Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (39-45), Month: May - August 2015, Available at: www.noveltyjournals.com

NEED AND APPLICATIONS OF DATA MINING

Sweta

Department of Computer Science And Engineering, Gulbarga, Karnataka, India.

Abstract: **The rapid growth of available data and the widespread use of database have created an immense need for data mining. It allows user to analyse data from many different dimensions, categorize it, and summarize the relationships identified for finding correlations or patterns among dozens of fields in large relational database to make processes more efficient, effective, predictable, and profitable. It is used to make critical business decisions include retail, telecommunication, sales and marketing, healthcare, insurance, finance, manufacturing and various other. This article discuss about data mining meaning, need for data mining, its major elements, advantages, various types of data used, and some data mining applications. The relation between data, information and knowledge in knowledge discovery is illustrated. Data mining finds applications in different areas like banking, financial, astrology, retail, telecommunication, education, bio-medical, agricultural, social media, spacial, tax governing fields, security and various others. Modern technologies allow storage of large amounts of raw data which invariably contains usable information, by using data mining techniques which can be used to analyse that data and uncover previously unknown rules and associations, hidden knowledge which once acquired and properly interpreted can be used in multitude of ways, such as astronomy, medical imaging, bio-informatics, combinatorial chemistry, remote sensing, and physics.**

Keywords: **data, data mining, educational data mining, information, knowledge, metadata, spatial data mining.**

I. INTRODUCTION

Data mining is the process of analysing data from different perspectives and summarizing it into useful information which can be used to increase revenue, cuts costs, or both. Data mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics and marketing. Web mining, a type of data mining used in customer relationship management takes advantage of the huge amount of information gathered by a Web site to look for patterns in user behaviour. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining the analysis step of the "Knowledge Discovery in Databases" an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. It involves database and data management aspects, data pre-processing model, inference considerations, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining consists of five major elements. Extract, transform, and load transaction data onto the data warehouse system. Store and manage the data in a multidimensional database system. Provide data access to business analysts and information technology professionals. Analyse the data by application software. Present the data in a useful format, such as a graph or table.

Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (39-45), Month: May - August 2015, Available at: www.noveltyjournals.com

Data, Information, and Knowledge:

Data:

Data is set of values of qualitative or quantitative variables restated, pieces of data are individual pieces of information. Data are any facts, numbers, or text that can be processed by a computer. Data is measured, collected and reported, and analysed and can be visualized using graphs or images. It refers to the fact that some existing information or knowledge is represented in some form suitable for better usage or processing. Data is collected and analysed to create information suitable for making decisions, while knowledge is derived from extensive amounts of experience dealing with information on a subject as shown in Fig .1.

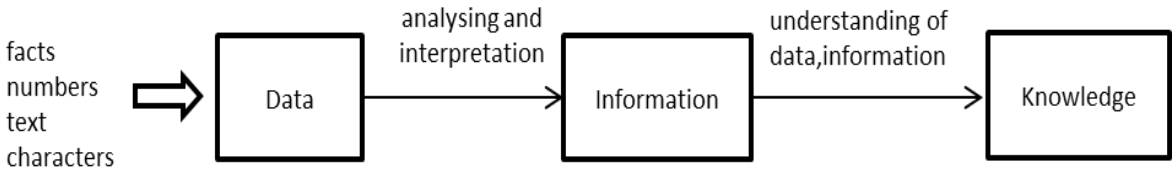

Fig.1.Relation between data, information and knowledge

Data is the least abstract, information the next least, and knowledge the most. Data becomes information after interpretation; the concept of information is closely related to notions of constraint, communication, control, data, form, instruction, knowledge, meaning, mental stimulus, pattern, perception, and representation. Raw data is a collection of numbers, characters, data processing commonly occurs by stages, and the "processed data" from one stage may be considered the "raw data" of the next. Experimental data is data that is generated within the context of a scientific investigation by observation and recording.

The organizations are growing with huge amount of data in different formats and different databases, like transactional data, metadata, multimedia data, spatial data, time series, relational data and various other types as shown in Fig.2.

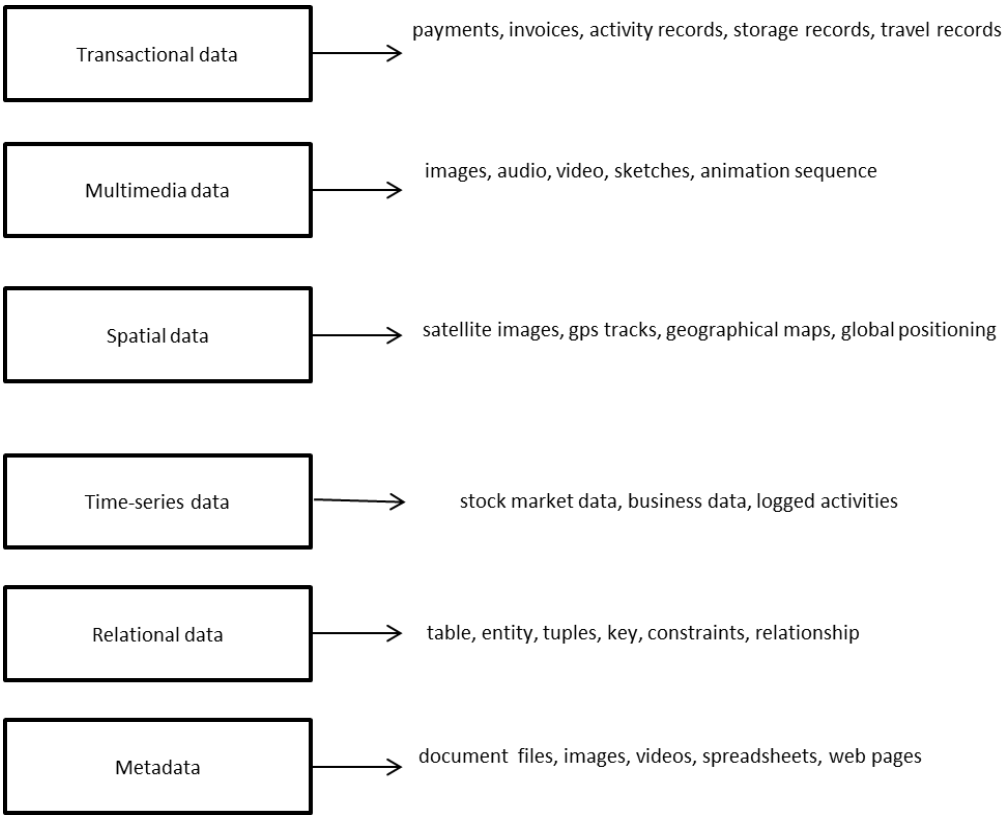

Fig.2.Different formats of data in data mining.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (39-45), Month: May - August 2015, Available at: www.noveltyjournals.com

Transactional data are data describing an event as a result of a transaction and is usually described with verbs. A transaction is a sequence of information exchange and related work such as database updating that is treated as a unit for the purposes of satisfying a request. It always has a time dimension, a numerical value and refers to one or more objects. Transactional data can be financial, logistical or work-related, involving everything from a purchase order to shipping status to employee hours worked to insurance costs and claims. It is grouped with associated master data and reference data and records a time and relevant reference data needed for a particular transaction record. Transactional data describes an internal or external event which takes place as the organization conducts business and can be financial, logistical or any business-related process involving activities such as purchases, requests, insurance claims, deposits, withdraws, and supports ongoing business operations and are included in the information and application systems that are used to automate an organization's key business processes such as online transaction processing systems. It is grouped with its associated and references master data such as product information and billing sources. Such as orders, sales, payroll, invoices, payments, plans, activity records, deliveries, storage records, travel records, etc.

Metadata is defined as the data providing information about one or more aspects of the data, such as means of creation of the data, purpose of the data, time and date of creation etc. It is stored and managed in a database, often called a metadata registry or metadata repository, which can make finding and working with particular instances of data easier. It is used for document files, images, videos, spreadsheets and web pages and provides information about a certain item's content.

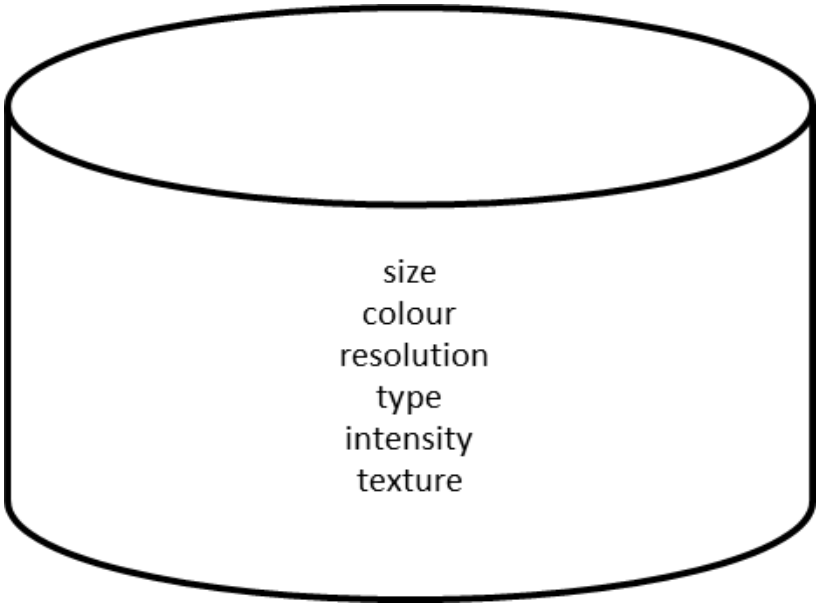

Fig.3.Metadata of image

The image may include metadata that describes about the colour depth, image size, type, texture, the image resolution and other data as shown in Fig.3. Web pages often include metadata in the form of metatags. The main purpose of metadata is to facilitate in the discovery of relevant information, more often classified as resource discovery. It helps organize electronic resources, provide digital identification, and helps support archiving and preservation of the resource to assists in resource discovery by allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information.

The multimedia data include one or more primary media data types such as text, images, graphic objects including drawings, sketches, illustrations, animation sequences, audio and video.These are typically the elements or the building blocks of generalized multimedia environments, platforms, or integrating tools. Spatial data store geographical information like maps, and global or regional positioning and is a database that is optimized to store and query data that represents objects defined in a geometric space. Time-series data contain time related data such stock market data or logged activities which usually have a continuous flow of new data coming in during real time analysis. Relational data is

a collection of tables, entity, tuples, keys, constraints, relation in relational database. Each table is a metaphorical representation of an entity or object that is in a tabular format consisting of columns and rows. Columns are the fields of a record or the attributes of an entity. The rows contain the values or data instances are called records or tuples. Relationships exist both among the columns within a table and among the tables. Spatial data store geographical information like maps, and global or regional positioning and is a database that is optimized to store and query data that represents objects defined in a geometric space. Time-series data contain time related data such stock market data or logged activities which usually have a continuous flow of new data coming in during real time analysis.

Information:

Data and information are interrelated. Data usually refers to raw data, or unprocessed data. Once the data is analysed, it is considered as information. Information is knowledge communicated regarding a particular fact or circumstance. Information is a sequence of symbols that can be interpreted as a message. Data represents 'values of qualitative or quantitative variables, belonging to a set of items, in the form of numbers, letters, or a set of characters. It is often collected via measurements and is represented by in a structure, such as tabular data, data tree, a data graph, etc. Information is the message that is being conveyed, whereas data are plain facts. Information is specific and the patterns, associations, or relationships among all data can provide information. The data is processed, organized, structured or presented in a given context which converts into information, knowledge.

Knowledge:

Knowledge is a familiarity, awareness or understanding of someone or something, such as facts, information, descriptions, or skills, which is acquired through experience or education by perceiving, discovering, or learning. It refers to a theoretical or practical understanding of a subject. It can be implicit as with practical skill or expertise or explicit as with the theoretical understanding of a subject it can be more or less formal or systematic. Information can be converted into knowledge about historical patterns and future trends. It is a combination of experience, values, contextual information, expert insight, and grounded intuition that provides an environment and framework for evaluating and incorporating new experiences and information. In organizations it often becomes embedded not only in documents or repositories, but also in organizational routines, practices and norms. Summary information on retail supermarket sales can be analysed in light of promotional efforts to provide knowledge of consumer buying behaviour using which manufacturer or retailer could determine which items are most susceptible to promotional efforts.

Knowledge Discovery in Databases is the process of searching for hidden knowledge in the massive amounts of data to extract knowledge from lower level data. Knowledge Discovery in Databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data and the goal is to distinguish from unprocessed data. Extraction of knowledge from raw data is accomplished by applying Data Mining methods. KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, pre-processing, sampling, and projections of the data prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

II. APPLICATIONS

Data mining is primarily used today in various fields with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics and the impact on sales, customer satisfaction, and profits. Data mining finds applications in different areas like banking, financial, astrology, retail, telecommunication, education, bio-medical, agricultural, social media, spacial, tax governing fields, security and various others as shown in Fig.4.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (39-45), Month: May - August 2015, Available at: www.noveltyjournals.com

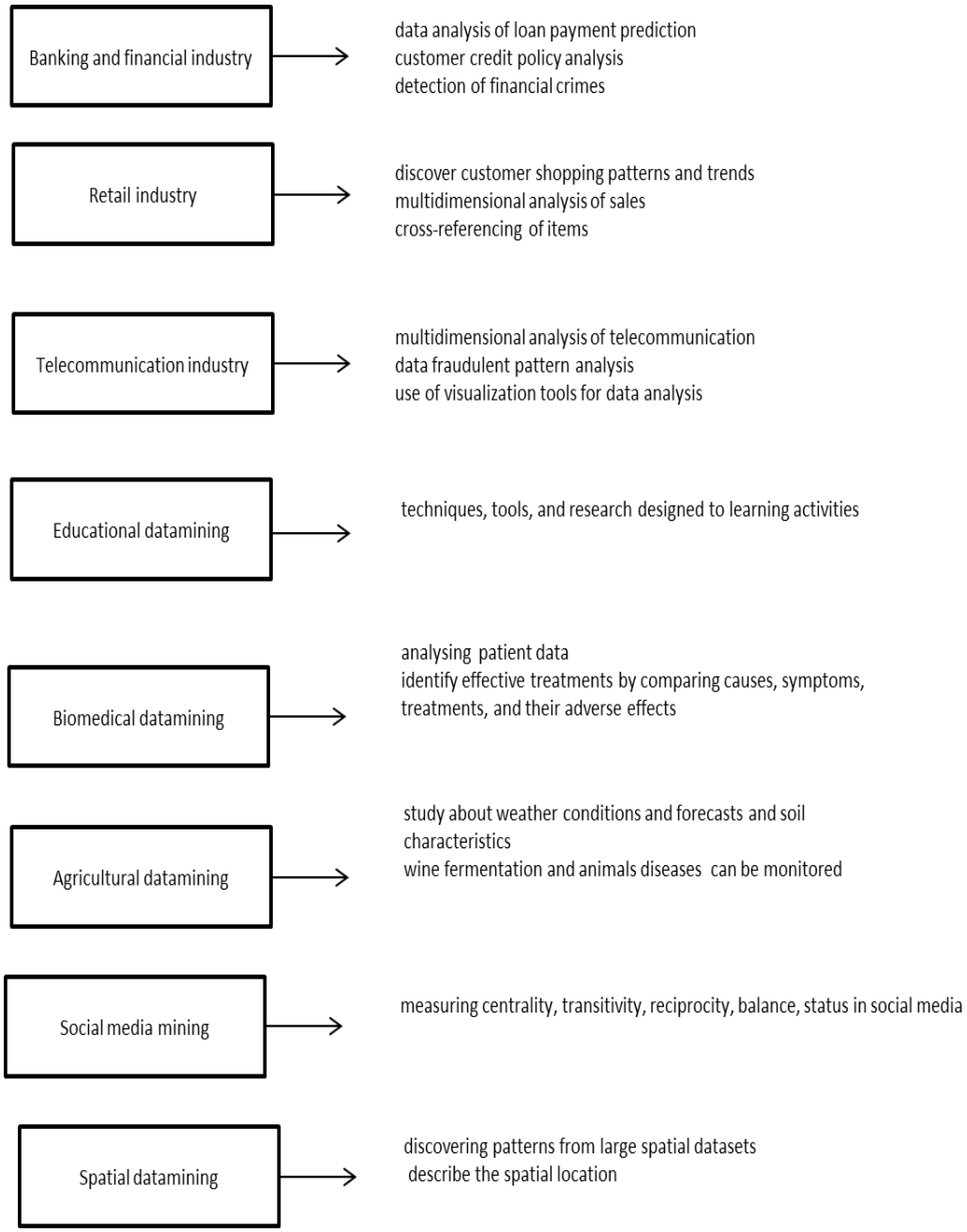

Fig.4.Application of data mining

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. The design and construction of data warehouse for multidimensional data analysis of loan payment prediction and customer credit policy analysis. Classification and clustering of customers for targeted marketing, for detection of money laundering and other financial crimes data mining is used. Several data mining techniques distributed data mining have been researched, modelled and developed to help credit card fraud detection and is used to identify customer loyalty by analysing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, total monetary value of all purchases and when was the last purchase. After analysing the dimensions, the relative measure is generated for each customer. By analysing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so it can plan and launch different special offers to retain those customers. Credit card spending by customer groups can be identified by using data mining. The hidden correlation between different financial indicators can be discovered by using data mining and it enables to identify stock trading rules.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (39-45), Month: May - August 2015, Available at: www.noveltyjournals.com

Retail industry involves large amount of data on sales and customer purchasing history, goods transportation, consumption and services. The quantity of data continues to expand rapidly, because of the increasing ease, availability and popularity of the business conducted on web. It provides a rich source for data mining and involves identification of customer behaviour, to discover customer shopping patterns and trends, improves the quality of customer service, achieve better customer retention and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business. Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction with design and construction of data warehouses based on the benefits of data mining, multidimensional analysis of sales, customers, products, time and region, analysis of effectiveness of sales campaigns, customer retention and product recommendation and cross-referencing of items.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Services include multidimensional analysis of telecommunication data, fraudulent pattern analysis, identification of unusual patterns, multidimensional association and sequential patterns analysis, mobile telecommunication services, and use of visualization tools in telecommunication data analysis.

Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to learning activities in educational settings. It concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students. It develops methods and applies techniques from statistics, machine learning, and data mining to analyse data collected during teaching and learning and tests learning theories and informs educational practice. Data collected from online learning systems can be aggregated over large numbers of students which contain many variables that data mining algorithms can explore for model building.

Biomedical data mining involves examining and analysing stored patient data, expert data miners can uncover important trends. Data mining in health care has become increasingly popular because it offers benefits to care providers, patients, healthcare organizations, researchers, and insurers. It is used to identify effective treatments and best practices, by comparing causes, symptoms, treatments, and their adverse effects, which analyse which course of actions, are most effective for specific patient groups. It can also identify clinical best practices to help develop guidelines and standards to provide a better healthcare service by identifying and track chronic diseases and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims.

Data mining in agriculture includes applications of data mining techniques to agriculture. It provides a lot of information on agricultural-related activities, which can then be analysed in order to study about weather conditions and forecasts. Data mining techniques are often used to study soil characteristics. Wine fermentation process can be monitored using data mining techniques and to monitor animals diseases.

Social Media Mining is the process of representing, analysing, and extracting actionable patterns from social media data and represents the virtual world of social media in a computable way, measures it, and designs models that can help to understand its interactions. It provides necessary tools to mine this world for interesting patterns, analyse information diffusion, provide effective recommendations, and analyse novel social behaviour in social media. It introduces basic concepts and principal algorithms suitable for investigating massive social media data and discusses theories and methodologies from different disciplines such as computer science, data mining, machine learning, social network analysis, network science, sociology, ethnography, statistics, optimization, and mathematics. It encompasses the tools to formally represent, measure, model, and mine meaningful patterns from large-scale social media data. The data mining here involves identifying communities on social networks, how it evolves, and evaluating identified communities, often without ground truth. Measuring the transitivity, reciprocity, balance, status, and similarity in social media, to simulate networks with specific characteristics, and analysing how information propagates in social media.

Data mining can aid law enforcers in identifying criminal suspects by examining trends in location, persons fingerprint, crime type, habit, and other related information for security purpose. Data mining can assist researchers by speeding up their data analysing process. Tax governing fields use data mining techniques to detect fraudulent transactions and single out suspicious tax returns.

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 2, Issue 2, pp: (39-45), Month: May - August 2015, Available at: www.noveltyjournals.com

III. CONCLUSION

The need, acceptance, applications of data mining has increased in the last few years. With rapid growing data with varied characteristics in large volumes, there is still need to find different areas that can make use of data mining benefits to produce accurate and practical forecasting methods, and developing analytics relevant for better decision-making. Data mining finds applications in different areas like banking, financial, astrology, retail, telecommunication, education, bio-medical, agricultural, social media, spacial, tax governing fields, security and various others. The increase in data in various fields results in need to analyse, extract, store, manage, multidimensional data analysis and to present the data into useful format. The security concerns must be properly managed to get better use of resources and to improve quality of service.

REFERENCES

[1] Venkatadri.M, Dr. Lokanatha C. Reddy: A Review on Data mining from past to the future, International Journal of Computer Applications (0975-8887) Volume 15-No.7, February 2011.

[2] Vitthal Yenkar, Prof.Mahip Bartere: Review on "Data Mining with Big Data", IJCSMC, Vol. 3, Issue. 4, April 2014, pg.97 – 102.

[3] Zhao Dan: Data Mining Applications in the Banking Industry in China (1998-2007).

[4] Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat ,Amit Khaparde: A Review Paper on Various Data Mining Techniques, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4, April 2014 ISSN: 2277 128X.

[5] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth: From Data Mining to Knowledge discovery in databases.

[6] L.Tanabe, UScherf, L.H. Smith, J.K. Lee,L. Hunter, J.N. Weinstein: MedMiner: An Internet text-mining tool for Biomedical information, with Application to gene expression profiling.

[7] S.L. Ting , C.C. Shum , S.K. Kwok , A.H.C. Tsang , W.B. Lee: Data Mining in Biomedicine: Current applications and further directions for research.
S.Lakshmi Prabha,Dr.A.R.Mohamed Shanavas: Educational data mining applications, Operations Research and Applications: An International Journal (ORAJ), Vol. 1, No. 1, August 2014.

[8] Jayanthi Ranjan: Applications of data mining techniques in pharmaceutical industry, Journal of Theoretical and Applied Information Technology

[9] K.Venkateswara Rao,A.Govardhan, K.V.Chalapati Rao: Spatiotemporal data mining: Issues, tasks and applications, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.1, February 2012.

[10] Wikipedia,The free Encyclopedia. https://en.wikipedia.org/wiki/Data_mining

[11] Wikipedia,The free Encyclopedia. https://en.wikipedia.org/wiki/Knowledge