3,500 transactions. While the basket-to-buy rate of the bookmark group is slightly lower than that of the baseline, the basket-to-buy rate of the referrer A group is about 40% higher than that of the baseline. Overall, the click-to-buy rate of the bookmark group is about 60% higher than that of the baseline, while the click-to-buy rate of the referrer A group shows the roughly the same number as the baseline. In addition, the average order value of all sessions in the bookmark group is about 60% higher than that of the baseline. As a side note, a linear relationship between the two metrics, the click-to-buy rates and the average order value of all sessions is observed for this data set.

Figure 5 shows the micro-conversions and average order values of sessions which are from two different ISPs. One group, labeled as A and colored yellow, consists of sessions from an ISP that provides connection through regular telephone modems, and the other, labeled as B and colored blue, consists of sessions from an ISP that provides service through television-based clients. Again, the number of sessions in each group is balanced for visual comparison. This figure shows that, for this specific period of time, the micro-conversions of Group B is about 50% lower than the other group. The difference in micro-conversions in the two session groups is clearly illustrated by the numbers of the lines connecting adjacent shopping steps colored blue and yellow, respectively. Note that, as in the previous example, a linear relationship is observed between two metrics, the click-to-buy rates and the average order value of all sessions.

In Figure 6, sessions are classified by their length in time. It is commonly believed that the increase in session length helps sales in online store. Web site owners make efforts to increase the “stickiness” of their sites. However, it is an unproven hypothesis that stickiness actually increases the sales in online stores. Visualization in Figure 6 is a simple-minded tackle on the issue. It shows the micro-conversions and average order values of two groups of sessions; short and long ones. Short sessions (colored yellow) spent less than 10 minutes in the site, while long sessions (colored blue) spent 30 to 60 minutes in the store. Sessions spent longer than 60 minutes were not included in the long session group in an effort to eliminate any effect by Web crawlers. The visualization and summary table in Figure 6 consistently show that short sessions give significantly lower conversion rates and average order value, while long sessions yield about 300% higher click-to-buy conversion rate and average order value than the baseline. These results validate the hypothesis that stickiness helps sales in online stores, at least to some extent. As mentioned earlier, however, these results are preliminary because the approach taken here is simple-minded in that it does not consider many different ways to improve the stickiness of a site (e.g., the use of special shopping metaphors,

merchandising cues, Web design features, and/or content). Further work is required for a better understanding of the aspect of online stores.

In Figure 7, sessions are classified by the number of shopping metaphors they used. In preparing data for this visualization, we noticed several interesting facts about shopping metaphor usage in the store: (1) the most popular shopping metaphor is browsing, followed by two different types of search, i.e., search by product numbers (by customers who already have paper catalogs) and search by keywords, and (2) 85% of sessions (about 30,000 out of 35,000) use only one shopping metaphor and only 15% uses two or more shopping metaphors. Two session groups displayed in Figure 7 are sessions that used only one metaphor (Group A colored blue), and ones that used more than one metaphor (Group B colored yellow). For visual comparison in Figure 7, again, the number of sessions in the two session groups are balanced. The visualization and summary table in Figure 8 consistently show that sessions in Group A yield significantly lower conversion rates and average order value, while sessions in Group B give about 300% higher click-to-buy rate and average order value than the baseline. Once again, the click-to-buy rate and average order value show a linear relationship. Visualization in Figure 7 may be seen to provide some auxiliary explanation for the results presented in Figure 6, because the number of shopping metaphors used by sessions may be related to the length of time they spent in the store.

To this point, we have discussed how to read parallel-coordinate visualizations of online shopping sessions. In Figures 8, 9, 10, and 11, we will discuss how to interpret the performance of products from scatterplot visualizations, as well as how these visualizations might indicate actions to increase a product's revenue. In each of these figures, the rectangles in the graph space represent products from two departments. Products from the Men's department are colored blue; those from the Women's department are yellow. (Note also that the data set presented in these figures is identical to that used in the previous visualizations, where there are a total of 143 products, 64 in Men's and 79 in Women's.) The shape of a product's rectangle in the scatterplot is significant. Width is an indicator of the product's price; height, its associated revenue in dollars. Thus, the taller the rectangle, the more revenue the product generates. The wider the rectangle, the more expensive that product is.

Figure 8 shows the performance (revenue in dollar amount) of individual products against the number of clickthroughs they had. Product revenue, in general, proportionally increases as a product gets more clickthroughs. This observation confirms the conventional wisdom that the more a product is exposed to shoppers, the more revenue it generates. Another observation from this figure is that the top four sellers in the store (two from the Men's department and the other two from Women's) commonly take a slim and tall shape. We might interpret that a low price positively affects the revenue of products in this store.

In Figures 8, 9, 10 and 11, we selected five products (three from the Men's department, M1, M2 and M3, and two from Women's, W1 and W2) to examine different behaviors, and provide examples of possible actions Web merchandisers can take to increase the performance of the store. M1 is a low-priced product and is the best seller of the store as shown in Figure 8. Also, it had the most clickthroughs. However, Figure 9 shows that M1 has a relatively low click-to-buy rate. Why does this happen? Figure 10 shows that M1 has a low click-to-basket rate, while Figure 11 shows that its basket-to-buy rate is relatively high. Thus, the click-to-basket rate is the anomaly. The relatively low click-to-basket rate might have been caused by various merchandising mistakes such as an inappropriate detailed information page of M1 or misleading links to the M1 page. Alternately, M1

may be so heavily promoted that much of its clickthrough is due to the strength of its promotion, not necessarily on its strength as a product. In this case, the merchandiser might take steps to enhance the M1 product page so that it stands up to its strong promotion and heavy traffic.

M2 shows a behavior similar to M1, but has different attributes. Figure 8 demonstrates that M2 is a high-priced product and generates relatively high revenue. Figure 9 shows that M2 had a high number of clickthroughs (about 240), but its click-to-buy rate is very low at about 0.5%. Figures 10 and 11 show that M2's click-to-basket rate is the source of problem, while its basket-to-buy rate is very high. Most customers who visited the detailed information page of M2 might have left the page when they found M2's high price. Or some of the links to M2's page might have been misleading. The Web merchandisers of this store will need to identify what caused M2's low click-to-basket rate to improve the performance of the store.

M3 shows another type of product behavior which can be read from the scatterplot visualization. Figure 8 shows that it is a medium-priced product, and generates medium-sized revenue. Figures 9 and 10 show that it yields relatively high click-to-buy and click-to-basket rates with a small number of clickthroughs. Figure 11 shows that M3 also has a very high basket-to-buy rate. The information from these figures consistently shows that M3 has a potential for good performance, although it was underexposed for the test period. The Web merchandisers of this store need to exposure M3 more to the shoppers in an effort to increase M3's clickthroughs and so the total revenue of the store.

W1 is the best seller among the products in the Women's department, and a low-priced product as shown in Figure 8. The number of clickthroughs W1 received is less than half of M1's. However, as Figures 9, 10, and 11 show, W1's micro-conversion rates (click-to-basket, basket-to-buy, and click-to-buy rates) are all higher than those of M1. We conclude that W1 has a potential for better performance if it is exposed to more customers and receives more clickthroughs.

W2 is a low-priced product that generates relatively high revenue, as shown in Figure 8. Figures 10 and 11 show that its performance is damaged by its low basket-to-buy rate, while its click-to-basket rate is the best in the store. The abandonment of W2 may have been caused by inappropriate merchandising at the check-out process of the product. The merchandisers of this store need to identify the causes of the low basket-to-buy rate and fix them up in an effort to increase the performance of W2 and the store.

In summary, sample visualizations presented in this section show results consistent with summary information about a few interesting, but unproven speculations in e-commerce. Also, the visualizations of a large number of individual clickstreams against multiple dimensions show a potential for identifying previously unknown patterns. Visualizations presented in this section are only a small subset of the visualizations that could be created from the available data and further work is required for a better understanding of Web merchandising. However, the visualizations presented in this section clearly demonstrated the applicability and usefulness of the interactive parallel coordinate system for understanding the effectiveness of merchandising efforts in online stores.

8. Related Work

Recently, there have been many efforts on Web server log analysis from both industry and academia. Quite a few commercial Web server log analysis software packages are available from various

companies [17]. While these packages differ in their specific reports available, they generally share several characteristics: static display, low-dimensionality of reports, lack of low-level details, relative lack of flexibility, and lack of integration of knowledge of site layout [7]. These packages focus on aggregate statistics, and are limited in visualizing user-level clickstream information. Also these packages are limited in reporting on the effectiveness of specific marketing and merchandising efforts because they primarily rely on information in Web server logs which are hard to interpret and extract data useful for measuring the business efforts.

In academia, there have been a number of research projects on the topic of Web server log display and visualization [5, 11, 13], which examined the use of log analysis for specific goals such as understanding of patterns in geographic origin of requests or caching performance. The tools used in these projects lack facilities for general-purpose, interactive exploration of log data. More recently, [7] presented a series of starfield visualizations of two-dimensional displays of access requests with color and size coding for additional attributes, and facilities for zooming and filtering. The work lacks the visualization of user-level clickstream information and does not relate Web usage data with their meaning in commerce. In [10], we provided a series of starfield visualizations of two-dimensional displays of users' shopping activities in online stores. This work is significantly extended in this paper by covering not just two steps of shopping activity but also a series of shopping steps, and by having session categorizing variables as part of visualizations.

Characterization and modeling of Web site access patterns has been an active area of research [3, 4, 6, 12, 16]. While these efforts often rely on Web log analysis, their focus is generally on modeling and data mining. Common data mining algorithms used in these studies include association rule generation, sequential pattern generation, and clustering. Some of these efforts showed how data mining techniques can be used to model Web sites in electronic commerce scenarios. Findings of shopping patterns in online stores by using this paper's work may augment the data models generated from data mining.

9. Concluding Remarks

In this paper, we have presented an interactive visualization system that provides users with abilities to actively interpret and explore clickstream data of online stores in the Web. This system is equipped with facilities for zooming, filtering, color coding, dynamic querying and data sampling. By using an information structure specifically devised for presenting online store clickstream data, we showed the potential use of parallel coordinates and scatterplot graphs for analyzing the effectiveness of various merchandising efforts in Web stores. The presented system visualizes the progression of sessions in the store, i.e., the conversions from one shopping step to another, and so provide insight into the effectiveness of each step's design. By associating the sessions with attributes that categorize them such as the referrers, host names, length, and the shopping metaphors and merchandising cues they used, the sessions and their conversions can be subdivided. The categorization of sessions help understand how sessions with different category values react to the site differently. An empirical study we performed with clickstream data from an online retailer validated the usefulness of the presented visualization system for understanding online store performance.

The work presented in this paper can be extended in several ways for future research. First, the underlying information structure used for displaying sessions can be extended to include richer information. The current structure uses a value unique to each session (i.e., timestamp) to represent

data points of the session for the shopping steps. This scheme is effective to visualize micro-conversions of sessions, but can be extended to visualize other metrics by adding more data such as different timestamp values for different steps and volume of the event.

Second, the four sequential shopping steps presented in this paper can be considered only a simple example of sequences of user interactions in a Web site that can be visualized by the proposed system. The sequence of user actions can be extended and/or varied. Different analysis purposes will require different sets of sequences to be visualized. For example, the analysis of new shopping paradigms in the Internet such as online auctions will require a different set of sequential steps visualized with parallel coordinates. The basic concept of the proposed parallel coordinate system is flexible enough to support different sequences of user actions in a Web site.

Third, more session category variables can be employed in the analysis with the proposed parallel coordinate system. This work, constrained by the limit of available data and labor-intensive process of data cleansing and preparation, studied only a small number of category variables, i.e., referrers, host names, stickiness, and the number of metaphors used. For a better understanding of Web merchandising, a richer set of variables including various shopping metaphors, merchandising cues, design features, and customer profile information need to be studied. Furthermore, the relationship among the category variables also need to be investigated for understanding their compound impact on the store performance.

Finally, more empirical studies with the proposed parallel coordinate system need to be performed covering a long time range (e.g., 6 months or one year), and over different types of online stores (e.g., in terms of products they carry, their business models, and the level of customer service they provide). Such work will help validate or invalidate various speculations about Web marketing and merchandising strategies in a rigorous way. Also, such work will help identify an “optimal” set of visualizations which will ideally provide necessary understanding of the effectiveness of an online store with minimal effort.

References

- [1] C. Alhberg and B. Schneiderman, “Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays,” *ACM CHI Conference on Human Factors in Computing Systems*, 1994, pp. 313-317.
- [2] B. Berman and J. R. Evans, *Retail Management: A Strategic Approach, 7th Edition*, Prentice-Hall, Inc., 1998.
- [3] A. G. Büchner and M. Mulvenna, “Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining,” *SIGMOD Record*, 27(4):54-61, December 1998.
- [4] M. S. Chen, J. S. Park, P. S. Yu, “Data Mining for Traversal Patterns in a Web Environment,” *Proc. of the 16th International Conference on Distributed Computing Systems*, 1996.
- [5] E. Chi, J. Pitkow, J. Mackinlay, P. Pirolli, R. Gossweiler, and S. Card, “Visualizing the Evolution of Web Ecologies,” *ACM CHI Conference on Human Factors in Computing Systems*, 1998, pp. 400-407.
- [6] R. Cooley, B. Mobasher, and J. Srivastava, “Data Preparation for Mining World Wide Web Browsing Patterns,” *Journal of Knowledge and Information Systems*, 1(1), 1999.

- [7] H. Hochheiser, and B. Schneiderman, "Understanding Patterns of User Visits to Web Sites: Interactive Starfield Visualizations of WWW Log Data," *Technical Report*, CS-TR-3989, Department of Computer Science, University of Maryland, 1999.
- [8] A. Inselberg and B. Dimsdale, "Parallel Coordinates A Tool for Visualizing Multivariate Relations," *Human-Machine Interactive Systems*, Plenum Publishing Corporation, 1991, pp. 199-233.
- [9] J. Lee, M. Podlaseck, E. Schonberg, R. Hoch, and S. Gomory, "Understanding Merchandising Effectiveness of Online Stores," To Be Published in *the International Journal of Electronic Commerce and Business Media*, January, 2000.
- [10] J. Lee, M. Podlaseck, E. Schonberg, R. Hoch, and S. Gomory, "Analysis and Visualization of Metrics for Online Merchandising," To Be Published in *Lecture Notes in Computer Science*, Springer-Verlag, 2000.
- [11] N. Papadakakis, E. P. Markatos, and A. E. Papathanasiou, "Palantir: a Visualization tool for the World Wide Web," *INET 98 Proceedings*, 1998.
- [12] J. Pitkow, "In Search of Reliable Usage Data on the WWW," *Technical Report*, College of Computing, Graphics, Visualization, and Usability Center, Georgia Tech, 1996.
- [13] J. Pitkow, and K. Bharat, "Webviz: a Tool for World Wide Web Access Log Analysis," *Proceedings of the 1st International Conference on the World Wide Web*, 1994.
- [14] D. Rabenhorst, "Interactive Exploration of Multidimensional Data," *Proceedings of the SPIE Symposium on Electronic Imaging*, 1994, pp. 277-286.
- [15] D. Rabenhorst, "IBM Diamond: Multi-Faceted Visualization of Multivariate Data," *IBM Research Web Site*, <http://www.research.ibm.com/people/d/Rabenhorst/gallery.html>.
- [16] L. Tauscher and S. Greenberg, "Revisitation Patterns in World Wide Web Navigation," *ACM CHI Conference on Human Factors in Computing Systems*, 1997, pp. 399-406.
- [17] T. Wilson, "Web Site Mining Gets Granular," *InternetWeek*, March 29, 1999.
- [18] P. Underhill, *Why We Buy: The Science of Shopping*, Simon & Schuster, 1999.

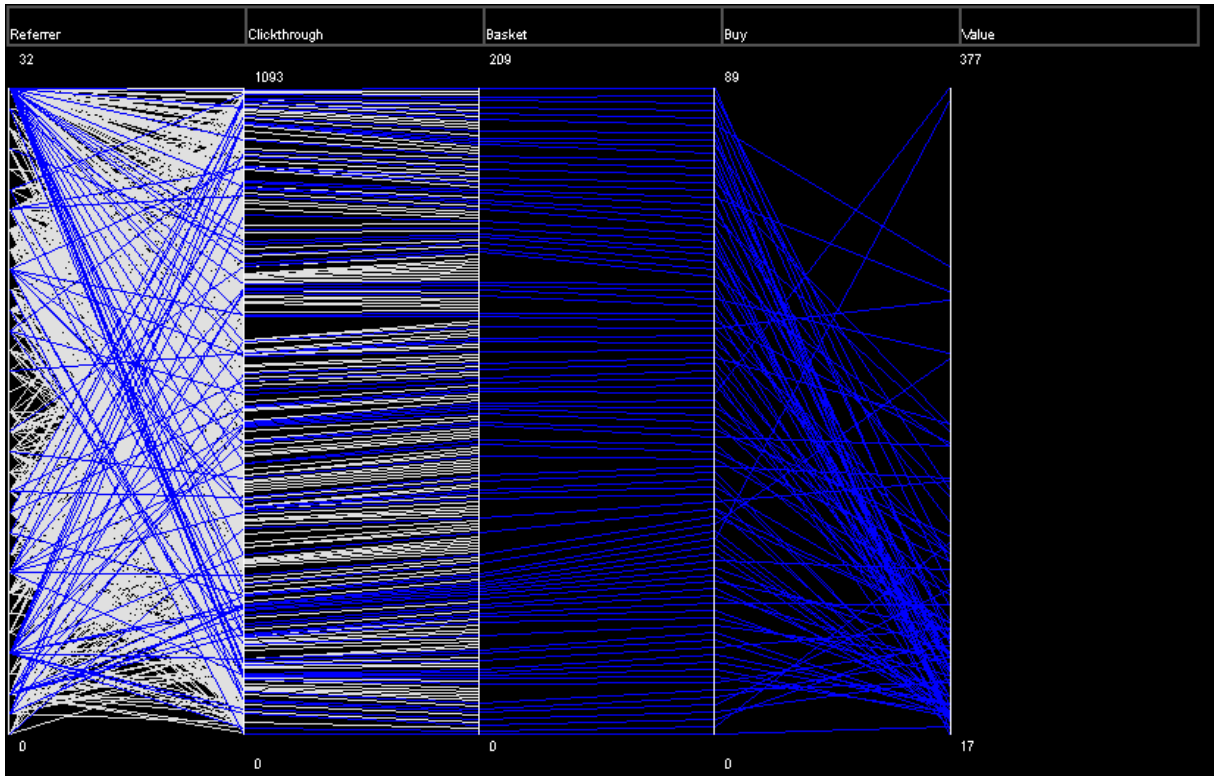


Figure 1. Visualizing micro-conversions of sessions with parallel coordinates

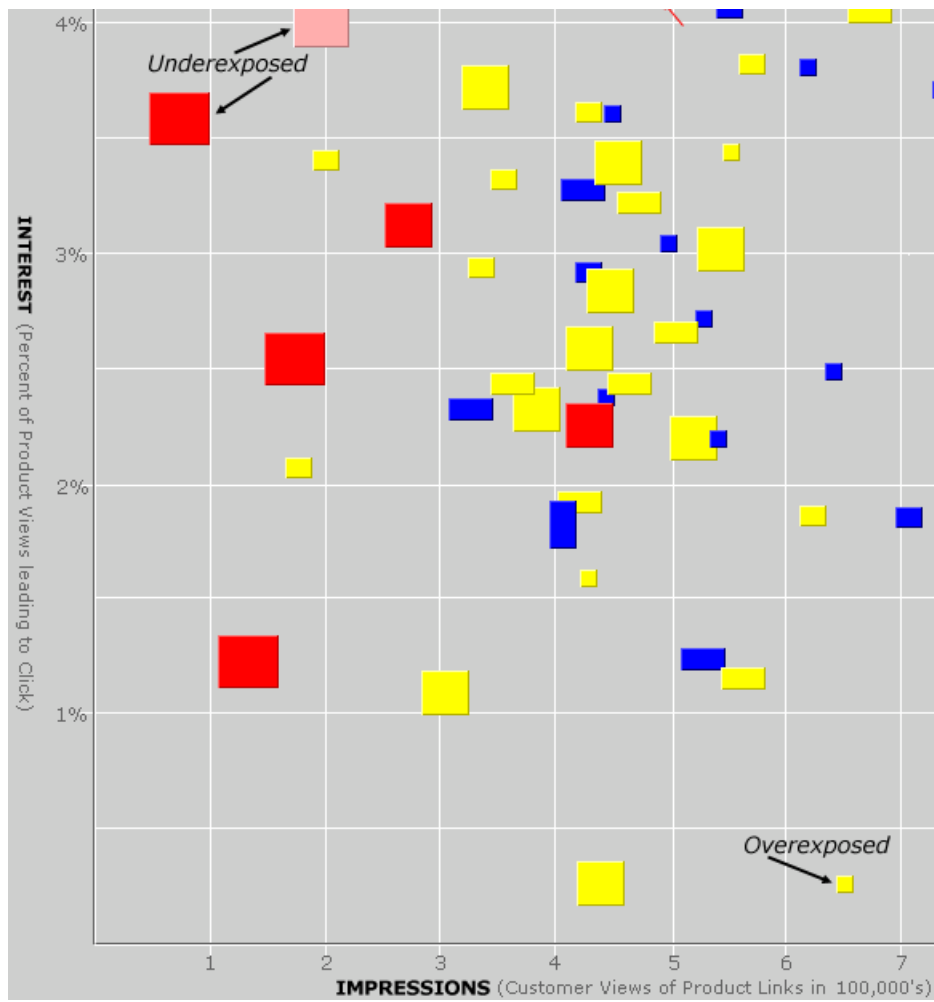


Figure 2. Visualizing micro-conversions of products with a scatterplot graph

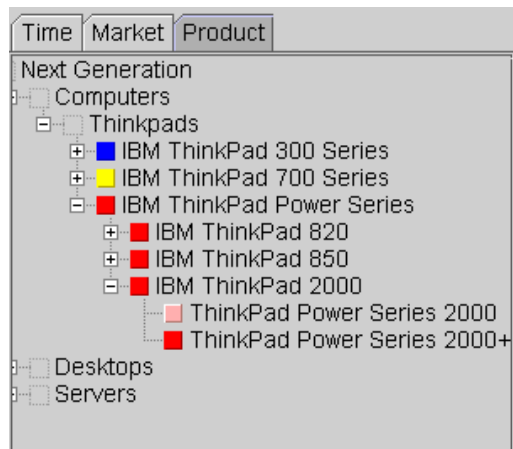


Figure 3. Tree control

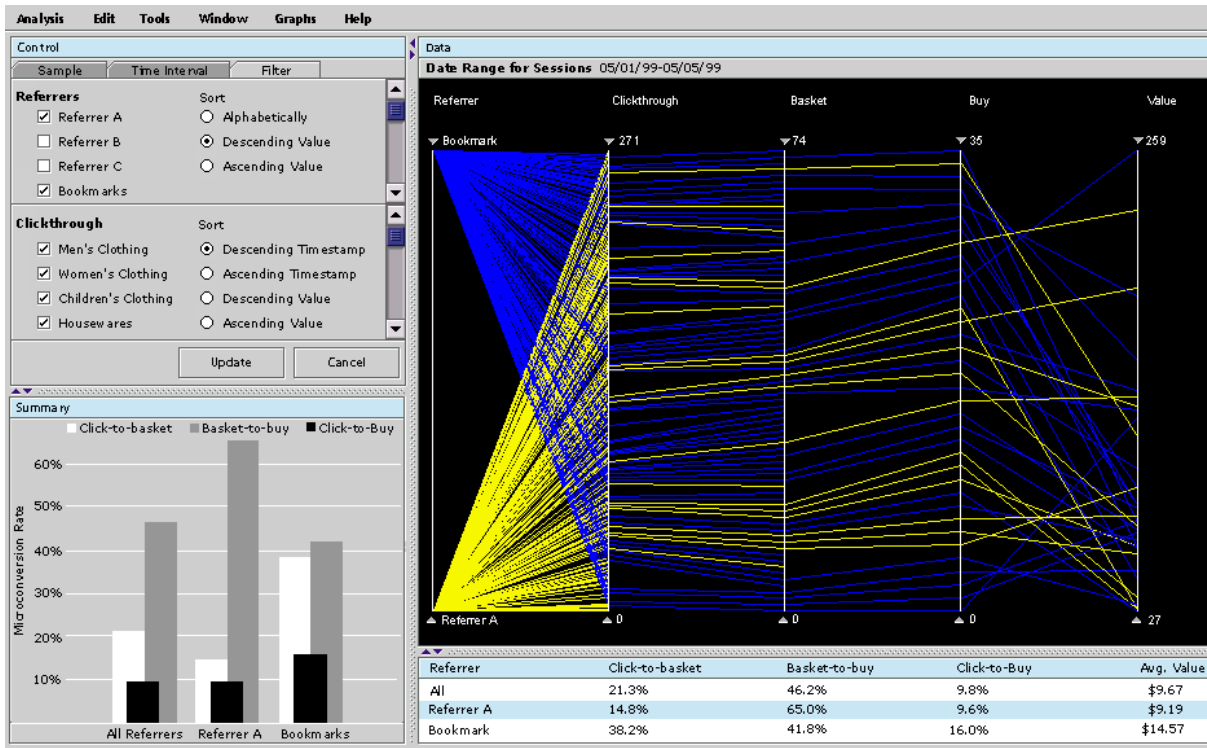


Figure 4. Micro-conversions of sessions categorized by referrer

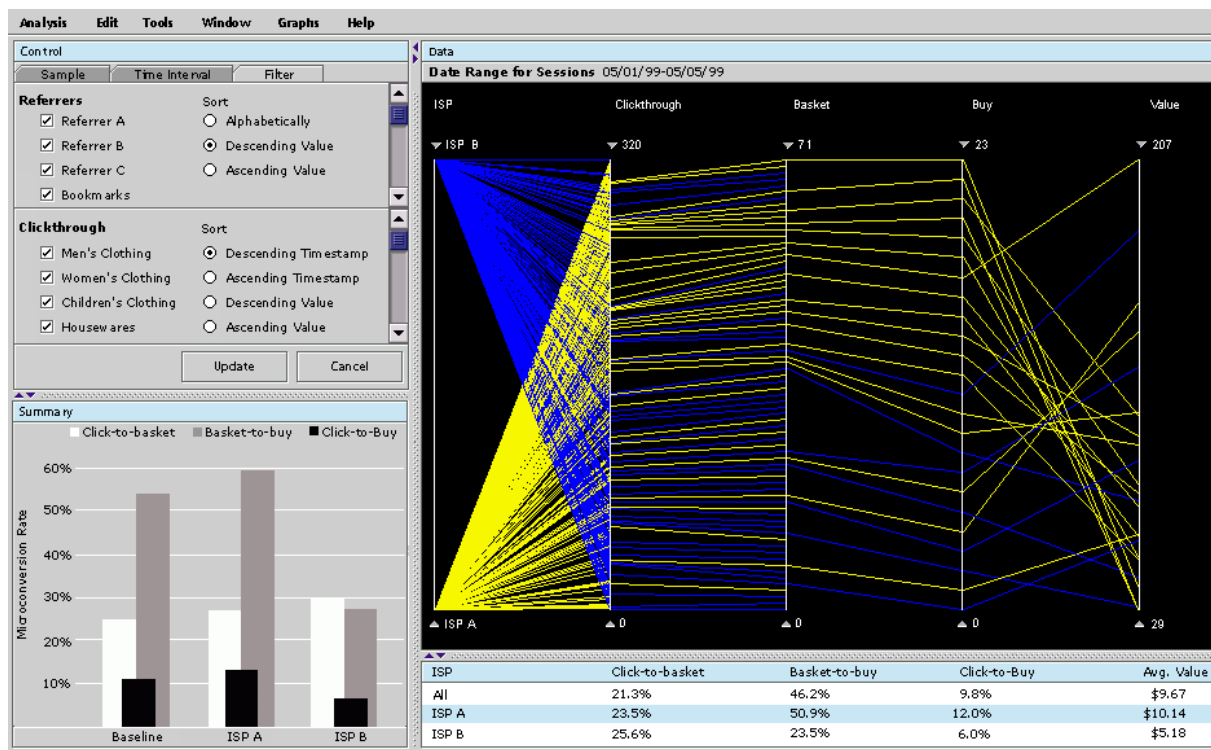


Figure 5. Micro-conversions of sessions categorized by ISP

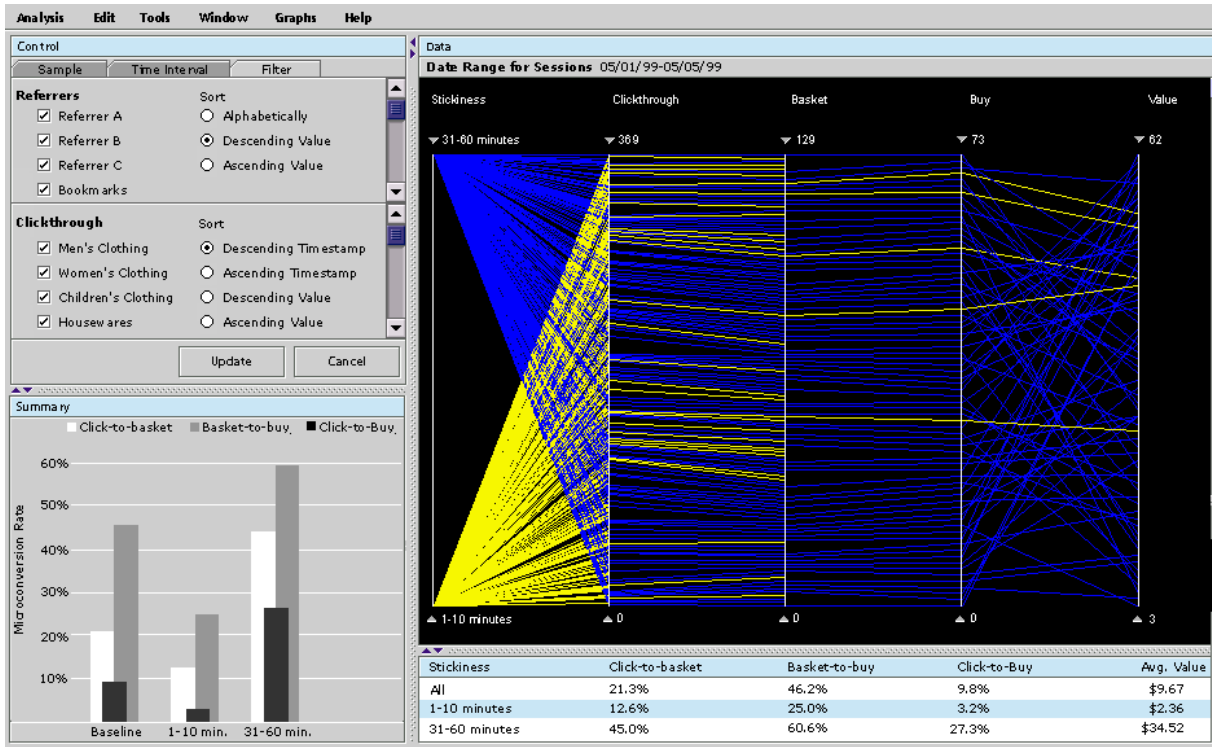


Figure 6. Micro-conversions of sessions categorized by length

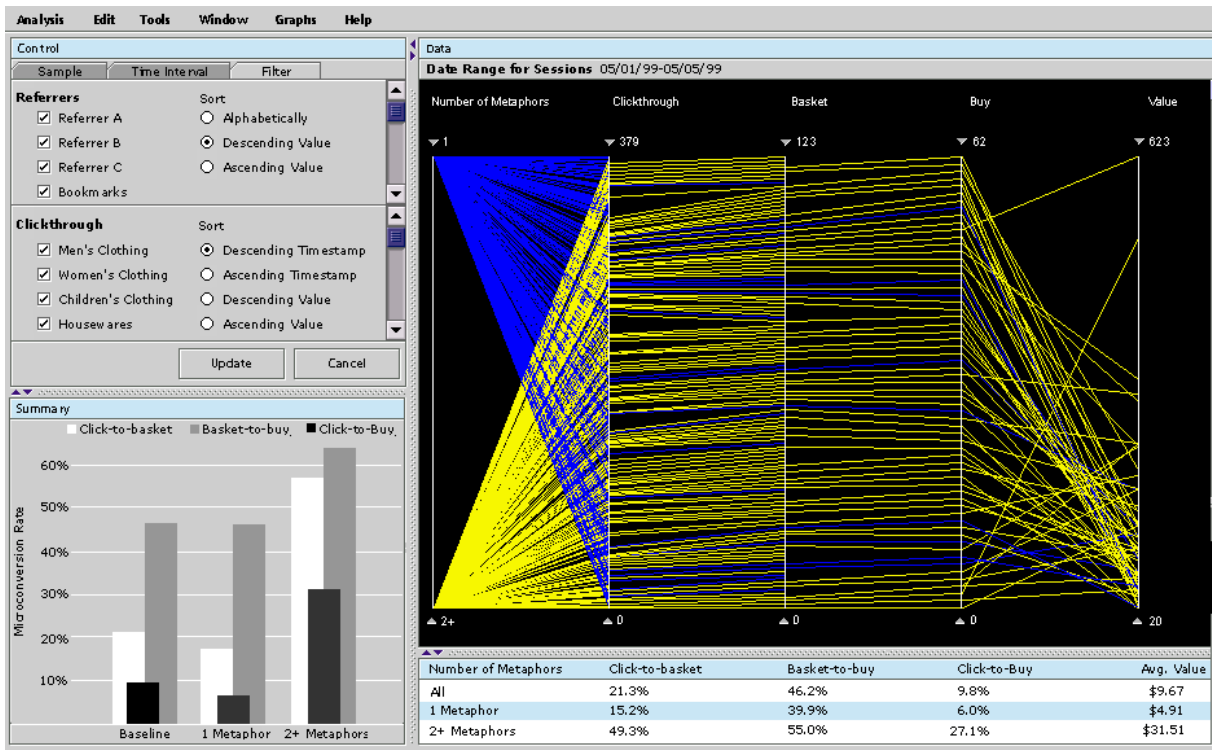


Figure 7. Micro-conversions of sessions categorized by number of shopping metaphors used

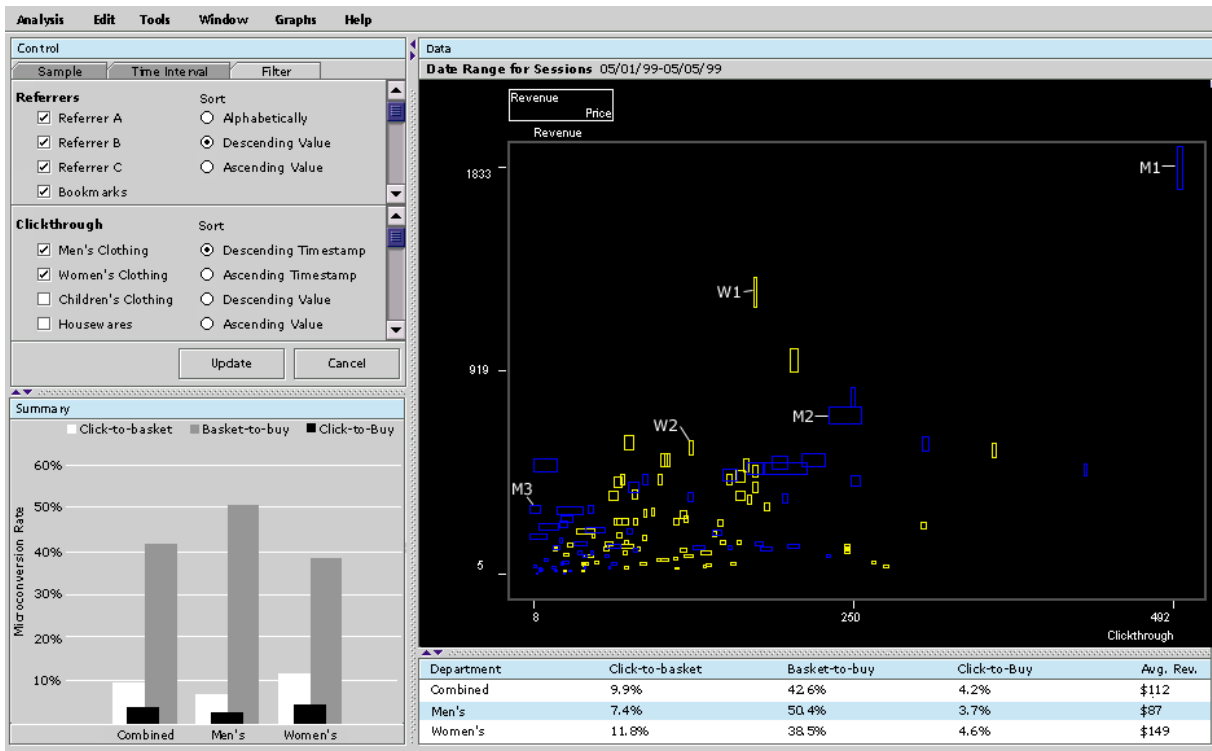


Figure 8. Clickthroughs vs. revenue of products

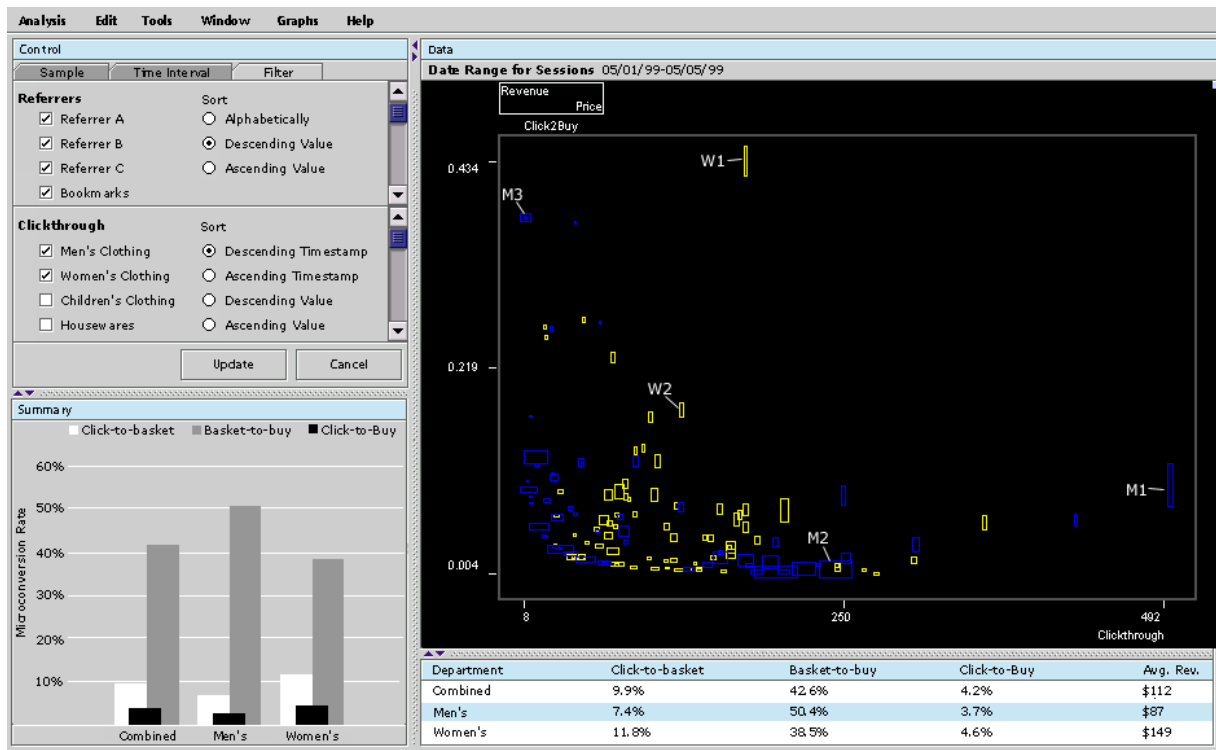


Figure 9. Click-to-buy rates of products

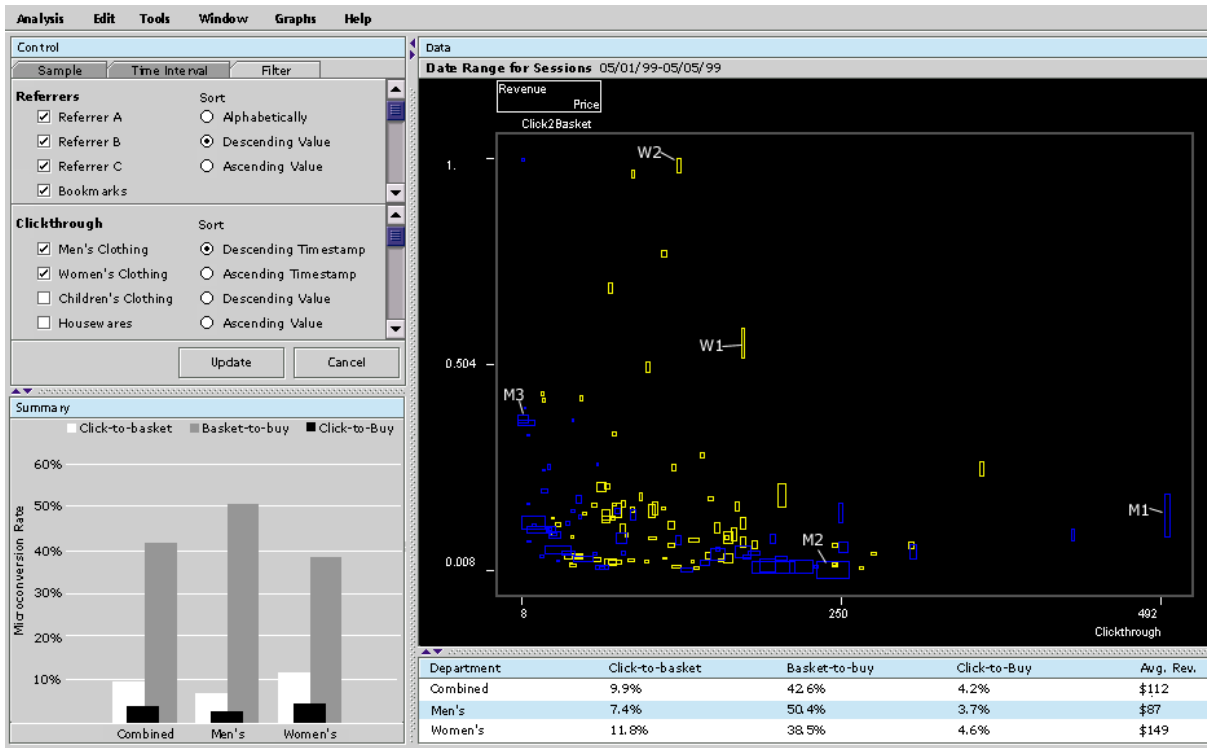


Figure 10. Click-to-basket rates of products

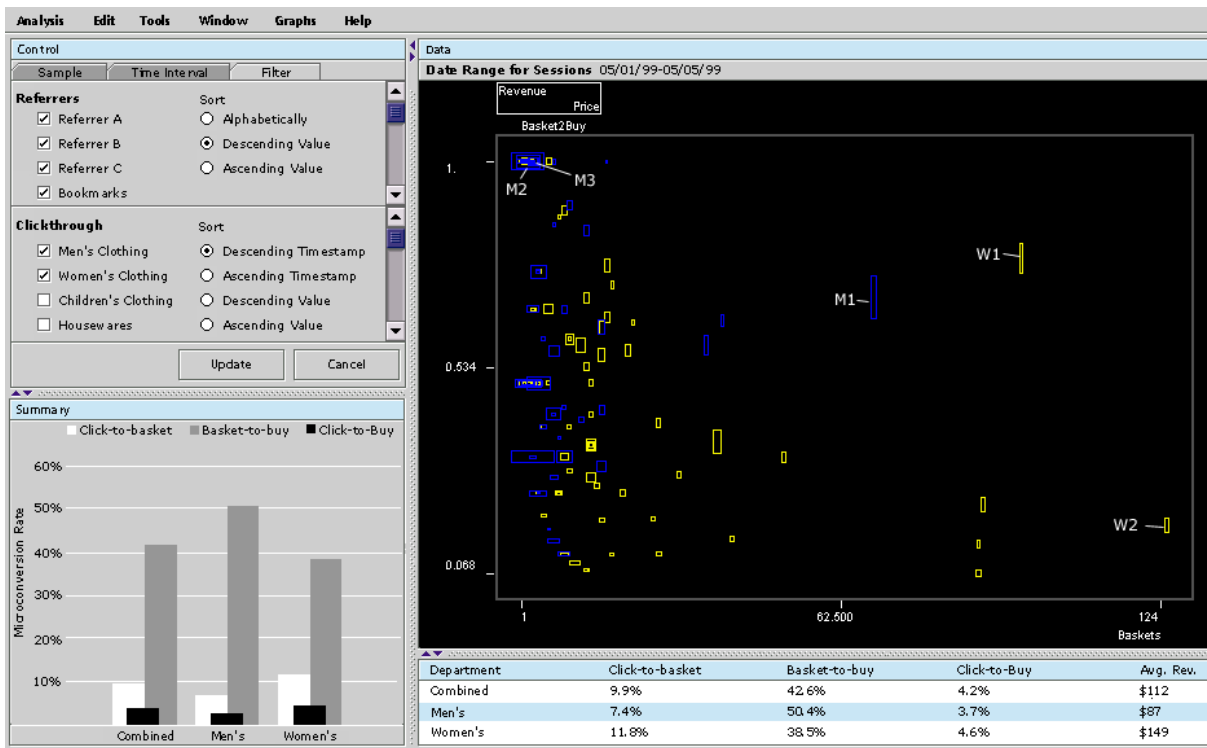


Figure 11. Basket-to-buy rates of products