12-31-2014

DRIP - Data Rich, Information Poor: A Concise Synopsis of Data Mining

Muhammad Obeidat
Southern Polytechnic State University

Max North
Southern Polytechnic State University, max@spsu.edu

Lloyd Burgess
Southern Polytechnic State University

Sarah North
Kennesaw State University, snorth@kennesaw.edu

Universal Journal of Management 3(1): 29-35, 2015
DOI: 10.13189/ujm.2015.030105 http://www.hrpub.org

DRIP – Data Rich, Information Poor: A Concise Synopsis of Data Mining

Muhammad Obeidat[1] , Max North[1,*] , Lloyd Burgess[1] , Richard Parker[1] , Sarah North[2]

[1]Management Information Systems, Visualization and Simulation Research Center, Business Administration Department, Southern Polytechnic State University, Marietta, GA 30060, USA
[2]Computer Science Department, College of Science and Mathematics, Kennesaw State University, Kennesaw, GA 30144, USA
*Corresponding Author : max@spsu.edu

Abstract As production of data is exponentially growing with a drastically lower cost, the importance of data mining required to extract and discover valuable information is becoming more paramount. To be functional in any business or industry, data must be capable of supporting sound decision-making and plausible prediction. The purpose of this paper is concisely but broadly to provide a synopsis of the technology and theory of data mining, providing an enhanced comprehension of the methods by which massive data can be transferred into meaningful information.

Keywords Data Mining , Massive Data

1. Introduction

DRIP – Data rich, information poor: "I don't know what I don't know." This seems to be a common complaint among organizations today – that they have spent the past several years compiling ever-increasing mountains of data to support their operations, but they do not feel that they are getting a good return. One could argue that valuable knowledge, knowledge that could lead to a scientific discovery or point out great business opportunities, is locked away somewhere in this mountain of data, and an organization's ability to analyze and understand it falls far behind its ability to accumulate it. Now, more than ever before, they are in need of tools to help them make sense of all of the data they are gathering. Over the past 10 years, a fairly new breed of tools has been evolving to help meet these challenges. These tools fall under a data analysis field called Knowledge Discovery in Databases (KDD). The process of searching through data via these tools is referred to as Data Mining, and it is the focus of this paper. Although it is not meant to be an exhaustive discussion on this subject, this paper will provide a basic foundation by offering some background and definition of data mining and placing data mining in context within the knowledge discovery framework, by exploring reasons why data mining is needed, by discussing why data mining has moved from hype to mainstream, and lastly, by offering closing statements regarding any social issues surrounding data mining and future expectations. Nevertheless, the growing volumes of data will continue to outpace human capabilities that can only be addressed using automated methods such as data mining. It will be exciting to see where data mining will take us over the next decades.

2. Basic Data Mining Definition

What is data mining? To help understand, one needs to put it into a context of the broader framework within which it exists. As previously mentioned, data mining is just one part of the process of discovering useful knowledge from data, referred to collectively as Knowledge Discovery in Databases (KDD). In brief, KDD outlines a process of knowledge discovery as a series of iterative steps: 1) Data selection, 2) Data preparation, 3) Data transformation, 4) Data Mining, 5) Interpretation/Evaluation, and 6) Knowledge (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). From this perspective, data mining is viewed as one step within a continuous process, and it is dependent upon other steps of data preparation before it can be implemented. This is not to minimize the importance of data mining, but instead to place emphasis on the significance of good preparation to make data mining as successful as it can be. The output of data mining is not an end in itself, but rather serves to support the end goal of informed decisions through knowledge discovery. Any further discussion regarding the process of KDD is beyond the scope of this paper; however, it is worth noting that KDD benefits from research in fields such as "databases, machine learning, pattern recognition, statistics, artificial intelligence and reasoning with uncertainty, knowledge acquisition for expert systems, data visualization, machine discovery, scientific discovery, information retrieval, and high-performance computing. KDD software systems incorporate theories, algorithms, and methods from all of these fields." (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Therefore, data mining benefits from each of these areas of research, applying these algorithms to "fuel the

KDD process" (Soman, Diwakar, & Ajay, 2006).

Many definitions for data mining exist, but most agree on a basic concept that data mining is a "knowledge discovery" technique whose goal is to reveal useful and actionable information that may lead to knowledge (Hermiz, 1999). It incorporates the "application of specific algorithms for extracting patterns (models) from data" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Similarly, Thearling (2000) defines it as "an automated detection of relevant patterns in a database." Gedam (2000) refers to data mining techniques as "machine learning algorithms that find buried patterns in databases and report or act on those findings." Lastly, Moxon (1996) holds that "data mining is a set of techniques used in an automated approach to exhaustively explore and bring to the surface complex relationships in very large datasets." The list of definitions goes on and on, but regardless of which definition is preferred, there is broad agreement and emphasis on key terms such as **automated**, **patterns**, and **relationships**.

Some of the tools used for data mining are

- Artificial neural networks - Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Decision trees - Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
- Rule induction - The extraction of useful if-then rules from data based on statistical significance.
- Genetic algorithms - Optimization techniques based on the concepts of genetic combination, mutation, and natural selection.
- Nearest neighbor - A classification technique that classifies each record based on the records most similar to it in an historical database.

3. Why does Data Mining Exist?

Data by itself is of little value unless we can extract useful information and use this information to help with decision making, whether it be in business or biology. Searching through data for information is not new. Therefore, to help answer the question, "why does data mining exist?" we must first consider the alternative. The traditional method of turning data into knowledge "relies on manual analysis and interpretation" (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). The larger the data set being analyzed, the more impractical the task. With advancements in database technologies, modern software vendors have provided tools such as query and reporting tools, statistical analysis packages, and on-line analytical processing (OLAP) to help with these large volumes of data. However, a shortcoming of these types of tools is that they require the analyst or researcher to drive the questioning themselves. These types of tools are "verification-based" tools, used to "verify or refute a hypothesis" (Moxon, 1996). Moxon goes on to say that the "effectiveness of this verification-based analysis is limited by a number of factors, including the ability of the analyst to pose appropriate questions and quickly return results, manage the complexity of the 'attribute space' (i.e., number of variables), and to just think 'out of the box.'"

The following is a summary of benefits across different industries taken directly from the website article entitled, *Data Mining: The Benefits*:

Retail/Marketing

- *Identify buying behavior patterns from customers.*
- *Find associations among customer demographic characteristics.*
- *Predict which customers will respond to mailing.*

Banking

- *Detect patterns of fraudulent credit card usage.*
- *Identify "loyal" customers.*
- *Predict customers that are likely to change their credit card affiliation.*
- *Determine credit card spending by customer groups.*
- *Find hidden correlations between different financial indicators.*
- *Identify stocks trading rules from historical market data.*

Insurance and Health Care

- *Claims analysis - determine which medical procedures are claimed together.*
- *Predict which customers will buy new policies.*
- *Identify behavior patterns of risky customers.*
- *Identify fraudulent behavior.*

Transportation

- *Determine the distribution schedules among outlets.*
- *Analyze loading patterns.*

Medicine

- *Characterize patient behavior to predict office visits.*
- *Identify successful medical therapies for different illnesses.*

4. Prominent Data Mining Applications

To assist in understanding the benefits of data mining listed in the prior section, we provide assorted current research reports for selected prominent data mining applications below.

5. Retail/Marketing

In an article in 2004, Kohaviet, Mason, and Zheng reviewed a popular architecture (Blue Martini Software) and presented several challenges and lessons learned from their data mining experiences (e-commerce section) over four

years. The researchers provided a worthy list of the lessons learned from their data mining experiences with retail data mining. The model building that might identify insight includes six steps: 1) Mine data at tight granularity levels; 2) Handle leaks in predictive models; 3) Improve scalability; 4) Build a simple model first; 5) Use data mining suites; and 6) Peel the onion and validate results. In addition, they provide a few guidelines for sharing insights, developing models, and closing the loop to improve business: 1) Represent models visually for better insight; 2) Understand the importance of the development context; and 3) Create actionable models to close the loop. Finally, the researchers offered their top lessons learned: 1) Integrate data collection into operations to support analytics and experimentation; ii) Do not confuse yourself with the target users; and 3) Provide simple reports and visualization before building more complex models.

Marketing has become more prevalent in today's busy society; customer reviews play a key role in identifying buying trends of a population and determining what products companies offer. There are two types of analysis used when looking at customer reviews: descriptive and predictive. The descriptive analysis explains the customer's behavior and activities when they are making a purchase of goods or services with a company. Predictive analysis will give the behaviors and activities expected of this same customer. Most companies mine this data into a relational model, so most of the data from the customer reviews go into tables to determine buying trends and predict future behavior. However, there is a significant percentage, approximately 20% of the data, which does not fit the structure (Yaakub M. R., Li, Alami, & Peng, 2012). This makes the data inconclusive, because one-fifth of it is ignored. The researchers in this paper attempted to explore a way to integrate the structured data and unstructured data by restructuring a traditional CRM (Customer Relations Management) system, picking up the descriptors that were not currently included in the structured analysis. By including this unique feedback, the new CRM system was able to combine the personal data of the customer with product information and the feedback on the product to give manufacturers new information that can lead to the improvement of a product (Yaakub M. R., Li, Algarni , & Peng, 2012).

New trends are occurring in data mining as users continue to find new avenues for the data. Today with a boom in mobile technology, assumptions cannot be made as to the location of the data. The user's location adds additional information to help in decision-making. Recommender systems have been in use for a while, but now the user is in need of adding more information to the system to give their customer increased recommendations (Shabib & Krogstie, 2011). Products and information are generally mined based on user profiles, but with the growing use of mobile devices, researchers state that location should now be viewed as a vital part of recommender systems. Recommender systems vary, but typically use one of the two common types of filters: either content base filters, which filter information of users based on their likes or dislikes, or collaborative filters, which rely on the commonalities of the user and the data to find matches. Collaborative filters are the most commonly used in recommenders because of their reduced complexity and ability to offer more information by recognizing similar attributes. In this study, the researchers used a collaborative filter and designed a system for MovieLens, a site that recommends movies to its customers. The design of the system model for MovieLens allowed their customers to use not only their profile but also their location. The recommendation system was able to give recommendations not only of movies in theaters at that time, but could be more specific as to movies within a specified radius to their proximity. The conclusion was that the recommender model with the location filter has great promise and could lead to other uses for data mining requiring more specific information for the client.

Banking

Applications of data mining have become essential to the success and growth of banking and related enterprises. As a specific illustration, banks and other financial institution have always had the difficult task of determining a customer's credit-worthiness when a customer applied for a credit card or loan. Credit-scoring systems have been designed to aid this analysis. Ince and Aktan (2009) prepared an empirical study to conclude whether existing credit scoring systems would offer the best information to determine whether a customer would default or pay as deemed by the contract. Using nine predictors about bank customers, including gender, age, marital status, education, occupation, job position, income, credit card held from other institutions, and Weka data mining software, the researchers performed their study. The researchers applied each of four different types of credit card scoring, discriminant analysis, logistic regression, decision trees (C5; CART) and neural networks to evaluate banking customers for credit. The researchers found that CART had above average results in classifying, but neural networks provided the better overall results for credit scoring (Ince & Aktan, 2009).

In the article, "Online Portfolio Selection: A Survey", Li and Hoi (2014) discuss the use of machine learning and data mining to create wealth in an investment portfolio. The authors' focus was in the area of online portfolios. They discuss the two currently predominate techniques used to optimize portfolio selection: mean-variance theory and capital growth theory. Capital growth theory is the theory used most in online trading, so it is the one discussed most by the authors. This theory describes analysis over a period of time that will provide the best rate of growth for a portfolio while at the end of trade providing a balanced portfolio. The study focused on three important ideas. First, rather than looking at each individual asset in the portfolio, the authors urge analysts to look at the portfolio as a *total asset*. Second, information about the portfolio must by up-to-the-minute and sequential, and decisions must be made immediately based upon the data received. Lastly, Li and Hoi recommend

understanding the structure of algorithms and current systems of portfolio selection used in online trading. The authors concluded that by looking at capital growth theory along with several algorithms used for portfolio management (including Benchmarks, Follow the Winner, Follow the Loser, Pattern-Matching-based approaches and Meta-Learning Algorithms), researchers could open a broader field of study to determine better systems to use of portfolio selection in online trading.

Insurance and Healthcare

While standard data mining has commonly been used in insurance and healthcare fields, there is a rise in innovative usage of the technology, due to the high error rates of claims processing. In an article, Kumar, Ghani, & Mei, (2010) report that healthcare industry has seen extreme rise in healthcare rates in the last years due to payment errors in processing claims. These errors have caused a need to add approximately an additional 33% of administrative staff to rework or reprocess these clams. Additional cost must be passed on to the consumer in the form of a raise in rates. As of 2010, it was estimated that rates had risen by 131% over the prior decade. Data mining shows the promise of helping to reduce these costs, becoming an important component of slowing the growth of health care premiums. The researchers discussed their work with Accenture Claims Administration, a group that works directly with the largest insurance companies in the United States, to create a system called the Rework Prevention Tool to predict which claims might produce errors in processing and generate a report to a company auditor to help explain the nature of these errors. When applied to two large US health insurance companies, the Rework Prevention Tool was able to produce a better response than systems presently in use, saving the insurers an estimated $15 - $20 million in costs annually. This program also provided additional implications for further research, so this type of data mining system could be used industry wide.

Transportation

Liu, Yu, Ding, Ni, Wang & Shannon (2010) presented a new way to calibrate transportation simulation models through data mining to see where infrastructure needed to be built or modified to keep up with growing traffic demands. In order for these models to be accurate, researchers determined they needed to be calibrated to existing traffic data. The authors of the paper deployed sensor technology, geographic networking, and knowledge discovery to update simulations. The researchers created a prototype simulation using CORSIM (Corridor-Microscopic Simulation) to act as the simulation engine. They then used pattern mining, a specific classification of data mining, to find the "controlling factors" in the data. These factors allowed them to calibrate the traffic models to have a more accurate model to make their transportation decisions.

Medicine

The cost of medicine is increasing rapidly worldwide; as a result, data mining has become a critical factor in medicine and related fields. In a selected study, Gheorghe and Petre (2014) discussed the importance of using data mining in combination with telemedicine. By incorporating the two approaches, the medical community can not only provide services more rapidly by determining patterns in patient health to explore and decide on best treatment for a patient, but they can also monitor and analyze patient progress to see if any further treatment is implicated. By accessing vast amounts of patient data and contrasting this information against available data, Gheorghe and Petre develop a compelling model for reduction of errors in the care of patients. Telemedicine lets the patient and the physician communicate information without being in the same place and leverages video conferencing (patient to doctor, doctor to doctor, and doctor to researcher) in order to diagnose, monitor, research and educate. Using data mining along with telemedicine empowers researcher and physicians with vast amounts of data to analyze. This methodology provides a clearer picture for the treatment of patients and illuminates trends and patterns of diseases and treatments of disease allowing doctors to give the most up-to-date treatment to their patients.

Another area of medicine that benefits from data mining techniques is emergency services. The approach reported in the study by Khan, Doucette, Fu, Jin & Cohen (2011) was to use an ontological approach, defined as semantic data mining. Mining using established inference rules determines all applicable relationships in the databanks and provides the user with a list of information that fulfills the rule. Data is mined based on the queries of the medical practitioner. The practitioner poses questions in response to the information retrieved after each question, pulling information from the patient's history, drug interactions, and descriptions of different types of maladies. In a three-layer system, the data mined is based on the patient name, qualities of the symptoms and an assigned value given to the symptoms. The study shows promise for use of data mining to speed up emergency diagnosis and treatments, though there is a need to further refine the queries.

Education

Many institutions of higher education are looking for ways to determine the success or failure of their incoming students in order to better supply both the student and faculty with the necessary materials and activities to make students successful in their education (Osmanbegovic & Suljic, 2012). The authors surveyed incoming first year students. The information obtained from the surveys along with the participating students' grades were included in the data to be disseminated using both unsupervised and supervised algorithms to discover any patterns. The study deployed data mining techniques to derive information that could be applied to help students become more successful and to aid teachers improve the quality of instruction. The implications for further work were that more divergent characteristics could be added to improve the accuracy of the outcome.

Calders and Pechenizkiy (2012) discuss Educational Data Mining (EDM), an area of research that includes professionals from the areas of education, computer science, psychology, psychometrics, and statistics with the goal of discovering better methods and techniques to teach students. The researchers looked at the possible uses of data mining in education and found that EDM could give educators actionable information by clustering students into academic and collaborative groups based on previous performances, comparing and contrasting the behaviors of students who graduate and those who do not, analyzing the curriculum for student outcomes, and predicting future outcomes of student success. The articles also assessed textbooks, offering suggestions for improvements to the text; examining on-line learning environments as supporting tools for learning; evaluated test questions and test scores; and discussed assessments and the knowledge levels of test takers. EDM has a wide amount of information that can assist education professionals by enabling better choice of educational materials, analysis of student achievement, and the ability to cluster students in collaborative groups and classwork.

Emergency Management

During a crisis, there is a tremendous need for communication, not only between organizations within the public sector themselves, but also between the public and private sectors. If this communication is linked, then emergency efforts can provide a more effective response to the situation and help in rebuilding areas affected by a disaster. Zheng, Shen, Tang, et al (2010) discussed the implications and created a data-mining program to enable the public and private sectors to work cohesively in a disaster. The program prototype, called the Business Continuity Information Network (BCIN), consisted of developing a partnership between the two sectors and linking them together so that as a disaster unfolds, the partners will have real-time information on affected areas, fatalities, logistics, and available resources through data mining. The researchers developed a program after interviewing both sectors; the program was developed and put into action for three disasters in the Miami-Dade County area of Florida. The program enabled participants to send instant messages between themselves, report updated information, share information, and allow situational browsing including identifying resources needed by the various emergency crews in real time as they become available.

6. From Hype to Mainstream

Gregory Shapiro, one of the early pioneers in the field of KDD, stated that data mining has followed what he terms the "hype curve." In a 2006 article, he remembers that years ago, data mining had so much potential, but other factors limited its ability to deliver to its fullest; expectations were high, so users soon became disappointed. However, with advances that exist now among other technologies in other areas of computing such as processing speed, parallel processing, data storage access, network speeds, etc., data mining is finally able to live up to users' expectations and more (Shapiro, 2006). Soman, et al. (2006) suggest that there are actually five (5) trends, external to the data mining tools themselves, that can be identified as increasing the popularity of data mining today – taking it from the world of hype and thrusting it into the mainstream of analysis tools used by leading businesses and researchers today:

- Data Trends – As already covered, data volume is exploding, and Soman et al. (2006) suggest that it has grown six to ten times larger over the past 20 years.
- Hardware Trends – The type of statistical and analytical tools used in the data mining process require significant processing power. Only in recent years has this type of power existed (e.g. parallel processing).
- Network Trends – The speeds of the next generation Internet will make it possible to perform analysis in distributed environments where centralization of all the data is not practical or even impossible.
- Scientific Computing Trends – Scientists are placing more emphasis and attention on simulation today. Data mining will serve as a "bridge" between theory, experiment, and simulation.
- Business Trends – In today's competitive marketplace, businesses will look to data mining to help them to "more accurately predict opportunities and risk" (Soman, et al., 2006).

To this point, we have explored and established that data mining is used widely in a variety of organizations, but often, data mining does not provide all of the necessary information. The data frequently contains only the minimum amount of data required for data mining, which may not be enough information for a business analyst to make correct assumptions from the data. Palpanas (2012) performed a case study in which he and his colleagues created algorithms to fit frameworks that gathered information through data mining, "organizing the data according to different dimensions where each dimension represents a particular aspect of the objects represented in the data mining results" (Palpanas, 2012). By utilizing a framework designed for this study, a business analyst was able to interpret the data mining information from different aspects, giving the researchers better ways to compare and analyze their data.

Chen, Pavlov, Berkhin, Seetharaman, & Meltzer (2009), authors of an article titled "Practical Lessons of Data Mining in Yahoo," point to the fact that they have been using data mining for some time to categorize shoppers and their buying and searching behaviors. Because online shopping has become such a large industry, data must be assimilated and analyzed to gain information and an understanding of the consumer. The authors stated that many companies fail to utilize data mining to gain insight management of their company and to learn from their past mistakes. Because of the enormous amounts of data that are available to

companies, they must learn to use data mining to gain information. Gaining information is a multistep process that begins with data preprocessing. Preprocessing allows data to be located and loaded into data warehouses, where data size and sampling are determined. Only then can an organization understand data distribution, look at the data to gain a sense of understanding, determine modeling goals and finally, evaluate the model. A model takes a group effort to achieve. Product managers and scientists must communicate and define the business problem before the scientist can design a model that will generate revenue, and the production manager must be trained to understand and implement the model properly. Production engineers must also work closely with the scientist to provide the scope and sequence of productions capabilities. Finally, the operations and reporting teams act in the final leg, when the model is built, tested, and instituted so they can give direct communication between the scientist, engineers, and operating team to determine whether the model works as designed or if changes must be made to provide optimal impact on company revenue. If all parties work together, data mining can provide all parties a working partnership in a successful business.

7. Implications for the Future

Quite often, information from data mining used to make decisions and research findings are frequently biased due to missing bits of data. Tremblay, Dutta, and Vandermeer (2010) attempted to explore the possibility of estimating the amount of missing data in a dataset to determine its validity. They stated there are two types of missing data: Missing at Random (MAR) and Missing Not at Random (MNAR). MAR is much harder to isolate, because a pattern in the missing data cannot be found, unlike MNAR, in which patterns can exist. In an exploratory study, these researchers conducted cluster analysis and created decision trees to determine the rules for data. From this, they were able to define the rules or patterns of the missing data. The decision tree seemed to have the best results, but only in smaller data groups. The study indicates that further testing need to be completed to work with larger data groups.

In an article by Silva and Antunes (2013), researchers described the need for more constraints in data streaming. The authors of this study defined data streaming as "an ordered sequence of instances that are continuously being generated and collected." It was their contention that in order to get the best extraction of pertinent information, there must have been constraints to identify the patterns in the data streaming. They experimented with the use of pattern-trees updated with regular frequency. Authors found that by updating and maintaining the pattern-tree, the data that was returned was smaller and more pertinent and that this process took less time to achieve.

Hoffman (2010) discussed ways in which companies could capitalize on employee productivity by data mining their communications. Hoffman talks about the link that companies are finding between their higher performing employees and the amount of communication the employee performs in her position. At present, Microsoft uses a technique to data mine their internal company communications to see which employees are sharing ideas and information. Google uses a similar approach to identify employees who may be underused and could be candidates to change positions or leave the company for more fulfilling jobs. IBM has probably harnessed the use of data mining within the company on the grandest scale. They have developed SmallBlue, a data mining system which analyzes and maps the electronic data of their employees creating a shorter "social Path" to other experts they may need to collaborate with. By linking 3 years of data, including 20 million emails and instant messages, 2 million blogs and database entries, and millions of data points from knowledge sharing and learning activities, IBM employees can reach out to find expert help with a problem, or a manager can assemble a dynamic project team by using information available in this data mining system.

8. Conclusions

Even though there is a tremendous amount of interest and excitement regarding the use of data mining today, there are still some areas for concern regarding data mining's use of personal data and the potential risk to personal privacy. Most privacy-related laws protect an individual's rights by declaring that data cannot be used for any purpose other than what was originally agreed to by the individual. With data mining, the use of the information is undetermined until after the data is mined, so this is an area where the laws will need to be reviewed for relevance. Another area of concern is the accuracy of the results from data mining. Inaccurate results may result in serious repercussions toward an individual or an organization. Laws may be challenged in this area as well (Papasliotis, 2005). With these types of concerns, Papasliotis suggests that there might even be an eventual need to regulate data mining to the point where projects need to be licensed, and that "by monitoring the use of new information generated by licensed data mining projects," a regulating body would have to be established and staffed by the government. Without proper regulatory safeguards in place, Papasliotis fears that there is too much of a risk that data mining could jeopardize our very freedoms and without being subject to the same "judicial or procedural constraints" that other forms of surveillance are limited (Papasliotis, 2005).

REFERENCES

[1] Calders, T., & Pechenizkiy, M. (2012.). Introduction to the special section on educational data mining. ACM SIGKDD

Explorations Newsletter, 13(2), 3-6.

[2] Chen, Y., Pavlov, D., Berkhin, P., Seetharaman, A., & Meltzer, A. (2009). Practical lessons of data mining at Yahoo! Proceedings of the 18th ACM conference on Information and knowledge management, 1047-1056.

[3] Data Mining: The Benefits. Retrieved (2008, December 4) from http://www.data-mining-software.net/Data-Mining-Business-Benefits.shtm

[4] Dublin, 50-56. Soman, K. P., Diwakar, S., & Ajay, V. (2006). Insight into Data Mining. New Delhi, India: Prentise-Hall of India Private Limited.

[5] Fayyad, U., Piatetsky-Shapiro, G., Smyth , P. (1996), The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM, 39(11) , 27-34.

[6] Gheorghe, H., & Petre, R. (2014). Integrating Data Mining Techniques into Telemedicine Systems. Informatica Economica. 18(1), 120-130.

[7] Gedam, D. R. (2000). Data Mining on the Web: There is Gold in that Mountain of Data. Web Techniques, January 5(1).

[8] Hermiz, K. B. (1999). Critical Success Factors for Data Mining Projects. DM Review, February.

[9] Hoffmann, L. (2010). Mine your business. Communications of the ACM. 53(6), 18-19.

[10] Ince, H., & Aktan, B. (2009). A Comparison of Data Mining Techniques for Credit Scoring In Banking: A Managerial Perspective. Journal of Business Economics & Management. 10(3), 233-240.

[11] Javid, S. (1999). Data Mining in the Next Millennium. DM Review, November.

[12] Khan, A., Doucette, J. A., Fu, L., Jin, C., & Cohen, R. (2011). An Ontological Approach to Data Mining For Emergency Medicine. Proceedings for the Northeast Region Decision Sciences Institute (NEDSI), 578-585.

[13] Kohavi, R., Mason, L., Parekh, R... & Zheng, Z. (2004). Lessons and Challenges from Mining Retail E-Commerce Data. *Journal of Machine Learning,* Kluwer Academic Publishers, 57(1-2), 83-113.

[14] Kumar, M., Ghani, R., & Mei, Z.-S. (2010). Data mining to predict and prevent errors in health insurance claims processing. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 65-74.

[15] Li, B., & Hoi, S. C. (2014). Online portfolio selection: A survey. ACM Computing Surveys (CSUR), 46(3), 1-36.

[16] Liu, H., Yu, Q., Ding, W., Ni, D., Wang, H., & Shannon, S. (2011). Feasibility Study for Automatic Calibration of Transportation Simulation Models. Proceeding of the 44th Annual Simulation Symposium, 87-94.

[17] Moxon, B. (1996). Defining Data Mining: the Hows and Whys of Data Mining and How It Differs from Other Analytic Techniques. DBMS, August.

[18] Osmanbegovic, E., & Suljic, M. (2012). Data Mining Approach For Predicting Student Performance. Economic Review: Journal of Economics & Business, 3-12.

[19] Palpanas, T. (2012). A Knowledge Mining Framework for Business Analysts. The DATA BASE for Advances in Information Systems, ACM SIGMIS Database, 43(1), 46-60.

[20] Papasliotis, I-E., 2005. Mining for data and personal privacy: reflections on an impasse. In Proceedings of the 4th international Symposium on information and Communication Technologies (Cape Town, South Africa, January 03 - 06, 2005). ACM International Conference Proceeding Series, 92(1). Trinity College, 50-56.

[21] Shabib, N., & Krogstie, J. (2011). The Use of Data Mining techniques in location-based Recommender System. Proceedings of the International Conference on Web Intelligence, Mining and Semantics, Article 28.

[22] Silva, A., & Antunes, C. (2013). Pushing constraints into data streams. BigMine '13 Proceedings of the 2nd International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, 79-86.

[23] Thearling, K. (2000). Data Mining and CRM: Zeroing in on Your Best Customers. DM Review, December.

[24] Tremblay, M. C., Dutta, K., & Vandermeer, D. (2010). Using Data Mining Techniques to Discover Bias Patterns in Missing Data. ACM Journal of Data and Information Quality. 2(1). Article 2.

[25] Wahlstrom, K., Roddick, J. F., On the impact of knowledge discovery and data mining, Selected papers from the second Australian Institute conference on Computer ethics, November 01, 2000, Canberra, Australia. 22-27.

[26] Yaakub, M. R., Li, Y., Algarni, A., & Peng, B. (2012). Integration of Opinion into Customer Analysis Model. IEEE/WIC/ACM International Conferences on Web Intelligence & Intelligent Agent Technology, 3, 164-168.

[27] Zheng, L., Shen, C., Tang, L., Li, T., Luis, S., Chen, S-C., & Hristidis, V. (2010). Using data mining techniques to address critical information exchange needs in disaster affected public-private networks. Proceedings of the 16th ACM SIGKDD International Conference: Knowledge Discovery & Data Mining, 125-134.