E-COMMERCE AND DATA MINING: INGREDIENTS AND CHALLENGES

Mr. Giri Prasad Madderla

Dr. Rajender Katla

Research Scholar, Department of Commerce & Business Management Kakatiya University, Warangal,TS.

Assistant Professor, Department of Commerce & Business Management, Kakatiya University, Warangal, TS.

Abstract

For successful data mining, several ingredients are needed and e-commerce provides all the right ones. Organizations conducting Electronic Commerce (e-commerce) can greatly benefit from the insight that data mining of transactional and click stream data provides. Many of the problems of dealing with web server log data can be resolved by properly architecting the commerce sites to generate data needed for mining. Even with a good architecture, however, there are challenging problems that remain hard to solve. The integrated e-commerce architecture can dramatically reduce the pre-processing, cleaning, and data understanding effort often documented to take 80% of the time in knowledge discovery projects. We detail the mining workbench with ingredients. We conclude with a set of challenges.

Keywords: E-commerce, Data mining, Web server, Customer interaction, Click stream

INTRODUCTION:

E-commerce is growing fast, and with this growth companies are willing to spend more on improving the online experience. The online business to consumer retail spending is increased which is showed in 1999 was $20.3 billion and estimated to grow to $144 billion by 2003 it is anticipated that it will reach to $500 by 2011. Global 500 companies will spend 85% more on e-commerce in 2010 than what they did decade back. Most of the existing sites are using primitive measures, such as page views, but the need for more serious analysis and personalization is growing quickly with the need to differentiate. In Measuring Web Success, the web analysts claim that "Leaders will use metrics to fuel personalization" and that "firms need web intelligence, not log analysis." Data mining tools aid the discovery of patterns in data. Until recently, companies that have concentrated on building horizontal data mining modeling tools, have had little commercial success. Data mining is fundamentally an applied field, driven more by a class of problems (e.g., classification, clustering, etc.) than by a specific set of methods. Nonetheless, most published work in the field focuses almost exclusively on data mining methods and algorithms. Most of the people assume that E-commerce is the killer-domain for data mining. It is ideal because many of the ingredients required for successful data mining are easily satisfied. Data mining technologies have been around for decades, without moving significantly beyond the domain of computer scientists, statisticians and hard-core business analysts. Data mining projects remain in the realm of research high potential reward, accompanied by high risk. Normally data mining have to take care of following elements which are normally found in e-commerce application too(Ron and Foster).

- Data with rich descriptions
- A large volume of data
- Controlled and reliable data collection
- The ability to evaluate results
- Ease of integration with existing process

Hence we can integrate the e-commerce along with data mining to get the maximum out of an available legacy system. It is also useful because of data records are plentiful, electronic collection provides reliable data, insight can easily be turned into action, and return on investment can be measured. To take advantage of this data mining must be integrated into the e-commerce systems with the appropriate data transformation bridges from the transaction processing system to the data warehouse and vice versa. Such integration can dramatically reduce the data preparation time, known to take about 80% of the time to complete an analysis. An integrated solution can also provide users with a uniform user interface and seamless access to metadata (Ansari, Ron, Mason et al).

At first we understand the business model of a traditional e-commerce. Most e-commerce start-ups had simple business models rooted in traditional brick sand-mortar perspectives. An archetypical change was the customers have to place orders over the Internet, or to sell advertisement space during web-browsing sessions. But emergent complexity soon began to dominate the evolution, and business biodiversity blossomed. In case of Amazon, eBay and other similar companies, the initial plan was to reduce costs by cutting out storefronts. The merchandising was displayed in the form digital catalog and the customers would need to place orders over the Internet and receive deliveries through couriers or post. On the other hand the web browser companies saw initially it as one of the income source through maintaing customer books and from advertisements. But sonly they found it as much more gain proportional business model then currently they practicing. Then they began R&D on the proposed business models which evolved new business opportunities, it resulted into applying data mining algorithms to recommend books or hotels to customers, or to target advertising more effectively, based upon activity histories for efficient e-commerce activities.

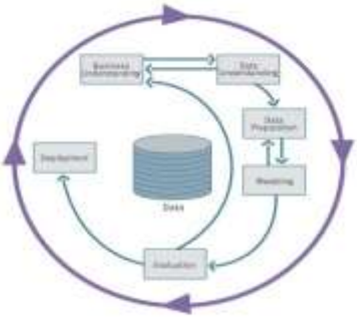

Fig1: **Ingredients of Data mining in E-commerce**

Later E-commerce has changed the face of business. It allows better customer management, new strategies for marketing, an expanded range of products, and more efficient operations. A key enabler of this change is the widespread use of increasingly sophisticated data mining tools. The Department of Commerce commissioned a study of 2003 economic data (U.S. Census Bureau, 2005). It showed that e-commerce, on a percentage basis, outperformed all four major economic sectors in 2002-2003. For the Manufacturing sector, 21.2% ($843 billion) of the total activity was classified as e-commerce. For the Merchant Wholesalers sector, e-commerce sales were 16.9% ($730 billion) of total sales; for Retail Trade it was 1.7% ($56 billion), and for Selected Service Industries it was 1% ($50 billion). The dominant component was Business-to-Business activity. These trends can only have increased since 2003 and expected to be continued in future too. Its growth was not been stopped major sector only later it resulted a growth in niche markets too. These ranged from on-line dating services to medical advice to financial advices.

With the involvement of data mining algorithms computer security became a necessary enabler, as well as Spam filters and technology such as Pay Pal to facilitate small-scale commercial purchases. This increased the confidence among the customers which resulted it to implementation of e-commerce every corner of the business like music file-sharing to online ticket booking.

INGREDIENTS

For a successful implementation of data mining in e-commerce application the following ingredients are to be considered. These will be shown in the following figure. It contains the most important component as Business Understanding, Data Understanding, Data Preparation, Modeling, evaluation and deployment. Connecting these components Deploy all expected Results.

❖ In the Business understanding phase first we need to understand all the business model and components which we need deploy. For the programmer must prepare a blue print of the proposed model and according to that he need to collect all the information related to it. This includes merchandising information (like products and price lists), content information (like web page templates, articles, images, and multimedia) and business rules (like personalized content rules, promotion rules, and rules for cross-sells and up-sells).

❖ In data understanding and data preparation the e-commerce business user must define the data and metadata associated with the business and we need to define components ability to define a rich set of attributes (metadata) for any type of data. For example, products can have attributes like size, color, and targeted age group, and can be arranged in a hierarchy representing categories like men's and women's, and subcategories like shoes and shirts. As another example, web page templates can have attributes indicating whether they show products, search results, or are used as part of the checkout process. Having a diverse set of available attributes is not only essential for data mining, but also for personalizing the customer experience.

❖ Later in the modeling phase we need to consider the Customer Interaction component also. This provides the interface between customers and the e-commerce business. The term customer interaction applies more generally to any sort of interaction with customers.

This interaction could take place through a web site, customer service, wireless application, or even a bricks-and-mortar point of sale system.

❖ For evaluation phase for the 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. Further details of the data collection architecture for the specific case of a web site are described in Section

❖ Later the developed model must be deployed as per the requirement of customer which will satisfy the needs of the end user. After deployment the developer must be needed to take feedback and re-modify as according the feedback.

CLICK STREAM LOGGING

“Stemming" refers to the preprocessing step in which a keyword is reduced to a generic form. If a user typed in “neural networks" they probably would not want to exclude documents that mention only “neural nets" or “neural network". The stemming process reduces the morphological variants of a word (plurals, tenses, gerunds, prefixes, and so forth) to a common form. Stemming generally improves the quality of the search, and has the additional benefit of reducing the number terms in the dictionary most e-commerce architectures rely on web server logs or packet sniffers as a source for click stream data. While both these systems have the advantage of being non-intrusive, allowing them to “bolt on” to any ecommerce application, they fall short in logging high level events and lack the capability to exploit metadata available in the application.

A typical web log contains data such as the page requested, time of request, client HTTP address, etc., for each web server request. For each page that is requested from the web server, there are a huge number of requests for images and other content on the page. Since all of these are recorded in the web server logs, most of the data in the logs relates to requests for image files that are mostly useless for analysis and are commonly filtered out. All these requests need to be purged from the web logs before they can be used. Because of the stateless nature of HTTP, each request in a web log appears independent of other requests, so it becomes extremely difficult to identify users and user sessions from this data.

Since the web logs only contain the name of the page that was requested, these page names have to be mapped to the content, products, etc., on the page. This problem is further compounded by the introduction of dynamic content where the same page can be used to display different content for each user. In this case, details of the content displayed on a web page may not even be captured in the web log. The mechanism used to send request data to the server also affects the information in the web logs. If the browser sends a request using the “POST” method, then the input parameters for this request are not recorded in the web log.

The click stream data collected from the application server is rich and interesting; however, significant insight can be gained by looking at subsets of requests as one logical event or episode. These are also known as business events. Some interesting business events that help with the analysis given above and are supported by the architecture are

- Add/Remove item to/from shopping cart
- Initiate checkout
- Finish checkout
- Search event
- Register event

The search keywords and the number of results for each of these searches that can be logged with the search events give marketers significant insight into the interests of their visitors and the effectiveness of the search mechanism.

CHALLENGES

In this section we describe several challenging problems based on our experiences in mining ecommerce data. The complexity and granularity of these problems differ, but each represents a real-life area where we believe improvements can be made. Except for the first two challenges, the problems deal with data mining algorithmic challenges.

1. Building systems that work in real applications: Although many instances have been demonstrated of specific data mining and data warehousing systems applied to specific data, many of these applications do not scale to complexity of the task. For example, several data mining algorithms have been applied to specific components of available bioinformatics data, but no integrated system has been developed allowing management and analysis of a significant subset of bioinformatics databases. A similar case can be made for other tasks, e.g., video collections of the BBC and CNN, analysis of information sources for financial institutions, Web click-stream data and associations between medical and social databases, integration of geo-spatial databases, and other E-applications.

2. Building systems that real people can use: Most data mining systems are operated by computer scientists specially trained in the domain of interest, or domain scientists specially trained in the operation of the data mining system Determining which data mining methods to use with a given problem and application. Numerous successful and unsuccessful applications of data mining methods have provided important data on the applicability of various methods to different tasks.

3. Non-traditional data mining and data warehousing tasks: With the advent of numerous, heterogeneous sources of structured, semi-structured and unstructured data, new data mining and data warehousing methods are necessary to integrate this data to support the extraction of knowledge relating different information sources.

4. Privacy Issues: The most important issue that should be encountered during the user profiling process is privacy violation. Many users are reluctant to giving away personal information either implicitly as mentioned before, or explicitly, being hesitant to visit Web sites that use cookies or avoiding disclosing personal data in registration forms. In both cases, the user looses anonymity and is aware that all of their actions will be recorded and used, in many cases without their consent. Additionally, even if a user has agreed to supply personal information to a site, through cookie technology such information can be exchanged between sites, resulting to its disclosure without the user's permission.

5. Tools & Applications: In this section we present some of the most popular Web sites that use methods such as decision tree guides, collaborative filtering and cookies in order to profile users and create customized Web pages.

6. Support and Model External Events: External events, such as marketing campaigns, and site redesigns change patterns in the data. The challenge is to be able to model such events, which create new patterns that spike and decay over time.

7. Support Slowly Changing Dimensions: Visitors' demographics change: people get married, their children grow, their salaries change, etc. With these changes, their needs, which are being modeled, change. Product attributes change: new choices may be available; packaging material or design change, and even quality may improve or degrade.

8. Identify Bots and Crawlers: Bots and crawlers can dramatically change click stream patterns at a web site. For example, Keynote provides site performance measurements. The Keynote about can generate a request multiple times a minute, 24 hours a day, 7 days a week, skewing the statistics about the number of sessions, page hits, and exit pages (last page at each session). Search engines conduct breadth first scans of the site, generating many requests in short duration. Internet Explorer 5.0 supports automatic synchronization of web pages when a user logs in, when the computer is idle, or on a specified schedule; it also supports offline browsing, which loads pages to a specified depth from a given page. These options create additional click streams and patterns. Identifying such bots to filter their click streams is a non-trivial task, especially for bots that pretend to be real users.

CONCLUSIONS

As we mentioned earlier that the e-commerce and data mining are the perfect blend of the successful business. The above ingredient of integration effectively solves several major problems associated with horizontal data mining tools including the enormous effort required in preprocessing of the data before it can be used for mining, and making the results of mining actionable. The tight integration between the three components of the architecture allows for automated construction of a data warehouse within the evolution component. The shared metadata across the three components further simplifies this construction, and, coupled with the rich set of mining algorithms and analysis tools (like visualization, reporting and OLAP) also increases the efficiency of the knowledge discovery process. The tight integration and shared metadata also make it easy to deploy results, effectively closing the loop. Finally we presented several challenging problems that need to be addressed for further enhancement of this architecture.

This paper discussed the experiences data mining in an industrial setting. A number of issues were raised if we solve them it will become a prototype model for the next generation e-commerce.

REFERENCES

1. Eric Schmitt, Harly Manning, Yolanda Paul and Sadaf Roshan, *Commerce Software Takes off, Forrester Report,* March 2000.

2. Ralph Kimball, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses,* John Wiley & Sons, 1996.

3. Gary M. Weiss, *Data Mining in the Real World: Experiences, Challenges and Recommendations.*

4. Suhail Ansari, Ron Kohavi, Llew Mason and Zijian Zheng: *Integrating E-Commerce and Data Mining: Architecture and Challenges http: //robotics .Stanford. EDU/~ronnyk /WEBKDD2000/index.html.*

5. J. Pitkow, In search of reliable usage data on the WWW, *Sixth International World Wide Web Conference*, 1997.

6. Shahana Sen, Balaji Padmanabhan, Alexander Tuzhilin, Norman H. White, and Roger Stein, The identification and satisfaction of consumer analysis-driven information needs of marketers on the WWW, *European Journal of Marketing,* Vol. 32 No. 7/8 1998.

7. Stephen Gomory, Robert Hoch, Juhnyoung Lee, Mark Podlaseck, Edith Schonberg, Analysis and Visualization of Metrics for Online Merchandizing, *Proceedings of WEBKDD'99*, Springer 1999.

8. Barry Becker, Ron Kohavi, and Dan Sommerfield, Visualizing the Simple Bayesian Classifier, *KDD Workshop on Issues in the Integration of Data Mining and Data Visualization*, 1997.

9. Michael J. A. Berry and Gordon Linoff*, Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, 2000.