Chapter 16

Data Mining and Data Warehousing:
Introduction to Data Mining and Data Warehousing

Sathiyamoorthi V
Sona College of Technology, India

ABSTRACT

It is generally observed throughout the world that in the last two decades, while the average speed of computers has almost doubled in a span of around eighteen months, the average speed of the network has doubled merely in a span of just eight months! In order to improve the performance, more and more researchers are focusing their research in the field of computers and its related technologies. Data Mining is one such research area. It extracts useful information the huge amount of data present in the database. The discovered knowledge can be applied in various application areas such as marketing, fraud detections and customer retention. It discovers implicit, previously unknown and potentially useful information out of datasets. Recent trend in data mining include web mining where it discover knowledge from web based information to improve the page layout, structure and its content.

INTRODUCTION

Data mining is the process of nontrivial extraction of implicit, previously unknown and potentially useful information from the raw data present in the large database (Jiawei et al. 2006). It is also known as Knowledge Discovery in Databases (KDD). Data mining techniques can be applied upon various data sources to improve the value of the existing information system. When implemented on high performance client and server system, data mining tools can analyze large databases to deliver highly reliable results. It is also described that the data mining techniques can be coupled with relational database engines (Jiawei et al. 2006). Data mining differs from the conventional database retrieval in the fact that it extracts hidden information or knowledge that is not explicitly available in the database, whereas database retrieval

extracts the data that is explicitly available in the databases through some query language. Based on the fact that, a certain degree of intelligence is incorporated in the system, data mining could further be viewed as a branch of artificial intelligence and thus, it could be treated as an intelligent database manipulation system. Dunham et al. (2006) have explained that data mining is an interdisciplinary field that incorporates concepts and techniques from several disciplines such as statistics, neural networks and machine learning in the process of knowledge discovery. Data warehousing is the location where it stores subject oriented and task relevant data for an organization decision support system. It contains data that are most important and relevant to decision making process. Hence, this chapter describes the functionality of data mining and data warehousing system with its applications. Also, it focuses on Web mining where it addresses the issues and challenges present in it. Finally, it describes the integration technique where data mining and data warehousing system can be combined for an effective functionality. As data mining is an interdisciplinary field, it uses algorithms and techniques from various fields such as statistics, machine learning, artificial intelligence, neural networks and database technology. The most commonly used methods that assist in data mining tasks are (Jiawei et al. 2006) given below:

- **Artificial Neural Networks (ANN):** A non-linear predictive model comprises of different layers namely input, hidden and output layers that learn through training and resemble biological neural network in a structure.
- **Decision Tree:** A tree structure comprises of nodes and branches and represents a set of decisions. A node in decision tree represents conditions and branches of outcome. These decisions generate rules for the classifications of a dataset. Specific decision tree method includes classification and regression trees.
- **Genetic Algorithm (GA):** This Evolutionary optimization technique uses operators such as genetic combination, mutation, and natural selection in a design-based concept of evolution. This can be applied to optimization problem that either maximize or minimize the given objective function.
- **Nearest Neighbor Method:** A technique that classifies each record in a dataset based on a combination of the classes of 'k' records that are most similar to its historical dataset. Sometimes called as the K-Nearest Neighbor (KNN) technique.
- **Rule Induction:** This is the extraction of useful if-then rules from the dataset.

BACKGROUND

It is described that data mining can be viewed as a crucial step in knowledge discovery process which is shown in Figure 1. It is composed of various phases such as:

- Pre-processing
- Data Mining
- Pattern Extraction
- Pattern Evaluation
- Knowledge Presentation

Related Content

Strategic Information Systems Planning (SISP): An Empirical Evaluation of Adoption of Formal Approaches to SISP in Australian Organizations
Zijad Pita, France Cheongand Brian Corbitt (2010). *International Journal of Strategic Decision Sciences (pp. 28-61).*
www.irma-international.org/article/strategic-information-systems-planning-sisp/44973

Developing a DSS for Allocating Gates to Flights at an International Airport
Vincent F. Yu, Katta G. Murty, Yat-wah Wan, Jerry Dannand Robin Lee (2009). *International Journal of Decision Support System Technology (pp. 46-68).*
www.irma-international.org/article/developing-dss-allocating-gates-flights/1744

Access and Use of Information by Primary Health Care Providers in Rural Uganda: A Qualitative Approach
Maria G. N. Musoke (2010). *International Journal of Decision Support System Technology (pp. 1-9).*
www.irma-international.org/article/access-use-information-primary-health/43906

Audit Quality and Readability of the Annual Reports
Ashutosh Deshmukhand Xin Zhao (2020). *International Journal of Strategic Decision Sciences (pp. 76-90).*
www.irma-international.org/article/audit-quality-and-readability-of-the-annual-reports/246324

New Features Extracted From Renal Stone NCCT Images to Predict Retreatment After Shock Wave Lithotripsy (SWL)
Toktam Khatibi, Mohammad Mehdi Sepehri, Mohammad Javad Soleimaniand Pejman Shadpour (2017). *Handbook of Research on Data Science for Effective Healthcare Practice and Administration (pp. 296-316).*
www.irma-international.org/chapter/new-features-extracted-from-renal-stone-ncct-images-to-predict-retreatment-after-shock-wave-lithotripsy-swl/186943