Method of Data Mining in E-Commerce

ZHANG Bao-zhong[1], ZHANG Dong-han[2], and HAN Ying[3]

(1.Economy and Management Collage, Hebei Polytechnic University Tangshan Hebei 063009;

2.International Business Department, Konkuk University Seoul Korea 100000;

3. Building and Architecture Collage, Hebei Polytechnic University Tangshan Hebei 063009)

Abstract An overview is given to the concepts of E-commerce, data, and data mining. The significance of data mining in E-commerce is summarized. Some suggestions are presented on customer profile data mining, recommendation systems data mining, Web personalization data mining, multimedia data mining, and buyer behavior data mining.

Key words data mining; E-commerce; implement

实施电子商务的数据采集方法

张宝忠[1]，张东汉[2]，韩 莹[3]

(1. 河北理工大学经济管理学院 河北 唐山 063009; 2. 韩国建国大学国际贸易学部 韩国 首尔;

3. 河北理工大学建筑工程学院 河北 唐山 063009)

【摘要】综合性地阐述了电子商务、数据、数据采集的概念，总结下数据采集在电子商务的重要性；并且针对电子商务在顾客资料、顾客建议、个性化要求、多媒体应用、购买者的行为等方面的数据采集问题，提出了有效的应用性。

关 键 词 数据采集； 电子商务； 实施

中图分类号 F253.9 **文献标识码** A doi:10.3969/j.issn.1001-0548.2009.z1.021

Electronic commerce is one of the most common business terms in use as we embark on the 21 century. The E-commerce means the application of advanced information technology to promote the effectiveness of the business relationship between trading partners. It is the new business style supported by advanced information technology to improve efficiency and effectiveness within the trading process, and it is a trend of business vision and enabling technology. It also can be described as a process of buying and selling products, services, and information over computer networks including the Internet.

The concept of data is any facts, numbers, or text that can be processed by a computer. At present, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes operational or transactional data such as sales, cost, inventory, payroll, and accounting, non-operational data such as industry sales, forecast data, and macro economic data, meta data-data about the data itself such as logical database design or data dictionary definitions.

The definition of data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases[1].

1 What can data mining do for E-Commerce

Data mining, in the recent years, is popularly used

Received date: 2009 ↖ 09 ↖ 15
收稿日期：2009 ↖ 09 ↖ 15

Biography：Zhang Baozhong , Konkuk University PHD student. Research interesting includ International Business and E-commerce .
作者简介：张宝忠(1979 ↖)，男，博士生，讲师，主要从事国际贸易与电子商务方面的研究.

in companies in retailing, financial management, communication and marketing organizations. It enables these companies to determine relationships among internal factors such as price, product positioning or staff skills and external factors such as economic indicators, competition and customer demographics. And it enables them to determine the impact on sales, customer satisfaction and corporate profits.

Using data mining, retailers could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

Such as, Blockbuster entertainment does data mining on its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures. WalMart pioneers huge amount of data mining to transform its supplier relationships. WalMart captures point-of-sale transactions from over 2 900 stores in 6 countries and continuously transmits this data to its massive 7.5 terabyte Teradata data warehouse. WalMart allows more than 3 500 suppliers to access data on their products and perform data analyses. These suppliers use this data to identify customer buying patterns at the store display level. They use this information to manage local store inventory and identify new merchandising opportunities.

The NBA explored a data mining application that can be used in conjunction with image recordings of basketball games. The advanced scout software analyzes the movements of players to help coaches orchestrate plays and strategies. Such as, an analysis of the play-by-play sheet of the game played between the New York Knicks and the Cleveland Cavaliers on January 6 2001 reveals that when mark price played the guard position, John Williams attempted four jump shots and made each one. Advanced Scout not only finds this pattern, but explains that it is interesting because it differs considerably from the average shooting percentage of 49.30% for the Cavaliers during that game[2].

The NBA uses the universal clock, a coach can automatically bring up the video clips showing each of the jump shots attempted by Williams with Price on the floor, without needing to comb through hours of video footage. Those clips show a very successful pick-and-roll play in which Price draws the Knick's defense and then finds Williams for an open jump shot.

2 How to implement data mining for E-Commerce

We survey articles that are very specific to data ming implementations in E-Commerce.

2.1 Implementation in customer profile data mining

Customers drive the revenues of any organization. To acquire new customers, to delight and to retain existing customers, and to predict buyer behavior will improve the availability of products and services and hence the profits. Thus the end goal of any data mining exercise in E-Commerce is to improve processes that contribute to delivering value to the end customer. Consider an on-line store like http:www.dell.com where the customers can configure a PC of their choice, place an order for the same, track its movement, as well as pay for the product and services. With the technology behind such a Web site, Dell has the opportunity to make the retail experience exceptional. At the most basic level, the information available in Web log files can illuminate what prospective customers are seeking from a site. Are they purposefully shopping or just browsing? Buying something they’re familiar with or something they know little about? Are they shopping from home, from work or from a hotel dial-up? The information available in log files is often used to determine what profiling can be dynamically processed in the background and indexed into the dynamic generation of HTML, and what performance can be expected from the servers and network to support customer service and make E-Business interaction productive.

The electric enterprises that has same business like Dell provide their customers access to details about all of the systems and configurations they have purchased, so they can incorporate the information into

their capacity planning and infrastructure integration. Back-end technology systems for the Website include sophisticated data mining tools that take care of knowledge representation of customer profiles and predictive modeling of scenarios of customer interactions[3]. For example, once a customer has purchased a certain number of servers, they are likely to need additional routers, switches, load balancers, backup devices etc. Rule-mining based systems could be used to propose such alternatives to the customers.

2.2 Implementation in recommendation systems data mining

Systems have also been developed to keep the customers automatically informed of important events of interest to them. The article by Jeng & Drissi discusses an intelligent framework called PENS that has the ability to not only notify customers of events, but also to predict events and event classes that are likely to be triggered by customers. The event notification system in PENS has the following components: event manager, event channel manager, registries and proxy manager. The event-prediction system is based on association rule-mining and clustering algorithms. The PENS system is used to actively help an E-Commerce service provider to forecast the demand of product categories better. Data mining has also been applied in detecting how customers may respond to promotional offers made by a credit card E-Commerce company. Techniques including fuzzy computing and interval computing are used to generate if-then-else rules.

Niuetal present a method to build customer profiles in E-Commerce settings, based on product hierarchy for more effective personalization. They divide each customer profile into three parts: basic profile learned from customer demographic data, preference profile learned from behavioral data and rule profile mainly referring to association rules. Based on customer profiles, the authors generate two kinds of recommendations, which are interest recommendation and association recommendation. They also propose a special data structure called profile tree for effective searching and matching.

2.3 Implementation in Web personalization data mining

Some companies can present a comprehensive overview of the personalization process based on Web usage mining. In this context, the author discusses a host of Web usage mining activities required for this process, including the preprocessing and integration of data from multiple sources, and common pattern discovery techniques that are applied to the integrated usage data. The goal of this paper is to show how pattern discovery techniques such as clustering, association rule-mining, and sequential pattern discovery, performed on Web usage data, can be leveraged effectively as an integrated part of a Web personalization system. The author observes that the log data collected automatically by the Web and application servers represent the fine grained navigational behavior of visitors. A data is captured by Web log depending on the goals of the analysis. E-Commerce data need to be transformed and aggregated at different levels of abstraction. E-Commerce data are also further classified as usage data, content data, structure data and user data. Usage data contain details of user sessions and page views. The content data in a site are the collection of objects and relationships that are conveyed to the user. For the most part, the data comprise combinations of textual material and images. The data sources used to deliver or generate data include static HTML/XML pages, images, video clips, sound files, dynamically generated page segments from scripts or other applications and collections of records from the operational database. Site content data also include semantic or structural meta-data embedded within the site or individual pages, such as descriptive keywords, document attributes, semantic tags or HTTP variables. Structure data represent the designer’s view of the content organization within the site. This organization is captured via the inter-page link-age structure among pages, as reflected through hyperlinks. Structure data also include the intra-page structure of the content represented in the arrangement of HTML or XML tags within a page. Structure data for a site are normally captured by an automatically generated site map which

represents the hyperlink structure of the site. The operational database for the site may include additional user profile information. Such data may include demographic or other identifying information on registered users, user ratings on various objects such as pages, products, or movies, past purchase or visit histories of users, as well as other explicit or implicit representations of a user's interests.

Once the data types are clear, data preparation is easily achieved by processes such as data cleansing, page view identification, use identification, session identification, the inference of missing references due to caching, and transaction identification. Ref.[4] proposes association rules, sequential and navigational patterns, and clustering approaches for personalization of transactions as well as Web pages.

2.4 Implementation in multimedia E-Commerce

Applications in virtual multimedia catalogs are highly interactive, as in E-malls selling multimedia content based products. It is difficult in such situations to estimate resource demands required for presentation of catalog contents. Hollfeld (2005) propose a method to predict presentation resource demands in interactive multimedia catalogs. The prediction is based on the results of mining the virtual mall action log file that contains information about previous user interests and browsing and buying behavior.

2.5 Implementation in buyer behavior in E-Commerce

For a successful E-Commerce site, reducing user-perceived latency is the second most important quality after good site-navigation quality. The most successful approach towards reducing user-perceived latency has been the extraction of path traversal patterns from past users access history to predict future user traversal behavior and required resources. However, this approach is suited for only non-E-Commerce sites where there is no purchase behavior. Vallamkondu & Gruenwald describe an approach to predict user behavior in E-Commerce sites. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users to predict the purchase and traversal behavior of future users.

Web sites are often used to establish a company's image, to promote and sell goods and to provide customer support. The success of a Web site affects and reflects directly the success of the company in the electronic market. Spiliopoulou & Pohle propose a methodology to improve the success of Web sites, based on the exploitation of navigation-pattern discovery. In particular, the authors present a theory, in which success on the basis of the navigation behavior of the site's users. They then exploit Web usage miner , a navigation pattern discovery miner, to study how the success of a site is reflected in the users' behavior. With WUM the authors measure the success of a site's components and obtain concrete indications of how the site should be improved.

In the context of Web mining, clustering could be used to cluster similar click-streams to determine learning behaviors in the case of E-learning, or general site access behaviors in E-Commerce. Most of the algorithms presented in the literature to deal with clustering Web sessions treat sessions as sets of visited pages within a time period and do not consider the sequence of the click-stream visitation. This has a significant consequence when comparing similarities between Web sessions. Wang & Zaiane propose an algorithm based on sequence alignment to measure similarities between Web sessions where sessions are chronologically ordered sequences of page accesses.

3 Conclusion

Data mining is largely used in several applications such as understanding consumer research marketing, product analysis, demand and supply analysis, E-Commerce, investment trend in stocks & real estates, telecommunications and so on. Data mining is based on mathematical algorithm and analytical skills to drive the desired results from the huge database collection. Data mining has great importance in today's highly competitive business environment. A new concept of business intelligence data mining has evolved now, which is widely used by leading corporate houses to stay ahead of their competitors. Business intelligence can help in providing latest information and used for competition analysis, market

research, economical trends, consume behavior, industry research, geographical information analysis and so on. Business intelligence data mining helps in decision-making.

参 考 文 献

[1] RALPH Y K. The data warehouse toolkit: practical techniques for building dimensional data warehouses[M]. New York: John Wiley & Sons, 2004: 93-103.
[2] RALPH Y K, LAURA H R, MARGY J R, et al. The data warehouse lifecycle toolkit: expert methods for de-signing, developing, and deploying data warehouses[M]. New York: J John Wiley & Sons, 2002: 302-313.
[3] ROBERT R C, BAMSHAD T M, JAIDEEP S S. Data preparation for mining world wide Web browsing patterns[J]. Knowledge and Information Systems, 1999: 302-313.
[4] GREGORY P S, RON Y B, TOM J K, et al. An overview of issues in developing industrial data mining and knowledge discovery applications[C]//Proceeding of the Second International Conference on Knowledge Discovery and Data Mining, [S.l.]: [s.n.],1996.

编 辑 黄 莘

(上接第103页)

参 考 文 献

[1] 徐 涛. PKI 在电子商务中的应用[J]. 商场现代化, 2007, 8: 162-163.
XU Tao. PKI application in E-commerce[J]. Market Modernization, 2007, 8: 162-163.
[2] ANDREW N, WILLIAM D, CELIA J, et al. Implementing and managing m-security[M]. [S. l.]: McGraw-Hill Osborne Media, 2002.
[3] 李 昕, 刘建辉. 一种电子商务信息安全保障机制[J]. 商场现代化, 2007, 2: 174.
LI Xin, LIU Jian-hui. A security mechanism of E-commerce[J]. Market Modernization, 2007, 2: 174.
[4] 谢冬青, 冷 健. PKI原理与技术[M]. 北京: 清华大学出版社, 2004.
XIE Dong-qing, LENG Jian. PKI principle and technology [M]. Beijing: Tsinghua University Press, 2004.
[5] 李 昕, 刘建辉. 基于PKI的电子商务信息安全性研究[J]. 中国管理信息化, 2007, 10(6): 68-70.
LI Xin, LIU Jian-hui. E-commerce security research based on PKI[J]. China Management Informationization, 2007, 10(6): 68-70.
[6] 周学广, 刘 艺. 信息安全学[M]. 北京: 机械工业出版社, 2004.
ZHOU Xue-guang, LIU Yi. Information security technology [M]. Beijing: China Machine Press, 2004.
[7] 李永先. 基于PKI 的电子商务安全研究[J]. 情报科学, 2004, 11: 1359-1361.
LI Yong-xian. Research and evaluation of the e-commerce security based PKI[J]. Information Science, 2004, 11: 1359-1361.
[8] 张一清. 电子商务中的安全基础设施PKI技术[J]. 商场现代化, 2006, 4: 89-90.
ZHANG Yi-qing. Security infrastructure PKI technology in E-commerce[J]. Market Modernization, 2006, 4: 89-90.
[9] 邢 莉. 电子商务中的安全体系研究[J]. 商场现代化, 2007, 7: 154-155.
XING Li. Security system research of e-commerce[J]. Market Modernization, 2007, 7: 154-155.
[10] 肖德琴. 电子商务安全保密技术与应用[M]. 广州: 华南理工大学出版社, 2005.
XIAO De-qin. Secrecy technique and application of E-Commerce[M]. Gguangzhou: South China University of Technology Press, 2005.

编 辑 税 红