A Brief Overview on Data Mining Survey

Adarsh Goyal[1] Anjali Jaiswal[2] Nawaz Kapadia[3] Varun Gadani[4] Atharva Dalvi[5]
[1,2,3,4,5]Student
[1,2,3,4,5]Department of Computer Engineering
[1,2,3,4,5]Thakur Polytechnic, Mumbai, Maharashtra, India

Abstract— This paper provides an introduction to the fundamental concept of information mining. which provides overview of information mining is employed to extract meaningful information and to develop significant relationships among variables stored in large data set/data warehouse. within the case study reported during this paper, an information mining approach is applied to extract knowledge from an information set. data processing is that the process of discovering potentially useful, interesting, and previously unknown patterns from an oversized collection of information. Data mining could be a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, AI, high-performance computing, and data visualization. We present techniques for the invention of patterns hidden in large data sets, that specialize in issues with reference to their feasibility, usefulness, effectiveness, and scalability. The automated, prospective analyses offered by data processing move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

Keywords: Data mining; Association rules; Clustering; k-means; Decision tree

I. INTRODUCTION

Data mining may be a process to extract the implicit information and knowledge which is potentially useful and folks don't know prior to, and this extraction is from the mass, incomplete, noisy, fuzzy and random data [2].The essential difference between the info mining and also the traditional data analysis (such as query, reporting and on-line application of analysis) is that the info mining is to mine information and find out knowledge on the premise of no clear assumption [1].In addition to industry driven demand for standards and interoperability, professional and academic activity have also made considerable contributions to the evolution of the methods and models; a commentary published during a 2008 issue of the International Journal of knowledge Technology and deciding summaries the results of a literature survey which traces and analyzes this evolution.[8]

Data mining is that the use of automated data analysis techniques to uncover previously undetected relationships among data items. data processing often involves the analysis of knowledge stored during a data warehouse. Three of the foremost data processing techniques are regression, classification and clustering. Data Mining, also popularly referred to as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data processing and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data processing is really a part of the knowledge discovery process. the subsequent figure (Figure 1.1) shows data processing as a step in an iterative knowledge discovery process.

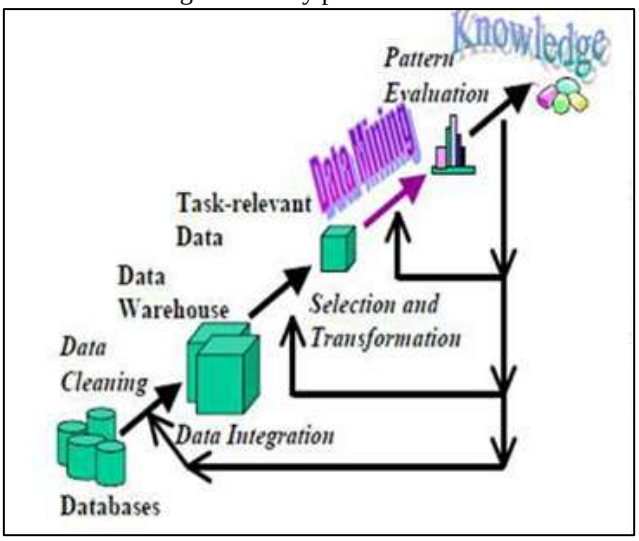


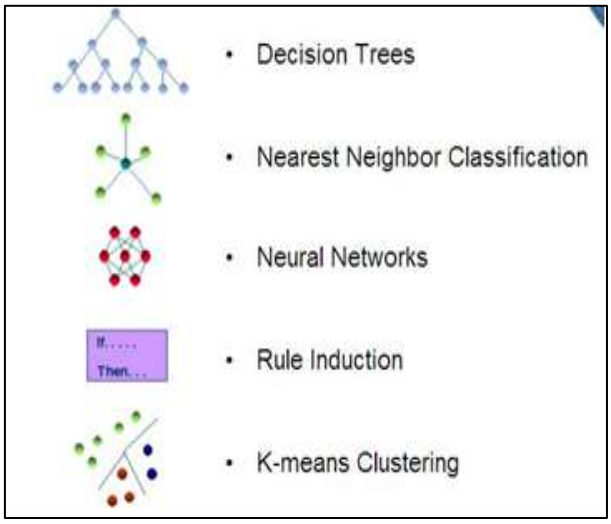
Fig. 1: Data mining is the core of Knowledge Discovery Process

The iterative process consists of the subsequent steps: Data cleaning: also referred to as data cleansing, it's a innovate which noise data and irrelevant data are aloof from the gathering. Data integration: at this stage, multiple data sources, often heterogeneous, is also combined in an exceedingly common source.

Data selection: at this step, the info relevant to the analysis is determined on and retrieved from the info collection. Data transformation: also referred to as data consolidation, it's a innovate which the chosen data is transformed into forms appropriate for the mining procedure. Data mining: it's the crucial step during which clever techniques are applied to extract patterns potentially useful. Pattern evaluation: during this step, strictly interesting patterns representing knowledge are identified supported given measures.

Knowledge representation: is that the final innovates which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to assist users understand and interpret the info mining results. It is common to mix a number of these steps together. as an example, data cleaning and data integration will be performed together as a pre-processing phase to come up with an information warehouse. Data selection and data transformation may be combined where the consolidation of the info is that the results of the choice, or, as for the case of information warehouses, the choice is completed on transformed data.

Data Mining is….



Data mining commonly involves four classes of tasks: [3].Clustering - is that the task of discovering groups and structures within the data that are in how or another "similar", without using known structures within the data. Clustering could be a data processing (machine learning) technique accustomed place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering. Classification - is that the task of generalizing known structure to use to new data. for instance, an email program might arrange to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.

Working with categorical data or a mix of continuous numeric and categorical data? Classification analysis might fit your needs well. this system is capable of processing a wider kind of data than regression and is growing in popularity. Regression - Attempts to search out a function which models the information with the smallest amount error. Regression is that the oldest and most well-known statistical technique that the information mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that matches the information. When you're able to use the results to predict future behavior, you just take your new data, plug it into the developed formula and you've got a prediction! the main limitation of this system is that it only works well with continuous quantitative data (like weight, speed or age). If you're working with categorical data where order isn't significant (like color, name or gender) you're at an advantage choosing another technique. Regression could be a data processing (machine learning) technique accustomed fit an equation to a dataset. the best style of regression, regression toward the mean, uses the formula of a line ($y = mx + b$) and determines the acceptable values for m and b to predict the worth of y based upon a given value of x. Advanced techniques, like multiple correlation, allow the utilization of over one input variable and permit for the fitting of more complex models, like a equation.

II. DATA MINING: CONVERGENCE OF THREE TECHNOLOGIES

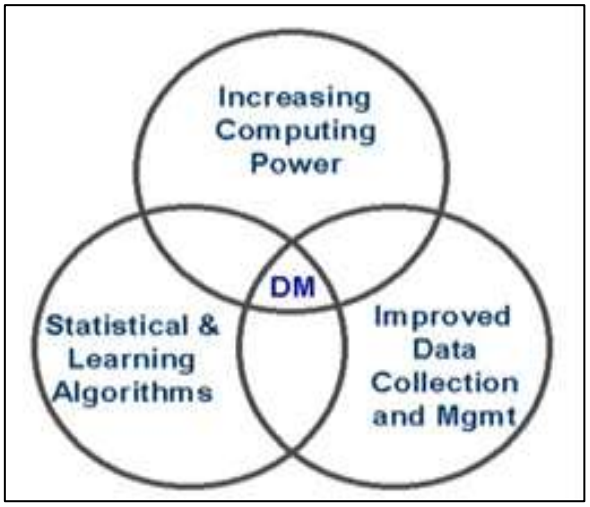
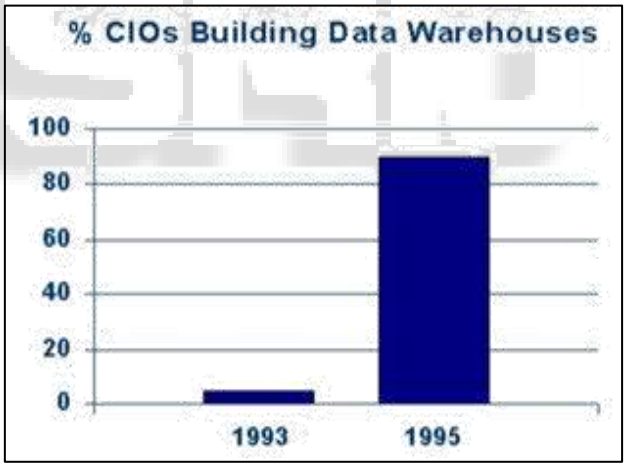


Fig. 2: Convergence of three technologies

- Increasing Computing Power
- Moore's law doubles computing power every 18 months
- Powerful workstations became common
- Cost effective servers (SMPs) provide parallel processing to the mass market
- Interesting tradeoff
- Small number of large analyses vs. large number of small analyses

A. Improved Data Collection



B. The Data Mining Process

Generally, data mining process is composed by data preparation, data mining, and information expression and analysis decision-making phases, the specific process as shown in fig.1 [5].

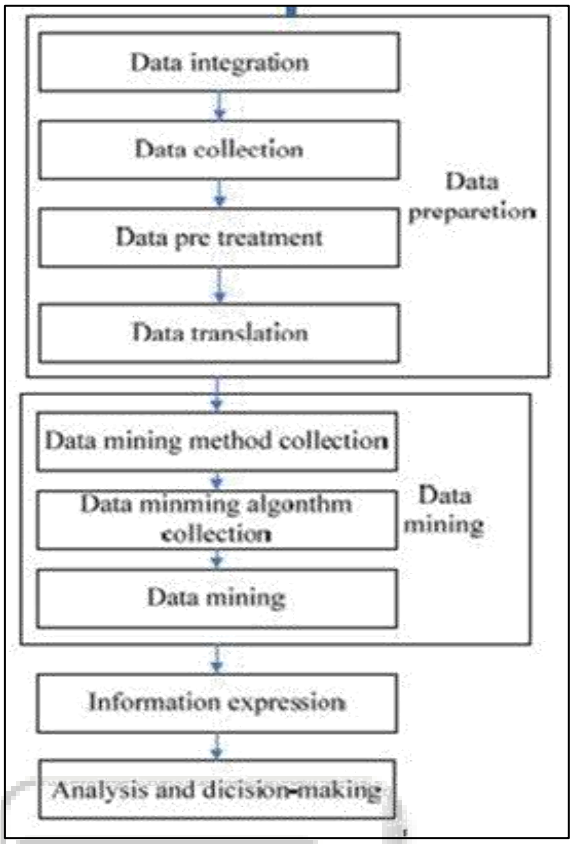


Fig. 3: General process of Data Mining

1) Data preparation

Data preparation generally consists of two processes: data collection and data collation. Data collectionis the primary step of knowledge mining, and therefore the data can come from the prevailing transaction processing systems, can also be obtained from the info warehouse; data collation is to eliminate noise or inconsistent data, it's the mandatory link of knowledge mining. the info obtained from the phase of the info collection may have a definite degree of "pollution", which refers to it within the data is also its own inconsistency, or some missing data, that the collation of the info is crucial [9] . At identical time, through data collation the info may be done on an easy generalization processing, thus on the idea of the initial data more rich data information are obtained, which is able to facilitate successive data processing step.

2) Data mining

Data mining is that the core stage of the whole process, it mainly uses the collected mining tools and techniques to manage the info, thus the foundations, patterns and trends are found.

3) Information expression

Information expression is to use visualization and knowledge information expression technology to produce the mined knowledge information for users, is a crucial means to indicate the info mining results. Clear and effective mining result information expression will greatly facilitate the accuracy and efficiency of the decision-making.

4) Analysis and decision-making

The ultimate goal of knowledge mining is to help the choice making. Decision-makers can analyze the results of knowledge mining and adjust the decision-making strategies combining with the particular situation.

C. Data mining architecture

There are three tiers within the tight-coupling data processing architecture:

1) Data layer: as mentioned above, data layer may be database and/or data warehouse systems. This layer is an interface for all data sources. data processing results are stored in data layer so it may be presented to end-user in style of reports or other reasonably visualization.
2) data processing application layer is employed to retrieve data from database. Some transformation routine may be performed here to remodel data into desired format. Then data is processed using various data processing algorithms.
3) Front-end layer provides intuitive and friendly programme for end-user to interact with data processing system. data processing result presented in visualization form to the user within the front-end layer.

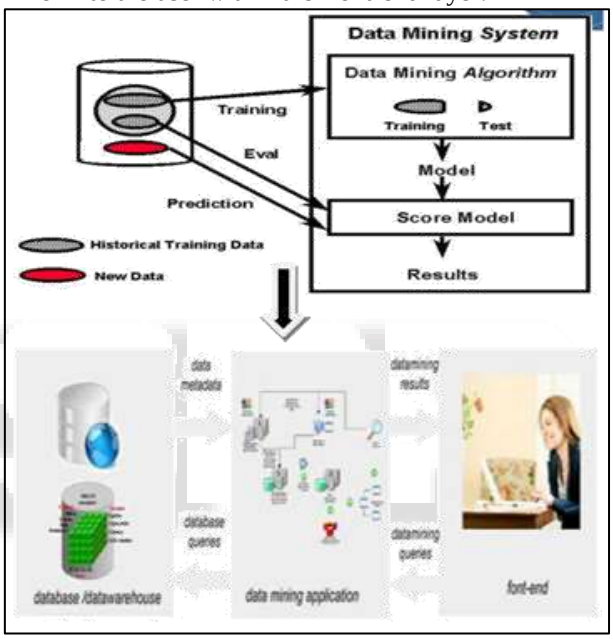


Fig. 4: Architecture of Data mining

In this article, we've discussed various data mining architectures, its advantages and disadvantages. And then we looked into a tight-couple data mining architecture – the most desired, high performance, high scalable data mining architecture.

Each data mining algorithm can be decomposed into four components:

1) Model or pattern structure
2) Interestingness measure (score function)
3) Search method
4) Data management strategy

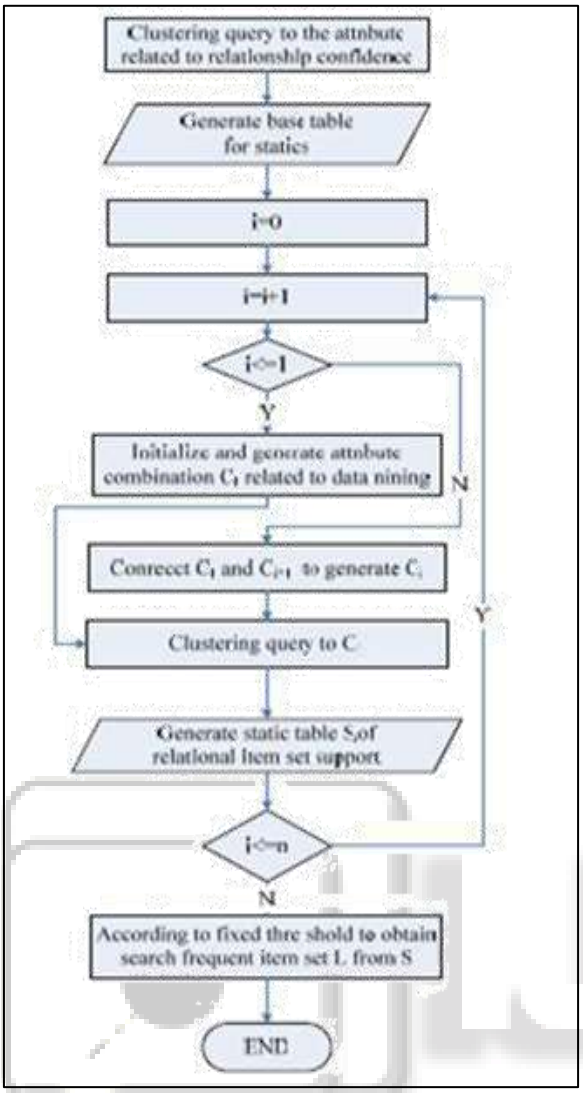


Fig. 5: Algorithm process

D. Data Mining Supported Decision Tree

Decision tree learning, employed in statistics, data processing and machine learning, uses a call tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that result in those class labels. In decision analysis, a call tree may be wont to visually and explicitly represent decisions and higher cognitive process. In data processing, a call tree describes data but not decisions; rather the resulting classification tree may be an input for higher cognitive process A decision web (DSS) may be a computer-based data system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of a company and help to form decisions, which can be rapidly changing and not easily laid out in advance. DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to assist decision makers compile useful information from a mixture of data, documents, personal knowledge, or business models to spot and solve problems and make decisions. Data mining requires data preparation which might uncover information or patterns which can compromise confidentiality and privacy obligations. a typical way for this to occur is thru data aggregation. Data aggregation is when the info are accrued, possibly from various sources, and put together in order that they'll be analyzed.[38] this is often not data processing intrinsically, but a result of the preparation of information before and for the needs of the analysis. The threat to a person's privacy comes into play when the info, once compiled, cause the info miner, or anyone who has access to the newly compiled data set, to be ready to identify specific individuals, especially when originally the info were anonymous.

E. Data mining supported neural network:

The data mining supported neural network consists by data preparation, rules extracting and rules assessment three phases, as shown in Fig. 2.

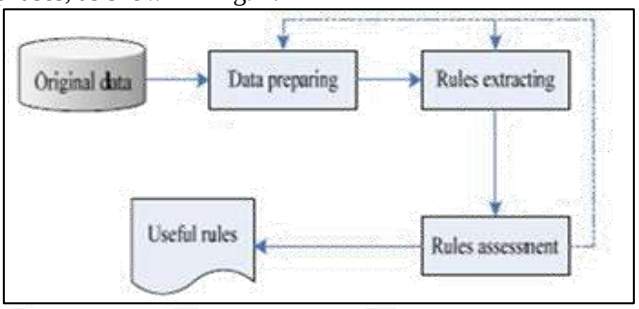


Fig. 6: Data mining process on neural network

There are seven common methods and techniques of knowledge mining which are the methods of statistical analysis, rough set, covering positive and rejecting inverse cases, formula found, fuzzy method, as well as visualization technology. Here, we focus on neural network method.

Neural network method is employed for classification, clustering, feature mining, prediction and pattern recognition. It imitates the neurons structure of animals, bases on the M-P model and Hebbien learning rule, so in essence it's a distributed matrix structure. Through training data processing, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected.

F. Data mining: K means clustering:

K-means clustering may be a data mining/machine learning algorithm wont to cluster observations into groups of related observations with none prior knowledge of these relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

G. The k-means Algorithm:

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation‟s proximity to the mean of the cluster. The cluster‟s mean is then recomputed and therefore the process begins again. Here‟s how the algorithm works:

1) The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2) Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3) Each cluster center is recomputed as the average of the points in that cluster.
4) Steps 2 and 3 repeat until the clusters converge. Convergence could also be defined differently depending upon the implementation, but it normally means either no observations change clusters when steps 2 and three are repeated or that the changes don't make a cloth difference within the definition of the clusters.

III. Application of Data Mining

- Data Mining in Agriculture
- Surveillance / Mass surveillance
- National Security Agency
- Quantitative structure-activity relationship
- Customer analytics
- Police-enforced ANPR in the UK
- Stellar wind (code name)
- Educational Data Mining

A. *Advantages of Data Mining*

1) Marketing / Retail

Data mining helps marketing companies to create models supported historical data to predict who will answer new marketing campaign like spam, online marketing campaign and etc. Through this prediction, marketers can have appropriate approach to sell profitable products to targeted customers with high satisfaction.

Data mining brings tons of benefit s to retail company within the same way as marketing. Through market basket analysis, the shop can have an appropriate production arrangement within the way that customers can purchase frequent buying products alongside pleasant. In addition, it also help the retail company offers a particular discount for particular products what is going to attract customers.

2) Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from previous customer‟s data with common characteristics, the bank and financial can estimate what are the god and/or bad loans and its risk level. In addition, data mining can help banks to detect fraudulent credit card transaction to help credit card‟s owner prevent their losses.

3) Manufacturing

By applying data processing in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers had a challenge that even the conditions of producing environments at different wafer production plants are similar, the standard of wafer are lot an equivalent and a few for unknown reasons even contain defects. Data mining has been applied to work out the ranges of control parameters that cause the assembly of golden wafer. Then those optimal control parameters are wont to manufacture wafers with desired quality.

4) Governments

Data mining helps agency by digging and analyzing records of monetary transaction to create patterns which will detect concealment or criminal activity.

Disadvantages of data mining

5) Privacy Issues

The concerns about the private privacy are increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs…. Because of privacy issues, people are scared of their personal information is collected and utilized in unethical way that potentially causing them tons of trouble. Businesses collect information about their customers in some ways for understanding their purchasing behaviors trends. However, businesses don't last forever, some days they'll be acquired by other or gone. At this point the private information they own probably is sold to other or leak.

6) Security issues

Security is a big issue. Businesses own information about their employee and customers including social security number, birthday, payroll and etc. However how properly this information is taken remains in questions. There are tons of cases that hackers were accesses and stole big data of consumers from big corporation like Ford Motor Credit Company, Sony… with so much personal and financial information available, the Mastercard stolen and fraud become an enormous problem.

7) Misuse of information/inaccurate information

Information collected through data processing intended for marketing or ethical purposes are often misused. This information is exploited by unethical people or business to require advantage of vulnerable people or discriminate against a gaggle of individuals.

In addition, data processing technique isn't perfectly accurate therefore if inaccurate information is employed for decision-making will cause serious consequence.

8) Challenges of Data Mining

- Scalability
- Dimensionality
  Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

9) Marketplace surveys

Several researchers and organizations have conducted reviews of knowledge mining tools and surveys of knowledge miners. These identify a number of the strengths and weaknesses of the software packages. They also provide a summary of the behaviors, preferences and views of knowledge miners.

IV. Future Enhancement

Over recent years data processing has been establishing itself together of the main disciplines in computing with growing industrial impact. Undoubtedly, research in data mining will continue and even increase over coming decades involve Mining complex objects of arbitrary type, fast, transparent and structured data preprocessing, Increasing usability. All aim at understanding consumer behavior, forecasting product

demand, managing and building the brand, tracking performance of consumers or products within the market and driving incremental revenue from transforming data into information and information into knowledge.

Although data processing remains in its infancy, companies during a wide selection of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to require advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data processing helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

V. CONCLUSION

Data mining may be a hot topic of the pc science research in recent years, and it's an extensive applications in various fields. Data mining technology is an application-oriented technology. It not only may be a simple search, query and transfer on the actual database, but also analyzes, integrates and reasons these data to guide the answer of practical problems and find the relation between events, and even to predict future activities through using the prevailing data. Data mining brings tons of advantages to businesses, society, governments also as individual. However, privacy, security and misuse of data are the large problem if it's not address correctly.

REFERENCES

[1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, l996, 8(6):866-883.
[2] R Agrawal ,T 1 mielinski, A Swami. Database Mining: A Performance Perspective[J]· IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.
[3] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf Retrieved 2008-12-17..
[4] Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery" International Journal of Information Technology and Decision Making, Volume 7, Issue 4 7: 639 – 682. doi:10.1142/S0219622008003204.
[5] Data mining:Ford, C.W.; Chia-Chu Chiang; Hao Wu; Chilka, R.R.; Talburt,J.R.; Information Technology: Coding and Computing, 2005. ITCC 2005 InternationalConference Volume: Digital Object Identifier: 10.1109/ITCC.2005.270 Publication Year: 2005 , Page(s): 122 - 127 Vol. 1
[6] Han, J. & M. Kamber, Data mining: concepts and techniques, San Francisco: Morgan Kaufman (2001).
[7] "Data mining tools", by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011
[8] "Data mining and ware housing". Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume:1, Publication Year: 2011 , Page(s): 1 – 5
[9] "The applied research on data mining in the financial analysis of university with the analysis of college students „arrears as an example" Chen Hongfei; Wang Xiaoyan; Business Management and Electronic Information (BMEI), 2011 International Conference on Volume:2DigitalObjectIdentifier: 10.1109/ICBMEI.2011.5917992Publication Year: 2011, Page(s): 633 - 636

