

Data Mining in Electronic Commerce: Benefits and Challenges

Mustapha Ismail, Mohammed Mansur Ibrahim, Zayyan Mahmoud Sanusi, Muesser Nat

Management Information Systems Department, Cyprus International University, Haspolat, Lefkoşa via Mersin, Turkey
Email: 20141064@student.ciu.edu.tr, 20143383@student.ciu.edu.tr, 20132066@student.ciu.edu.tr, mnat@ciu.edu.tr

Received 31 October 2015; accepted 25 December 2015; published 28 December 2015

Copyright © 2015 by authors and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Huge volume of structured and unstructured data which is called big data, nowadays, provides opportunities for companies especially those that use electronic commerce (e-commerce). The data is collected from customer's internal processes, vendors, markets and business environment. This paper presents a data mining (DM) process for e-commerce including the three common algorithms: association, clustering and prediction. It also highlights some of the benefits of DM to e-commerce companies in terms of merchandise planning, sale forecasting, basket analysis, customer relationship management and market segmentation which can be achieved with the three data mining algorithms. The main aim of this paper is to review the application of data mining in e-commerce by focusing on structured and unstructured data collected through various resources and cloud computing services in order to justify the importance of data mining. Moreover, this study evaluates certain challenges of data mining like spider identification, data transformations and making data model comprehensible to business users. Other challenges which are supporting the slow changing dimensions of data, making the data transformation and model building accessible to business users are also evaluated. A clear guide to e-commerce companies sitting on huge volume of data to easily manipulate the data for business improvement which in return will place them highly competitive among their competitors is also provided in this paper.

Keywords

Data Mining, Big Data, E-Commerce, Cloud Computing

1. Introduction

Data mining in e-commerce is all about integrating statistics, databases and artificial intelligence together with some subjects to form a new idea or a new integrated technology for the purpose of better decision making. Data

mining as a whole is believed to be a good promoter of e-commerce. Presently, applying data mining to e-commerce has become a hot cake among businesses [1]. Data mining in cloud computing is the process of extracting structured information from unstructured or semi unstructured web data sources. From business point of view, the core concept of cloud computing is to render computing resources in form of service to the users who need to buy whenever they are in demand [2]. The end product of data mining creates an avenue for decision makers to be able to track their customers' purchasing patterns, demand trends and locations, making their strategic decision more effective for the betterment of their business. This can bring down the cost of inventory together with other expenses and maximizing the overall profit of the company.

With the wide availability of the Internet, 21st century companies highly utilize online tools and technologies for various reasons. Therefore, today many companies buy and sell through e-commerce and the need for developing e-commerce applications by an expert who takes responsibility for running and maintaining the services is increasing. When businesses grow, the required resources for e-commerce maintenance may increase more than the level the enterprise can handle. Based on that regard, data mining can be used to handle e-commerce enterprise services and explore patterns for online customers so companies can boost sales and the general productivity of the business [3]. However, the cost of running such services is a challenge to almost all e-commerce companies. Therefore cloud computing becomes a game changer in the way and manner companies transact their businesses by offering a comprehensive scalable and flexible services over the Internet. Cloud computing provides a new breakthrough for enterprises, offering a service model that includes network storage, new information resource sharing, on-demand access to information and processing mechanism. It is possible to provide data mining software via cloud computing which gives e-commerce companies opportunity to centralize their software management and data storage with absolute assurance of reliability, efficiency and protected services to their users which in turn cut their cost and increase their profit [4].

Cloud computing is a technology that has to do with accessing products and services in the cloud without shouldering the burden of hosting or delivering these services. It can be also viewed as a "model that enhances a flexible on-demand network access to a shared pool of configurable computing resources like networks, servers, storage applications and services that can speedily provisioned and released with minimal management effort or service provider interaction". In the aspect of cloud computing everything is considered as a service. There are three service delivery models of cloud computing namely: Infrastructure as a Service (IaaS) which is responsible for fundamental computing resources like, storage, processing, networks and also some standardized services over the networks. The second is the Platform as a Service (PaaS) which gives abstractions together with the services for developing, testing, hosting and of course maintaining the applications in the complex and developed environment. The third one is the Software as the Service (SaaS). The entire application or service is delivered over the web through a browser or via application programming interface (API). With service model the consumers only need to focus on administering users to the system.

One of the most important applications of cloud computing is the storage capability. Cloud storage has the capability to cluster different types of storage equipment by employing cluster system, grid technology or distributed system in the network to provide external data storage and access services by the use of software application. Cloud computing in e-commerce is the idea of paying bandwidth and storage space on the scale that depends on the usage. It is much more on the utility on-demand basis whereby a user pays for less with pay per use models. Most e-commerce companies welcome the idea as it eliminates the high cost of storage for large volume of business data by keeping it in the cloud data centers. The platform also gives opportunity to use e-commerce business applications e.g. B2B and B2C with smaller investment. Some other advantages of cloud computing for e-commerce include the following: cost effective, speed of operations, scalability and security of the entire service [3] [4].

The association between cloud computing and data mining is that cloud is used to store the data on the servers and data mining is use to provide client server relationship as a service and information being collected based on ethical issues like privacy and individuality are violated [5].

Considering the importance of data mining for today's companies, this paper discusses benefits and challenges of data mining for e-commerce companies. Furthermore, it reviews the process of data mining in e-commerce together with the common types of database and cloud computing in the field of e-commerce.

2. Data Mining

Data mining is the process of discovering meaningful pattern and correlation by sifting through large amounts of

data stored in repositories. There are several tools for this data generation, which include abstractions, aggregations, summarization and characteristics of data [6]. In the past decade, data mining has change the e-commerce business. Data mining is not specific to one type of data. Data mining can be germane to any type of information source, however, algorithms and tactics may differ when applied to different kind of data. The challenges presented by different type of data varies. Data mining is being used in many form of databases like flat file, data warehouses, object oriented databases and etc.

This paper concentrates on relational databases. Relational database consists of a set of tables containing either values of entity attributes or values of attributes from entity relationship. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key [6]. The most commonly used query language for relational database is SQL, which allows to manipulate and retrieve data stored in the tables. Data mining algorithms using relational database can be more versatile than data mining algorithms specifically written for flat files. Data mining can benefit from SQL for data selection, transformation and consolidation [7].

There are several core techniques in data mining that are used to build data mining. Most common techniques are as follows [8] [9]:

1) Association Rules

Association rule mining is among the most important methods of data mining. The essence of this method is extracting interesting correlation and association among sets of items in the transactional databases or other data pools. Association rules are used extensively in various areas. A typical association rule has an implication of the form $A \rightarrow B$ where A is an item set and B is an item set that contains only a single atomic condition [10].

2) Clustering

This is the organisation of data in classes or it refers to a collection of objects by grouping similar objects to form more than one class of methods. Moreover, clustering class labels are unidentified and it is up to the clustering algorithm to discover acceptable classes. Clustering is sometimes called unsupervised classification. The reason was classification is not dictated by given class labels. Clustering is the process of grouping a set of physical or abstract object into classes of similar object [10].

3) Prediction

Prediction has attracted substantial attention given the possible consequences of successful forecasting in a business context. There are two types of predictions. The first one is predicting unavailable data values and the second one is as soon as classification model is form on a training set, the class label of the object can be predicted based on the attribute values of the object. Prediction is more often referred to the forecast of missing numerical values [10].

3. Some Common Data Mining Tools

1) Weka

To have accurate data mining result require the right tool for the dataset you are mining. Weka however, gives the ability to put into reality the learning methods algorithms. The tool has lots of benefits as it's include all the standard data mining procedures like data pre-processing, clustering, association, classification, regression and also attribute selection. It has both the Java and non-Java version together with visualization application, and the tool is free to users to customize it to their own specification [11] [12].

2) NLTK

It is mainly for language processing task with pool of different language processing tools together with machine learning, data mining and sentiment analysis, data scrapping and different language processing tasks. NLTK tool require a user to install the tool on his systems and have access to the full package. It is built in python and a user can build application on top and can play around with the tool to his own specification. All the three mentioned tools above are open source [11].

3) Spider Miner

A data mining tool that does not require a user to write a code, written in Java programming language. Part of SpiderMiner tool capability is that, it provides a thorough analytics via template-based frameworks. It is very flexible tool and user friendly offered as a service, and apart from data mining function, the tool can visualize, predict, data pre-processing, deployment statistical modelling and of course evaluation functions. In the tool

there learning schemes, algorithms and models from WEKA and R script which makes the tool to be more powerful [12]. All the three mentioned tools above are open source.

4. Data Mining in E-Commerce

Data mining in e-commerce is a vital way of repositioning the e-commerce company for supporting the enterprise with the required information concerning the business. Recently, most companies adopt e-commerce and being in possession of big data in their data repositories. The only way to get the most out of this data is to mine it to increase decision making or to enable business intelligence. In e-commerce data mining there are three important processes that data must pass before turning into knowledge or application. **Figure 1** shows the steps for data mining in e-commerce.

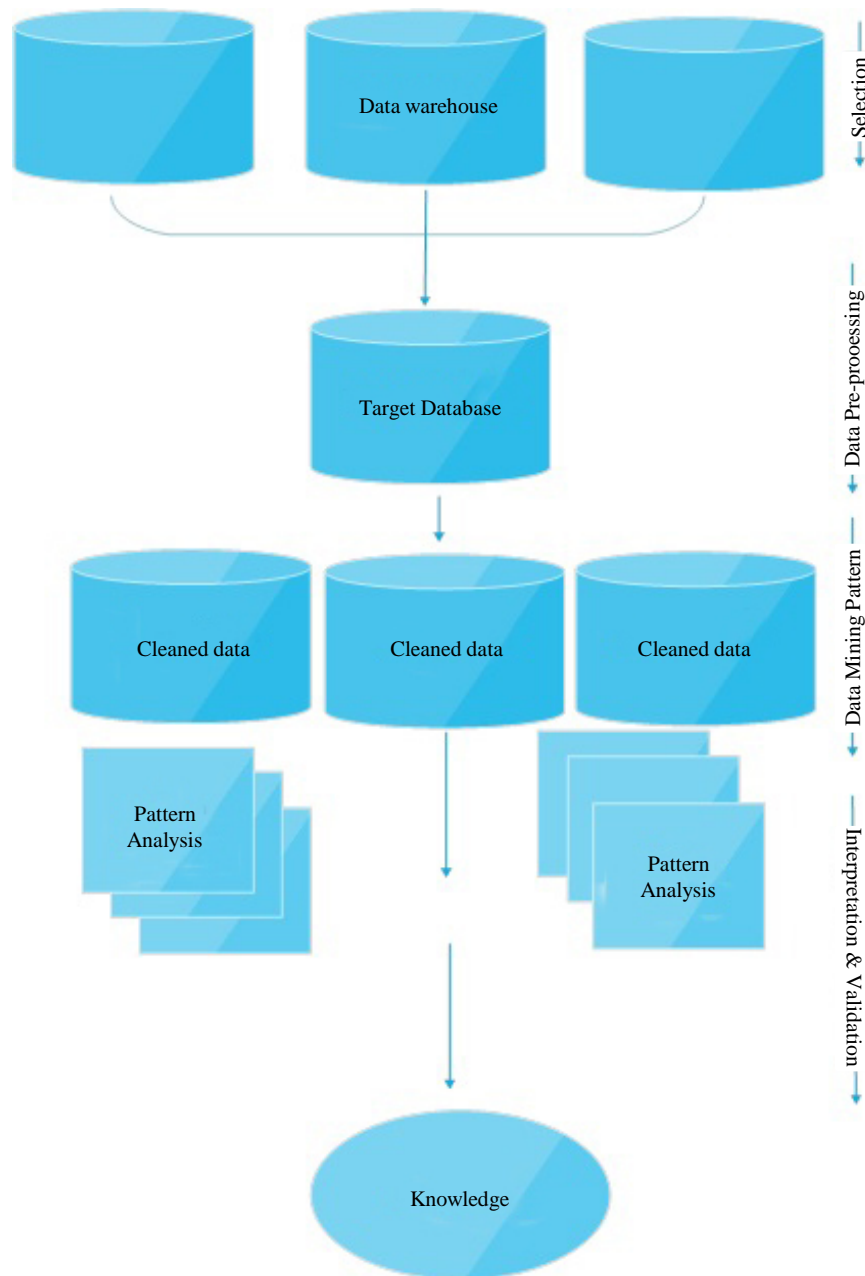


Figure 1. Data mining process in e-commerce [16].

The first and easier process of data mining is data preprocessing and it is actually a step before the data mining, whereby, the data is cleaned by removing the unwanted data that has no relation with the required analysis. Hence, the process will boost the performance of the entire data mining process and the accuracy of the data will also be high and the time needed for the actual mining will be minimised reasonably. Usually this happens if company already have an existing target data warehouse, but if not then the process will consume at least 80% of the selection, cleaning and transformation of data termed as preprocessing [13].

Mining pattern is the second step and it actually refers to techniques or approach used to develop a recommendation rules, or developing a model out of a large data set. It can also be referred as techniques or algorithms of data mining. The most common patterns used in e-commerce are prediction, clustering and association rules.

The purpose of third step which is pattern analysis is to verify and shade more light on the discovered model in order to give a clear path for the startup up for applying of the data mining result. The analysis lay much emphasis on the statistics and rules of the pattern used, by observing them after multiple users have accessed them [14].

However all this has to do with how iterative the overall process is, and the interpretation of visual information you get at each sub step. Therefore, in general data mining process iterates from the following five basic steps, which are:

- Data selection: This step is all about identifying the kind of data to be mined, the goals for it and the necessary tool to enable the process. At the end of it the right input attributes and output information in order to represent the task are chosen.
- Data transformation: This step is all about organising the data based on the requirements by removing noise, converting one type of data to another, normalising the data if there is need to, and also defining the strategy to handle the missing data.
- Data mining step per se: Having mined the transformed data using any of the techniques to extract pattern of interest, the miner can also make data mining method by performing the proceeding steps correctly.
- Result interpretation and validation: For better understanding of data and it synthesised knowledge together with its validity span, the robustness is check by data mining application test. The information retrieved can also be evaluated by comparing it with the earlier expertise in the application domain.
- Incorporation of the discovered knowledge: This has to do with presenting the result of discovered knowledge to decision maker so that it is possible to compare or check/resolve for conflict with an earlier extracted knowledge where a new discovered pattern can be applied [15].

5. Benefits of Data Mining in E-Commerce

Application of data mining in e-commerce refers to possible areas in the field of e-commerce where data mining can be utilised for the purpose of enhancements in business. As we all know while visiting an online store for shopping, users normally leave behind certain facts that companies can store in their database. These facts represent unstructured or structured data that can be mined to provide a competitive advantage to the company. The following areas are where data mining can be applied in the field of e-commerce for the benefits of companies:

1) Customer Profiling

This is also known as customer-oriented strategy in e-commerce. This allows companies to use business intelligence through the mining of customer's data to plan their business activities and operations as well as develop new research on products or services for prosperous e-commerce. Classifying the customers of great purchasing potentially from the visiting data can help companies to lessen the sales cost [17]. Companies can use users' browsing data to identify whether they purposefully shopping or just browsing or buying something they are familiar with or something new. This helps companies to plan and improve their infrastructure [18].

2) Personalization of Service

Personalization is the act to provide contents and services geared to individuals on the basis of information of their needs and behavior. Data mining research related to personalization has focused mostly on recommender systems and related subjects such as collaborative filtering. Recommender systems have been explored intensively in the data mining community. This systems can be divided into three groups: Content-based, social data mining and collaborative filtering. These systems are cultured and learned from explicit or implicit feedback of users and are usually represented as the user profile. Social data mining, in considering the source of data that are created by the group of individuals as part of their daily activities, can be important source of important information for companies. Contrarily, personalization can be achieved by the aid of collaborative filtering, where

users are matched with particular interest and in the same vein the preferences of these users to make recommendations [19].

3) Basket Analysis

Every shoppers' basket has a story to tell and market basket analysis (MBA) is a common retail, analytic and business intelligence tool that helps retailers to know their customers better. There are different ways to get the best out of market basket analysis and these include:

- Identification of product affinities; tracking not so apparent product affinities and leveraging on them is the real challenge in retail. Walmart customers purchasing Barbie dolls shows an affinity towards one of three candy bars, obscure connection such as this can be discovered with an advanced market basket analytics for planning more effective marketing efforts.
- Cross-sell and up-sell campaigns; these shows the products purchased together, so customers who purchase the printer can be persuaded to pick up high quality paper or premium cartridges.
- Planograms and product combos; are used for better inventory control based on product affinities, developing combo offers and design effective user friendly planograms in focusing on products that sells together.
- Shoppers profile; in analyzing market basket with the aid of data mining over time to get a glimpse of who your shoppers really are, gaining insight to their ages, income range, buying habits, likes and dislikes, purchase preferences, leveraging this and giving the customer experience [19].

4) Sales Forecasting

Sales forecasting involves the aspect of the time an individual customer spend to buy an item and in this process trying to predict if the customer will buy again. This type of analysis can be used to determine a strategy of planned obsolescence or figure out complimentary products to sell. In sales forecasting, cash flow can be projected into three which include the pessimistic, optimistic and the realistic. This helps to have a plan on the adequate amount of capital available to endure the worst possible scenario that is if sales do not go actually as planned [19].

5) Merchandise Planning

Merchandise planning is useful for both online and offline retail companies. In the case of online business, merchandise planning will help to determine stocking options and the inventory warehousing, while in the case of offline companies, business that are looking to boost by adding stores can assess the required amount of merchandise they will be adequately needing by having a foresight at the exact layout of the current store [20].

Using the right approach to merchandise planning will definitely lead to answers on what to do with:

- Pricing: the aspect of database mining will help determining the suited best price of products or services in the processes of revealing customer sensitivity.
- Deciding on products; data mining provides e-commerce businesses with the aspect of which products customers actually desire, which includes the aspect of intelligence on competitor's merchandise.
- Balancing of stocks; in mining the retail database, it helps determine the right and specific amount of stocks needed *i.e.* not too much and not too less, throughout the business year and also during the buying seasons.

6) Market Segmentation

Customer segmentation is one of the best uses of data mining. From the lots of data gotten, it can be broken down into different and meaningful segments like income, age, gender, occupation of customers, and this can be used when either the companies are running email marketing campaigns or SEO strategies. The aspect of market segmentation can also help a company identify its own competitors. This provided information alone can help the retail company identify that the periodic respondents are usually not the only ones pointing the same customer money as the present company is [21].

Segmenting the database of a retail company will improve the conversion rates as the company can focus there promotion on a close-fitted and highly wanted market. This also helps the retail company to understand the competitors that are involved in each and every segment in the process permitting the customization of products that will actually satisfy the target audience in a generic way [21].

6. Challenges of Data Mining in E-Commerce

Besides the benefits data mining provides challenges for e-commerce companies, which are as follows:

1) Spider Identification

As it is commonly known main aim of data mining is to convert data into useful knowledge. Main source of

data for e-commerce companies is web pages. Therefore, it is critical for e-commerce companies to understand how search engines work to follow how quickly things happen, how they happen and when changes will show up in the search engines. Spiders are software programs that are sent out by the search engine to find new information. These spiders can also be called as bots or crawlers. It is a software program that search engine uses to request pages and download them, it comes as a surprise to some people, however what the search engine does is they use a link of an existing website to find a new website and request a copy of that page to download it to their server. This is what the search engines use to run the ranking algorithm against and that is what shows up in the search engine result page. Therefore, the challenge here is that the search engines need to download a correct copy of the website. E-commerce website needs to be readable and seeable and the algorithm is applied to the search engines database. Tools are needed to have the mechanisms to enable them automatically remove unwanted data that will be transformed to information in order for data mining algorithm to provide reliable and sensible output [22].

2) Data Transformations

In this case data transformation pose a challenge for data mining tools. Today, the data needed to transform can only be gotten from two different sources, one of which an active and operational system for the data warehouse to be built and secondly it should include some activities that involves assigning new columns, binning data and also aggregating the data as well. In the first process, it is needed to be modified infrequently that is only when there is a change in the site and lastly the set of the transformed data gives a significantly great challenge in the data mining process [22].

3) Scalability of Data Mining Algorithms

With yahoo which has over 1.2 billion page views in a day with the presence of large amount of data, scalability arises with significant issues;

- Due to the large amount of data size gathered from the website at a reasonable time, the data mining algorithm can handle or process it as much as it's needed especially because of the scale nonlinearly.
- The models that are generated tends to be too complicated for individuals to understand how it is interpreted [22].

4) Make Data Mining Models Comprehensible to Business Users

The results of data mining should be clearly understood by business users, from the merchandisers who are in charge of decision making to the creative designers that design the sites to marketers to spend advertising money. The challenge is to design and define extra model types and a strategic way to present them to business users, what regression models can we come up with and how can we present them? (Even linear regression is usually hard for business users to understand.) How can we present nearest-neighbor models, for example? How can we present the results of association rule algorithms without overwhelming users with tens of thousands of rules? [22].

5) Support Slowly Changing Dimensions

The demographic aspect of visitors change, in that they may get married, there is an increase in salaries or income, the rapid growth of their children, needs which are the bases on which it is modelled changes. Thus, the products attributes also change, in terms of new choices may be available, the design and the way the products or service is packaged and also the increase or degrade of quality. These attribute that change over time are often known as "Slowly Changing Dimensions". In this case the main challenge here is to keep track of those changes and in the same vein providing support for the identified change in the analysis [2].

6) Make Data Transformation and Model Building Accessible to Business Users

Having the ability to provide definite answers to questions by individual business users, this requires the aspects of data transformations but with the technical understanding of the tools used in the analysis. Many commercials report designers and also online analytical processing (OLAP) tools are basically hard to understand by business users. In this case, two preferred solutions are (I) provision of templates, (e.g. online analytical processing cubes and recommended transformations for mining) for the expected questions and (ii) provision of the experts via consultation or even a service organization. This mentioned challenge basically is to find a way to enrich the business users to as to be able to analyze the information themselves without and hiccups [2].

7. Summary and Conclusion

Data mining for e-commerce companies should no longer be privilege but requirement in order to survive and

remain relevant in the competitive environment. On one hand, data mining offers number of benefits to e-commerce companies and allows them to do merchandise planning, analyze customers' purchasing behaviors and forecast their sales which in turn would place them over other companies and generate more revenue. On the other hand, there are certain challenges of data mining in the field of e-commerce such as spider identification, data transformation, scalability of data mining algorithms, making data mining model comprehensible to business users, support slow changing dimensions and making data transformation and model building accessible to business users.

The data collected about customers and their transactions, which are the greatest assets of e-commerce companies, needs to be used consciously for the benefits of the companies. For such companies, data mining plays an important role in providing customer-oriented services to increase customer satisfaction. It has become apparent that utilizing data mining tools is a necessity for e-commerce companies in this global competitive environment.

Although the complexity and granularity of the mentioned challenges differ, e-commerce companies can overcome these problems by using and applying the right techniques. For example, developing e-commerce website in a way that search engines can read and access the latest version of the website, help companies to overcome the search engine spider identification problem.

Another hot topic in e-commerce data mining is cloud computing which is also covered in this paper. While the need of data mining tools is growing every day, the ability of integrating them in cloud computing becomes more stringent. It is obvious that making good use of the cloud computing technology in e-commerce helps effective use of resources and reduces costs for companies that enable efficient data mining.

References

- [1] Cao, L., Li, Y. and Yu, H. (2011) Research of Data Mining in Electronic Commerce. *IEEE Computer Society*, Hebei.
- [2] Bhagyashree, A. and Borkar, V. (2012) Data Mining in Cloud Computing. *Multi Conference (MPGINMC-2012)*. <http://reserach.ijcaonline.org/ncrtc/number6/mpginme1047.pdf>
- [3] Rao, T.K.R.K., Khan, S.A., Begun, Z. and Divakar, Ch. (2013) Mining the E-Commerce Cloud: A Survey on Emerging Relationship between Web Mining, E-Commerce and Cloud Computing. *IEEE International Conference on Computational Intelligence and Computing Research*, Enathi, 26-28 December 2013, 1-4. <http://dx.doi.org/10.1109/iccic.2013.6724234>
- [4] Wu, M., Zhang, H. and Li, Y. (2013) Data Mining Pattern Valuation in Apparel Industry E-Commerce Cloud. *IEEE 4th International Conference on Software Engineering and Service Science (ICSESS)*, 689-690.
- [5] Srinivva, A., Srinivas, M.K. and Harsh, A.V.R.K. (2013) A Study on Cloud Computing Data Mining. *International Journal of Innovative Research in Computer and Communication Engineering*, **1**, 1232-1237.
- [6] Carbone, P.L. (2000) Expanding the Meaning and Application of Data Mining. *International Conference on Systems, Man and Cybernetics*, **3**, 1872-1873. <http://dx.doi.org/10.1109/icsmc.2000.886383>
- [7] Barry, M.J.A. and Linoff, G.S. (2004) On Data Mining Techniques for Marketing, Sales and Customer Relationship Management. Indianapolis Publishing Inc., Indiana.
- [8] Pan, Q. (2011) Research of Data Mining Technology in Electronic Commerce. *IEEE Computer Society*, Wuhan, 12-14 August 2011, 1-4. <http://dx.doi.org/10.1109/icmss.2011.5999185>
- [9] Verma, N., Verma, A., Rishma and Madhuri (2012) Efficient and Enhanced Data Mining Approach for Recommender System. *International Conference on Artificial Intelligence and Embedded Systems (ICAIES2012)*, Singapore, 15-16 July 2012.
- [10] Kamba, M. and Hang, J. (2006) Data Mining Concept and Techniques. Morgan Kaufmann Publishers, San Fransisco.
- [11] News Stack (2015). <http://thenewstack.io/six-of-the-best-open-source-data-mining-tools/>
- [12] Witten, I.H. and Frank, E. (2014) The Morgan Kaufmann Series on Data Mining Management Systems: Data Mining. 2nd Edition, Publisher Morgan Kaufmann, San Francisco, 365-528.
- [13] Liu, X.Y. And Wang, P.Z. (2008) Data Mining Technology and Its Application in Electronic Commerce. *IEEE Computer Society*, Dalian, 12-14 October 2008, 1-5.
- [14] Zeng, D.H. (2012) Advances in Computer Science and Engineering. Springer Heidelberg, NewYork.
- [15] Ralph, K. and Caserta, J. (2011) The Data Warehouse ETL Toolkit: Practical Techniques for Extraction, Cleaning, Conforming and Delivering Data. Wiley Publishing Inc., USA.
- [16] Michael, L.-W. (1997) Discovering the Hidden Secrets in Your Data—The Data Mining Approach to Information. *In-*

-
- formation Research*, **3**. <http://informationr.net/ir/3-2/>
- [17] Li, H.J. and Yang, D.X. (2006) Study on Data Mining and Its Application in E-Business. *Journal of Gansu Lianhe University (Natural Science)*, No. 2006, 30-33.
- [18] Raghavan, S.N.R. (2005) Data Mining in E-Commerce: A Survey. *Sadhana*, **30**, 275-289. <http://dx.doi.org/10.1007/BF02706248>
- [19] Michael, J.A.B. and Gordon, S.L. (1997) *Data Mining Techniques: For Marketing and Sales, and Customer Relationship Management*. 3rd Edition, Wiley Publishing Inc., Canada.
- [20] Wang, J.-C., David, C.Y. and Chris, R. (2002) Data Mining Techniques for Customer Relationship Management. *Technology in Society*, **24**, 483-502.
- [21] Christos, P., Prabhakar. R. and Jon, K. (1998) A Microeconomic View of Data Mining. *Data Mining and Knowledge Discovery*, **2**, 311-324. <http://dx.doi.org/10.1023/A:1009726428407>
- [22] Yahoo (2001) Second Quarter Financial Report. Yahoo Inc., California.