Int.J.Curr.Res.Aca.Rev.2024; 12(11): 92-97

EXCELLENT PUBLISHERS

International Journal of Current Research and Academic Review

ISSN: 2347-3215 (Online) Volume 12 Number 11 (November-2024)

Journal homepage: http://www.ijcrar.com

doi: https://doi.org/10.20546/ijcrar.2024.1211.008

Data Mining and Its Applications in Content of Higher Education M.P. India

Kaushal Kishor Dwivedi* and Daniel Glance Danny

Department of Computer Applications, Govt.Geetanjali P.G.Girls College Bhopal, M.P, India

**Corresponding author*

Abstract

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies beyond their expectations. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

Article Info

Received: 24 September 2024
Accepted: 30 October 2024
Available Online: 20 November 2024

Keywords

data, information, knowledge, On-line analytic processing (OLAP), multidimensional databases, data warehouses.

Introduction

We are in an age often referred to as the information age. In this information age, because we believe that information leads to power and success, and thanks to sophisticated technologies such as computers, satellites, Etc., we have been collecting tremendous amounts of information. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this amalgam of information. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). The efficient database management systems have been very important assets for management of a large corpus of data and especially for effective and efficient retrieval of particular information from a large collection whenever needed. The proliferation of database management systems has also contributed to recent massive gathering of all sorts of information.

Today, we have far more information than we can handle: from business transactions and scientific data, to satellite pictures, text reports and military intelligence. Information retrieval is simply not enough anymore for decision-making. Confronted with huge collections of data, we have now created new needs to help us make better managerial choices. These needs are automatic

summarization of data, extraction of the "essence" of information stored, and the discovery of patterns in raw data [Ref. 2].

The task of extracting previously unknown patterns from a data source is called *data mining*. The field has grown rapidly in the last few years as the amount of data stored in corporate databases has exploded. Companies believe that many of their customers follow patterns of behavior that the company does not yet recognize. Data mining techniques have much wider applicability than searching for patterns within customer data. The reminder of this paper is structured as follows. Section 2 gives an overview on how data mining can be done and its working. Section 3 is the issue of data mining. Section 4 represents the applications of data mining. The conclusion is done at the last section.

Data Mining

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining consists of five major elements: [Ref. 1]

- 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.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Consider the scenario; one Midwest grocery chain used the data mining capacity of Oracle software to analyze local buying patterns. They discovered that when men bought diapers on Thursdays and Saturdays, they also tended to buy beer. Further analysis showed that these shoppers typically did their weekly grocery shopping on Saturdays. On Thursdays, however, they only bought a few items. The retailer concluded that they purchased the beer to have it available for the upcoming weekend. The grocery chain could use this newly discovered information in various ways to increase revenue. For example, they could move the beer display closer to the diaper display. And, they could make sure beer and diapers were sold at full price on Thursdays.

Data mining or *information retrieval* win the race to make the Web more accessible. What is going on behind the scenes at sites like Goggle, Yahoo, and MSN? All of the big search engines have the same basic components:

- Web spider (or robot) that searches for web pages
- Index in which the text of the web pages are cached
- Relevance ranking function which orders documents with respect to a search query

Data Mining and Knowledge Discovery

With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making [Ref. 7].

Data Mining, also popularly known 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 mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure (Figure 1.1) shows data mining as a step in an iterative knowledge discovery process.

1. **Data cleaning**: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
2. **Data integration**: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. **Data selection**: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. **Data transformation**: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

5. **Data mining**: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. **Pattern evaluation**: in this step, strictly interesting patterns representing knowledge are identified based on given measures
7. **Knowledge representation**: is the final phase in which the discovered knowledge is visually represented to the users. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

Data Can Be Mined

In principle, data mining is not specific to one type of media or data. Data mining should be applicable to any kind of information repository. However, algorithms and approaches may differ when applied to different types of data. Indeed, the challenges presented by different types of data vary significantly. Data mining is being put into use and studied for databases, including relational databases, object-relational databases and object-oriented databases, data warehouses, transactional databases, unstructured and semi structured repositories such as the World Wide Web, advanced databases such as spatial databases, multimedia databases, time-series databases and textual databases, and even flat files. Here are some examples in more detail: [Ref.4]

Flat files

Flat files are actually the most common data source for data mining algorithms, especially at the research level. Flat files are simple data files in text or binary format with a structure known by the data-mining algorithm to be applied. The data in these files can be transactions, time-series data, scientific Measurements, etc.

Relational Databases

Briefly, a relational database consists of a set of tables containing either values of entity attributes, or values of attributes from entity relationships. Tables have columns and rows, where columns represent attributes and rows represent tuples. A tuple in a relational table corresponds to either an object or a relationship between objects and is identified by a set of attribute values representing a unique key. In Figure 1.2 we present some relations *Customer*, *Items*, and *Borrow* representing business activity in a fictitious video store Our Video store. These relations are just a subset of what could be a database for the video store and is given as an example for Our Video store.

The most commonly used query language for relational database is SQL, which allows retrieval and manipulation of the data stored in the tables, as well as the calculation of aggregate functions such as average, sum, min, max and count. For instance, an SQL query to select the videos grouped by category would be:

Select count (*) From Items Where type=video Group By category

Data mining algorithms using relational databases can be more versatile than data mining algorithms specifically written for flat files, since they can take advantage of the structure inherent to relational databases. While data mining can benefit from SQL for data selection, transformation and consolidation, it goes beyond what SQL could provide, such as predicting, comparing, detecting deviations, etc.

Working of Data Mining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought: [Ref.3, 6]

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.
- **Sequential patterns**: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.
- 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. Analyze the data by application software. Present the data in a useful format, such as a graph or table.

Issues of Data Mining

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous.

Before data mining develops into a conventional, mature and trusted discipline, many pending issues still have to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way. [Ref.3]

Security and social issues

Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behaviors understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information.

Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

User interface issues

The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs

Mining methodology issues

These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs, the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality*.

Performance issues

Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm.

In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining.

Figure.1

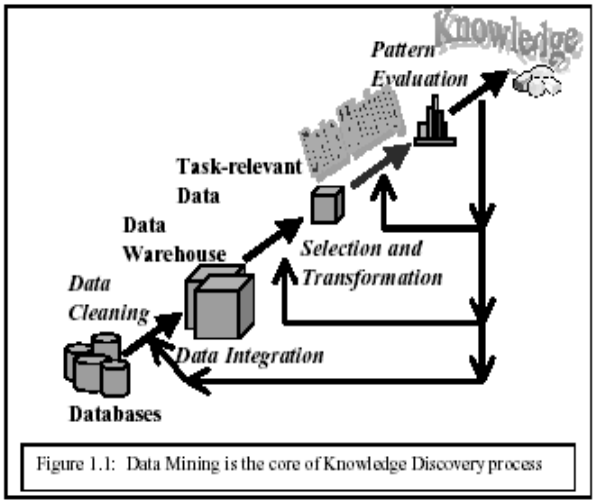

Figure 1.1: Data Mining is the core of Knowledge Discovery process

Data source issues

There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting.

The current practice is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant.

Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources.

Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data-mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

Applications of Data Mining

Data mining techniques can be applied to a large variety of decision-making context in business. In particular, areas of significant payoffs are expected to include the following [Ref.1, 2]

Marketing

Applications include analyses of consumes behavior based on buying patterns, determination of marketing strategies including advertising, store location, and targeted mailing, segmentation of customers, stores or products, and design of catalogues, store layouts, and advertising campaigns.

Finance

Application include analyses of credit worthiness of clients, segmentations of account receivable

Manufacturing

Applications involves optimization of resources like machines, manpower and materials, optimal design of manufacturing processes, shop-floor layouts, and product design, such as for automobiles based on customer requirements.

Healthcare

Application include an analysis of effectiveness of certain treatments, optimization of processes within a hospital, relating patient wellness data with doctor qualification and analyzing side effect of drugs.

Conclusions

Data Mining is the extraction of hidden predictive information from large databases. This is a new powerful new technology with great potential to help companies focus on the most important information in data warehousing. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. "The automated, prospective analyses offered by data mining move beyond the analyzes of past events provided by retrospective tools typical of decision support systems. Data mining is important to large systems because it finds things in large data repositories that you did not know existed. "A simple metaphor would be finding two needles in a haystack that match. The haystack is the database, the individual lengths of the hay represent your data fields, and the needles represent data fields with a relationship worth more to you than all the hay put together

Acknowledgement

We thank to Reader Sanjay Jain, computer Application department in our institute that provided the required facilities and all of the staffs that helped us to complete this work.

References

Abraham Silberschat Henry F. Korthz, S. Sudarshan Database System Concepts (2003). McGraw-Hill Science Engineering ISBN: 0072958863

Chen, M. S., J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996

Hand, D., H. Mannila, P. Smyth (2001). *Principles of data mining*. Mit press, cambridge, ma. Isbn 0-262-08290-x.

Kantardzic and Mehmed (2003). *Data mining: concepts, models, methods, and algorithms*. John wiley & sons. Isbn 0471228524

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm

http://www.bitscapesolutions.com/data-mining.htm1

Imielinski, T. and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

How to cite this article:

Kaushal Kishor Dwivedi and Daniel Glance Danny. 2024. Data Mining and Its Applications in Content of Higher Education M.P. *Int.J.Curr.Res.Aca.Rev.* 12(11), 92-97. doi: https://doi.org/10.20546/ijcrar.2024.1211.008