Path Breaking Case Studies in E-commerce using Data Mining

Rupesh Sanchati, P.C. Patidar, Gaurav Kulkarni

Abstract— The e-commerce domain can provide all the right ingredients for successful data mining and claim that it is a killer domain for data mining. The architecture of various e-commerce sites has supported data collection, transformation, and data mining since its inception. With click-streams being collected at the application-server layer, high-level events being logged, and data automatically transformed into a data warehouse using meta-data, common problems plaguing data mining using weblogs (e.g., sessionization and conflating multi-sourced data) are obviated, thus allowing one to concentrate on actual data mining goals. The paper briefly reviews the architecture of integrated E-Commerce with Data Mining, discusses some case studies and puts forward some conclusions inferred from the same. While the conclusions are drawn from the case studies from the retail e-commerce domain, they are also equally applicable to other data mining domains, as well.

Index Terms— Web analytics, retail e-commerce, Simpson’s paradox, Timeout Analysis, bot analysis.

I. INTRODUCTION

The advent of e-commerce revolutionized every industry. Every aspect of commerce, from sales pitch to final delivery, could be automated and made available 24 hours a day, all over the world. B2B solutions carried this one step further, allowing vertical partnerships and co-branding. Businesses found a new incentive to bring their data into the digital age. And dynamic content allowed the first truly personalized, interactive websites to come into being, all through the magic of e-commerce. Meanwhile, away from the digital storefront, data warehouses were springing up in the machine rooms of industry - gargantuan information repositories for the collection of every bit of business trivia. The sources of data were of the old realm of business: point-of-sale terminals, inventory databases, transaction records. Attempts to understand the data, first with statistical tools, later with OLAP systems, met with limited success - until the introduction of Data Mining. Using machine learning algorithms, Data Mining software finds hidden patterns in the data, and uses them to form new rules and predict the future behavior of customers - turning that mountain of data into valuable knowledge and untapped business opportunities.

Using a website as a data collection tool is now commonplace, because of its interactivity, simplicity, and unobtrusiveness. So naturally one would want to analyze this data with the best data mining techniques available. The results of the data mining - the rules which say which customers are likely to buy what products at the same time, or who is about to switch to your competitor - would ideally then be integrated into your dynamic website - thus providing an automated, end-to-end, targeted marketing and e-CRM tool.

II. LITERATURE REVIEW

A. Data Mining and retail e-commerce-

Not withstanding several notable successes, data mining projects remain in the realm of research: high potential reward, accompanied by high risk. The risk stems from several sources. It has been reported by many researchers and has been our experience, that the data mining or algorithmic modeling phase of the knowledge discovery process occupies at most 20% of the effort in a data mining project. Unfortunately, the other 80% contains several substantial hurdles that without heroic effort may block the successful completion of the project.

So, why is e-commerce different? In short, many of the hurdles are significantly lower. As compared to ancient or shielded legacy systems, data collection can be controlled to a larger extent. We now have the opportunity to design systems that collect data for the purposes of data mining, rather than having to struggle with translating and mining data collected for other purposes. Data are collected electronically, rather than manually, so less noise is introduced from manual processing. E-commerce data are rich, containing information on prior purchase activity and detailed demographic data. In addition, some data that previously were very difficult to collect now are accessible easily. For example, e-commerce systems can record the actions of customers in the virtual “store,” including what they look at, what they put into their shopping cart and do not buy, and so on. Previously, in order to obtain such data companies had to trail customers (in person), surreptitiously recording their activities, or had to undertake complicated analyses of in-store videos. It was not cost-effective to collect such data in bulk, and correlating them with individual customers is practically impossible. For e-commerce systems massive Amounts of data can be collected inexpensively.

Unlike many data mining applications, the vehicle for capitalizing on the results of mining—the system—already is
B. Integrating E-Commerce and Data Mining: Architecture

In this section we give a high level overview of architecture for an e-commerce system with integrated data mining. In this architecture there are three main components, Business Data Definition, Customer Interaction, and Analysis. Connecting these components are three data transfer bridges, Stage Data, Build Data Warehouse, and Deploy Results. The relationship between the components and the data transfer bridges is illustrated in Figure 1.

Figure: 1 Architecture of Integrated Data Mining with E-Commerce

In the Business Data Definition component the e-commerce business user defines the data and metadata associated with their business. This data includes merchandising information (e.g., products, assortments, and price lists), content information (e.g., web page templates, articles, images, and multimedia) and business rules (e.g., personalized content rules, promotion rules, and rules for cross-sells and up-sells). From a data mining perspective the key to the Business Data Definition component is the ability to define a rich set of attributes (metadata) for any type of data.

The Customer Interaction component provides the interface between customers and the e-commerce business. This interaction could take place through a web site (e.g., a marketing site or a web store), customer service (via telephony or email), wireless application, or even a bricks-and-mortar point of sale system. For effective analysis of all of these data sources, a data collector needs to be an integrated part of the Customer Interaction component. To provide maximum utility, the data collector should not only log sale transactions, but it should also log other types of customer interactions, such as web page views for a web site.

The Analysis component provides an integrated environment for decision support utilizing data transformations, reporting, data mining algorithms, visualization, and OLAP tools. The richness of the available metadata gives the Analysis component significant advantages over horizontal decision support tools, in both power and ease-of-use.

The Stage Data bridge connects the Business Data Definition component to the Customer Interaction component. This bridge transfers (or stages) the data and metadata into the Customer Interaction component. Having a staging process has several advantages, including the ability to test changes before having them implemented in production, allowing for changes in the data formats and replication between the two components for efficiency, and enabling e-commerce businesses to have zero down-time.

The Build Data Warehouse bridge links the Customer Interaction component with the Analysis component. This bridge transfers the data collected within the Customer Interaction component to the Analysis component and builds a data warehouse for analysis purposes. The Build Data Warehouse bridge also transfers all of the business data defined within the Business Data Definition component (which was transferred to the Customer Interaction component using the Stage Data bridge).

The last bridge, Deploy Results, is the key to “closing the loop” and making analytical results actionable. It provides the ability to transfer models, scores, results and new attributes constructed using data transformations back into the Business Data Definition and Customer Interaction components for use in business rules for personalization.

III. PROPOSED FRAMEWORK

After analyzing various retail e-commerce sites, we propose some analyses that would be useful in practice. In each of the following subsections we describe the lessons learned from path breaking case studies.

Case 1: Bot Analysis

Web robots, spiders, crawlers, and aggregators, which we collectively call bots, are automated programs that create traffic to websites. Bots include search engines, such as Google, web monitoring software, such as Keynote and Gomez, and shopping comparison agents, such as mySimon. Because such bots crawl sites and may bring in additional
human traffic through referrals, it is not a good idea for websites to block them from accessing the site. In addition to these “good bots,” there are e-mail harvesters, which try to look for e-mails that are sold as e-mail lists, offline browsers (e.g., Internet Explorer has such an option), and many experimental bots by students and companies trying out new ideas.

2. Just because the traffic is increasing immediately after registering with search engines, one should not get overwhelmed, because substantial part of that might be bot traffic.

3. Many commercial web analytic packages include basic bot detection through a list of known bots, identified by their user agent or IP. However, such lists must be updated regularly to keep track of new evolving and mutating bots.

Case 2: Session Timeout Analysis

Enhancing the user browsing experience is an important goal for website developers. One hindrance to a smooth browsing experience is the occurrence of a session timeout. A user session is determined by the application logic to have timed out (ended) after a certain predefined period of inactivity.

Figure 2 shows the impact of different session timeout thresholds set at 10-minute intervals on two large clients.

**Observations:**
1. Both account for 5 to 40% of sessions. Due to the volume and type of traffic that they generate, bots can dramatically skew site statistics.
2. Even when the human traffic is fluctuating substantially, the bot traffic still remains the same.
3. After registering with search engine the external bot traffic increases substantially, as expected.

**Lesson:**
1. Accurately identifying bots and eliminating them before performing any type of analysis on the website is critical.
shopping cart

Lesson:

1. The software save the shopping cart automatically at timeout and restore it when the visitor returns.
2. Clients must determine the timeout threshold only after careful analysis of their own data.
3. Setting the session timeout threshold too high would mean that fewer users would experience timeout thereby improving the user experience.
4. A larger number of sessions would have to be kept active (in memory) at the website thereby resulting in a higher load on the website system resources.
5. Setting an appropriate session timeout threshold involves a trade-off between website memory utilization (which may impact performance) and user experience. So maintain a right balance.

Case 3: Simpson’s paradox

On a few occasions it becomes difficult to present insights that are seemingly counter-intuitive. For instance, when analyzing a client’s data we came across an example of Simpson’s paradox (Simpson, 1951). Simpson’s paradox occurs when the correlation between two variables is reversed when a third variable is controlled.

We were comparing customers with at least two purchases and looking at their channel preferences, i.e., where they made purchases. Do people who shop from the web only spend more on average as compared to people who shop from more than one channel, such as the web and physical retail stores.

Observations:

1. The line chart in Figure 3 shows that for each group of shoppers who shopped once, twice, three times, four times, five times, and more than five times respectively, the average spending per customer on the web-only channel is more than the average spending per customer on multiple channels.
2. However, the bar chart in Figure 3 shows that the average spending per customer for multi-channel customers is more than that of the web-only channel.

Figure 3: Average yearly spending per customer for multi-channel and web-only purchasers by number of purchases (left), and average yearly spending per customer for multi-channel and web-only purchasers (right).

Lesson:

1. Explain counter-intuitive insights - The reversal of the trend in the above case is happening because a weighted average is being computed and the number of customers who shopped more than five times on the web is much smaller than the number of customers who shopped more than five times across multiple channels. Such insights must be explained to business users.

Figure 4: Clarification of Simpson’s paradox

Case 4: Search Effectiveness Analysis

Significant time and effort is spent in designing forms that are aesthetically pleasing. The eventual use of the collected form data for the purpose of data mining must also be kept in mind when designing forms.

Observation:

1. On the basis of average sales per visit, it can be said that Customers that search are worth two times as much as customers that do not search.
2. Failed searches hurt sales severely.
shown in Figure 3 that orders seem to “follow” visits by five hours. It turned out different servers were being used to log clickstream (visits) and transactions (orders), and these servers’ system clocks were off by five hours. One was set to GMT and the other to EST.

IV. CONCLUSION AND FUTURE WORK

We reviewed the integrated architecture of Data Mining with E-Commerce, which provides powerful capabilities to collect additional click stream data not usually available in web logs, while also obviating the need to solve problems usually bottlenecking analysis (and which are much less accurate when done as an afterthought), such as sessionization and conflating data from multiple sources. We believe that such architectures where click streams are logged by the application server layer are significantly superior and have proven themselves with various E-commerce sites.

Our focus on Business to Consumer (B2C) e-commerce for retailers allowed us to drill deeper into business needs to develop the required expertise and design out-of-the-box reports and analyses in this domain. Further, we believe that most lessons will generalize to other domains outside of retail e-commerce.

The top 3 lessons are:
1. Accurately identifying bots and eliminating them before performing any type of analysis on the website is critical.
2. Setting an appropriate session timeout threshold involves a trade-off between website memory utilization (which may impact performance) and user experience. So maintain a right balance.
3. Counter-intuitive insights must be explained to business users in depth.

E-commerce is still in its infancy, with less than a decade of experience. Best practices and important lessons are being learned every day. The Science of Shopping is well developed for bricks and mortar stores. Although the techniques show very promising results, the topic is still in its infancy. Surviving this topic, we listed some challenges that constitute promising research directions.

REFERENCES


[5] Blue Martini Software. (2003a). Blue Martini Business Intelligence at Work: Charting the Terrains of MEC Website Data

Lesson:
1. Design forms with data mining in mind.
2. Create custom pages for often searched keywords.
3. Do not allow empty search.

Case 5: Data Auditing

Data cleansing is a crucial prerequisite to any form of data analysis. Even when most of the data are collected electronically, as in the case of e-commerce, there can be serious data quality issues.

Consider following graph that shows distribution of visits and orders by hour-of-day for a real website.

Observation:

The above graph shows an interesting pattern in visit and order. Orders follow visits by five hours, while we are expecting visits and orders to be close to each other in time.

Lesson:
1. Data cleansing is a crucial prerequisite to any form of data analysis.
2. We’ve found serious data quality issues in data warehouses that should contain clean data, especially when the data were collected from multiple channels, archaic point-of-sale systems, and old mainframes. As

Figure 5: Effectiveness of search.

Figure 6: Distribution of visits and orders by hour-of-day.