www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 689-691

e-ISSN : 2583-1062
Impact Factor : 7.001

INTRODUCTION TO DATA MINING AND DATA WAREHOUSE

Er. Gurjit Kaur[1], Er. Neelam Rani[2]

[1]Assistant Professor, Computer Science and Engineering Department, Sant Baba Bhag Singh University, Jalandhar, Punjab, India.

[2]Assistant Professor, Computer Science and Engineering Department, Sant Baba Bhag Singh University, Jalandhar, Punjab, India.

ABSTRACT

The amount of data being generated and stored is growing exponentially, due in large part to the continuing advances in computer technology. In this paper, discuss how the modern field of data mining can be used to extract useful knowledge from the data that surround us.

Keywords: Data mining, Data warehouse, Rule learning, OLAP, ETL.

1. INTRODUCTION

Generally, 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, Information, and Knowledge

Data

Data is unprocessed facts and figures without any added interpretation or analysis. "The price of crude oil is $80 per barrel."

Information

Information is data that has been interpreted so that it has meaning for the user. "The price of crude oil has risen from $70 to $80 per barrel" gives meaning to the data and so is said to be information to someone who tracks oil prices. The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

Knowledge

Knowledge is a combination of information, experience and insight that may benefit the individual or the organisation. "When crude oil prices go up by $10 per barrel, it's likely that petrol prices will rise by 2p per litre" is knowledge. Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

2. DATA WAREHOUSES

A data warehouse is a relational database that is designed for query and analysis rather than for transaction processing. It usually contains historical data derived from transaction data, but it can include data from other sources. It separates analysis workload from transaction workload and enables an organization to consolidate data from several sources. In addition to a relational database, a data warehouse environment includes an extraction, transportation, transformation, and loading (ETL) solution, an online analytical processing (OLAP) engine, client analysis tools, and other applications that manage the process of gathering data and delivering it to business users. A common way of introducing data warehousing is to refer to the characteristics of a data warehouse as set forth by William Inmon: Subject Oriented Integrated Nonvolatile Time Variant

Subject Oriented- Data warehouses are designed to help you analyze data. For example, to learn more about your company's sales data, you can build a warehouse that concentrates on sales. Using this warehouse, you can answer questions like "Who was our best customer for this item last year?" This ability to define a data warehouse by subject matter, sales in this case, makes the data warehouse subject oriented.

Integrated- Integration is closely related to subject orientation. Data warehouses must put data from disparate sources into a consistent format. They must resolve such problems as naming conflicts and inconsistencies among units of measure. When they achieve this, they are said to be integrated.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 689-691
e-ISSN : 2583-1062
Impact Factor : 7.001

Nonvolatile- Nonvolatile means that, once entered into the warehouse, data should not change. This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

Time Variant- In order to discover trends in business, analysts need large amounts of data. This is very much in contrast to online transaction processing (OLTP) systems, where performance requirements demand that historical data be moved to an archive. A data warehouse's focus on change over time is what is meant by the term time variant. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining

3. TYPES OF DATA MINING

Which type of data mining technique you should use really depends on the type of business problem that you are trying to solve. The most important objective of any data mining process is to find useful information that is easily understood in large data sets. There are a few important classes of tasks that are involved with data mining:

Anomaly or Outlier Detection

Anomaly detection refers to the search for data items in a dataset that do not match a projected pattern or expected behaviour. Anomalies are also called outliers, exceptions, surprises or contaminants and they often provide critical and actionable information. An outlier is an object that deviates significantly from the general average within a dataset or a combination of data. It is numerically distant from the rest of the data and therefore, the outlier indicates that something is out of the ordinary and requires additional analysis. Anomaly detection is used to detect fraud or risks within critical systems and they have all the characteristics to be of interest to an analyst, who can further analyse the anomalies to find out what's really going on. It can help find extraordinary occurrences that could indicate fraudulent actions, flawed procedures or areas where a certain theory is invalid. Important to note is that in large datasets, a small amount of outliers is common. Outliers may indicate bad data but may also be due to random variation or may indicate something scientifically interesting. In all cases, additional research is required.

Association Rule Learning

Association rule learning enables the discovery of interesting relations (interdependencies) between different variables in large databases. Association rule learning uncovers hidden patterns in the data that can be used to identify variables within the data and the co-occurrences of different variables that appear with the greatest frequencies. Association rule learning is often used in the retail industry when finding patterns in point-of-sales data. These patterns can be used when recommending new products to others based on what others have bought before or based on which products are bought together. If this is done correctly, it can help organisations increase their conversion rate. A well-known example is that thanks to data mining, Walmart, already in 2004, discovered that Strawberry Pop-tarts sales increase by seven times prior to a hurricane. Since this discovery, Walmart places the Strawberry Pop-Tarts at the checkouts prior to a hurricane.

Clustering Analysis

Clustering analysis is the process of identifying data sets that are similar to each other to understand the differences as well as the similarities within the data. Clusters have certain traits in common that can be used to improve targeting algorithms. For example, clusters of customers with similar buying behaviour can be targeted with similar products and services in order to increase the conversation rate. A result from a clustering analysis can be the creation of personas. Personas are fictional characters created to represent the different user types within a targeted demographic, attitude and/or behaviour set that might use a site, brand or product in a similar way. The programming language R has large variety of functions to perform relevant cluster analysis and is therefore especially relevant for performing a clustering analysis.

Classification Analysis

Classification Analysis is a systematic process for obtaining important and relevant information about data, and metadata - data about data. The classification analysis helps identifying to which of a set of categories different types of data belong. Classification analysis is closely linked to cluster analysis as the classification can be used to cluster data. Your email provider performs a well-known example of classification analysis: they use algorithms that are capable of classifying your email as legitimate or mark it as spam. This is done based on data that is linked with the email or the information that is in the email, for example certain words or attachments that indicate spam.

Regression Analysis

Regression analysis tries to define the dependency between variables. It assumes a one-way causal effect from one variable to the response of another variable. Independent variables can be affected by each other but it does not mean

that this dependency is both ways as is the case with correlation analysis. A regression analysis can show that one variable is dependent on another but not vice-versa. Regression analysis is used to determine different levels of customer satisfactions and how they affect customer loyalty and how service levels can be affected by for example the weather.

4. DIFFERENT LEVELS OF ANALYSIS ARE AVAILABLE

Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure. Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution. Decision trees: Tree-shaped structures that represent [sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID) . CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID. Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique. Rule induction: The extraction of useful if-then rules from data based on statistical significance. Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

5. CONCLUSION

In many commercial fields, data mining is crucial for identifying trends, forecasting, knowledge discovery, etc. Finding patterns to determine future trends in business growth is made easier by data mining techniques and algorithms like classification, clustering, etc. Since data mining has a broad range of applications in practically every business where data is collected, it is regarded as one of the most significant database and information systems frontiers as well as one of the most promising multidisciplinary breakthroughs in information technology.

6. REFERENCES

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.
[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.
[3] Z. Haiyang, "A Short Introduction to Data Mining and Its Applications", IEEE, 2011
[4] Rohit Arora,Suman “Comparative Analysis of Classification Algorithms on Different Datasets using WEKA“ International Journal of Computer Applications (0975 – 8887) Volume 54– No.13, September 2012
[5] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.
[6] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
[7] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.