DATA MINING AND E-COMMERCE: METHODS, APPLICATIONS, AND CHALLENGES

Faculty of Computer Science and Information Systems
University Technology Malaysia
81300 Skudai, Johor

hamid_rastegari@yahoo.com, mohdnoor@utm.my

Abstract: Electronic commerce processes and data mining tools have revolutionized many companies. Data that businesses collect about customers and their transactions are the greatest assets of that business. Data mining is a set of automated techniques used to extract buried or previously unknown pieces of information from large databases, using different criteria, which makes it possible to discover patterns and relationships. This paper discusses the important role of business based on data mining knowledge development to detection the relation of data mining and electronic commerce. And express some applications and challenges in this case.

Keywords: Data mining, e-commerce, web mining, business intelligence, web personalization.

1. INTRODUCTION

In today's business world there is an abundance of available data and a great need to make good use of it. In the first, data must be organized by data base tools and data warehouses, and then it needs an instrument for knowledge discovery. Data mining can be defined as the art of extracting non-obvious, useful information from large databases. This emerging field brings a set of powerful techniques which are relevance for companies to focus their efforts in taking advantage of their data.

Data mining tools generate new information for decision makers from very large databases. The various mechanisms of this generation include abstractions, aggregations, summarizations, and characterizations of data [1]. These forms, in turn, are the result of applying sophisticated modeling techniques from the diverse fields of statistics, artificial intelligence, database management and computer graphics.

Having a huge amount of data, make some problems for detection of hidden relationships among various attributes of data and between several snapshots of data over a

period of time. These hidden patterns have enormous potential in predictions and personalization in e-commerce. Data mining has been pursued as a research topic by at least three communities: the statisticians, the artificial intelligence researchers, and the database engineers [2].

Although much work has been done to date, more studies need to be conducted to as various subjects in a variety of e-commerce problems. The purpose of this paper is a present of data mining methods and expression application of data mining in business. It is a briefing of works that have been done in this area. This study can be useful for future work.

2. APPLICATIONS DATA MINING IN E-COMMERCE

In this section, we survey articles that are very specific to data mining implementations in e-commerce. The salient applications of data mining techniques are presented first. Later in this section, architecture and data collection issues are discussed.

2.1 Customer Profiling

It may be observed that customers drive the revenues of any organization. Acquiring new customers, delighting and retaining existing customers, and predicting 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 customer can configure a PC of his/her 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 detect what prospective customers are seeking from a site.

Companies 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. 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 Recommendation Systems

Systems have also been developed to keep the customers automatically informed of important events of interest to them. The article by Jeng & Drissi [3] 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 activated 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 [4]. Techniques including fuzzy computing and interval computing are used to generate if-then-else rules.

Niu et al present a method to build customer profiles in e-commerce settings, based on product hierarchy for more effective personalization [5]. 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 Web Personalization

Mobasher presents a comprehensive overview of the personalization process based on web usage mining [6]. 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.

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(s). Site content data also include semantic or structural metadata 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 linkage 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(s) 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 users' interests.

2.4 Customer 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 to prefetch the 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 [7]. The core of their approach involves extracting knowledge from integrated data of purchase and path traversal patterns of past users (obtainable from web server logs) 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 [8]. In particular, the authors present a theory, in which success is modeled on the basis of the navigation behavior of the site's users. They then exploit web usage miner (WUM), 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 [9].

3. BUSINESS INTELLIGENCE

Data mining is about finding useful patterns in data. This word useful can be unpacked to expose many of the key properties of successful data mining. The patterns discovered by data mining are useful because they extend existing business knowledge in useful ways. But new business knowledge is not created "in a vacuum"; it builds on existing business knowledge, and this existing knowledge is in the mind of the business expert. The business expert therefore plays a critical role in data mining, both as an essential source of input (business knowledge) and as the consumer of the results of data mining. The business expert not only uses the results of data mining but also evaluates them, and this evaluation should be a continual source of guidance for the data mining process. Data mining can reveal patterns in data, but only the business expert can judge their usefulness. It is important to remember that the data is not the business, but only a dim reflection of it. This gap, between the data and the business reality it represents, is called the chasm of representation to emphasize the effort needed to cross it.

Patterns found in the data may fail to be useful for many different reasons. They may reflect properties of the data, which do not represent reality at all, for example when an artifact of data collection, such as the time a snapshot is taken, distorts its reflection of the business. Alternatively, the patterns found may be true reflections of the business, but they merely describe the problem that data mining was intended to solve – for example arriving at the conclusion that "purchasers of this product have high incomes" in a project to market the product to a broader range of income groups. Finally, patterns may be a true and pertinent reflection of the business, but nevertheless merely repeat "truisms" about the business, already well known to those within it. It is all too easy for data mining, which is insufficiently informed by business knowledge to produce useless results for reasons like the above. To prevent this, the business expert must be at the very heart of the data mining process, spotting "false starts" before they consume significant effort. The expert must either literally "sit with"

the data miner, or actually perform the data mining. In either case, the close involvement of the business expert has far-reaching consequences for the field of data mining.

4. WEB TRANSACTIONS

All transaction on the web are gathered into web log file. This file can be saved on the server side. Web log server files are the primary means of collecting data and include transactions that the user performs, session level attributes, customer attributes, product attributes and abstract attributes. Session level analysis could highlight the number of page views per session, unique pages per session, time spent per session, average time per page, fast vs. slow connection etc. Additionally, this could throw light on whether users went through registration, if so, when, did the users look at the privacy statement; did they use search facilities, etc. The user level analysis could reveal whether the user is an initial or repeat or recent visitor/purchaser; whether the users are readers, browsers, heavy spenders, original referrers etc. [10].

The view of web transactions as sequences of page views allows one to employ a number of useful and well-studied models which can be used to discover or analyze user navigation patterns. One such approach is to model the navigational activity in the website as a Markov chain [11]. In the context of web transactions, Markov chains can be used to model transition probabilities between page views. In web-usage analysis, they have been proposed as the underlying modeling machinery for web prefetching applications or to minimize system latencies.

Hu&Cercone present a new approach called on-line analytical mining for web data [12]. Their approach consists of data capture, web house construction, and pattern discovery and pattern evaluation. The authors describe the challenges in each of these phases and present their approach for web usage mining. Their approach is useful in determining the most profitable customers, the difference between buyers and non-buyers, identification of website parts that attract most visits, parts of website that are session killers, parts of the site that lead to the most purchases, identifying the typical path of customers that leads to a purchase or otherwise etc. The web house is akin to the data warehouse.

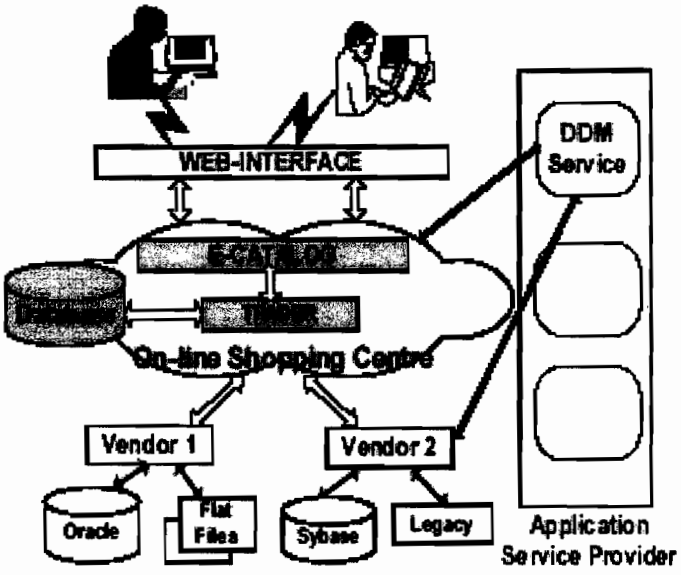
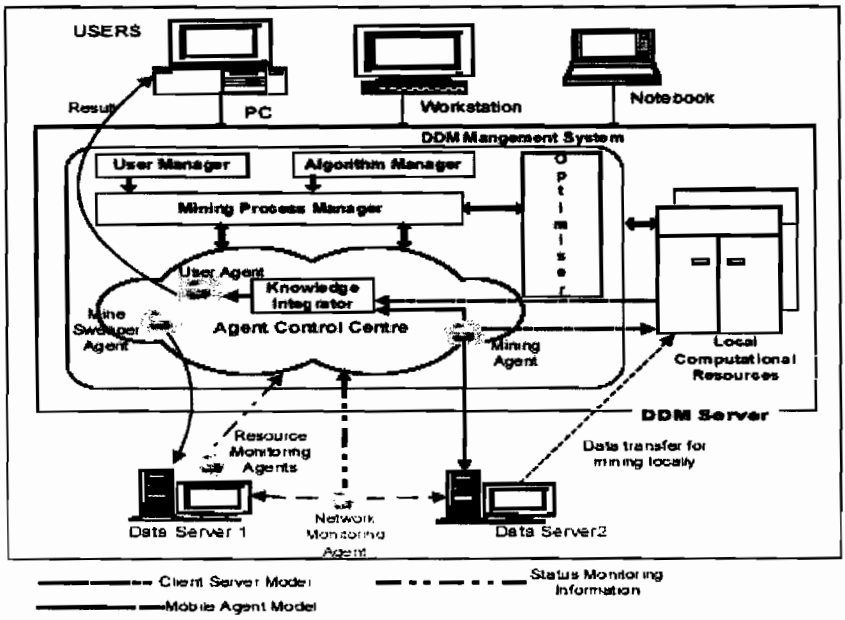


Figure 1: Distributed data mining in the e-commerce

5. AN ARCHITECTURE FOR DATA MINING

In a B2B e-commerce setting, it is very likely that vendors, customers and application service providers (ASP) (usually the middlemen) have varying data mining requirements. Vendors would be interested in data mining tailored for market basket analysis to know customer segments. On the other hand, end customers are keen to know updates on seasonal offerings and discounts all the while. The role of the ASP is then to be the common meeting ground for vendors and customers. Krishnaswamy et al propose a distributed data mining architecture that enables a data mining to be conducted in such a naturally distributed environment. A framework for the role of distributed data mining in the e-commerce is illustrated in figure 1 [13]. Figure 2 shows the components of the hybrid DDM architecture. The proposed distributed data mining system is intended for the ASP to provide generic data mining services to its subscribers. In order to support the robust functioning of the system it possesses certain characteristics such as heterogeneity, costing infrastructure availability, presence of a generic optimization engine, security and extensibility. Heterogeneity implies that the system can mine data from heterogeneous and distributed locations. The proposed system is designed to support user requirements with respect to different distributed computing paradigms (including the client-server and mobile agent based models). The costing infrastructure refers to the system having a framework for estimating the costs of different tasks. This implies that a task that requires higher computational resources and/or faster response time should cost the users more on a relative scale of costs. Further, the system should be able to optimize the

distributed data mining process to provide the users with the best response time possible (given the constraints of the mining environment and the expenses the user is willing to incur). The authors have indeed designed and implemented such a framework.



F
igure 2: The components of the hybrid DDM architecture

Maintaining security implies that in some instances, the user might be mining highly sensitive data that should not leave the owner's site. In such cases, the authors provide the option to use the mobile-agent model where the mining algorithm and the relevant parameters are shipped to the data site and at the end of the process the mobile agent is destroyed on the site itself. The system is extensible to provide for a wide range of mining algorithms [13]. The authors provide a facility wherein the user can register their algorithms with the ASP for use in their specific distributed data mining jobs.

6. CASES IN E-COMMERCE DATA MINING

In this section, we first present an interesting application of data mining in e-commerce. We then present some important lessons learnt by some authors while implementing data mining in e-commerce.

.

6.1 Distributed Spatial Data

In various e-commerce domains involving spatial data (real estate, environmental planning, precision agriculture), participating businesses may increase their economic returns using knowledge extracted from spatial databases. However, in practice, spatial data is often inherently distributed at multiple sites. Due to security, competition and a lack of appropriate knowledge discovery algorithms, spatial information from such physically dispersed sites is often not properly exploited. Lazarevic et al develop a distributed spatial knowledge discovery system for precision agriculture[14]. In the proposed system, a centralized server collects proprietary site-specific spatial data from subscribed businesses as well as relevant data from public and commercial sources and integrates knowledge in order to provide valuable management information to subscribed customers. Spatial data mining software interfaces this database to extract interesting and novel knowledge from data [15]. Specific objectives include a better understanding of spatial data, discovering relationships between spatial and nonspatial data, construction of spatial knowledge-bases, query optimization and data reorganization in spatial databases. Knowledge extracted from spatial data can consist of characteristic and discriminate rules, prominent structures or clusters, spatial associations and other forms.

Challenges involved in spatial data mining include multiple layers of data, missing attributes and high noise due to a low sensibility of instruments and to spatial interpolation on sparsely collected attributes. To address some of these problems, data are cleaned by removing duplicates, removing outliers and by filtering through a median filter with a specified window size [16]. The goal of precision agriculture management is to estimate and perform site-specific crop treatment in order to maximize profit and minimize environmental damage. Through a knowledge discovery (KDD) process, Lazarevic et al propose learning algorithms that perform data modeling using data sets from different fields in possibly different regions and years[16]. Each dataset may contain attributes whose values are not manageable (e.g. topographic data), as well as those attributes that are manageable (e.g. nutrient concentrations).

In order to improve prediction ability when dealing with heterogeneous spatial data, an approach employed in the proposed system by Lazarevic et al is based on identifying spatial regions having similar characteristics using a clustering algorithm[16]. A clustering algorithm is used for partitioning multivariate data into meaningful subgroups (clusters), so that patterns within a cluster are more similar to each other than are patterns belonging to

different clusters. Local regression models are built on each of these spatial regions describing the relationship between the spatial data characteristics and the target attribute.

6.2 Data Mining Applied to Retail E-Commerce

Kohavi et al have attempted a practical implementation of data mining in retail ecommerce data. They share their experience in terms of lessons that they learnt [10]. They classify the important issues in practical studies, into two categories: business-related and technology related. We now summarize their findings on the technical issues here.

(1) Collecting data at the right level of abstraction is very important. Web server logs were originally meant for debugging the server software. Hence they convey very little useful information on customer-related transactions. Approaches including sessionising the web logs may yield better results. A preferred alternative would be having the application server itself log the user related activities. This is certainly going to be richer in semantics compared to the state-less web logs, and is easier to maintain compared to state-full web logs.

(2) Designing user interface forms needs to consider the data mining issues in mind. For instance, disabling default values on various important attributes like Gender, Marital status, Employment status, etc., will result in richer data collected for demographical analysis. The users should be made to enter these values, since it was found by Kohavi et al that several users left the default values untouched [10].

(3) Certain important implementation parameters in retail e-commerce sites like the automatic time outs of user sessions due to perceived inactivity at the user end, need to be based not purely on data mining algorithms, but on the relative importance of the users to the organization. It should not turn out that large clients are made to lose their shopping carts due to the time outs that were fixed based on a data mining of the application logs.

(4) Generating logs for several million transactions is a costly exercise. It may be wise to generate appropriate logs by conducting random sampling, as is done in statistical quality control. But such a sampling may not capture rare events, and in some cases like in advertisement referral based compensations, the data capture may be mandatory. Techniques thus need to be in place that can do this sampling in an intelligent fashion.

(5) Auditing of data procured for mining, from data warehouses, is mandatory. This is due to the fact that the data warehouse might have collated data from several disparate systems with a high chance of data being duplicated or lost during the ETL operations.

(6) Mining data at the right level of granularity is essential. Otherwise, the results from the data mining exercise may not be correct.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented how web mining (in a broad sense, data mining applied to ecommerce) is applicable to improve the services provided by e-commerce based enterprises. Specifically, we first discussed some popular tools and techniques used in data mining. Statistics, AI and database methods were surveyed and their relevance to data mining in general was discussed. We then presented a host of applications of these tools to data mining in e-commerce. Later, we also highlighted architectural and implementation issues. We now present some ways in which web mining can be extended for further research. With the growing interest in the notion of semantic web, an increasing number of sites use structured semantics and domain ontology as part of the site design, creation, and content delivery. The notion of Semantic Web Mining was introduced by Berendt et al [17]. The primary challenge for the next-generation of personalization systems is to effectively integrate semantic knowledge from domain ontology into the various parts of the process, including the data preparation, pattern discovery, and recommendation phases. Such a process must involve some or all of the following tasks and activities [6].

(1) Ontology learning, extraction, and preprocessing: Given a page in the web site, we must be able extract domain-level structured objects as semantic entities contained within this page.

(2) Semantic data mining: In the pattern discovery phase, data mining algorithms must be able to deal with complex semantic objects.

(3) Domain-level aggregation and representation: Given a set of structured objects representing a discovered pattern, we must then be able to create an aggregated representation as a set of pseudo objects, each characterizing objects of different types occurring commonly across the user sessions.

(4) Ontology-based recommendations: Finally, the recommendation process must also incorporate semantic knowledge from the domain ontology.

Some of the challenges in e-commerce data mining include the following [18].

- Crawler/bot/spider/robot identification: Bots and crawlers can dramatically change clickstream patterns at a web site. For example, some websites like (www.keynote.com) provide site performance measurements. The Keynote bot 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 tools need to have

mechanisms to automatically sieve such noisy data in order for data mining algorithms to yield sensible and pragmatic proposals.

- Data transformations: There are two sets of transformations that need to take place first data must be brought in from the operational system to build a data warehouse, and second data may need to undergo transformations to answer a specific business question, a process that involves operations such as defining new columns, binning data, and aggregating it. While the first set of transformations needs to be modified infrequently (only when the site changes), the second set of transformations provides a significant challenge faced by many data mining tools today.

- Scalability of data mining algorithms: With a large amount of data, two scalability issues arise: (i) most data mining algorithms cannot process the amount of data gathered at web sites in reasonable time, especially because they scale nonlinearly; and (ii) generated models are too complicated for humans to comprehend.

Episode mining involves mining not one-time events, but mining for a historical pattern of events. Episode-mining methods rely on extensions of rule-mining methods. Alternate approaches could be explored here. Support vector machines have taken the centre stage of late, in learning linear and nonlinear relationships[19]. Their applications in episode mining could be an exciting area for further work.

REFERENCES

[1] P. L. Carbone, "Expanding the meaning of and applications for data mining," *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics,* vol. 3, pp. 1872-1873, 2000.

[2] N. R. S. Raghavan, "Data mining in e-commerce: A survey," *Sadhana - Academy Proceedings in Engineering Sciences,* vol. 30, pp. 275-289, 2005.

[3] J.-J. Jeng and Y. Drissi, "PENS: A predictive event notification system for e-Commerce environment," *Proceedings - IEEE Computer Society's International Computer Software and Applications Conference,* pp. 93-98, 2000.

[4] X. Z. Zhang, "Building personalized recommendation system in E-Commerce using association rule-based mining and classification," in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*, 2007, pp. 4113-4118.

[5] L. Niu, X. W. Yan, C. Q. Zhang, and S. C. Zhang, "Product hierarchy-based customer profiles for electronic commerce recommendation," *Proceedings of 2002 International Conference on Machine Learning and Cybernetics,* vol. 2, pp. 1075-1080, 2002.

[6] B. Mobasher, "Web usage mining and personalization," *Practical Handbook of Internet Computing,* 2004.

[7] S. Vallamkondu and L. Gruenwald, "Integrating purchase patterns and traversal patterns to predict http requests in e-commerce sites," *IEEE Int. Conf. on e-commerce,* pp. 256-263, 2003.

[8] M. Spiliopoulou and C. Pohle, "Data mining to measure and improve the success of web sites," *J. Data Mining and Knowledge Discovery,* 2000.

[9] W. Zhu, J. Chen, and J. Yin, "Application of data mining in E-business," *Jisuanji Gongcheng/Computer Engineering,* vol. 28, p. 73, 2002.

[10] R. Kohavi, "Lessons and Challenges from Mining Retail E-Commerce Data," 2004.

[11] R. R. Sarukkai, "Link prediction and path analysis using markov chains," *Proceedings of the 9th International World Wide Web Conference,* 2000.

[12] X. Hu and N. Çercone, "An OLAM framework for web usage mining and business intelligence reporting," *IEEE International Conference on Plasma Science,* vol. 2, pp. 950-955, 2002.

[13] S. Krishnaswamy, A. Zaslavsky, and S. W. Loke, "An architecture to support distributed data mining services in e-commerce environments," in *Advanced Issues of E-Commerce and Web-Based Information Systems, 2000. WECWIS 2000. Second International Workshop on*, 2000, pp. 239-246.

[14] A. Lazarevic, X. Xu, T. Fiez, and Z. Obradovic, "Clustering-regression-ordering steps for knowledge discovery in spatial databases," *Proc. IEEE/INNS Int'l Conf. on Neural Neural Networks,* 1999.

[15] K. Koperski, J. Adhikary, and J. Han, "Spatial data mining: Progress and challenges," *J. Data Mining Knowledge Discovery,* 1996.

[16] E. Lazcorreta, F. Botella, and A. Ferna?ndez-Caballero, "Towards personalized recommendation by two-step modified Apriori data mining algorithm," *Expert Systems with Applications,* vol. 35, pp. 1422-1429, 2008.

[17] B. Berendt, A. Hotho, and G. Stumme, "Towards semantic web mining," *Proceedings of the International Semantic Web Conference,* vol. 2342, pp. 264-278, 2002.

[18] R. Kohavi, R. M. Henne, and D. Sommerfield, "Practical guide to controlled experiments on the web: Listen to your customers not to the hippo," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 959-967.

[19] S. Haykin, *Neural Networks: A Comprehensive Foundation,* 1994.