International Core Journal of Engineering Volume 5 Issue 10, 2019
ISSN: 2414-1895 DOI: 10.6919/ICJE.201909_5(10).0021

Overview of Data Mining in the Era of Big Data

Chenggang Zhen [a], Cong Jiang [b]
School of North China Electric Power University, Baoding 071000, China;
[a]zhencg@163.com, [b]846823191@qq.com

Abstract

The era of big data has come into our life, the acceleration of mass data growth, people with the naked eye to observe the work in the data pile with the growth of data becomes more and more laborious, data mining technology came into being.This paper analyzes the current situation of big data mining, expounds the relevant concepts, characteristics, process and relevant algorithms of data mining, and analyzes the future development direction and trend of data mining technology.

Keywords

Data mining; The Apriori algorithm; Clustering algorithm; Decision tree mining algorithm.

1. Introduction

Data mining is considered as an emerging topic across knowledge fields and disciplines, which enables us to analyze and dig out useful information from the original simple query of data, providing technical support for our decision-making behavior.Due to the huge amount and variety of data in real life, in order to meet the needs of people to obtain useful knowledge from the massive data, all walks of life are devoted to the research of this hot topic -- data mining.

Since the first KDD&Data Mining international academic conference was held in Montreal, Canada in 1995, the annual international academic conference has been held since then.With the efforts of excellent people from all walks of life, data mining has developed rapidly and achieved considerable research results.Many data mining software products have been applied in some European and north American countries.At present, widely used data mining systems include Intelligent Miner from IBM, SetMiner from SGI, Clementine from SPSS, Enterprise Miner from SAS, Warehouse Studio from Sybase, etc.In our country, mining technology is also improving.

2. Data Mining Related Knowledge

2.1 Concept of Data Mining

Data mining is a process of processing a large number of fuzzy and incomplete data to dig out useful information and knowledge with special relationship.The main task of this process is to find the relevant patterns hidden in the data.Data mining has a wide range of applications.Data mining takes statistical analysis as the starting point, adds the idea of sampling, estimation and hypothesis testing in the statistical field, and integrates the theory and related technology of machine learning.

2.2 Characteristics of data mining

A large number of enterprises and institutions have attached great importance to the data mining technology to fully mine the valuable information in the data, they hope to be able to operate it in their business process, thus saving a lot of manpower and material resources.In fact, the idea of applying data mining technology to business has been around for a long time. In short, it is to

ISSN: 2414-1895 DOI: 10.6919/ICJE.201909_5(10).0021

automatically help users and optimize their decisions based on experience.What is different now is that there is a huge amount of data to support it.

The characteristics of data mining are:(1) find rich structural relationships among data;(2) analyze the non-relational data and make further speculation based on it;(3) obtain valuable information;(4) able to mine distributed and heterogeneous data under the network structure, and then conduct effective integration with the operational system to achieve real-time dynamic correspondence between the data mining system and the language model system, effectively combine and provide decision support;(5) to further explore larger, more complex and higher-dimensional data sets, and increase the plasticity of the system through the perfect combination of data mining system and language model system;(6) data sources can be mined from more sources by adding data mining and mobile computing.

2.3 Characteristics of data mining

Data mining is to mine valuable knowledge from massive data and analyze the obtained information to make optimized decisions.Data mining knowledge is a process, it needs to really consider the data mining technology and practical experience with the actual situation and needs of the enterprise, practice for many times, modify repeatedly, to achieve certain results.The process of data mining mainly includes:(1)determine the target.Select the appropriate sample set according to the target;(2)data preprocessing.Conduct research on the quality of data to remove some abnormal, duplicate and wrong data;(3)data transformation.Feature extraction to ensure that these features can describe the data;(4)data modeling.Training data and training model are obtained.Analyze the results and improve them.According to the corresponding method, the mining results are analyzed to optimize the organizational structure.

3. Data Mining Algorithm

To be specific, current data mining technologies mainly involve neural network, decision tree, genetic algorithm, mathematical statistical analysis, association rules and clustering analysis.

3.1 Association Rule Mining Algorithm

In the field of data mining, association rule algorithm should be an important research method, which is widely used in various industries. Apriori algorithm is the classical algorithm of association rule algorithm.

Apriori algorithm: based on the frequent item set of association rules, the Apriori algorithm is obtained by corresponding calculation. Other algorithms in data mining are basically derived from this algorithm, so Apriori algorithm is the foundation of all association rule algorithms.In Apriori algorithm, there are many concepts involved, including: item set, frequent item set, item set frequency, association rule. The collection of all items is collectively called itemset, where itemset transactions occur, and the concept of itemset frequency is the frequency at which itemset transactions occur in the itemset.In the itemset frequency, if the itemset can meet the requirement of minimum support, we can call it frequent itemset. After finding the frequent item sets of the transaction database, the relationships between the data items are discovered, and the association rules are formed.

Apriori algorithm is based on the Apriori algorithm principle. In the trading process, the more often two kinds of items appear in pairs, the more correlated they are, indicating that they are strongly correlated. Apriori algorithm needs to traverse the database for many times. After the first time, it needs to count the number of times of each item and delete some items that do not meet the requirements of minimum support. Before the second traversal, the objects obtained from the first traversal should be combined in pairs. Then, the number of combinations is counted, and some combinations that do not meet the requirements are deleted. Then, the second iteration is repeated until no new combinations are formed, which represents the end of the implementation process.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201909_5(10).0021

3.2 Decision Tree Mining Algorithm

The reason why decision tree is widely used in the field of data analysis and mining is that it does not need to know much background knowledge in the construction process, and this method is quite handy for exploring data mining and high-dimensional data.The advantage of decision tree is that typical rules can be directly applied to the negative aspects of business optimization without sorting, and the rules are simple and easy to understand, and the extreme value of data has little influence on them.

The concrete process of decision tree classification: establishing decision tree;Simplified decision tree;Decision tree model generated by prediction;Use model classification;Extract the classification rules.

ID3 algorithm is a decision tree algorithm of great practical value. Its biggest feature lies in the criteria for selecting independent variables: selecting the highest information entropy gain as the conditional attribute of node segmentation.For non-terminal successor nodes, the same selection criteria are used to segment training samples, so that the information required to classify these processed nodes is minimal.

3.3 Clustering Algorithm

Clustering algorithm is widely used in data analysis, customer copy, artificial intelligence and other fields.Different from classification, cluster analysis deals with unknown classes.Clustering is the process of grouping many data according to corresponding standards or forming multiple clusters.There are many common features in the group, and there are many differences between the groups. As far as possible, the data points in the same group are as similar as possible.Will be heterogeneous data points with as much heterogeneity as possible.

(1) Hierarchical method

Hierarchical method is the hierarchical decomposition of a given set of data objects, which can be divided into aggregation and splitting.The way it condenses is from the bottom up;Splitting is the opposite, from the top down.The disadvantage of the hierarchical clustering method is that once it starts, it cannot finish, nor can it correct the wrong decision.Of course, we often use two methods to improve it: first, the chameleon method, when dividing, analyzes the connection between objects;In the second method, the aggregation algorithm is used first, and then iterative repositioning is used to compensate for the shortcomings of the hierarchical method.

(2) Grid-based method

It transforms space into a finite number of meshes based on specific rules and performs cluster analysis on the meshes.The advantage of this method is that it is fast and its processing time is only related to the number of units in each dimension of the quantized space, independent of the data object.

(3) Partition based method

Given the data set D of n data objects, k partitions are constructed through multiple partition, each partition is 1 cluster, most of which are based on distance.The most classical algorithms are k-means algorithm and k-medoids algorithm.

4. Conclusion

With the advent of the era of big data, the rapid growth of data volume and the diversification of database, all walks of life will need to apply data mining technology. This paper also analyzes the characteristics of various data mining algorithms, each algorithm has its own suitable scenario, also has its own advantages and disadvantages. We apply the process in practice, and use a variety of methods to solve the problem. In addition, the results of data mining are not absolutely correct, and we need to combine more data and factors for a more comprehensive analysis to make a more scientific decision.

ISSN: 2414-1895 DOI: 10.6919/ICJE.201909_5(10).0021

References

[1] Z.W. Zhang, J.N. Wang: Crane Design Manual (China Railway Press, China 1998), p.683-685. (In Chinese)[1]Dong xingchao. Development and research of data collection and application system based on clustering optimization [D]. Zhejiang university,2019.

[2] Tao zhi. Design and implementation of cache system based on association rules [D]. Huazhong university of science and technology,2018.

[3] Yan dongming. Research on key technologies of information association based on data mining [D]. Changchun university of science and technology,2018.

[4] Wu xiaodong, zeng yuzhu. Data mining of college student performance based on Apriori algorithm [J]. Journal of langfang normal university (natural science edition),2019,19(01):31-36.

[5] Wang jinqing. Improvement and application of decision tree C4.5 algorithm [D]. Xi 'an university of technology,2017.

[6] Zang zhaojie. Research on k-medoids clustering algorithm based on Spark [D]. Dalian university,2018.

[7] Zou yi. Review of data mining technology [J]. Information communication,2016(12):164-165.

[8] Zou zhiwen, zhu jinwei. Research and review of data mining algorithms [J]. Computer engineering and design,2005(09):2304-2307.

[9] Chen shan. Analysis of classical algorithms for data mining [J]. Electronic technology and software engineering,2019(15):128-129.

[10] Yin Ming, wang wenjie, zhang guangyu, jiang jijiao. A maximum frequent itemset mining algorithm based on adjacency table [J]. Journal of electronics and information technology, 2019,41(08):2009-2016.