International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228683

A Review on Data Mining Concepts and Tools

Dr. L. C. Manikandan[*1], **Dr. R. K. Selvakumar**[2]

[*1]CSE, Valia Koonambaikulathamma College of Engineering & Technology, Trivandrum, Kerala, India
[2]CSE, CVR College of Engineering, Hyderabad, Telungana, India

Article Info
Publication Issue :
Volume 8, Issue 6
November-December-2022
Page Number : 600-605
Article History
Accepted: 05 Dec 2022
Published: 20 Dec 2022

ABSTRACT

Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It is well-known that data mining is a significant sub-field of knowledge management and that it is one of the most important information exploration phases in the database cycle. Over the ensuing decades, the practice of data mining will increase in market and learning companies. An introduction to the fundamentals of data mining is provided in this paper.

Keywords: Data mining, Web-Mining, Data Warehouse, KDD, OALP

I. INTRODUCTION

The availability and interchange of information over the internet have increased during the past ten years. As more businesses and organizations started to gather data about their own operations, database technologists looked for effective ways to retrieve, store and manipulate data, while the machine learning community concentrated on methods for creating, learning, and acquiring knowledge from the data [1]. Data mining is an interdisciplinary subfield of computer science and estimations with a general target to remove data (with canny methodologies) from a data set and change the data into a fathomable structure for extra usage. Data mining is the analysis adventure of the "knowledge discovery in databases" technique or KDD [2, 3].

II. DATA MINING

Data mining refers to extracting or mining Knowledge (usable data) from large amounts of data. The process of extracting a few valuable nuggets from a large quantity of raw material is known as mining

A. Knowledge Discovery from Databases

Finding usable information and patterns in data is a process known as knowledge discovery in databases (KDD). The KDD process produces information and patterns, which are then extracted using algorithms. A pattern means that the data (visual or not) are correlated that they have a relationship and that they are predictable [5, 7]. Knowledge discovery as a process is depicted in Figure 1.

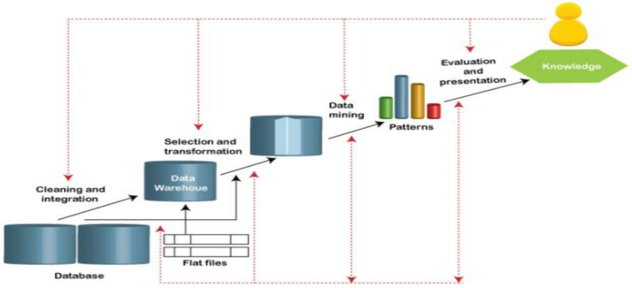

Figure 1. Knowledge Discovery Process

Databases, data warehouses, the Internet, other information sources, or dynamically streaming data can all be used as sources of information.

a. **Data cleaning:** To eliminate ambiguous and noisy data.
b. **Data integration:** Multiple data sources may be combined.
c. **Data selection:** Data relevant to the analysis task are retrieved from the database.
d. **Data transformation:** By executing summary or aggregation processes, data is converted or consolidated into forms suitable for mining.
e. **Data mining:** An essential process where intelligent methods are applied in order to extract data patterns
f. **Pattern evaluation:** Pattern evaluation is defined as identifying patterns representing knowledge based on given measures.
g. **Knowledge presentation:** Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Steps 1 through 4 are under data pre-processing, where data are prepared for mining. Data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

III. ARCHITECTURE OF A DATA MINING SYSTEM

Data mining system's architecture is depicted in Figure 2.

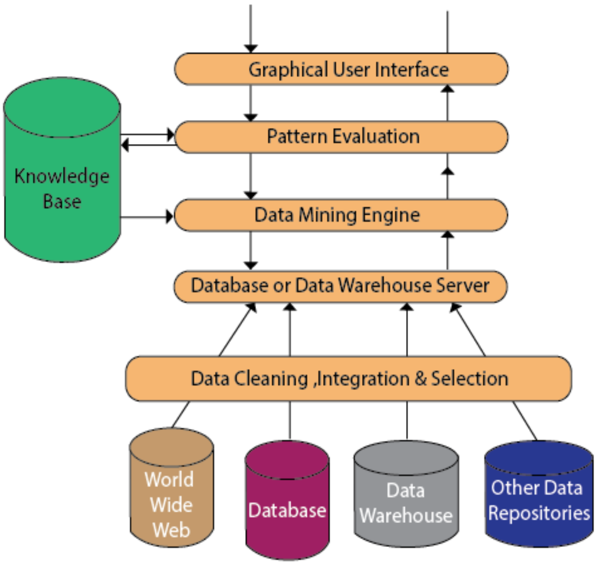

Figure 2. Architecture of data mining system

Data mining system components are:

a. **Database, data warehouse, World Wide Web or other information repository:** This consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
b. **Database or data warehouse server:** Database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
c. **Knowledge base:** This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns.
d. **Data mining engine:** It is essential to the data mining system and ideally consists of a set of functional modules for tasks such as characterization, classification, prediction, cluster analysis, and evolution analysis.
e. **Pattern evaluation module:** The main duty of the pattern evaluation module is to measure the pattern's investigation using a threshold value. It works in conjunction with the data mining engine to narrow the search to intriguing patterns.
f. **User interface:** This module handles User and data mining system communication. Without realising

how difficult the procedure is, this module enables the user to use the system quickly and effectively. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

IV. DATA MINING APPLICATIONS

Here is the list of areas where data mining is widely used:

a. **Classification:** Classification is a data mining function that assigns items in a collection to target categories or classes.

b. **Estimation:** Predict the attribute of a data instance. Example: estimate the percentage of marks of a student, whose previous marks are already known.

c. **Prediction:** Predictive model predicts a future outcome rather than the current behavior. Example: Predict next week's closing price for the Google share price per unit.

d. **Market basket analysis:** Data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves analyzing large data sets, such as purchase history, to reveal product groupings, as well as products that are likely to be purchased together.

e. **Business intelligence:** Business intelligence is a collection of applications and techniques used to transform data into actionable information.

f. **Business data analytics:** Data analysis is a method that can be used to investigate, analyze, and demonstrate data to find useful information.

g. **Bioinformatics:** Applications of data mining to bioinformatics include gene finding, protein function domain detection, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing etc.

h. **Web mining:** Web mining is the application of data mining techniques to discover patterns from the World Wide Web. It uses automated methods to extract both structured and unstructured data from web pages

i. **Text mining:** Text data mining is the process of transforming unstructured text into a structured format to identify meaningful patterns and new insights.

V. DATA MINING MODELS

Data mining models are strategies that are typically used to display information and different applications of information to certain concerns and issues [6, 7]. Figure 3 illustrates the data mining models.

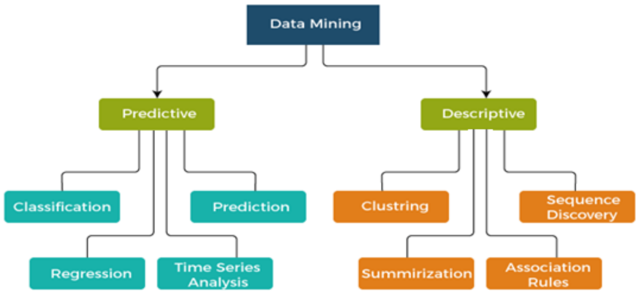

Figure 3. Data mining models

A. Predictive Model

A predictive model forecasts data values based on known findings gleaned from various data sets. The inclusion of additional historical data can be used to create predictive modeling. Predictive model tasks are given below:

a. **Classification:** Classification maps the data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data.

b. **Regression:** It is used for appropriate data. It is a technique that verifies data values for a function. (i). Linear Regression is associated with the search for the optimal line to fit the two attributes so that one attribute can be applied to predict the other.

(ii). Multi-Linear Regression involves two or more than two attributes and data are fit to multidimensional space.

c. **Time Series Analysis:** The value of an attribute is examined as it varies over time. Values usually are obtained as evenly spaced time points (daily, weekly, hourly, etc.).

d. **Prediction:** Prediction is used to identify data value based on the description of another corresponding data value. The prediction in data mining is known as Numeric Prediction. Generally, regression analysis is used for prediction. Example: In credit card fraud detection, data history for a particular person's credit card usage has to be analyzed. If any abnormal pattern was detected, it should be reported as 'fraudulent action'.

B. Descriptive Model

Using descriptive models, data correlations or patterns are identified. Instead of making new predictions, a descriptive model is used to study the features of the data being studied. Descriptive model tasks are given below:

a. **Clustering:** Clustering means grouping a set of objects, so that objects in the same group called a cluster are more similar than those in other group's clusters.

b. **Summarization:** Summarization holds a data set in more depth which is easy to understand form. The summarization briefly characterizes the contents of the database.

c. **Association Rules:** Association rules determine a causal relationship between huge sets of data objects. Example: a list of items you purchase at the grocery store for the past six months data, and it calculates a percentage at which items are purchased together. For example, what are the chances of you buying milk with cereal?

d. **Sequence Discovery:** It is used to determine sequential patterns in data, based on a time sequence of actions. Example: Heart pulse

VI. TECHNOLOGIES FOR DATA MINING

Several techniques used in the development of data mining methods [7] are shown in Figure 4. Some of them are mentioned below:

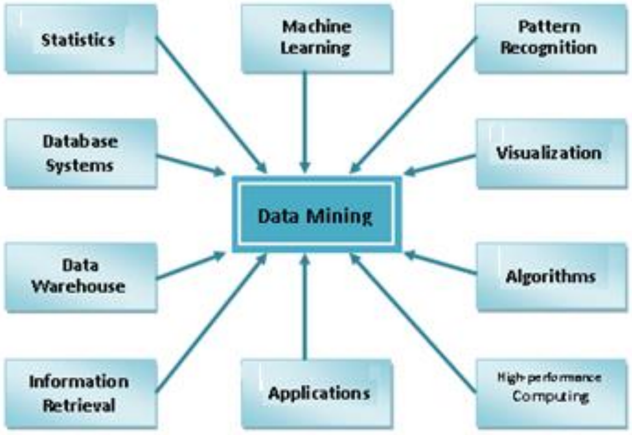

Figure 4. Technologies for Data Mining

a. **Statistics:** Statistics uses the mathematical analysis to express representation, model and summarize experimental data or real world observations. Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

b. **Machine learning:** Investigates how computers can learn (or improve their performance) based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. Methods for machine learning are:

(i) Supervised learning: It is based on the classification. In this method, the desired outputs are included in the training dataset.

(ii) Unsupervised learning: It is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

(iii) Semi-supervised learning: Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions.

(iv) Active learning: It is a powerful approach in analyzing the data efficiently. The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

c. **Information retrieval:** Finding relevant information from a large document.

d. **Database systems and data warehouse:** Databases are used for the purpose of recording the data as well as data warehousing. Online Transactional Processing (OLTP) uses databases for day to day transaction purpose. To remove the redundant data and save the storage space, data is normalized and stored in the form of tables. Data warehouses are used to store historical data which helps to take strategical decision for business. It is used for online analytical processing (OALP), which helps to analyze the data.

e. **Data Visualization:** Data can be presented in visual ways through charts, graphs, maps, diagrams, and more. This is a primary way in which data scientists display their findings.

f. **Pattern recognition:** pattern recognition and data mining is to extract useful information or knowledge from large data sets. Pattern Recognition aims at extracting, matching, quantifying and predicting patterns in measurements.

VII. DATA MINING TOOLS

The Most Popular Data Mining Tools are briefly explained below [4].

a. **Rapid miner:** One of the best predictive analytic systems was created by the firm called Rapid Miner, and it is called Rapid Miner. Rapid Miner has model-based frameworks that enable quick delivery with fewer errors.

b. **Mahout:** Mahout is an Apache open source machine learning library. The algorithms it uses can be broadly categorised as machine learning or collective intelligence. This can signify a variety of things, but for Mahout it mostly refers to the grouping and categorization that recommender engines currently use.

c. **Orange:** Orange is capable of reading files in their native tab-delimited format and filling them with data from any of the main common spreadsheet file types, including CSV and Excel. A header row containing the names of the characteristic (columns) is where native design begins. The attribute type, which can be continuous, discrete, temporal, or string, is indicated in the second title row.

d. **Weka:** WEKA offers cultural algorithm implementations that you can easily adapt to your dataset. There are certain incremental techniques in WEKA that can be utilised to handle very huge datasets. You can drag boxes that represent learning algorithms on the Knowledge Flow interface.

e. **DataMelt:** DataMelt is a tool that uses open-source software in conjunction with a cogent user interface and tools that are competitive with those found in commercial products to generate data study situations. In element physics, where data mining is a key duty, DataMelt has its origins.

VIII. CONCLUSION

This research paper focused on the fundamental concepts of data mining, data mining system components, data mining models, data mining technologies and data mining tools. This would be very helpful for young readers and researchers to understand the basic concepts of data mining.

IX. REFERENCES

[1] Ashish Barhate, Sumit Gupta et al, "Study of Data Mining Concepts", International Journal of New Innovations in Engineering and Technology, Vol. 9, Issue 1, October 2018.

[2] Ankita, "Review Paper on Data Mining Concepts and Its Techniques", Journal of Advances and Scholarly Researches in Allied Education, Vol. 14, Issue 2, January 2018.

[3] R Agrawal, T Mielinski, A Swami, "Database Mining: A Performance Perspective [J]", IEEE Transactions on Knowledge and Data Engineering, Vol.12, pp. 914-925, 1993.

[4] Shilpa.N, Ammulu.T and Prameela.N, "Data Mining Concepts and Techniques", Journal of Emerging Technologies and Innovative Research, Vol. 9, Issue 12, December 2022.

[5] Ralf Mikut and Markus Reischl, "Data mining tools", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011.

[6] Ming-Syan Chen, Jiawei Han and Philip S Yu, "Data Mining: An Overview from a Database Perspective[J]", IEEE Transactions on Knowledge and Data Engineering, 8(6), pp. 866-883, 1996.

[7] Dunham M H, "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, 2003.

AUTHOR PROFILE

Dr. L. C. Manikandan is working as Professor at Valia Koonambaikulathamma College of Engineering and Technology, Thiruvananthapuram, Kerala INDIA. He has received his Ph.D. and M.Tech. Degree in Computer and Information Technology from Manonmaniam Sundaranar University, M.Sc., and B.Sc. degree in Computer Science from Bharathidasan and Manonmaniam Sundaranar University. He has 18 years of teaching experience in reputed institutions. He has published several research papers in various reputed international journals and published three textbooks. He carries out research in Digital Image Processing, Video Surveillance and Video coding.

Dr.R.K.Selvakumar received his Ph.D degree in Computer Engineering from Manonmaniam Sundaranar University, India. He has completed the Master degrees M.Sc.(Maths), DOEACC-C(CT), M.C.A., M.Phil.(CS) and M.Tech.(C&IT). He is working as Professor at CVR College of Engineering, Hyderabad, Telungana, INDIA. He carries out research in Digital Image Processing, Video Surveillance, Video coding, Digital Watermarking and Soft Computing.

Cite this Article

Dr. L. C. Manikandan, Dr. R. K. Selvakumar, "A Review on Data Mining Concepts and Tools", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 600-605, November-December 2022. Available at doi : https://doi.org/10.32628/CSEIT228683
Journal URL : https://ijsrcseit.com/CSEIT228683