IJARSCT

ISSN (Online) 2581-9429

Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 6, June 2022

Overview of Application of Data Mining

Parveen Kumar
B. Tech (CSE) Student
Dronacharya College of Engineering, Gurgaon, Haryana, India

Abstract: *Internet of Things (IoT) has been growing rapidly due to recent advancements in communications and sensor technologies. Interfacing an every object together through internet looks very difficult, but within a frame of time Internet of Things will drastically change our life. The enormous data captured by the Internet of Things (IoT) are considered of high business as well as social values and extracting hidden information from raw data, various data mining algorithm can be applied to IoT data. In this paper, we survey systematic review of various data mining models as well as its application in Internet of Thing (IoT) field along with its merits and demerits. At last, we discussed challenges in IoT.*

Keywords: Data Mining Process, Data Preprocessing Phase, Association Analysis, Time Series Analysis, Data Mining Applications.

I. INTRODUCTION

Data mining discovers patterns within data, using predictive techniques. These patterns play a very important role in the decision making because they emphasize areas where business processes require improvement. Using the data mining solutions, organizations can increase their profitability, can detect fraud, or may enhance the risk management activities. The models discovered by using data mining solutions are helping organizations to make better decisions in a shorter amount of time.

Data mining is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It has been defined as:

The automated analysis of large or complex data sets in order to discover significant patterns or trends that would otherwise go unrecognised.

The process of digging through data to discover hidden connections and predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn't coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), artificial intelligence (human-like intelligence displayed by software and/or machines) and machine learning (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of big data and affordable computing power.

1.1 The Data Mining Process

The data mining process is a pipeline containing many phases such as data cleaning, feature extraction, and algorithmic design. In this section, we will study these different phases. The workflow of a typical data mining application contains the following phases:

1. **Data Collection:** Data collection may require the use of specialized hardware such as a sensor network, manual labor such as the collection of user surveys, or software tools such as a Web document crawling engine to collect documents. While this stage is highly application-specific and often outside the realm of the data mining analyst, it is critically important because good choices at this stage may significantly impact the data mining process. After the collection phase, the data are often stored in a database, or, more generally, a data warehouse for processing.
2. **Feature Extraction and Data Cleaning:** When the data are collected, they are often not in a form that is suitable for processing. For example, the data may be encoded in complex logs or free-form documents. In many cases, different types of data may be arbitrarily mixed together in a free-form document. To make the data suitable for processing, it is essential to transform them into a format that is friendly to data mining algorithms, such as multidimensional, time series, or semistructured format. The multidimensional format is the most common one,

DOI: 10.48175/568

in which different fields of the data correspond to the different measured properties that are referred to as features, attributes, or dimensions. It is crucial to extract relevant features for the mining process. The feature extraction phase is often performed in parallel with data cleaning, where missing and erroneous parts of the data are either estimated or corrected. In many cases, the data may be extracted from multiple sources and need to be integrated into a unified format for processing. The final result of this procedure is a nicely structured data set, which can be effectively used by a computer program. After the feature extraction phase, the data may again be stored in a database for processing.

1.2 The Data Preprocessing Phase

The data preprocessing phase is perhaps the most crucial one in the data mining process. Yet, it is rarely explored to the extent that it deserves because most of the focus is on the analytical aspects of data mining. This phase begins after the collection of the data, and it consists of the following steps:

1. **Feature Extraction:** An analyst may be confronted with vast volumes of raw documents, system logs, or commercial transactions with little guidance on how these raw data should be transformed into meaningful database features for processing. This phase is highly dependent on the analyst to be able to abstract out the features that are most relevant to a particular application. For example, in a credit-card fraud detection application, the amount of a charge, the repeat frequency, and the location are often good indicators of fraud. However, many other features may be poorer indicators of fraud. Therefore, extracting the right features is often a skill that requires an understanding of the specific application domain at hand.
2. **Data cleaning:** The extracted data may have erroneous or missing entries. Therefore, some records may need to be dropped, or missing entries may need to be estimated. Inconsistencies may need to be removed.
3. **Feature selection and transformation:** When the data are very high dimensional, many data mining algorithms do not work effectively. Furthermore, many of the high dimensional features are noisy and may add errors to the data mining process. Therefore, a variety of methods are used to either remove irrelevant features or transform the current set of features to a new data space that is more amenable for analysis. Another related aspect is data transformation, where a data set with a particular set of attributes may be transformed into a data set with another set of attributes of the same or a different type. For example, an attribute, such as age, may be partitioned into ranges to create discrete values for analytical convenience.

II. ASSOCIATION ANALYSIS

Association rule mining focuses on the market basket analysis or transaction data analysis, and it targets discovery of rules showing attribute value associations that occur frequently and also help in the generation of more general and qualitative knowledge which in turn helps in decision makin.

1. For the first catalog of association analysis algorithms, the data will be processed sequentially. The a priori based algorithms have been used to discover intratransaction associations and then discover associations; there are lots of extension algorithms. According to the data record format, it clusters into 2 types: Horizontal Database Format Algorithms and Vertical Database Format Algorithms; the typical algorithms include MSPS and LAPIN-SPAM. Pattern growth algorithm is more complex but can be faster to calculate given large volumes of data. The typical algorithm is FP-Growth algorithm.
2. In some area, the data would be a flow of events and therefore the problem would be to discover event patterns that occur frequently together. It divides into 2 parts: event-based algorithms and event-oriented algorithms; the typical algorithm is PROWL.

In order to take advantage of distributed parallel computer systems, some algorithms are developed, for example, Par-CSP.

III. TIME SERIES ANALYSIS

A time series is a collection of temporal data objects; the characteristics of time series data include large data size, high dimensionality, and updating continuously. Commonly, time series task relies on 3 parts of components, including representation, similarity measures, and indexing.

DOI: 10.48175/568

1. One of the major reasons for time series representation is to reduce the dimension, and it divides into three categories: model based representation, nondata-adaptive representation, and data adaptive representation. The model based representations want to find parameters of underlying model for a representation.. In data adaptive representations, the parameters of a transformation will change according to the data available and related works including representations version of DFT /PAA and indexable PLA .
2. The similarity measure of time series analysis is typically carried out in an approximate manner; the research directions include subsequence matching and full sequence matching.
3. The indexing of time series analysis is closely associated with representation and similarity measure part; the research topic includes SAMs (Spatial Access Methods) and TS-Tree.

IV. DATA MINING APPLICATIONS

4.1 Data Mining in e-Commerce

Data mining enables the businesses to understand the patterns hidden inside past purchase transactions, thus helping in planning and launching new marketing campaigns in prompt and cost effective way. e-commerce is one of the most prospective domains for data mining because data records, including customer data, product data, users' action log data, are plentiful; IT team has enriched data mining skill and return on investment can be measured. Researchers leverage association analysis and clustering to provide the insight of what product combinations were purchased; it encourages customers to purchase related products that they may have been missed or overlooked. Users' behaviors are monitored and analyzed to find similarities and patterns in Web surfing behavior so that the Web can be more successful in meeting user needs. A complementary method of identifying potentially interesting content uses data on the preference of a set of users, called collaborative filtering or recommender systems, and it leverages user's correlation and other similarity metrics to identify and cluster similar user profiles for the purpose of recommending informational items to users. And the recommender system also extends to social network, education area, academic library, and tourism.

4.2 Data Mining in Industry

Data mining can highly benefit industries such as retail, banking, and telecommunications; classification and clustering can be applied to this area. One of the key success factors of insurance organizations and banks is the assessment of borrowers' credit worthiness in advance during the credit evaluation process. Credit scoring becomes more and more important and several data mining methods are applied for credit scoring problem. Retailers collect customer information, related transactions information, and product information to significantly improve accuracy of product demand forecasting, assortment optimization, product recommendation, and ranking across retailers and manufacturers. Researchers leverage SVM, support vector regression, or Bass model to forecast the products' demand.

4.3 Data Mining in Health Care

In health care, data mining is becoming increasingly popular, if not increasingly essential. Heterogeneous medical data have been generated in various health care organizations, including payers, medicine providers, pharmaceuticals information, prescription information, doctor's notes, or clinical records produced day by day. These quantitative data can be used to do clinical text mining, predictive modeling, survival analysis, patient similarity analysis, and clustering, to improve care treatment and reduce waste. In health care area, association analysis, clustering, and outlier analysis can be applied.

Treatment record data can be mined to explore ways to cut costs and deliver better medicine. Data mining also can be used to identify and understand high-cost patients and applied to mass of data generated by millions of prescriptions, operations, and treatment courses to identify unusual patterns and uncover fraud.

REFERENCES

[1]. Shen Bin, Liu Yuan, and Wang Xiaoyi. Research on data mining models for the internet of things. In Image Analysis and Signal Processing (IASP), 2010 International Conference on, pages 127– 132. IEEE, 2010.

[2]. Olaiya Folorunsho and Adesesan Adeyemo. Application of data mining techniques in weather prediction and climate change studies. 4, 02 2012.

DOI: 10.48175/568

[3]. Jeu Young Kim, Hark-Jin Lee, Ji-Yeon Son, and Jun-Hee Park. Smart home web of objects-based iot management model and methods for home data mining. In Network Operations and Management Symposium (APNOMS), 2015 17th Asia-Pacific, pages 327–331. IEEE, 2015.

[4]. Saral Nigam, Shikha Asthana, and Punit Gupta. Iot based intelligent billboard using data mining. In Innovation and Challenges in Cyber Security (ICICCS-INBUSH), 2016 International Conference on, pages 107–110. IEEE, 2016.

[5]. Alexander Muriuki Njeru, Mwana Said Omar, Sun Yi, Samiullah Paracha, and Muhammad Wannous. Using iot technology to improve online education through data mining. In Applied System Innovation (ICASI), 2017 International Conference on, pages 515–518. IEEE, 2017.

[6]. S. Ranka and V. Singh, “CLOUDS: a decision tree classifier for large datasets,” in Proceedings of the 4th Knowledge Discovery and Data Mining Conference, pp. 2–8, 1998.

[7]. R. M. Alguliev, R. M. Aliguliyev, M. S. Hajirahimova and C. A. Mehdiyev, "MCMR: Maximum coverage and minimum redundant text summarization model", Expert Systems with Applications Vol. 38, No. 12, pp.: 14514-14522, 2011.

DOI: 10.48175/568