Mining Gold from E-Commerce Transactions: Challenges[1]

By
S. R. Dalal, D. Egan, Y. Ho, C. Lochbaum, M. Rosenstein
Telcordia Technologies 445 South Street, Morristown, NJ 07960

Abstract

E-Commerce is a powerful growing force in the economy that even in its infancy generated more than $160 billion in revenues in 1998. E-Commerce will fundamentally change marketplaces, capturing large percentages of many business-to-business and business-to-consumer markets, and in some cases entirely displacing traditional markets. As commerce moves from the physical to the electronic space, data related to commerce that previously had been expensive or even impossible to collect becomes available almost for free.

This paper begins with examples of present day e-commerce, and e-commerce scenarios possible in the not too distant future. The data generated in the course of these scenarios reveal much about customers' pre-purchase behavior. We can observe, for example, how customers shop, the items they inspect before making their purchase decisions, and their search strategies for products and services. Demographic and psychographic data, as well as traditional customer data stored in legacy systems in different geographical locations can augment these observations. By using these data, a business can facilitate marketing, sales, and customer service through highly targeted advertising, recommendations based on preferences, online rating and discounting, and real time responses to customer inquiries.

To realize this vision, data based on Web transactions must be mined intelligently. We describe case studies showing that Web server log data cannot be taken at face value without risking gross misunderstanding of a Web site's traffic. A good understanding of a site's traffic is required for advertising purposes, and the computation of important e-commerce indices. We next describe Latent Semantic Indexing (LSI), a technology that aids natural language processing. We describe data mining experiments that use LSI to explore two applications important to e-commerce: automating the handling of customer inquiries, and automating recommendations of products based on text descriptions of their content.

Finally, we consider some future challenges for e-commerce data mining. One set of challenges involves augmenting Web e-commerce session data in real time with additional data captured in different contexts, data provided by third parties, and data integrated from legacy systems. A second major challenge concerns the privacy of user data. Surveys show that users already are greatly concerned about privacy, and advanced e-commerce data mining is likely to generate even more concerns. We conclude with a discussion of technologies that may enable the convenience afforded by e-commerce data mining, while allowing the user to own and distribute personal data.

1. Introduction

Thirty years have passed since the invention of the Internet was devised as a means of communications among defense agencies and research universities. With the development of World Wide Web and browsers, the focus of the Internet has shifted to transactions among businesses and consumers. These transactions are typically known as e-commerce transactions. There are many definitions of e-commerce. For example Electronic Commerce, or simply e-commerce, can be defined as "the buying and selling of information, products, and services via computer networks" including the "support for any kind of business

[1] Please address all correspondence to S. Dalal, sid@research.telcordia.com

transactions over a digital infrastructure"[2]. These may include transfer of ownership, products, or services. Whatever the definition for electronic commerce, it is clear that it is responsible for revolutionizing our concept of commerce by creating a digital economy.

To make the concept of e-commerce transactions more concrete we list few a examples of them.
- a consumer buys a music CD from an online music store
- a consumer buys and downloads software online
- a consumer sells a family heirloom at an auction site to another consumer
- a retailer orders online goods from a distribution center for just in time (JIT) inventory management
- a business orders merchandise from another business using Electronic Data Interchange (EDI)

In 1998, the total Internet economy generated around 1.2 million jobs in the USA, resulting in total revenues of US$301 billion (Barua, Pinnell, Shutter & Whinston, 1999). Taking out the Infrastructure (e.g., network suppliers, hardware suppliers) and the Application (e.g., browser suppliers, database suppliers) layers, around 750,000 jobs were attributed to electronic commerce activities resulting in US$160 billion in revenues. More impressive is the compound annual growth rate statistic. From 1995 to 1998, the compound annual growth rate of the Internet economy was 174.5% compared to a 2.8% growth rate of the overall US economy. According to the US Department of Commerce, e-commerce growth is going to become even more significant given that at present (1999) there are 5 million web sites with 80 million Americans connected to the web (200 million world-wide) and 800 million web pages.

Broadly speaking, the e-commerce marketplace can be categorized into four segments: 1) Consumer to Consumer, 2) Business to Consumer, 3) Business to Business, and 4) Consumer to Business. For example, the auction sites like eBay are examples of the Consumer to Consumer segment, while online shopping sites like Amazon.com, CDNow, etc., are examples of the Business to Consumer segment. The Consumer to Business segment is the least developed, but includes individuals selling antiques and other collectibles to dealers, or items to be sold on consignment by dealers. In fact, some eBay transactions are almost certainly of this kind. Finally, the Cisco site is an example of a Business to Business e-commerce site where businesses can buy routers and other network equipment. In fact, the Cisco Systems Inc. site is currently the world's largest e-commerce site, selling more than $32 million in products every day[3]. Given the importance of supply chain management, JIT, and migration towards virtual business enterprises, it is expected that the Business to Business segment will be the fastest growing segment.

2. Need and Complexity of Measurements

Over the last several centuries, we have developed the statistical science of gathering information about economic activity including markets and transactions. The information is used by governments for the purpose of taxation, and by businesses to identify their customers, run promotions, manage their suppliers, and create new needs. However, as commerce moves from the physical to the digital space, data that previously had been expensive or even impossible to collect becomes available almost for free.

These new datasets reveal much about customers' pre-purchase behavior. We can observe, for instance, how customers shop, the items they inspect before making their purchase decisions, and their search strategies for products and services. To fully utilize this data, it must be coupled with the traditional information stored in legacy systems residing in different physical and geographical locations. Linking traditional demographic, and buying-history information with shopping behavior allows the e-commerce merchant to better serve clients' needs by recommending appropriate goods and services. This coupling generates substantial new opportunities for making purchase decisions more accurate, quicker and easier. To realize this vision, a significant challenge is to develop an integrated view of a customer. This view allows the merchant to accurately recommend the right bundle of products, services, prices, and discounts. To put this in perspective consider the following scenario.

---

[2] Bloch, Pigneur, Segev
[3] Cisco System, Inc.

A consumer goes online looking for a new science fiction book. He goes to the web site of a particular bookstore. As soon as he is on the web site of the bookstore, the bookstore flashes an ad, and the home page gives a number of options (e.g., book, music, videos, etc.). The consumer clicks on their book section. In return, the bookstore recommends a new book by a particular author. It also flashes reviews of the book, and a quick summary. To clinch a quick sale and promote its music sales, it plays the music from a new movie based on that book, and offers a 10% instant rebate on it if it is bought with the book. The consumer buys the book with a credit card, and also buys the music, which he downloads. Besides the purchase price of the book, he pays a shipping and handling charge. Two days later, since he did not receive the book in 24 hours, he contacts the bookstore's online customer care center where he is informed that his package is currently at the delivery service's center hub in Tennessee and will be in his town at 3 pm and is given the tracking number. He gets his book at 4 pm.

This typical transaction generates an enormous amount of data. As soon as the customer enters the bookstore's site, the transaction is logged onto the bookstore's server. The information transmitted includes his specific IP address, time of day, and the website which he previously visited. This gives a clue to the bookstore as to the consumer's current state of mind. Further, since he has visited the bookstore previously (known from the cookie placed on his PC), his past profile and purchase behavior are also analyzed in real time. Based on the analysis, the bookstore decides to flash an appropriate ad, and also realizes that he is more likely to buy science fiction. A collaborative filtering system compares the past purchase data with the data from other customers, and finds a group of similar customers who have recently bought a new book. Based on their comments (in text) and their purchase behavior, it flashes the most appropriate book to the consumer. The system also figures out that the person is more likely to buy music, and increases the probability of a sale by offering a rebate while the music is played in background for the consumer. Once the consumer decides to buy the book and the music, an online discounting engine checks for a package deal on the book and the music. Finally, the consumer receives a bill with shipping and handling charges.

Besides generating the consumer related data, "backend systems" at the bookstore authenticate the consumer's credit card, and send the request for that book to the central warehouse of the publisher by an EDI (electronic data interchange) along with the order number. With further processing, information is sent to the music server of the music producing company about the music download and to the delivery company for pickup and delivery to the consumer. Finally, arrangements are made to collect the funds from the consumer's bank, send out an acknowledgement by e-mail to the consumer including the package number, and transmit all this information to a customer care center. When the customer care center is contacted on line by the customer, a trouble management system reads the customer complaint in English, parses it, realizes the kind of trouble (see Section 6.2), contacts the shipping company, finds out where the package is, and instantly informs the customer of the status of the package and expected time of delivery.

The above example covers many aspects of e-commerce. There are many players playing multiple roles in this simple transaction. For example, the bookstore is a seller and a buyer. As in any measurement program, we need to determine from whom to collect needed transaction data, from the buyer or the seller? For example, from a statistical perspective, one could estimate e-commerce retail sales of books by surveying households (the buyers). This would require a very large sample and be very expensive. Alternatively, one could survey on-line bookstores (the sellers), this would be a much more cost-effective data collection strategy that could provide timely, high quality estimates from a very small sample.

The above example also points out that any given business-to-consumer transaction will involve a large number of related business-to-business transactions. Growth in transactions is expected because as e-commerce expands, related business-to-business transactions will become more distributed; participants will concentrate on performing their highest-valued activities and rely increasingly on third parties for lower-value-added activities. The measurement challenges include, for example, avoiding double counting the value of related transactions.

3. Basic measurements- Use of log files to collect data

There are many specific metrics to measure e-commerce transactions. They range from traffic, delays, consumer interactions, their other interests, etc. They are used for improving the web site's performance, usability, tracking customers for advertising and promotional purposes, etc. The basic unit of the measurement is a ***hit.*** *"Hit"* means a single request from a web browser for a single item from a web server; thus, in order for a web browser to display a page that contains 3 graphics, 4 "hits" would occur at the server: one for the HTML page, and one for each of the 3 graphics. Since much of the information about hits and other relevant information can be reconstructed from the information in log files, we spend a little time exploring log files. A detailed discussion on metrics is deferred to Section 4.

The common activity log format is the standard for most Web servers where the Web site may reside. It exists as a text file in the log directory. Numerous log formats exist, each of which can be used to analyze Web site traffic. The common log will report the following information on every visitor and what activity they performed at the web site: host/ip, RFC name, logname, datestamp, retrieval code, and bytes.

The first field in the record is the host/IP. It indicates the hostname, or RFC name, of the visitor or an IP number if domain name server lookup is not enabled for that visitor. Visitors can have .com, .edu, .gov, .org, .net, .uk or many other extensions, indicating whether they originate from a company, an ISP, a university, another country, etc. The next important field is the datestamp, which is useful when running the log figures through analysis software. The datestamp information can be used to graph peak activity throughout the day, week and month. This will show whether traffic picked up in response to a promotion, weekly or seasonal cycles, etc. Next, the retrieval method is posted. Most visitors retrieve multiple files which make up a single Web page. If a banner ad also appeared on a page and was retrieved with other files, an appropriate counting device would register an ad impression. The code field indicates whether a retrieval was successful, and the bytes field indicates the size of the file retrieved. This size data is important because many ISPs charge a Web site based on traffic activity as measured in bytes. Other fields in a log file show which browser and computing platform were used by the visitor.

Along with the log file, another file, Referrer URL file gives additional information about where the visitor came from. Often, the referring URL will be a search engine or directory. In this case, the field will also show what "keyword phrase" the visitor used to find a particular site in the search engine listings. This can be used to fine-tune the keyword usage in your home page. If the referring URL field reports NULL, then the visitor typed in the URL directly or has it saved in a bookmark or hotlist file. Knowing the difference between a visitor who was "referred" by another site, and a visitor who came on their own will help in discriminating new visitors from repeat visitors. "NULL" in the referrer field could also mean that the visitor typed in the URL from a printed listing in a magazine or newspaper. Some direct marketing companies use different URLs that ultimately take the user to the same site. The URLs placed in various ads, mailings, and media can be used to track the effectiveness of marketing campaigns. Besides this information, using cookies enables Webmasters to log even more detailed information about how individual users are accessing a site.

In addition to the traffic activity, logs can show other information relevant to advertisers. This could include high levels of activity from students and other visitors who are or are not prospects for the advertiser's product, attention span (e.g. short visit times may indicate that a site gets many visitors who show a low interest level), and visits by "bots" which count an impression that was not actually viewed by a human.

4. Basic Measurements: Web Metrics

Many current e-commerce activities are related to advertising. It is generating over $3 billion per year at the present time and is expected to grow to over $7 billion per year by 2001. Typically advertising is done by putting a banner ad on a web page visible to a user. A banner ad often includes a link to the sponsor's web site. The ad can be targeted by a number of factors. They include: the user's domain, the time of day, the site the user is coming from, demographics (if the user has registered previously), content of the page the user is browsing, etc. Since advertising occupies such an important place in e-commerce, we give a

table of commonly used web metrics for measuring effectiveness of ads in Table 1. Besides these, there are many other traditional metrics being used for measuring effectiveness, such as recall [4].

One of the critical problems in advertising is to select the best of a number of potential ads to target consumers. Since the response rate is a random variable, and the difference between a good ad and others could be high, a banner ad needs to be selected from a number of competing ads by using a statistical sampling approach. Dalal and Srinivasan (1976) have proposed an optimal Bayesian sampling method for selecting the best ad by pre-testing when the response variable is dichotomous (e.g. recall).

Table 1: Table of commonly used web metrics in measuring ad effectiveness and their explanation.[5]

Metrics	Explanation/Definition
Ad Clicks	Number of times users click on an ad banner.
Ad Click Rate	Sometimes referred to as "click-through," this is the percentage of ad views that resulted in an ad click.
Ad Views (Impressions)	Number of times an ad banner is downloaded and presumably seen by visitors. If the same ad appears on multiple pages simultaneously, this statistic may understate the number of ad impressions, due to browser caching. Corresponds to net impressions in traditional media. There is currently no way of knowing if an ad was actually loaded. Most servers record an ad as served even if it was not.
Bandwidth	How much information (text, images, video, sound) can be sent through a connection. Usually measured in bits-per-second. Used to calculate the amount of time an ad will take to download.
Browser Caching	To speed surfing, browsers store recently used pages on a user's disk. If a site is revisited, browsers display pages from the disk instead of requesting them from the server. As a result, servers under-count the number of times a page is viewed.
Click through	The percentage of ad views that resulted in an ad click.
CPC	Cost-per-click for a specific banner ad. Cost usually runs in the range of $.10 -.$20 per click.
Conversion Rate	The percent of visitors who become bonafide buyers.
CPM	CPM is the cost per thousand for a particular site. A Web site that charges $15,000 per banner and guarantees 600,000 impressions has a CPM of $25 ($15,000 divided by 600).
Gross Exposures (Hits)	Each time a Web server sends a file to a browser, it is recorded in the server log file as a "hit." Hits are generated for every element of a requested page (including graphics, text and interactive items). If a page containing two graphics is viewed by a user, three hits will be recorded - one for the page itself and one for each graphic. Webmasters use hits to measure their server's work load. Because page designs vary greatly, hits are a poor guide for traffic measurement.
Page Views	Number of times a user requests a page that may contain a particular ad. Indicative of the number of times an ad was potentially seen, or "gross impressions." Page views may overstate ad impressions if users choose to turn off graphics (often done to speed browsing).
Unique Users	The number of different individuals who visit a site within a specific time period. To identify unique users, Web sites rely on some form of user registration or identification system.
Valid Hits	A further refinement of hits, valid hits are hits that deliver all information to a user. Excludes hits such as redirects, error messages and computer-generated hits.
Visits	A sequence of requests made by one user at one site. If a visitor does not request any new information for a period of time, known as the "time-out" period, then the next request by the visitor is considered a new visit. For example, I/PRO uses a 30-minute time-out period.

[4] Dalal & Srinivasan (1976), Management Science,
[5] Extracted from a number of Internet Dictionaries

5. Fallacies in looking at log files- some measurement issues

We have discussed the use of log files to generate different web metrics. These metrics are supposed to measure traffic, transactions, etc. However, in spite of the promise, these metrics have limitations which can skew measurements in a particular direction and bias the results. We now discuss these issues here.

One of the key metrics of interest is the number of unique visitors. Since this metric is not directly available from the log files, analysis tools try to derive it by using IP addresses, registration information, etc. This metric is thus, highly susceptible to what is being metered in the log file. For example, since most sites have repeat visitors, and most visitors view more than a single page of information, so 100,000 impressions could actually represent 10,000 to 40,000 unique visitors. Further, most "Hits" and "page hits" are not favored as measurement units because a hit is registered every time any text or graphic file is delivered, whether advertising is displayed or not. Thus, the ratio of hits to page hits could easily be 10 or more.

There are many other ways over-counting can occur. For example, an intelligent agent, or "bot," could automatically visit a site every hour and inflate a visitor count. Such intelligent agent may include Spiders" and "Crawlers," which are software programs that visit virtually every page on the Web to create indexes for search engines, and performance measurement companies, that measure download time by sending traffic to a site. There are also "rogue agents" which have been deliberately designed to inflate traffic. A rogue agent acts as a "visitor" to a web site and sends a fictitious http_referer and a fake IP address. Rogue bots are rare. Besides bogus visitors, there are other legitimate users who increase the number of ad impressions. For example, Internet bottlenecks can cause the visitor to request the same page repeatedly until the full page with graphics is actually displayed. But the traffic counting software is counting all requests for the full page, thereby over-counting impressions. Another possibility is that, if the tracking software is counting requests for a text page, rather than the ad banner, this could also result in over-counting the actual ad impressions . For example, "virtual includes", and filters can also cause over counting. "Virtual includes" are techniques that are used to capture only a desired image, such as a weather map, without a user having to visit a page and view banner advertising. Filter software such as "Junkbuster" enables a user to strip banners and cookies from accessed pages in order to speed up page delivery. Users may also disable graphics on their browsers to only show text with a similar effect.

Besides over-counting there are other factors which can cause under-counting. This under-counting can be caused by "caching", "site mirroring" and "firewalls." With caching and site mirroring, the page displaying the ad is sometimes delivered once, then stored locally on the visitor's computer or the ISP's computer. When the page is viewed later -- by the same visitor or other visitors -- it is retrieved from local memory rather than from the Web site. It is faster for surfers, but prevents the tracking software from counting it as an impression delivered from the Web server. A firewall with proxy server is used by corporate intranets for security, but it also can cause impressions to drop because it's possible for 10 or 10,000 corporate employees to be identified by the same corporate visitor address. Further, many ISPs assign the same IP string to each new user as the previous user logs off. Proxy servers and local caching also hide portions of IP strings from the server log access files. So, during a very busy time period, more than one unique visitor could visit a site with the same reported IP string, and the log will consider it a single visitor.

There are several possible ways to improve the data from Web log files, none totally satisfactory. The most expensive and comprehensive solution is to use statistical surveys (see Section 5.2), and subscription information to supplement the estimates. Another solution that many sites use to correct their traffic measurement is to set a cookie that uniquely identifies each visitor. This is assigned, and stored in the browser's cookie file the first time one visits the site. On subsequent visits, the unique cookie helps identify the visitor in conjunction with the IP address. Cookies are only partially helpful in identifying unique visitors because they can be turned off or the visitor may use different computers and browsers with

differing cookies, which makes them a less than perfect confirmation tool. A different solution related to cache is Cache-Busting. "Cache-Busting" refers to the ability by ad management systems such as MatchLogic and Imgis to "bust" cached banner ads (to "view" inside a cache and count the number of ads delivered from the cache rather than from a server). This improves accuracy of reporting delivered advertising impressions.

We summarize the above discussion in the following table.

Table 2: Table of Factors and methods affecting the reliability of "Unique Users" metric

Effect	Factors	Method
Under-count	Local Caching	Frequently used pages are stored in local cache. User doesn't go to the designated server, and thus, server undercounts
	Proxy Servers/Firewalls	Gives same IP address to a number of users- thus, # of IP addresses are undercounting # of users
	Site Mirroring	Frequently used pages are mirrored in another server. User doesn't go to the designated server, and thus, server undercounts
	Dynamically generated IP	IP addresses are dynamically generated- thus, same user can get different addresses and vice versa
Over-count	Crawlers	Agents for search engines which go on visiting pages for indexing purposes
	Rogue Bots	Software which mechanically generates traffic and IP addresses to inflate count
	Graphics Off on browser	Only text part of a graphic ad is seen- but, registered as a visit
	Performance Measurement Tools	To measure download time, and reliability, goes on polling sites throughout the day.
	Filters/Intelligent Agents/Virtual Includes	Filters and agents are fetching only a specified part of a web page (e.g. text), thus, users do not see ads
	Internet bottleneck	When download time is long, user goes on repeatedly requesting the same page

5.1 Why the 4am Thursday Crowd Isn't Buying (a sidebar)

Here we consider a case study of some of the practical problems of interpreting Web log data.

Figure 1 shows traffic levels at a web site taken from its server logs. The graph shows summary view of the data by hour from April 1999. In all aspects but one, the data show the expected pattern of traffic for a United States based business-to-business site with the business day starting around 8am EDT and a long decreasing tail from customers across the three major U.S. timezones. The anomaly is the traffic spike at 4am. As might be readily expected, the spike does not represent actual customers. Further analysis and inspection of the web server logs revealed that the indexing spider for the site caused the spike. Each day at 4am, the spider gathered up web pages to be indexed and used by the site's search engine.

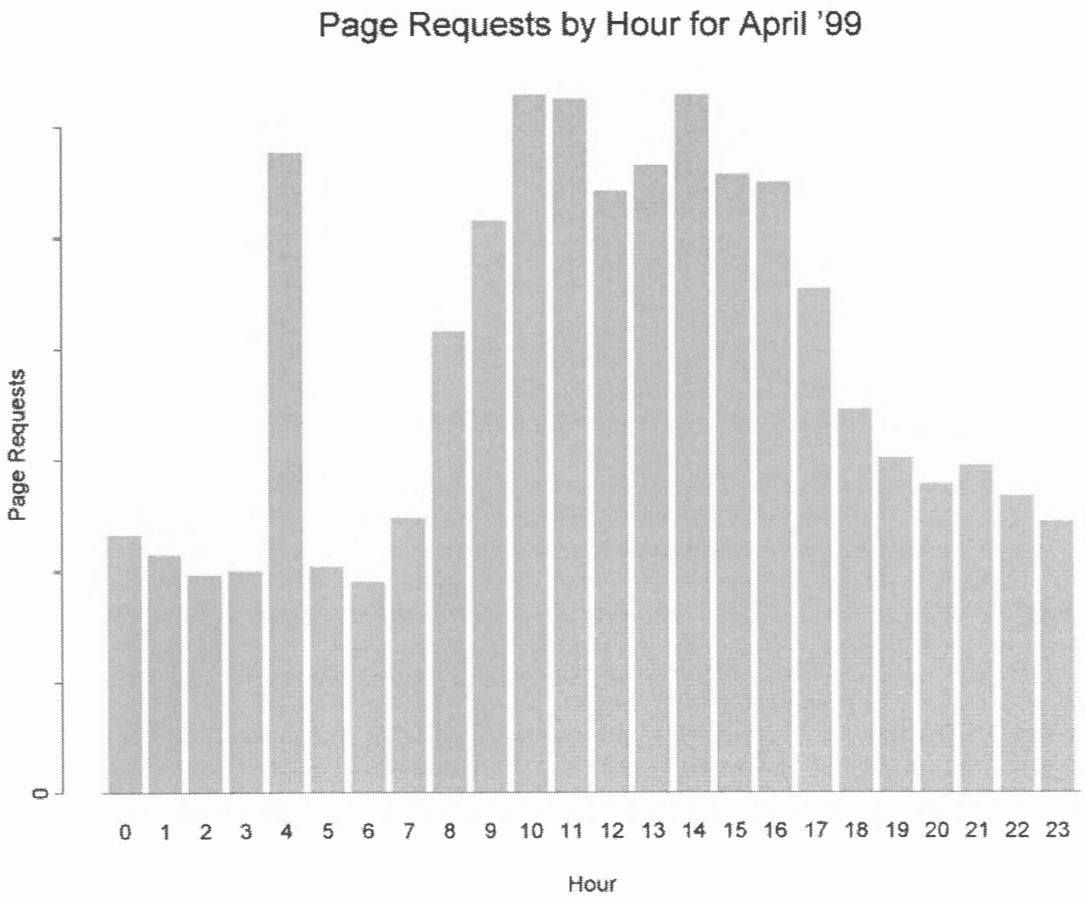

Figure 1. Traffic per hour of day based on a Web site's server logs.

That in itself is unremarkable, but what is remarkable is that the spike showed up at all in the analysis. The analysis software that produced this view was configured to exclude all hits to the site from clients within the company, so this spike (while actually very interesting!) should have been filtered out. Why it wasn't provides a significant lesson for analyzing server log data.

Typically server logs record the IP address that originated a request on the server. To translate a name such as mymachine.mycompany.com into a numerical IP address requires a Domain Name Server (DNS) lookup. The analysis tool allows filtering by host, and the filter in effect was *.mycompany.com, meaning to exclude any host with a name that ended with mycompany.com. Due to an oversight, the machine on which the indexing spider ran was never assigned a name. It had an IP address, but no name. When the analysis tool attempted to look up the IP address of the indexing spider machine, no name was returned, so that address was not filtered out, and all the hits from the spider were reported from the analysis tool. Accurately enumerating all the possible subnets for a large company can take significant effort, but to actually remove all the sources of intracompany hits requires configuring the analysis tool with that information.

Figure 2 shows traffic levels at this same web site over the same period, but here a summary of traffic by day is shown. Again, except for one possible exception, this shows the expected traffic levels of a business-to-business site, with weekday traffic dominating weekend traffic. For some reason, Thursday is about 15% higher than the other weekdays. Though this might not be significant, it seemed worth a further look.

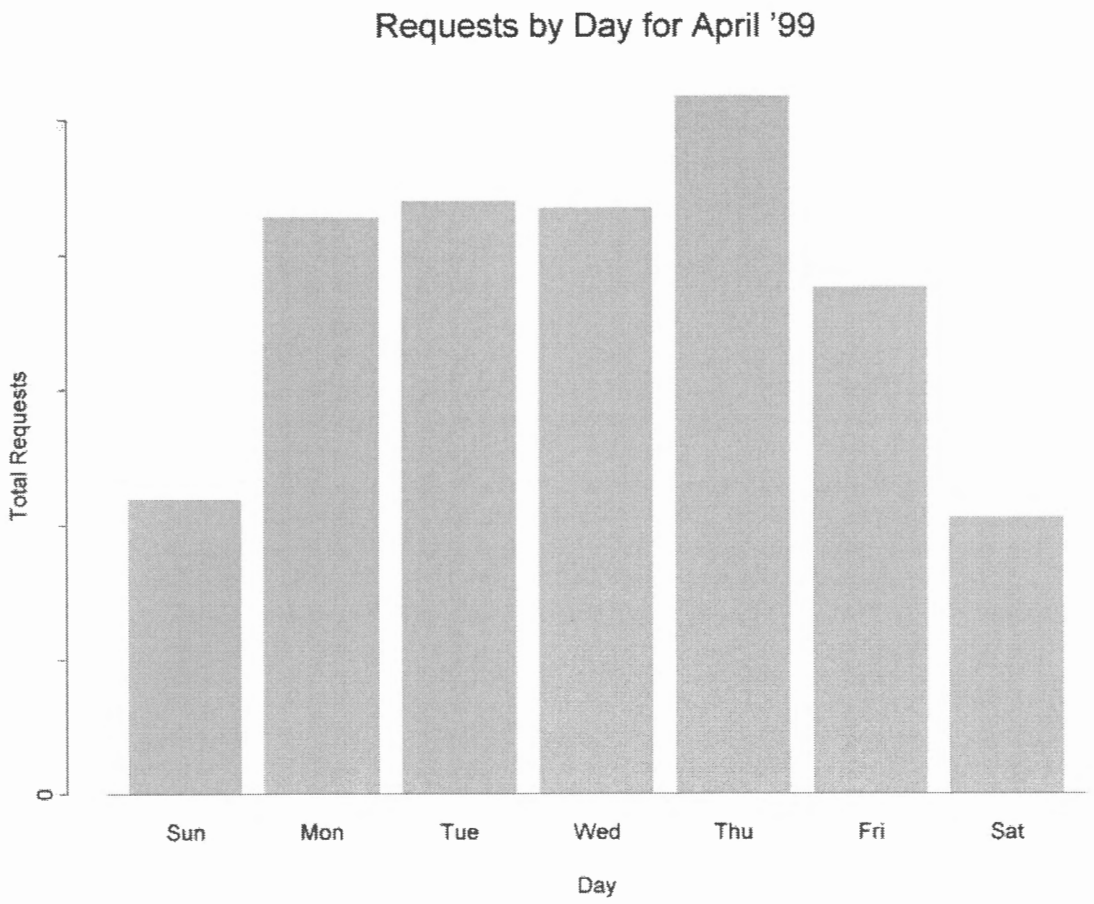

Figure 2. Traffic by day of week based on a Web site's server logs.

As it turns out, April 28 (a Thursday) was a great traffic day - twice the traffic of the other three Thursdays in April. Unfortunately, there were no customers hooked onto this traffic. Just from the ferocity of the requests, it is clear that a human did not generate this traffic. The user_agent field of these requests clearly indicates that a Proxy Server, almost certainly one that had gotten out of control generated this traffic. Note that this was not apparently an attack, since the Proxy Server was located at an important client. This might be of interest for web traffic management, but for business analysis it should be filtered out. At this point it is clear that for understanding the business of this customer, the 4am Thursday crowd never really existed!

The total height of the bars in Figure 3 show the traffic at this site for the week of April 11 with the Spider and Proxy hits removed. The height of the dark gray bars is actual traffic. The difference, depicted in light gray, is the result of hits generated by a firm hired to monitor the performance and availability of the site. This monitoring accounts for about 30% of traffic on weekdays, and well over half on weekends.

Due to a miscommunication, the group analyzing the site was unaware of this monitoring activity by another organization. This substantial amount of activity was not originally noticed, since the amount of monitoring was ramped up over time, masking the traffic in an expected increase in total traffic over time. Also the monitoring was conducted from a large set of machines around the Internet, so the individual load of each monitoring machine closely mimicked the traffic behavior of large customer sites. What gave the game away, again, was the user_agent field in the log, which annotated this monitoring service. Notice with this distorting influence, it appears that weekend traffic is about half of weekday traffic, which would be unusual for a business-to-business site, while in fact it was closer to 1/5.

Though this site may be an extreme example, understanding all this extraneous traffic is crucial for understanding the site. All the business parameters such as buy-to-browse ratio are radically distorted with the original set of data. Not until the analyst has some handle on traffic originating from real customers does it make sense to try and do more detailed modeling.

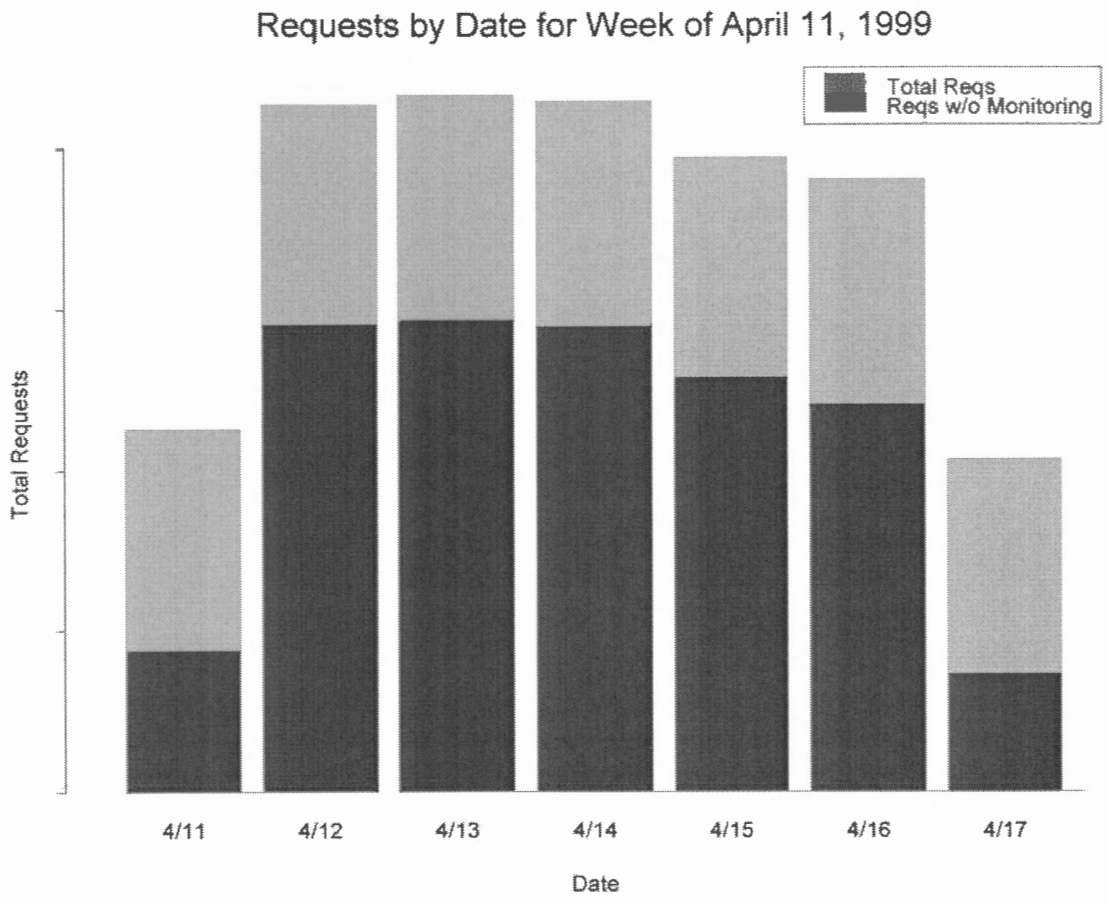

Figure 3. Traffic per day of week based on a Web site's server logs.

5.2 An Alternative Method for Measuring Web Traffic: User Panels

Monitoring the activity of samples of users with specially instrumented browsers is one alternative method for measuring Web traffic that avoids many of the pitfalls of analyzing Web transaction logs. For example, Media Metrix, Inc. employs a sample of more than 50,000 users with browser software that tracks and reports the users' activities. This user panel is constructed to be representative of the demographics of the entire population of Web users in the United States. The proportion of the panel that visits a certain site in a given month is the basis for a projected estimate of the total number of unique visitors to the site that month. The demographics of the panel members visiting the site is useful to advertisers. Because this method begins with known individual users and tracks them closely, it is not susceptible to the over- and under-counting problems summarized in Table 2.

Extrapolating results from a sample user panel is not the entire solution, however. One problem with this approach is that, while it may generate reliable traffic estimates for the most heavily visited Web sites, it is not useful for the vast majority of sites. If one were to rank order sites by the estimated number of monthly unique visitors and then graph the result, the outcome would be a curve that falls very sharply, and then has a very long tail. In the September, 1999 Media Metrix, Inc. report, the most heavily visited site had an estimated 52,000,000 unique visitors, the tenth ranked site had something over 10,000,000 visitors, and the fiftieth site had just over 3,000,000 visitors. The estimated number of unique visitors per site declines very gradually from that point. What this means is that a site could have substantial traffic, say 500,000 unique visitors per month, and yet this amounts to less than 1% of all visitors. Estimating a fraction of 1% by means of a national panel is extremely difficult.

Sites with less than 1% of the universe of visitors still may be very important and have needs to understand their traffic. Sites with 500,000 or fewer unique visitors per month include many large metropolitan online newspapers and other publications, many Fortune 500 companies, major universities, and the majority of government sites. Sites operated by mid-sized and small companies, sites focused on local interests, and specialty publishing sites may have 50,000 or fewer unique visitors per month—a traffic count virtually impossible to gauge via a national panel.

Estimating the entire population of Web users—a figure that is necessary to project panel data to the population—is also difficult. The best methods presently use a modified area probability design to recruit large samples of respondents who are interviewed face-to-face. The population of Web users is a rapidly moving target that is growing at an estimated 12.5% per year, and one whose demographics are changing in the direction of becoming more "mainstream." Thus, surveys must be updated constantly. There are issues concerning the age of respondents to be included, as well as whether and how to measure respondents' usage at home, school, and work. The result is that estimates for the entire population of Web users at any particular time can differ by tens of millions.

A final problem with using panel estimates and target projections for Web traffic is that this method is inherently indirect. Even if this method were to generate very accurate traffic reports, coupled with the demographics of the panel members for a large range of Web sites, it would still not directly facilitate the one-to-one customer relationships that the Web makes possible. Web site operators need to know in real time precisely who is searching for information, making selections, and conducting transactions in order to provide the most effective marketing and customer service functions.

6. Text Mining Using Latent Semantic Indexing (LSI)

6.1 Natural Language Processing Via LSI

The popularity and accessibility of the web has made it the number one choice for the interface between users and a centralized system. The scenario we described in Section 2 is no longer a science fiction. As a matter of fact, many e-commerce companies are working towards the goal of building automatic on-line customer care centers where only minimum human involvement is required . Customers will be entering their complaints and/or questions on-line. The text will be stored and processed. A query is then issued to retrieve possible solutions and/or answers. Let us revisit the scenario described in Section2. The customer inquired whereabouts of his book, his question was stored and processed. Based on the processed text information, the system decided that there's a high possibility that the customer wanted to know the whereabouts of his book. After receiving a confirmation from the customer, it sent out a query to trace the package. It turned out that the book was lost on its way to the delivery service's center hub in Tennessee. The system then reissued a delivery request to the delivery service and informed the customer of a new order number and estimated arrival time. Since computers are replacing humans in performing tedious tasks like these, they need to be equipped with the ability to process natural language. A Telcordia patented technology called Latent Semantic Indexing (LSI) was invented for Natural Language Processing (NLP) when the web was just in its infancy. Basically, an English document can be represented by a term frequency vector; elements of the vector represent counts of corresponding words that appear in the document. Since there are about 50,000 words in English, a collection of 1,000 documents can be represented by a term by document matrix (TDM) as large as 50,000 by 1,000. We will refer to a column of TDM as term frequency vector – frequency of a term which appears in the corresponding document, and a row of TDM as document frequency vector – how many times a document in the collection contains the corresponding term. To be able to efficiently process and extract global information from these documents, we need to be rid of inconsistencies and ambiguities that individual documents have shown and only maintain the most important underlying structure of the original TDM. LSI does just that. It takes the full TDM, performs a Singular Value Decomposition (SVD), and selects k most influential singular vectors to give a lower rank approximation to the original TDM. More specifically, let X_{td} be a t by d TDM (t terms and d documents. Then SVD will decompose X_{td} into the product of three matrices,

$$X_{td} = T_{dm} S_{mm} D^t_{md},$$

Where T (D) has orthogonal unit length column vectors referred to as the left (right) singular vectors and S is a diagonal matrix of singular values in decreasing order. Here m is the rank of X_{td}. The first k largest singular values are retained while others are set to 0. When three matrices are multiplied back to obtain an approximation to X_{td} only the corresponding k left and right singular vectors remain. Therefore the rank of X_{td} is effectively reduced from m to k**.** Usually k is much smaller than m. Let $\hat{X}_{td}$ denote the lower rank approximation to X_{td}. Then

$$\hat{X}_{td} = \hat{T}_{tk} \hat{S}_{kk} \hat{D}^t_{kd}.$$

But what is a reasonable numeric representation for a document? What is a reasonable numeric representation for a term? Answers to these questions are similar. We will try to answer the first one, and the answer to the second one will follow right through with similar arguments. Suppose two documents are similar, then the pattern of their term frequency vector will be similar. If we take the inner product of their term frequency vectors, we obtain a larger value than if they were dissimilar. Therefore, the ijth element of $X^t_{dt} X_{dt}$ can serve as a measure of the similarity between ith and jth documents. But

$$X_{dt}^t X_{dt} = (D_{dm} S_{mm} T_{mt})(T_{tm} S_{mm} D_{md}^t) = (D_{dm} S_{mm})(S_{mm} D_{md}^t),$$

so we can treat rows of $D_{dm}S_{mm}$ as a numeric representation for the documents when we want to compare between documents. After dimension reduction, rows of $\hat{D}_{dk}\hat{S}_{kk}$ are approximated numeric representation for the documents . A new document or query can be projected into the LSI document vector space by $\hat{d}^* = d^*\hat{T}_{tm}\hat{S}_{mm}^{-1}$ where d^* is the term frequency vector of the new document or query.

Similarly, rows of $T_{tm}S_{mm}$ can serve as the numeric representation of terms when we want to compare between terms. Therefore, rows of $\hat{T}_{tk}\hat{S}_{kk}$ are approximation to the numerical representation for the terms and a new term can be projected into the LSI term vector space by $\hat{t}^* = t^*\hat{D}_{dm}\hat{S}_{mm}^{-1}$, where t^* is the document frequency vector for the new term.

Since we have numerical representations for all documents (terms), we can measure similarity between documents (terms) by some similarity measure, such as cosine of the angle between two LSI document (term) vectors.

6.2 LSI Applied to Customer Trouble Reports

For the rest of this section, we will discuss two real world LSI applications we developed at Telcordia Technologies. We first applied LSI to local loop Customer Trouble Reports. Although this application has very little to do with Web Mining directly, it suggests a methodology for handling natural language text queries (e.g. questions or complaints emailed by customers) in an automated way.

A Customer Trouble Report (CTR) in telecommunications represents a customer's complaint on their telephone service ranging from poor quality to complete outage. The goal of our project on mining Customer Trouble Reports (CTRs) was to help telecommunications companies to identify systematically possible trouble causes soon after a customer files a trouble report. The importance of correctly identifying trouble causes is two fold (1) to avoid unnecessary dispatches, and (2) to increase customer satisfaction.

A CTR database usually contains hundreds of fields to characterize a CTR. Among them, the narrative field stores information on customers' description of the problems. It is usually of limited length, in our case 40 characters long, nonetheless it carries more information than any other fields in the database. Previous studies on CTRs either ignored the narrative field, or analyzed it in an ad hoc manner. The contribution of our research lies in the systematic approach to extract information from text using LSI. With this dimension reduction technique, we are able to represent each narrative record in the CTR database by 46 numerical values. For the historic data set, we know exactly the true trouble cause for each CTR, but to be able to predict trouble cause for a new CTR, we can only use variables/predictors that are available at the time when a CTR comes in. The predictors we selected are the 46 LSI dimensions, class of service and outage status at time of report. A multinomial logistic regression model is fitted to the historic data to derive estimates for the model parameters. When a new CTR arrives, it is first projected into the 46 dimensional LSI document vector space. Given the values of the predictors (46 LSI dimensions, class of service and outage status), we calculate the probabilities of the trouble causes for this specific CTR based on the fitted multinomial logistic regression model. Naturally, the trouble cause with largest estimated probability, referred to later as the ***target probability***, is identified as the potential trouble cause and appropriate resources will be allocated to fix the problem. But if the target probability is not discriminating (there are 36 trouble causes for the CTR database we analyzed), it is of little use in making clear-cut decisions. Fortunately, this is not the case in CTR mining. We first randomly took 90% of the historical database as training set and the rest as test set. We then applied the method described above to the training set and verified results on the test set. Our methodology achieves low classification error rates -- around 8% -- for both training and test sets. Also, among the correctly classified CTRs, more than 90% of them have target probabilities greater than 0.9. Therefore, it is very easy to identify a single trouble cause for a given

CTR. Interestingly enough, for the 8% that were incorrectly classified, the methodology gave low target probabilities to warn us that additional processing or manual intervention might be necessary.

To show how the usage of text data has improved the result, we also performed a similar classification but without using the narrative field. Let us refer to our methodology as the complete classifier and the one without using narrative field as the simple classifier. The classification error rates from the simple classifier were more than 55% for both training and test sets. For all the cases including those that were correctly and incorrectly classified, the target probabilities were centered at 0.4 with a second mode centered at 0.7. Again, a good classifier should estimate/predict high target probabilities when it correctly classifies and low target probabilities when it incorrectly classifies. The simple classifier does neither.

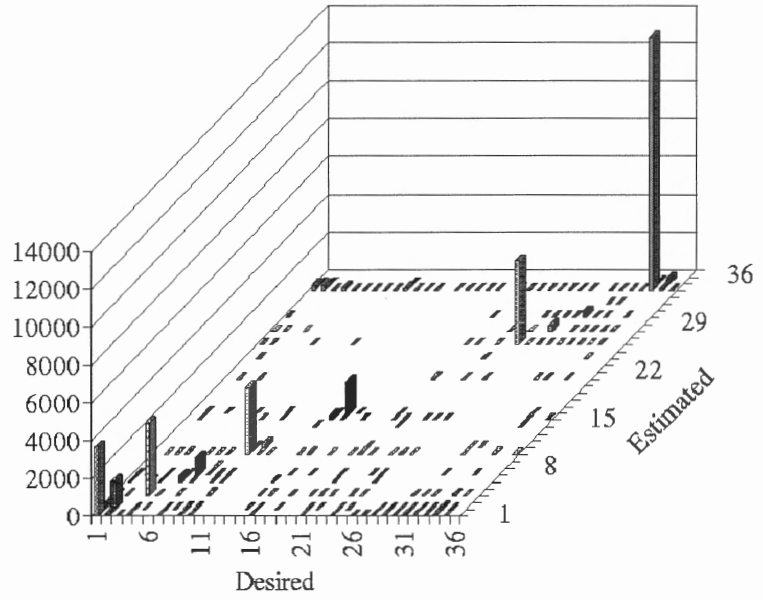

Figure 4. Confusion matrix from the complete classifier – training set

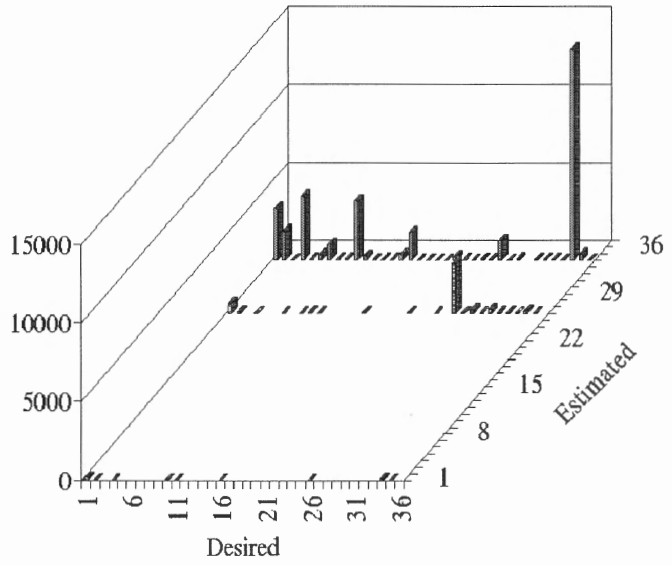

Figure 5. Confusion matrix from the simple classifier – training set

Figure 4 (Figure 5) gives a visualization of the confusion matrix from the complete (simple) classifier –A perfect classifier has a diagonal confusion matrix where the desired is exactly the same as the estimated. In

practice, this is not possible. However, we do have a very close to diagonal confusion matrix from the complete classifier, whereas the confusion matrix from the simple classifier is far from perfect. In figure 6 (Figure 7), we also show the histogram of the target probabilities for those CTRs that were correctly classified by the complete (simple) classifiers. Ideally, the target probabilities should be as close to unity as possible. The complete classifier again demonstrates its superiority over the simple classifier. We note that all the plots we show here are on the training set. Plots on the test set are similar to those obtained with the training set.

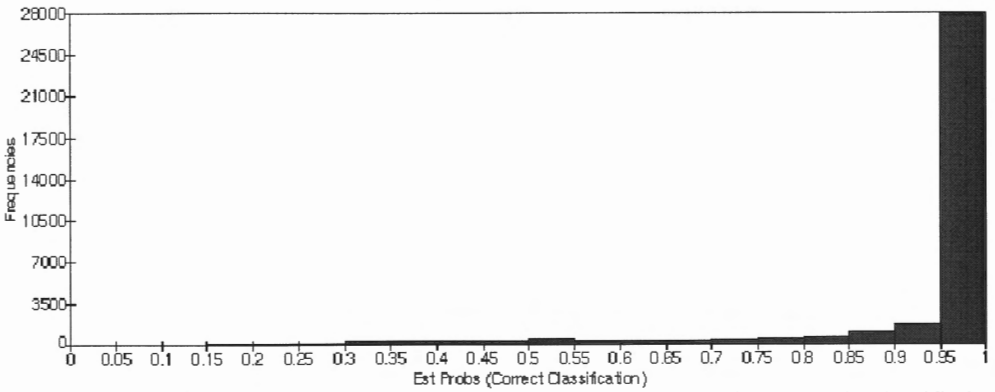

Figure 6. Target probabilities from the complete classifier for correctly classified cases

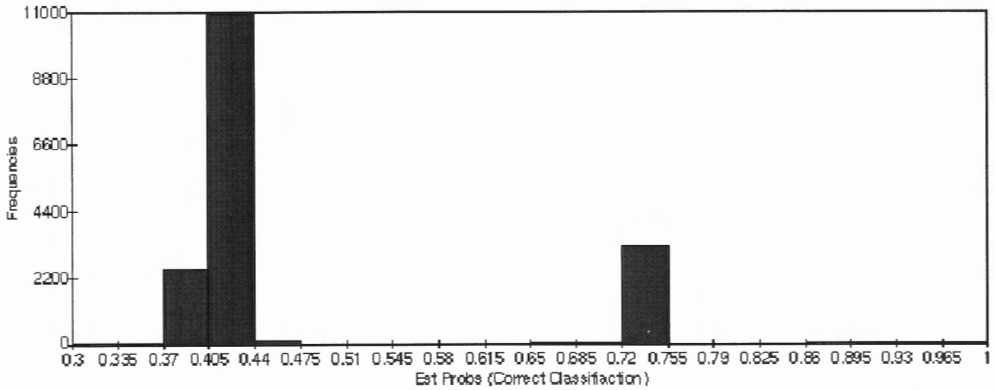

Figure 7. Target probabilities from the simple classifier for correctly classified cases

6.3 LSI Applied to Recommending Content

The second application of LSI is based on work with the Telcordia Information Superstore. In our Information Superstore we sell telecommunications documents online. We used LSI to analyze the abstracts of all documents to classify them into categories. Since items we sell online are constantly changing, we rerun LSI periodically to make sure the LSI document vector space is sufficient to describe items in our store. We have built a document recommender using the same technology. Immediately after a customer puts an item into the shopping cart, we show the top 5 documents which are most similar in LSI measure to the item they just put in the shopping cart.

Note that this form of content-based recommending is distinct from collaborative filtering in which recommendations are based on the similarities of preferences among sets of users. These two forms of recommending ultimately might be combined, if they are found to make unique contributions to good recommendations. At this point we have only preliminary results of an evaluation study, but the results

appear promising. Since the inception of this recommendation system, we boosted the monthly items sold by 9 percent and the monthly dollar total sold by 6 percent.

With the encouragement of this first success with LSI, we plan to analyze the e-mail archive from customers in an attempt to build an automatic e-mail reply system. The system will automatically analyze customers' e-mails and answer their questions. In a typical e-mail, customers usually described the kind of problems they need answered, for example " If I want to be a Local Exchange Carrier, what documents should I purchase?" In some e-mails, customers just like to know the status of their orders. By using LSI, we can first group e-mails into similar groups. For the group that asks for recommendation, we can project the content of the e-mail into the LSI document vector space and return customer with a list of relevant items in the store. For others, such as status inquiry group, we can query against the status database and inform the customer of the status of their orders. Eventually, the automatic e-mail reply system may evolve into an online customer care center where responses to customers are automated and returned in real time.

7. Emerging Challenges for E-Commerce Data Mining

7.1 Enhancement of Session Data

This paper has focused on measurement issues related to a user's interactions with an e-commerce site in the course of a single user session, or perhaps multiple sessions at the same site. In the future, data from sessions of this kind can be enhanced in several useful ways. Here we discuss three kinds of enhancements: tracking across contexts, enhancing with third-party data sources, and integrating Web-based data with legacy data sources.

Tracking across contexts refers to the capability to record users' interactions across different Web sites and sessions. In the limit, one could imagine compiling a complete history of a user's experience on the Web, essentially a concatenation of the log data from each site visited. Advertisers could use data of this kind to schedule ads far more effectively than the scheduling possible today. Not only could advertisers limit the number of exposures of a given message, but they could also schedule and space those messages in optimal ways to promote retention according to principles of human learning and memory (see Landauer & Bjork, 1978). Tracking across contexts might also support very sophisticated kinds of affiliate marketing and detailed data mining concerning the effectiveness of advertised messages. An analysis could begin with an exposure to an advertisement, then track the clickthrough on a second, related ad on a subsequent day on a different site, then note that the user gathered information about the advertised product, and finally that the user purchased the product after checking several sites for the best price.

Data from a Web session also might be enhanced by third-party data sources. Direct marketing companies are able to create a detailed picture of a consumer, beginning with some known information, and then enhancing that record with data taken from the census, various geographically-oriented data sources, credit references, etc. One such company claims to be able to generate up to 1500 data attributes for most US citizens. Presently these data are used primarily to generate mailing lists. In the future, it may be possible to take a single known piece of information about a Web user, say an email address, and expand that information to a rather complete picture. This expanded picture of a customer would enable e-commerce sites to create highly targeted marketing, sales, and customer service experiences, including upselling, cross selling, and product bundling opportunities tailored to the individual.

A third kind of enhancement to Web session data would integrate Web data with other data available about the same user from legacy sources. Consider a large bank, retailer, or telephone company. Each of these companies has several channels with which they interact with a given customer. A customer might walk into a company building and interact with a service representative or sales person. A customer might make a phone call to order a new product, or ask for information or some other assistance. The same customer might use a company credit card or sponsored bankcard. And now the same customer can interact with the company's Web site to search for information, order products and services, or seek assistance. There is much to be gained by having a complete picture of the customer available to support all of these customer

interactions. Indeed, the multiple channels for dealing with a customer should give traditional companies that open a Web channel (the "clicks and mortar" companies) a significant advantage over the "pure play" companies that operate only on the Web.

These opportunities to enhance user session data entail certain technical challenges. One set of challenges are rather familiar issues with the entire service chain of mining customer data. As discussed in Sections 3-5, issues concerning data acquisition, validation, warehousing, structuring, analysis, and mining have some unique twists when Web data are considered. Integrating Web and other sources of customer data obtained across multiple channels in a large business can be a nontrivial technical and organizational problem. Integrating third-party sources of data, an activity that is already part of established data mining practice, has some special difficulties when the data are integrated with Web transaction data. These challenges, while requiring detailed understanding of Web data and clever applications, probably do not require major technical breakthroughs to solve.

The areas of real time data mining and cross context tracking do represent challenges that may require breakthrough technologies. Implicit in many of the stated advantages to enhancing session data is the possibility that the data can be mined in real time or "close" to real time. Ideally, enhancement of a Web session's data with legacy data sources, third-party sources, or cross-context sources would be available instantly to support real time decisions concerning marketing, sales, and customer service. At this point, the technology to support real time updates, querying, and retrieval from large remote databases on the scale suggested here is not yet technically feasible.

The ability to track users and collect and reconcile data obtained across many Web contexts is also extremely limited. As noted previously, "cookies" provide one crude mechanism for the purpose, but cookies—which can identify a given browser instance but not the person using the browser—can provide misleading information and can be easily defeated. It is likely that the issue of tracking users across contexts will become even more technically challenging as the number of Web-enabled devices expands so that a given user might have a laptop for travel, one PC at home, another at the office, and a wireless microbrowser for mobility support.

7.2 Privacy

Just describing the kind of e-commerce data mining that is possible today, and the some possible future enhancements to session data quite naturally raises serious issues regarding the privacy of users. Web users in the US consistently register high levels of general concern about privacy while online. Westin (1998) found that 81% of users were concerned about threats to privacy while online, while Ackerman, Cranor, and Reagle (1999) found that only 13% of their survey respondents were "not very" or "not at all" concerned about online privacy. The latter study also probed respondents about the types of personal data they would feel comfortable providing online. Respondents were "always" or "usually" comfortable in providing preference information such as favorite TV show (82%) and favorite snack (80%), but far less so in providing medical information (18%), income (17%), phone number (11%), credit card number (3%) and social security number (1%).

Privacy protection mechanisms have progressed on two fronts so far. One is an effort to standardize the types of personal information and transmission of that information between a user and a Web service. The World Wide Web Consortium's Platform for Privacy Preferences (P3P) is defining a way to communicate personal data in machine readable form conforming to an XML standard (see Reagle & Cranor, 1999). The standard would allow users to indicate privacy preferences, and Web services to indicate which information would be collected, how it would be used, whether it would be shared, how it would be stored, etc. A second approach is essentially self-regulation. Web sites post their privacy policies (as a majority of e-commerce sites do now), and may subscribe to a service such as TRUSTe or BBBOnline to certify that the policy meets certain specifications. The privacy policies may be monitored and verified on an ongoing basis by the service providing certification.

It should be noted that the enhancements to session data described above would require a very thorough analysis and treatment of privacy issues before being implemented. For example, Ackerman, Cranor, and Reagle (1999) found that only 44% of their survey respondents said that they would definitely or probably agree to using a persistent identifier that would enable customized advertising across many Web sites.

7.3 Data Ownership and Distribution

The creation of a technology or network service that satisfies many privacy concerns and allows users to control the use of personal data represents a significant technical challenge as well as a possible new business opportunity. Surveys have shown that users wish to control access to personal data, and determine how and by whom it is used. There are legitimate reasons why a user might not want to share data such as a mailing address or phone number, even though that same user has a great incentive to make an online purchase.

Consider the following scenario: A person who is housebound has a prescription for purchasing a drug. The person likes the convenience of purchasing the drug at an online pharmacy and wishes to have the drug delivered. For reasons of privacy, the person does not want the purchase to be traceable to a particular address, phone number, or credit card. This is just one of a large number of scenarios in which people would choose the convenience of shopping on the Web if their privacy could be protected. This particular scenario is complicated by the fact that the person with the prescription also must first be authenticated before purchasing the prescription drug online. It is possible to imagine many other scenarios, including some business-to-business transactions that would have the same kind of requirements.

Components of the technologies to support the ownership of personal data by users and the releasing of that data while maintaining strict privacy exist today. We have already discussed the P3P specification that goes part of the way to resolving privacy concerns. Encryption technologies can help safeguard the transmission of information. Beyond these, "anonymizers" and "psuedonymizers" may play an important role. The simplest of these technologies removes all means of tracing a transmission back to a particular user. A more advanced application of these technologies might create an intermediary allowing users to establish one or more Web personae with which to conduct e-commerce.

The value proposition for any technology or service that would give the user the kind of control over personal data described in the scenario above would have to be examined closely. Would enough users find such technology or service valuable enough to purchase software or subscribe to the service? As increasing amounts of personal data including medical records, preferences, and financial data become necessary for e-commerce, the value to users of a mechanism to hold personal data and distribute it in a private fashion will increase greatly.

A technology or service that enables scenarios like the one presented would alter dramatically the "balance of power" in favor of the Web consumer. Currently, online retailers are trying to acquire as much customer information as possible to "capture customers" and promote "stickiness." The rationale is that if a specific retailer can get customers to share a lot of personal data, the customers should find it so easy to shop with that retailer that they would not shop anywhere else. If personal data were owned by users and could be distributed easily by users, the situation would change. Online retailers would compete, not on the basis of who owns the most customer data, but rather on the basis of what is done with customer data to provide the best customer service, selection, and price.

8. References

Ackerman, M.S., Cranor, L.F. and Joseph Reagle (1999). Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences. In *Proceedings of the ACM Conference on Electronic Commerce*, (Denver, CO, November 3-5, 1999), ACM, New York, pp. 1-8.

Bura, A., Pinnell, J., Shutter, J. & Andrew B. Whinston (1999). Measuring the Internet Economy: An Exploratory Study. Center for Research in Electronic Commerce, The Graduate School of Business, University of Texas at Austin.

Landauer, T.K. & Robert A. Bjork (1978). Optimal rehearsal patterns and name learning. In M.M. Gruneberg, P.E. Morris, and R.M. Sykes (Eds.) *Practical aspects of memory*. London: Academic Press, 1978, 625-632.

Reagle, Joseph and Lorrie Faith Cranor (1999). The platform for privacy preferences. *Communications of the ACM* 42(2): 48-55.

Westin, Alan F. (1998). *E-commerce & Privacy: What Net Users Want*. Hackensack, NJ: Privacy & American Business.