Design of customer marketing big data processing system based on data mining clustering technology

Jingzhe Wang

School of Information Engineering, Zhengzhou University, Henan Province, Zhengzhou, 450000

346591653@163.com

Keywords: Data mining; Hadoop; MapReduce; HBase

Abstract. Data mining technology brings together databases, artificial intelligence, machine learning, statistics, visualization, parallel computing in different fields, to build their own methodology. Use database technology for front-end data processing, application of machine learning methods to extract useful knowledge from the data processed, the data and to analyze the characteristics and trends behind the final data given about the overall characteristics and trends. Use of visualization techniques the human observation and intelligence into the system, with an intuitive graphical information mode, association or trend data presented to decision-makers, enabling users to interactively analyze data. Customer relationship management is an important means to maintain market competitiveness and indispensable part. The introduction of data mining technology to achieve the goal of high-quality customer relationship management, give full play to the role of customer relationship management. At present, many foreign companies in order to gain a competitive advantage, actively engaged in the study of human and material resources and applications, and achieved a better return on investment.

Introduction

With the rapid development of technology, the Internet has become the world's largest database, driven by emerging applications, we have entered the era of big data [1, 2]. How to extract the valuable information from large-scale rules, provide a reference for people's work, life and play an active role in promoting, has become a hot topic of data mining community. When conventional computer architectures has been unable to deal with these data, the emergence of cloud computing will undoubtedly provide us with a very good solution, its low-cost, unlimited scalable storage and computing power are leading to the Internet under a brilliant [3].

Given the numerous deficiencies and significant advantages of cloud computing alone computing, this study uses cloud computing technology, cloud storage, cloud database technology and data mining technology combined method, stored as text data mining and database mining, for example, based on data mining platform Hadoo theory, operation mechanism distributed program, and tap the theoretical model constructed from robust distributed data [4, 5]. It can be made stable, efficient data mining methods. In this study, greatly reducing programming time and calculations, saving a certain amount of manpower, material and provide a scientific basis for the data mining staff working on Hadoop platform for accurate and rapid excavation work in large-scale data provided solid protection [6, 7].

Data mining is a customer relationship management engine. Although customer relationship management is essentially talking about is a management and not the result of technological progress, but technological progress does offer the best opportunity for its development. In the past due to technical limitations, open enterprise information system deficiencies caused between the cross-system integration is not easy to achieve, so the full understanding of customers, grasp the customer's characteristics and needs only an ideal. However, under conditions of rapid development of network technology, through cross-platform integration, coupled with increasingly sophisticated data warehouse and data mining technology, allowing companies to more effectively control customer behavior and needs. If the business profits as its goal, customer relationship management is

to achieve the most useful tool for this aim, data mining is the best engine of this tool. Data mining is the process of looking for hidden information from large amounts of data.

Data mining technology and clustering algorithm

Data mining is digging out implicit, previously unknown, potentially useful for decision-making and rules of knowledge from large data. These rules contain a specific relationship between a set of objects in the database, revealing some useful information for management decision-making, marketing, financial forecasting to provide evidence. For example, a supermarket manager from a large number of transaction data found that many fathers in the purchase of beer is like way back some disposable diapers, so the two commodities in very close location for fathers later, the results of these two commodities sales have increased a lot.

Data mining system has four main modules: user interface, data preparation, (also known as data preprocessing), interpretation and evaluation of mining and pattern (Figure 1).

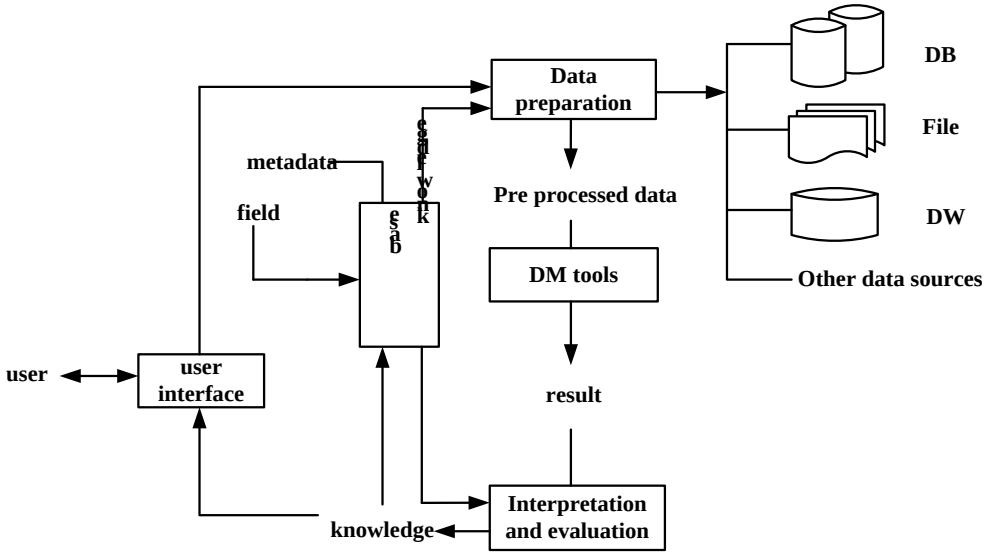

Figure 1. Data mining architecture

K-modes clustering algorithm is a classical algorithm, is an extension of K-means clustering algorithm. It has in dealing with discrete properties better performance, using a new dissimilarity measure, replace the cluster center with all the attributes one per cluster, and use the calculation method based on the frequency of the mode clusters of isolated point insensitive. Before delving into K-modes algorithm, first look at a few definitions:

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \quad \text{among} \quad \delta(x_j,y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\neq y_j)\end{cases} \tag{1}$$

Wherein, d (X, Y) can be seen as the difference between the two samples, the maximum difference is m, the minimum difference is zero.

The objective function of the algorithm is:

$$P(W,Q)=\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{l=1}^{n}w_{i,l}\delta(x_{i,j},q_{l,j}) \tag{2}$$

Where k is the number of clusters, n is the number of samples, m is the number of attributes. Q is the cluster center matrix:

$$w_{i.l}=\begin{cases}0 & \text{Sample i belongs to the cluster l}\\ 1 & \text{Sample i does not belong to the cluster l}\end{cases} \tag{3}$$

Running K-means clustering algorithm is iterative, after the data have finished running again by the end of the need to decide whether the conditions required for the next iteration of work.

Hadoop platform build

Hadoop can run in three modes in different computer environments: stand-alone mode, pseudo-distributed mode and full distributed mode (Yang 2011). Stand-alone mode to start a separate process in a single node, multiple independent pseudo-distributed mode starts processes on a single node. Both methods are based on single distributed computing simulation mode, can not be regarded as a true distributed computing, we can only learn to work and function testing. Fully distributed mode using multiple cluster nodes to build, is a substantive distributed computing, cloud computing platform has all the features and functions, this study used Hadoop platform to build a fully distributed mode.

See URL information of MapReduce, the console node enter: http: // localhost: 50030. HDFS information, view the website: http: // localhost: 50070. URL Manager lists the operational status of the cluster and the information shown in Figure 2. As can be seen from Figure 2, the cluster is currently two TaskTracker, you can run 4 Map 4 reduce tasks or tasks.

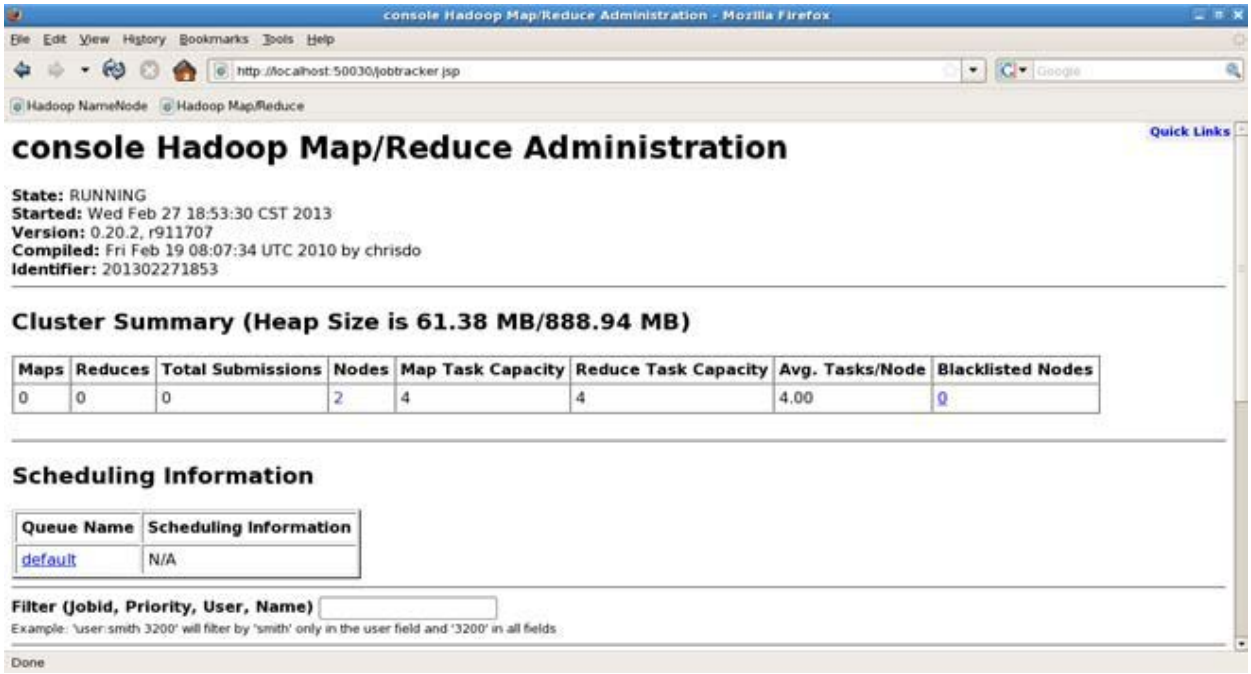

Figure 2. URL Manager lists the operational status of the cluster and the information

Experiments and results

Customer relationship management applications

Data mining is a specific analysis and presentation of data and extract actionable, implicit and novel information to solve business problems process. Commercial banks customer relationship management, a major role in data mining is carried out to integrate customer information, revealing potential relevance and laws provide routine reports for policy makers, customer segmentation and communication, customer behavior anomaly analysis, to provide customers with personalized service, so as to enhance the competitiveness of banks.

Customer relationship management of commercial banks, the data mining is mainly divided into two types: driven verification and discovery-driven, Figure 3 shows the form of data mining tools and techniques.

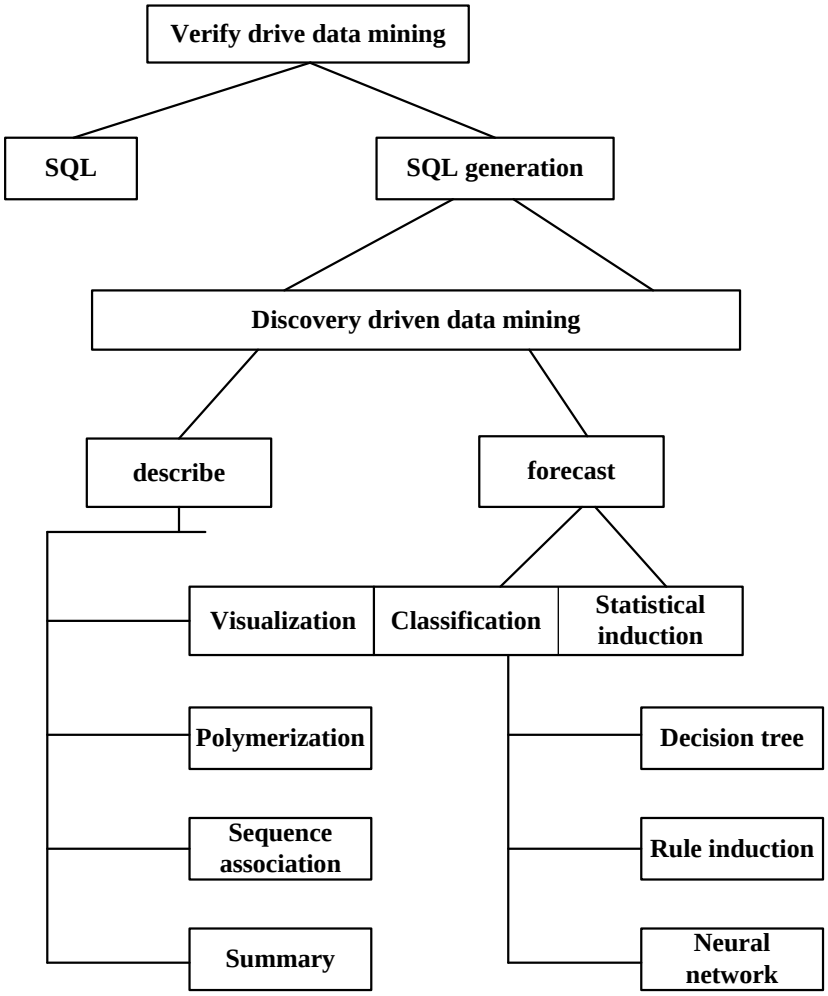

Figure 3. The form of data mining tools and techniques

Data mining is in many ways the commercial banks customer relationship management, such as customer evaluations and customer segmentation, customer behavior analysis, customer communication and personalized service, and so have a wide range of applications. Verification-driven data mining involves the use of common technologies, such as Structured Query Language (SQL) programming or SQL generator, query and online analytical processing tools to verify the hypothesis, especially online analytical processing tools presented to the user is a multidimensional view of data, OLAP cubes are usually three-dimensional data slices display a three-dimensional plane. As there is a time dimension cube, product dimension, income maintenance, it is easy to display graphics on the screen and slice. But to add a dimension (such as adding branches dimension), the graph is difficult to imagine, is not easy to draw on the screen. To break the three-dimensional obstacles, we must understand the difference between logical and physical dimension of peacekeeping. OLAP multidimensional analysis view is to break the three-dimensional physical concept, it uses a rotating, nesting, slice, drill and high-dimensional visualization techniques, multi-dimensional view of the structure of the display on the screen, allowing users to intuitively understand and analyze the data, make decisions stand by.

Summary

Distributed naive Bayes classification algorithm and distributed K-modes clustering algorithm running time compared to the stand-alone version can be increased by at least 3 times, the use of flow reading method, memory consumption remained at the MB level, for high computer performance is concerned, the level of TB can theoretically handle more volume and data. This article discusses the system of customer relationship management and data mining technology, and to data mining technology in customer relationship management depth discussions. We believe that the modern

enterprise competition is based on competition and information services, data mining technology can effectively discover useful information and knowledge from large amounts of customer data, and thus can effectively improve the quality of customer relationship management, to improve the competitiveness of enterprises purpose.

References

[1] Bing Liu. “Web Data Mining-Exploring Hyperlinks Contents, and Usage Data”. Chicago: Springer. 2006, pp. 21-45.

[2] David Cheung. “Advances in Knowledge Discovery and Data Mining.” 5th Pacific-Asia Conference: Hong Kong, 2001,pp. 82-85.

[3] C.Apte, SM Weiss. “Data mining With decision trees and decision rules.” Futre Generation Computer Systems November. 1997, pp. 197-210.

[4] Srivastava J. “Web usage mining:Discovery and application of usage patterns from web data. “ SIGKDD Explorations, 2000, Vol 17(1): pp. 12-13.

[5] C.C.Aggarwal, P.S.Yu. “Data Mining Techniques for Associations.” Clustering and Classification. 2002, pp.22-67.

[6] Anu Vajdyanathan. “Malcolm Shorc and Mark Billingburst.” Data in Social Network Analysis. 2008, pp. 23-25.

[7] Monika Henzinger. “Link Analysis in Web Information Retrieval.” IEEE Data Engineering Bulletin. Sep 2000, pp. 3-8.