International Journal of advanced studies in Computer Science and Engineering
IJASCSE, Volume 3, Issue 2, 2014

Industry Wide Applications of Data Mining

Eze, Udoka Felista (PhD)[1] , Adeoye, Olufemi Sunday[2], Ikemelu, Chinelo Rose-keziah[3]
[1]Department of Information Technology, Federal University of Technology, Owerri, Nigeria.
[2]Department of Computer Science, University of Uyo, Nigeria
[3]Department of Computer Science, Nwafor Orizu College of Education, Onitsha, Nigeria.

ABSTRACT - The field of data mining has been growing in leaps and bounds, and has shown great potential for the future. Companies in a wide range of industries –including retail, finance, health care, telecommunications, transportation, and aerospace – are already using data mining tools and techniques to take advantage of historical data. This paper examines the wide application domain of data mining in industry where data is generated. It is then discovered that data mining is one of the most important frontiers in database and information systems and also one of the most promising interdisciplinary development in Information Technology.

Keywords - Data Mining, Telecommunications, Retail Industry, Financial Sector, Biological Data Analysis, Intrusion Detection

1. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection and customer retention, to production control and science exploration [8]. Data mining being a young discipline has attracted several definitions. Simply stated, data mining refers to extracting or mining knowledge from large amounts of data. From a data warehouse perspective, data mining can be viewed as an advanced stage of on-line analytical processing (OLAP).

However, data mining goes far beyond the narrow scope of summarization-style analytical processing of data warehouse systems by incorporating more advanced techniques for data analysis. Data mining involves an integration of techniques from multiple disciplines such as database and data warehouse technology, statistics, machine learning, high performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial or temporal data analysis [8]. As a young research field, data mining has made broad and significant progress since its early beginnings in the 1980s. Today, data mining is used in a vast array of areas, and numerous commercial data mining systems are available.

Data mining is primarily used by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among “internal” factors such as price, product positioning, or staff skills, and “external” factors such as economic indicators, competition, and consumer demographics. It enables them to determine the impact on sales, customer satisfaction, and corporate profits. Besides, it enables them to “drill down” into summary information to view detail transactional data.

Data mining is a synonym for another popularly used term “Knowledge Discovery in Databases” or KDD. Knowledge discovery is an iterative process consisting of data cleaning, to remove noisy and inconsistent data, data integration, to combine multiple heterogeneous or homogeneous data sources, data selection, to consider only data relevant to the task and data transformation where data is transformed into forms appropriate for mining functions such as aggregation or summarization.

Then data mining algorithms are employed to extract interesting and meaningful patterns from the data and present the knowledge to the domain expert in an informative manner. Based on this, it is intuitive that the typical data mining system has a multi-tiered architecture as shown in Fig 1.

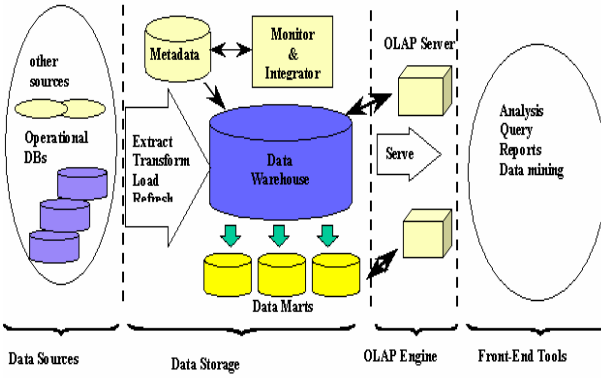

Figure 1: Architecture of a Data Mining System
Source: CSE 300: Topics in Biomedical Informatics, Data Mining and its Applications and Usage in Medicine, Radhika, Spring 2008

Data from a set of databases, data warehouses, spreadsheets or other information repositories form the first tier. Data cleaning and integration techniques maybe performed on the data to make it more tuned for the user queries. A database or data warehouse server is then responsible for fetching the relevant data from the database based on the user's mining request. A knowledge base supports the data mining engine that processes the user queries. This is the domain knowledge that will guide the search or evaluate the resulting patterns for knowledge.

It can include concept hierarchies which are used to organize attributes and attribute values into different levels of abstraction. Domain knowledge can also include additional interestingness constraints and threshold values as well as metadata describing the data from multiple heterogeneous sources. The data mining or OLAP engine consists of a set of modules which contain the algorithms for different types of mining techniques such as association rule mining, classification and prediction or clustering. The front end of the system will contain the pattern evaluation module and the graphical user interface which will represent the mined data in easily to understand visualized forms such as graphs and figures.

II. THE DATA MINING PROCESS

Data mining is an iterative process that typically involves the following phases:

Problem definition

A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required.

Data exploration

Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data.

Data preparation

Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed.

Modeling

Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built.

Evaluation

Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract

Business explanations and evaluate the following questions:

Does the model achieve the business objective?

Have all business issues been considered?

At the end of the evaluation phase, the data mining experts decide how to use the data mining results.

Deployment

Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets. The Intelligent Mine products assist you to follow this process. You can apply the functions of the Intelligent Miner products independently, iteratively, or in combination. The following figure shows the phases of the Cross Industry Standard Process for data mining (CRISP DM) process model.

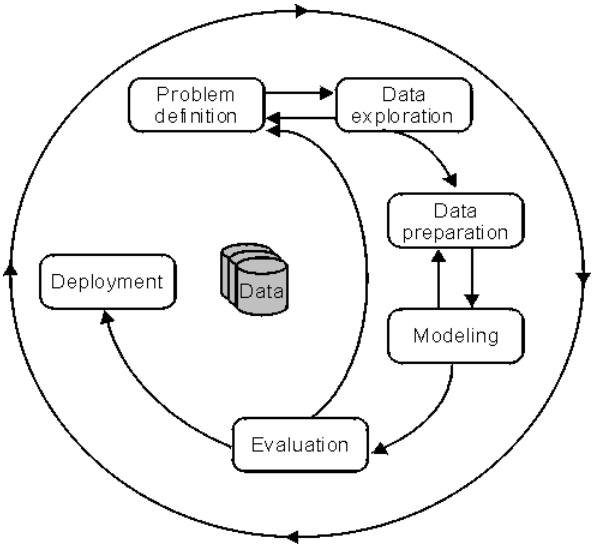

Fig. 2 Phases of the CRISP-DM process (Chapman et al, 2000)

Intelligent Mine (IM) Modeling helps you to select the input data, explore the data, transform the data, and mine the data. With IM Visualization you can display the data mining results to analyze and interpret them. With IM Scoring, you can apply the model that you have created with IM Modeling [16].

III. APPLICATIONS OF DATA MINING

3.1Data mining in telecommunications

The telecommunication industry generates and stores a tremendous amount of data. These data include call detail data, which describes the calls that traverse the telecommunication networks, network data, which describes the state of the hardware and software components in the network and customer data, such as billing information, as well as information obtained from outside parties such as credit score information, which describes the

Telecommunication customers:

This information can be quite useful and often is combined with telecommunication-specific data to improve the results of data mining. For example, while call detail data can be used to identify suspicious calling patterns, a customer's credit score is often incorporated into the

analysis before determining the likelihood that fraud is actually taking place.

The amount of data is so great that manual analysis of the data is difficult, if not impossible. The need to handle such large volumes of data led to the development of knowledge-based expert systems. These automated systems performed important functions such as identifying fraudulent phone calls and identifying network faults. The problem with this approach is that it is time consuming to obtain the knowledge from human experts (the "knowledge acquisition bottleneck") and, in many cases, the experts do not have the requisite knowledge. The advent of data mining technology promised solutions to these problems and for this reason the telecommunications industry was an early adopter of data mining technology [7].

Data mining in telecommunication industry helps to understand the business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources, and improve the quality of service. A large class of Data Mining algorithms developed for this purpose includes CART, C4.5, neural networks, and Bayesian classifiers, among others. One of the assumptions made by these algorithms, which are carried over into data mining applications is that of clean data.

The companies in the telecommunication industry face the problem of churning - the process of customer turnover. This is a major concern for the companies having many customers who can easily switch to other competitors. Data mining helps to do appropriate credit scoring and to combat churns in the telecom industry. Data mining can be used in chum analysis to perform two key tasks; predict whether a particular customer will churn and when it will happen; understand why particular customer churn [14].

3.2 Data mining in retail industry

The retail industry is a major application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service. The quantity of data collected continues to expand rapidly, especially due to the increasing ease, availability, and popularity of business conducted on the web or e-commerce [8]. Today, many stores also have websites where customers can make purchases online. Retail data provide a rich source for data mining. Data mining can help spot sales trends, develop smarter marketing campaigns, and accurately predict customer loyalty.

Specific uses of data mining in retail industry include but are not limited to:
Market Segmentation - identify the common characteristics of customers who buy the same products from your company.
Customer Churn - predict which customers are likely to leave your company and go to a competitor.
Fraud Detection - identify which transactions are most likely to be fraudulent.
Direct Marketing - identify which prospects should be included in a mailing list to obtain the highest response rate.
Interactive Marketing - predict what each individual accessing a web site is most likely interested in seeing.
Market Basket Analysis - understand what products or services are commonly purchased together; e.g. beer and diapers
Trend Analysis - reveal the difference between a typical customer this month and last.
Data mining technology can generate new business opportunities by:
Automated prediction of trends and behaviours –

Data mining automates the process of finding predictive information in a large database. Questions that traditionally required extensive hands-on analysis can now be directly answered from the data. A typical example of a predictive problem is target marketing. Data mining uses

data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings.

Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Automated discovery of previously unknown patterns: Data mining tools sweep through databases and identify previously hidden patterns. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Using massively parallel computers, companies dig through volumes of data to discover patterns about their customers and products.

3.3 DATA MINING IN FINANCIAL SECTOR

Data mining is worthwhile in the banking industry. Date mining assists the banks in order to search for hidden pattern in a group and determine unknown relationship in the data. Bank has detail data about all the clients. The client data contains personal data that describes the financial status and the financial behavior before and by the time the client was given the credit. Forecasting stock market, currency exchange rate, bank bankruptcies, understanding and managing financial risk, trading futures, credit rating, loan management, bank customer profiling, and money laundering analyses are core financial tasks for data mining [11]. Some of these tasks such as bank customer profiling have many similarities with data mining for customer profiling in other fields [3].

Stock market forecasting includes uncovering market trends, planning investment strategies, identifying the last time to purchase the stocks and what stocks to purchase. Financial institutions produce huge datasets that build a foundations for approaching these enormously complex and dynamic problems with data mining tools.

Data mining as a process of discovery useful patterns, correlations has its own niche in financial modeling. Similarly to other computational methods almost every data mining method and technique can be used in financial modeling.

3.4 DATA MINING IN HEALTH CARE.

When it comes to the challenges faced by hospital and other medical facilities there is no better prescription than data mining for health care. The unpredictable ebb and flow of Medicare funding have made it absolutely imperative for health care facilities to provide the best service at all times. Data mining for healthcare is useful in evaluating the effectiveness of medical treatments. Through comparing and contrasting various causes, symptoms and treatment methodologies, data mining ca produce an analysis of which treatment correct specific symptoms most effectively.

Data mining can also help physicians discover which medications are the most cost-efficient while still working effectively. Other data mining applications could be used to associate the most common side-effects of a medication, to collate typical symptoms in order to improve the accuracy of diagnosis, or to discover proactive steps to reduce the risk of affliction.

In order to improve healthcare management, data mining applications are able to work to identify and track high-risk patients in order to design appropriate interventions as a means to lower the number of admissions, re-admissions, and claims. In other cases, data mining for healthcare has been used to decrease patient length-of-stay,

avoid medical complications, improve patient outcomes, hospital infection control and early warning systems, etc. By understanding patient preferences, patterns, and characteristics, you can significantly improve their level of satisfaction.

Data mining is also helpful in funding patterns amongst patients surveyed as a means of setting reasonable wait time expectations, discovering what patients want from their health care providers and finding ways to improve services. Another advantage to data mining for health care is the ability to detect and decrease insurance fraud. Data mining applications are able to establish norms and then identity any abnormal patterns and claims in order to eliminate inappropriate prescriptions or referrals, and fraudulent medical claims. With the right tolls, the data mining for health care has significantly impact the Medicare quality rating and overall budget [17].

3.5 DATA MINING IN BIOLOGICAL DATA ANALYSIS

Biological data are data or measurements collected from biological sources, which are often stored or exchanged in a digital form. Biological data are commonly stored in files or databases. Examples of biological data are DNA base-pair sequences, and population data used in ecology. DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four bases—adenine (A), guanine (G), cytosine (C), and thymine (T) - in a strand of DNA.

The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery. These four nucleotides are combined to form long sequences or chain that resembled a twisted ladder. Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as diagnostic, biotechnology, forensic biology, and biological systematic. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species.

It has been revealed that human beings have around 100,000 genes. A gene is usually composed of hundreds of individual nucleotides arranged in a particular order. There are almost an unlimited number of ways that the nucleotides can be ordered and sequenced to form distinct genes. It is challenges to identify particular gene sequence patterns that play roles in various diseases. Since many interesting sequential pattern analysis and similarity search techniques have been developed in data mining, data mining has become a powerful tool and contributes substantially to DNA analysis in the following ways.

Semantic integration of heterogeneous, distributed genomic and proteomic databases: Due to the highly distributed, uncontrolled generation and use of a wide variety of DNA data, the semantic integration of such heterogeneous and wide variety of distributed genome databases become an important task for systematic DNA coordinated analysis of DNA databases. This has promoted the development of integrated data warehouses and distributed federated databases to store and manages the primary and derived genetic data. Data cleaning and data integration methods developed in data mining will help the integration of genetic data and the construction of data warehouses for genetic data analysis.

Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequence: One of the most important search

problem in genetic analysis is similarity search and comparison among DNA sequences. Gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes. This can be done by first retrieving the gene sequences from the two tissue classes, and then finding and comparing the frequency in the diseased samples than in the healthy samples might indicate the genetic factors of the disease; on the other hand, those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. The analysis of frequent sequential patterns is important in the analysis of similarity and dissimilarity in genetic sequences.

Association and Path Analysis: Currently, many studies have focused on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes and the study of interactions and relationships between them.

While a group of genes may contribute to a disease process, different genes may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible develop pharmaceutical

interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analysis is expected to play an important role in genetic studies.

Visualization tools in genetic data analysis: Complex structures and sequencing patterns of genes are most effectively presented in graphs, trees, cuboids, and chains by various kinds of visualization tools. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization therefore plays an important role in biomedical data mining.

3.6 DATA MINING IN OTHER SCIENTIFIC APPLICATIONS

Scientific data mining is defined as data mining applied to scientific problems, rather than database marketing, finance, or business-driven applications. Scientific data mining distinguishes itself in the sense that the nature of the datasets is often very different from traditional market-driven data mining applications. The datasets now might involve vast amounts of precise and continuous data, and accounting for underlying system nonlinearities can be extremely challenging from a machine learning point of view [6][13]. Scientific data mining is an interactive and iterative process involving data pre-processing, search for patterns, knowledge evaluation, and possible refinement of the process based on input from domain experts or feedback from one of the steps.

The pre-processing of the data is a time-consuming, but critical, first step in the data mining process. It is often domain and application dependent; however, several techniques developed in the context of one application or domain can be applied to other applications and domains as well. The pattern recognition step is usually independent of the domain or application

Large-scale scientific data mining is a very challenging field, making it a source of several open research problems. In order to extend data mining techniques to large-scale data, several barriers must be overcome. The extraction of key

features from large, multi-dimensional, complex data is a critical issue that must be addressed:

first, prior to the application of the pattern recognition algorithms. The features extracted

must be relevant to the problem, insensitive to small changes in the data, and invariant to scaling, rotation, and translation. In addition, we need to select discriminating features through appropriate dimension reduction techniques. The pattern recognition step poses several challenges as well. For example, is it possible to modify existing algorithms, or design new ones, that are scalable, robust, accurate, and interpretable? Further, can these algorithms be applied effectively and efficiently to complex, multi-dimensional data? And, is it possible to implement these algorithms efficiently on large-scale multiprocessor systems so that a scientist can interactively explore and analyze the data?

While these problems must be overcome for large-scale data mining to be applied in any domain, certain additional concerns must be addressed for scientific data. For example, data from science applications are often available as images, a format that is known to pose serious challenges in the extraction of features. Further, problems in knowledge discovery may be such that the class of interest occurs with low probability, making random sampling inapplicable and traditional clustering techniques ineffective. In many cases, there may be a scarcity of labeled data in a classification problem and several iterations of the data mining process may be required to obtain a reasonable sized training set.

Some applications, such as remote sensing, may need data fusion techniques to mine the data collected by several different sensors, at different resolutions. Another key feature in which data mining applied to science applications differs from its commercial counterpart is that high accuracy and precision are required in prediction and description in order to test or refute competing theories. These problems, specific to scientific data sets, preclude the direct application of software and techniques that have been developed for commercial applications.

3.7 DATA MINING IN INTRUSION DETECTION

Intrusion refers to any kind of action that threatens integrity, confidentiality, or availability of network resources. In today's world where nearly every company is dependent on the Internet to survive, it is not surprising that the role of network intrusion detection has grown so rapidly. While there may still be some arguments as to what is the best way to protect a company's networks (i.e. firewalls, patches, intrusion detection, training,…) it is certain that the intrusion detection system (IDS) will likely maintain an important role in providing for a secure network architecture.

In order for one to determine how data mining can help advance intrusion detection it is important to understand how current IDS work to identify an intrusion. There are two different approaches to intrusion detection: (i) misuse detection and (ii) anomaly detection.

Misuse detection is the ability to identify intrusion based on a known pattern for the malicious activity. These known patterns are referred to as signatures. The second approach, anomaly detection, is the attempt to identify malicious traffic based on deviations from established normal network traffic patterns. Most, if not all, IDS which can be purchased today are based on misuse detection [9]. Current IDS products come with a large set of signatures which have been identified as unique to a particular vulnerability or exploit. While the ability to develop and use signatures to detection attacks is a useful and viable approach there are shortfalls to only using this approach. These shortfalls include Variants, False positives, False negatives, and Data overload.

Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. By identifying bounds for valid network activity, data mining will aid an analyst

in his/her ability to distinguish attack activity from common everyday traffic on the network.

VARIANTS: Since anomaly detection is not based on pre-defined signatures, the concern with variants in the code of an exploit are not as great since we are looking for abnormal activity versus a unique signature. An example might be a remote procedure call (RPC) butter overflow exploit whose code has been modified slightly to evade an IDS using signatures. With anomaly detection, the activity would be flagged since the destination machine has never seen an RPC connection attempt and the source IP was never seen connecting to the network.

FALSE POSITIVES: In regards to false positives there has been some work to determine if data mining can be used to identify recurring sequence of alarms in order to help identify valid network activity which can be filtered out.

FALSE NEGATIVES: Detecting attacks for which there are no known signatures. By attempting to establish pattern for normal activity and identifying that activity which lies outside identified bounds, attacks for which signatures have not been developed might be detected. An extremely simple example of how this would work would be to take a web server and develop a profile of the network activity seen to and from the system. Let us say the web server is locked down and only connections to ports 25 and 110 are ever seen to the server. Thus, whenever a connection to a port other than 25 or 110 is seen the IDS should identify that as an anomaly. This example could be extended to profiling not only individual hosts, but entire networks, users traffic based on days of the week or hours in a day, and the list goes on.

DATA OVERLOAD: The area where data mining is sure to play vital role is in the area of data reduction. With current data mining algorithms there exists the capability to identify or extract data which is most relevant and provide analysts with different "views" of the data to aid in their analysis [9].

IV. CONCLUSION

Data mining is a promising discipline and has wide applicability. It can be applied in various domains. Data mining as the confluence of multiple intertwined disciplines, including statistic, machine learning pattern recognition, database system, information retrieval world wide web, visualization, and many other application domains, has made great progress in the past decade. In this paper we have discussed the various industry wide applications of data mining. We also believe that more other areas of application will evolve in the future.

REFERENCES:

[1] Arun K. Pujari (2001): Data Mining Techniques, University Press, Chicago.

[2] Alex Berson, Smith J. Stephen (2004): Data Warehousing Data Mining and OLAP, McGraw-Hill Education, India.

[3] Berka, P. PKDD Discovery challenge on financial Data, IN: Proceeding of the First International Workshop On Data Mining Lessons Learned, (DMLL-2002) 8-12 July 2002, Sydeney, Australia.

[4] Boris Kovalerchuk: Data Mning for Financial Applications, Central Washington University, USA

[5] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000), CRISP-DM 1.0, Chikago, IL. SPSS.

[6] Curt M. Breneman, Kristin P. Bennett, Mark Embrechts, Steven Cramer, Minghu Song, and Jinbo Bi [2003] "Descriptor Generation, Selection and Model Building in Quantitative Structure-Property Analysis" Chapter 11 in

QSAR Developments, James N. Crawse, Ed., John Wiley.

[7] Gary M. Weiss (2013): Data Mining in Telecommunications, Department of Computer and Information Science, Fordham University.

[8] Jiawei Han and Michelien Kumber (2006): Data Mining (Concepts and Techniques), Morgan Kaufmann Publishers, London.

[9] Manh Phung,(October 24, 2000): Intrusion Detection FAQ: Data Mining in Intrusion Detection.

[10] Margaret H. Dunliam (2006): Data Mining Introductory and Advanced Topics, Pearson Education.

[11] Nakhaeizadeh, G, Steurer, E., Bartmae, K; Banking and Finance, In: Klosgen WL, Zytkow J. Handbook of data mining and knowledge discovery, oxford university press, Oxford, 2002, 771 -780

[12] Raph Kimball (2010): The Data Warehouse Lifecycle Booklet, Wiley Student Edition.

[13] Robert H. Kewley, and Mark J. Embrechts [2000] "Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks," IEEE Transactions on Neural Networks, Vol.11 (3), pp. 668-679.

[14] S. Sumathi, S. N. Sivanandam (2006): Introduction to Data Mining and its Applications, Springer Berlin Heidelberg, New York.

[15] Data Mining and Text Mining (UIC 583) @ Politecnico di Milano Unpublished.

[16] Data Mining at a Glance @ Source:http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=%2Fcom.ibm.im.easy.doc%2Fc_dm_process.html

[17] http://www.ikanow.com/blog/02/21/-datamining-for-healthcare-proven-remecy-for-an-ail-industry.