ORIENTAL JOURNAL OF
COMPUTER SCIENCE & TECHNOLOGY
An International Open Free Access, Peer Reviewed Research Journal
Published By: **Oriental Scientific Publishing Co., India.**
www.computerscijournal.org

ISSN: 0974-6471
December 2013,
Vol. 6, No. (4):
Pgs. 413-419

Applications and Trends in Data Mining

S.V.S. GANGA DEVI

Professor in MCA, K.S.R.M. College of Engineering, Kadapa - 516 003, India.

(Received: October 15, 2013; Accepted: October 25, 2013)

ABSTRACT

The advent of Computing Technology has significantly influenced our lives and two major impacts of this effect are Business Data Processing and Scientific Computing. During the early years of the development of computer techniques for business, computer professionals were concerned with designing files to store the data so that information could be efficiently retrieved. There were restrictions on storage size for storing data and on the speed of accessing the data. Needless to say, the activity was restricted to a very few, highly qualified professional. Then came an era when Database management System simplified the task. The responsibility of intricate tasks, such as declarative aspects of the programs was passed on to the database administrator and the user could pose his query in simpler languages such as query languages. Thus almost any business-small, medium or large scale began using computers for day-to-day activities.

Now what is the use of all this data? Up to the early 1990's the answer to this was "NOT much". No one was really interested in utilizing data, which was accumulated during the process of daily activities. As a result a new discipline in computer science, Data Mining gradually evolved.

Key words: Applicaiton, Computer technology, Business data, Processing

INTRODUCTION

Data mining is the process of extraction of interesting (nontrivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. It is the set of activities used to find new, hidden or unexpected patterns in data of unusual patterns in data. Data Mining is the exploration and analysis of large sets, in order to discover meaningful patterns and rules. The key idea is to find effective ways to combine computers power to process data with the human eye's ability to detect patterns. The techniques of data mining are designed for work best with large data sets.

In this paper a few application domains of Data Mining (such as finance ,the retail industry and telecommunication) and Trends in Data Mining are discussed. Section 2 describes the Applications of Data mining, Section 3 Trends in Data mining and Section 4 the Conclusion was described.

Applications

Data mining is many and varied fields of applications.

Data Mining for Financial Data Analysis

Financial data collected in the banking and financial industry are often relatively complete,

reliable, and of high quality, which facilitates systematic data analysis and data mining. A few examples of data mining in the Financial Data Analysis are outlined as follows:

Detect patterns of fraudulent credit card use

Identify loyal customers

Predict customers likely to change their credit card affiliation

Determine credit card spending by customer groups

Find hidden correlations between different financial indicators

Identify stock trading rules from historical market data.

Design and Construction of data warehouses for multidimensional data analysis and data mining

Data warehouses need to be constructed for banking and financial data. Multidimensional data analysis methods should be used to analyze the general properties of such data warehouses, data cubes, multi feature and discovery-driven data cubes, characterization and class comparisons, and outlier analysis all play important roles in financial data analysis and mining.

Loan payment prediction and customer credit policy analysis

Data mining methods, such as attribute selection and attribute relevance ranking, may help identify important factors such as loan-to-value ratio, duration of the loan, debt ratio, payment to-income ratio, customer income level, education level, residence region, and credit history and eliminate irrelevant ones for loan payment prediction and customer credit rating.

Classification and clustering of customers for targeted marketing

Classification technique is used to identify the most crucial factors that may influence a customer's decision regarding banking. Customers with similar behaviors regarding loan payments may be identified by multidimensional clustering techniques. These data mining techniques helps to identify customer groups, associate a new customer with an appropriate customer group, and facilitate targeted marketing. Multiple data analysis tools such as data visualization tools, linkage analysis tools, classification tools, clustering tools, outlier analysis tools and sequential pattern analysis tools can then be used to detect unusual patterns, such as large amount of cash flow at certain periods, by certain groups of customers and also may identify important relationships and patterns of activities for further examination.

Data Mining for the 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. Retail data mining can help to:

1. Identify buying patterns from customers
2. Discovers customer shopping patterns and trends
3. Find associations among customer demographic characteristics
4. Predict response to mailing campaigns Improve the quality of customer service
5. Achieve better customer retention and satisfaction
6. Reduces the cost of business
7. Market basket analysis

A few examples of data mining in the retail industry are outlined as follows

Design and construction of data warehouses based on the benefits of data mining

The outcome of preliminary data mining exercises can be used to help guide the design and development of data warehouse structures. This involves deciding which dimensions and levels to include and what preprocessing to perform in order to facilitate effective data mining.

Multidimensional analysis of sales, customers, products, time, and region

The retail industry requires timely information regarding customer needs, product sales, trends, and fashions, as well as the quality, cost, profit, and service of commodities. It is therefore important to provide powerful multidimensional analysis and visualization tools and multi feature data cube to facilitate analysis on aggregates with complex conditions.

Analysis of the effectiveness of sales campaigns

The retail industry conducts sales campaigns using advertisements, coupons, and various kinds of discounts and bonuses to promote products and attract customers. Multidimensional analysis can be used for careful analysis of the effectiveness of sales campaigns to improve company profits. Association analysis may disclose which items are likely to be purchased together with the items on sale.

Customer retention – analysis of customer loyalty

Sequential pattern mining can be used to investigate changes in customer consumption or loyalty and suggest adjustments on the pricing and variety of goods in order to help retain customers and attract new ones. Product recommendations can also be advertised on sales receipts, in weekly flyers or on the web to help improve customer service aid customers in selecting items, and increase sales.

Data mining for the Telecommunication Industry

The integration of telecommunication, computer network, internet, and numerous other means of communication and computing is underway. With the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand for data mining in order to help.

1. To understand the business involved
2. To identify telecommunication patterns
3. To catch fraudulent activities
4. To make better use of resources
5. To improve the quality of service.

The following are a few scenarios for which data mining may improve Telecommunication services:

Multidimensional analysis of telecommunication data

The multidimensional analysis of telecommunication data using OLAP and visualization tools can be used to identify and compare the data traffic, system workload, resource usage, user group behavior, and profit.

Fraudulent pattern analysis and the identification of unusual patterns

The multidimensional analysis, cluster analysis, and outlier analysis are used to (1) identify potentially fraudulent users and their a typical usage patterns; (2) detect attempts to gain fraudulent entry to customer accounts; and (3) discover unusual patterns.

Multidimensional association and sequential pattern analysis

The discovery of association and sequential patterns in multidimensional analysis can be used to promote telecommunication service.

Mobile telecommunication services

Mobile Telecommunication, Web and information services, and mobile computing are becoming increasingly integrated and common in our work and life. One important feature of mobile telecommunication data is its association with spatiotemporal information. Data Mining will likely play a major role in the design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.

Use of visualization tools in telecommunication data analysis

Tools for OLAP visualization, linkage visualization, association visualization, clustering, and outlier visualization have been shown to be very useful for telecommunication data analysis.

Data Mining for biological Data Analysis

Biological data mining has become an essential part of a new research field called bioinformatics. The identification of DNA or amino acid sequence patterns that play roles in various biological function, genetic diseases, and evolution is challenging. Biological data mining helps to:

1. Characterize patient behaviour to predict office visits
2. Identify successful medical therapies for different illnesses
3. Develop effective genomic and proteomic data analysis tools.

DNA sequences form the foundation of the genetic codes of all living organisms. All DNA sequences are comprised of four basic building blocks, called nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). These four nucleotides (or bases) are combined to form long sequences or chains that resemble a twisted ladder. A genome is the complete set of genes of an organism. The human genome is estimated to contain around 20,000 to 25,000 genes. Genomic is the analysis of genome sequences.

Proteins are essential molecules for any organism. They perform life functions and make up the majority of cellular structures. There are 20 amino acids, each of the amino acids is coded for by one or more triplets of nucleotides making up DNA. The end of the chain is coded for by another set of triplets. Thus, a linear string or sequence of DNA is translated into a sequence of amino acids, forming a protein .

DNA sequence	**CTA**	**CAC**	**ACG**	**TGT**	**AAC**
Amino acid sequence	L	H	T	C	N

Fig : A DNA sequence and corresponding amino acid sequence

A proteome is the complete molecules present in a cell tissue, or organism. Proteomics is the study of proteome sequences.
Data mining may contribute to biological data analysis in the following aspects:

Semantic integration of heterogeneous, distributed genomic and proteomic databases

Data cleaning, data integration, reference reconciliation, classification, and clustering methods will facilitate the systematic and coordinated analysis of genome and biological data and also integrates the biological data and the construction of data warehouses for biological data analysis.

Alignment, indexing, similarity search, and comparative analysis of multiple nucleotide / protein sequences

Multiple sequence alignment is considered a more challenging task. Methods that can help include (1) reducing a multiple alignment to a series of pair wise alignment and then combining the result, and (2) using hidden Markov Models or HMMs. Multiple sequence alignments can be used to identify highly conserved residues among genomes, and such conserved regions can be used to build phylogenetic trees to infer evolutionary relationships among species.

Discovery of structural patterns and analysis of genetic networks and protein pathways

In biology, protein sequences are folded into three-dimensional structures which interact with each other based on their relative position and the distances between which forms basis of genetic networks and protein pathways. Powerful and scalable data mining methods are developed to discover approximate and frequent structural patterns and to study the regularities and irregularities among such interconnected biological networks.

Association and path analysis: Identifying co-occurring gene sequences and linking genes to different stages of disease development

Association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples and also facilitate the discovery of groups of genes and the study of interactions and relationships between them. Path analysis develops pharmaceutical interventions that target the different stages of disease development.

Visualization tools in genetic data analysis

The visually appealing of biological structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization and visual data mining therefore play an important role in biological data analysis.

Data Mining in other Scientific Applications

Vast amounts of data have been collected from scientific domain (including geosciences, astronomy, and meteorology) using sophisticated telescopes, multispectral high resolution remote satellite sensors, and global positioning systems to analyze complex data set. Some of the emerging scientific applications of data mining are:

Data warehouses and data preprocessing

Data warehouses are critical for information exchange and data mining. Scientific applications requires methods for integrating data from heterogeneous sources, for identifying events, for the efficient computation of sophisticated spatial aggregates and the handling of spatial-related data streams.

Mining Complex data types

Scientific data sets are heterogeneous in nature, typically involving semi-structured and unstructured data. Robust methods are needed for handling spatiotemporal data, related concept hierarchies, and complex geographic relationships.

Graph-based mining

In graph modeling, each object to be mined is represented by a vertex in a graph, and edges between vertices represent relationships between objects. The success of graph-modeling, however, depends on improvement in the scalability and efficiency of many classical data mining tasks, such as classification, frequent pattern mining, and clustering.

Visualization tools and domain-specific knowledge

High-level graphical user interface and visualization tools are required for scientific data mining systems to guide researcher and general users in searching for patterns, interpreting and visualizing discovered patterns and using discovered knowledge in their decision making.

Text Mining and Web Mining

Text mining is the process of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be established. Using text mining however, we can easily derive certain patterns in the comments that may help identify a common set of customer perceptions not captured by the other survey questions.

An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. It enhances the web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer. Web mining is especially exciting because it enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyze the data across one or multiple sites. For example the search engines work on the principle of data mining.

Data Mining for Intrusion Detection

An intrusion can be defined as any set of actions that threaten the integrity, confidentiality or availability of a network resource. An intrusion detection system for a large complete network can typically generate thousands or millions of alarms per day representing an overwhelming task for the security analysts.

Anomaly detection builds models of normal network behaviour (called profiles), which it uses to detect new pattern that significantly deviate from the profiles.

The following are areas in which data mining technology may be applied or further developed for intrusion detection.

Development of data mining algorithms for intrusion detection

Data mining algorithms can be used for misuse detection and anomaly detection. Anomaly detection builds models of normal behavior and automatically detects significant deviations from it. Supervised or unsupervised learning can be used. The techniques used must be efficient and scalable, and capable of handling network data of high volume, dimensionality, and heterogeneity.

Association and Correlation analysis, and aggregation to help select and build discriminating attributes

Association and correlation mining can be applied to find relationships between system attributes describing the network data for intrusion detection. Thus it is necessary to study what sequences of events are frequently encountered together, find sequential patterns, and identify outliers.

Distributed data mining

Distributed data mining methods may be used to analyze network data from several network location in order to detect these distributed attacks.

Visualization and querying tools

Visualization tools detects anomalous patterns using associations, clusters, and outliers techniques. These tools are more precise and require for less manual processing and input from human experts.

Higher Education

An important challenge that higher education faces today is predicting paths of students and alumni. Which student will enroll in particular course programs? Who will need additional assistance in order to graduate? Mean while additional issues, enrollment management and time-to-degree, continue to exert pressure on colleges to search for new and faster solutions. Institutions can better address these students and alumni through the analysis and presentation of data. Data mining has quickly emerged as a highly desirable tool for using current reporting capabilities to uncover and understand hidden patterns in vast databases.

Trends

The diversity of data, data mining tasks, and data mining approaches poses many challenging research issues in data mining. The development of efficient and effective data mining methods and systems, the construction of interactive and integrated data mining environments, the design of data mining languages, and the application of data mining techniques to solve large application problems are important tasks for data mining researchers and data mining system and application developers. Some of the trends in data mining that reflect the pursuit of these challenges are:

Application exploration

Data mining is increasingly used for the exploration of applications in other areas, such as financial analysis, telecommunications; biomedicine, wireless security and science.

Scalable and interactive data mining methods

Constraint-based mining handles huge amounts of data efficiently with added control by allowing the specification and use of constraints to guide data mining systems in their search for interesting patterns.

Web database systems

The Web database systems will ensure data availability, data mining portability, scalability, high performance, and an integrated information processing environment for multidimensional data analysis and exploration.

Standardization of data mining language

A standard data mining language will facilitate the systematic development of data mining solutions, improve interoperability among multiple data mining systems and functions, and promote the education and use of data mining systems in industry and society.

Visual data mining

Visual data mining is an effective way to discover knowledge from huge amounts of data.

New methods for mining complex types of data

New methods should be adopted for mining complex types of data to bridge a huge gap between the needs for these applications and the available technology.

Biological data mining

Mining DNA and protein sequences, mining high dimensional microarray data, biological pathway and network analysis, link analysis across heterogeneous biological data, and information integration of biological data by data mining are interesting topics for biological data mining research.

Data mining and software engineering

Further development of data mining methodologies for software debugging will enhance software robustness and bring new vigor to software engineering.

Web Mining

Due to vast amount of information available

on the Web. Web content mining, Web log mining, and data mining services on the Internet becomes one of the most important and flourishing subfields in data mining.

Distributed data mining

Advances in distributed data mining methods are expected to work in distributed computing environments.

Real-time or time-critical data mining

Many applications involving stream data (such as e=commerce, Web mining, stock analysis, intrusion detection, mobile data mining, and data mining for counter terrorism) require dynamic data mining models to be built in real time. Graph modeling is also useful for analyzing links in Web structure mining.

Multi-relational and multi-database data mining

Multi-relational data mining methods search for patterns involving multiple tables from a relational database. Multi-database mining searches for patterns across multiple databases.

Privacy protection and information security in data mining

Privacy protection and information security is to be provided by the data mining system.

Need of data mining

The massive growth of data from terabytes to perabytes is due to the wide availability of data in automated form from various sources as WWW, Business, Science, Society and many more. But we are drowning in data but deficient of knowledge data is useless, if it cannot deliver knowledge. That is why data mining is gaining wide acceptance in today's world. A lot has been done in this field and lot more need to be done.

CONCLUSION

Since data mining is a young discipline with wide and diverse applications, there is still a nontrivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. A few application domains of Data Mining (such as finance, the retail industry and telecommunication) and Trends in Data Mining which include further efforts towards the exploration of new application areas and new methods for handling complex data types, algorithms scalability, constraint based mining and visualization methods, the integration of data mining with data warehousing and database systems, the standardization of data mining languages, and data privacy protection and security.

REFERENCES

1. Data Mining Concepts and Techniques – Jiawei Han & Micheline Kamber.
2. DataMining: concepts,Models,Methods and Algorithms, Wiley-Interscience,Hoboken,Nj (2003).
3. Modern Data Warehousing, Mining and Visualization Core Concepts by George M. Marakas.