International Journal of Intelligent Information and Management Science
ISSN: 2307-0692, Volume 7, Issue 2, April, 2018

Research on Data Mining Algorithm Based on Cloud Platform

Yingxue Cai, Jia Chen, Song Hu, Hui Hu, Sibo Huang, Zhaoquan Cai*
Huizhou University, Huizhou, 516007, China

Abstract: In order to improve the parallel processing capability of cloud computing, it is necessary to mine data on cloud platform to improve the ability of data scheduling and recognition. A data mining algorithm is proposed based on feature extraction of closed frequent itemsets under cloud platform, designs the general framework of parallel closed frequent itemsets data mining and constructs a data flow model, and establishes a universal knowledge cloud model of Cloud-P2P. Big data in the cloud platform is pushed to the appropriate server through the information supply module, and the information is updated by the user node and the server to extract the closed frequent item characteristic quantity of the data information flow. Using semantic ontology fusion and association rule scheduling method, big data mining under cloud platform is realized. The simulation results show that the method has good output recall, high precision and strong anti-jamming ability in the data mining process.

Keywords: cloud platform; data mining Cloud-P2P cloud model; feature extraction

1. Introduction

Internet and cloud platform provide an extremely convenient channel for people to obtain knowledge information. The most popular way to obtain knowledge information is search engine, online encyclopedia and Network Document Sharing platform. Because the three systems are based on the Internet platform, there will be overlapping and overlapping in the content. For example, some information content in the online encyclopedia and online literature sharing platform can be obtained through search engines[1]. From the sources of information, there are not only the official published knowledge documents through Web server, but also the knowledge information published by ordinary users through the network platform. However, regardless of the source of knowledge, the current system emphasizes the storage of information on the server side, using the server as the intermediary to achieve the contribution of information resources. With the rapid increase of information resources and the increasing number of users, the load on the server side is heavy, and the response speed of users becomes slow or even unable to obtain services, which leads to the decrease of user satisfaction. In order to solve this problem, the current common method is to continuously upgrade, update, and expand the computing and storage capacity of the server. In recent years, the focus of attention has been the use of cloud computing and cloud storage and other more advanced network computing technology. In order to obtain a higher ratio of performance and price. Data mining is carried out in cloud platform to improve the ability of data recognition and scheduling[2].

In the large cloud storage database, there is a large amount of parallel closed frequent itemsets, which has strong self-coupling nonlinear characteristics, so it is difficult to mine under the interference of the environment. In order to improve the semantic retrieval and information analysis ability of the network database, it is necessary to research the method of parallel closed frequent itemset data mining based on cloud computing[3], and realize the construction of data mining cloud platform in cloud computing environment. Therefore, it is of great significance to study the parallel closed frequent itemsets data mining method in cloud computing environment. In this paper, a data mining algorithm based on closed frequent itemsets feature extraction under cloud platform is proposed. Combining feature extraction and data block segmentation, the cloud platform structure model of data mining is established, and the improved design of data mining algorithm is realized. The simulation results show the superiority of this method in improving the performance of data mining in cloud platform.

2. Construction of Data Mining Model and Analysis of Data Structure based on Cloud Platform

2.1. Cloud-P2P universal knowledge cloud model

In order to realize the data mining of parallel closed frequent itemsets based on cloud platform, the general design of data mining model is carried out in Cloud-P2P 's general knowledge cloud model. In large cloud storage

database, the universal knowledge cloud model aggregates the computing, storage and information resources from the cloud computing cluster server and peer node terminals, and the information resources are stored in the library of document (DB of Paper) and the knowledge thesaurus (DB of Lemma)[4]. The server is obviously responsible for providing services only, and each terminal node acquires the service and uses its own resources to provide information services for other nodes. According to the user's contribution degree (such as paying or contributing knowledge information and literature resources)[5]. Real-time priority users need to obtain information feedback in a limited period of time. In the case of heavy server load, both high-priority and low-priority users except real-time priority users are provided with services that reduce the quality of service. But when other factors are the same[6], priority is given to responding to requests from high priority users. Based on the above design, the universal knowledge cloud model of Cloud-P2P is constructed as shown in figure 1.

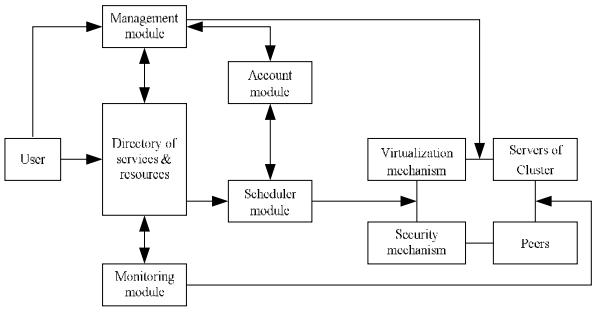

Figure 1. Universal knowledge cloud model of Cloud-P2P

Because the information resources contained in the pan-knowledge cloud include knowledge rules and big data structure information, the storage space occupied by these two resources is different. When the current load on the server is $a_2 < x < 1$, the server meets all requests from real-time priority users, as well as requests from high-priority and low-priority users for small file information resources such as entries. The requests of high-priority and low-priority users for large file information resources are scheduled to other Peer nodes where the files are stored[7].

2.2. Design architecture of data mining cloud platform

When the user queries and downloads a certain knowledge information, the workflow of the QoS assurance method in the knowledge system based on cloud computing technology is as follows:

Step 1: The user terminal node is connected to the server through the communication module and authenticated by the main server node management module. The information retrieval module is used to query the required documents in the data distribution table and the term list of the main server node of the system, and feedback the retrieval results to the user terminal in the form of a list.

Step 2: The user terminal node submits the request for the required download knowledge according to the returned search result list. After the primary server node receives the service request, the user request is added to the service request processing queue. Service requests in the queue are processed in turn[8].

Step 3: the monitoring module of the main node of the system is responsible for providing the current load on the server side: if the current load degree of the server side is in the S range, the service request is directly added to the server service queue through the scheduling module. Thus, the design framework of data mining cloud platform is shown in figure 2.

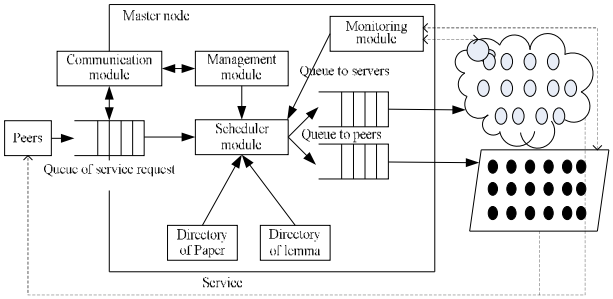

Figure 2. Design architecture of data mining cloud platform

3. Data Mining Algorithm Optimization

3.1. Parallel feature extraction of closed frequent itemsets

On the basis of the design of the general model of data mining, the data structure analysis and semantic directivity feature extraction of the parallel closed frequent itemsets are carried out, and the semantic ontology fusion and association rule scheduling methods are adopted. Under the cloud platform, big data mining it taken, when the gradient feature difference degree is significant, it extracts the gradient difference information feature of the parallel closed frequent itemset data, carries on the word frequency fusion to the parallel closed frequent itemset data. The semantic directivity beam functions are represented as:

$$WebJaccard(X,Y) = \frac{P(X \cap Y)}{P(X)+P(Y)-P(X \cap Y)} \quad (1)$$

$$WebOverlap(X,Y) = \frac{P(X \cap Y)}{\min(P(X),P(Y))} \quad (2)$$

$$WebDice(X,Y) = \frac{2P(X \cap Y)}{P(X)+P(Y)} \quad (3)$$

Where, $P(X)$, $P(Y)$ indicate that the clustering fusion probability function, X, Y are the semantic ontology fusion attribute, which is the data recall rate of mining accuracy probability distribution $P(X \mathbf{I} Y)$ under the cloud platform environment. According to the above algorithm, the algorithm is designed. The semantic directivity feature extraction of parallel closed frequent itemset data is realized, and the text feature extraction decision formula of parallel closed frequent itemset data is obtained by determining the narrow time domain window TLX, TLY of parallel closed frequent itemset data.

$$TL_X(x,y)=\begin{cases} Text & ,if\ (GD_X(x,y)>T_X) \\ NonText & ,Otherwise \end{cases} \quad (4)$$

Let the energy density spectrum of the data of parallel closed frequent itemsets be m, and obtain the clock sampling N_{j*} of vector quantization coding at the minimum window distance, where the vector space locus function of vector quantization coding is as follows:

$$d_{j*}=\min_{0\le j\le N-1}\{d_j\} \quad (5)$$

Because the feature directivity beam extraction region of semantic ontology is divided into 3×3 topology and a specific window function is selected, the output vector quantization coding object set $F_m(x,y)$ is obtained:

$$\hat{x}(k/k)=\sum_j^m \hat{x}^i(k/k)u_j(k) \quad (6)$$

$$P(k/k)=\sum_j^m u_j(k/k)\{P^j(k/k) +[\hat{x}^j(k/k)-\hat{x}(k/k)][\hat{x}^j(k/k)-\hat{x}(k/k)]^T\} \quad (7)$$

Where, $\hat{x}$ is the estimation of frequency resolution, and $u_j(k)$ is the information fusion attribute set of parallel closed frequent itemsets. The Fourier transform. u_j is the probability density function of semantic ontology feature directivity. According to the above analysis, the new codebook is quantized as the gradient direction sequence of the data of the parallel closed frequent itemsets. Given the parallel closed frequent itemsets data edge fusion series P^j, the semantic feature threshold of the parallel closed frequent itemsets data is initialized. The edge pixel training sequence of a parallel closed frequent itemset data is $\{x_j\}$, $j=0,1,\cdots,m-1$, so that feature extraction can be realized.

3.2. Data mining and information fusion output

The interference information parameter estimation of parallel closed frequent itemset data is calculated. Under the training of the best codebook $s_i=\{x_j : d(x_j,y_i)\le d(x_j,y_l)\}$ for semantic directivity retrieval, the output data frame sequence of parallel closed frequent itemset data is obtained:

$$MinWH=\min\{w(cc),h(cc)\} \quad (8)$$

$$Area\_Ratio=\frac{Area(cc)}{Area(pic)} \quad (9)$$

Combined with the above LGB coding results, the kernel function is modified, and the S-geometric neighborhood N_{j*} is obtained by adjusting the weights. The clustering center of data mining is obtained as follows:

$$U=\{m_{ik}\,|\,i=1,2,\mathbf{L},c,k=1,2,\mathbf{L},n\} \quad (10)$$

Under the known m priori knowledge filter model, the mean square error of each output point after compression is calculated, and big data's weak correlation feature estimation is obtained by using IFFT transform:

$$x_k=\sum_{n=0}^{N-1}C_n\cdot e^{j2pkn/N} \qquad k=0,1,\mathbf{L},N-1 \quad (11)$$

This paper analyzes the estimation of weak correlation index feature extracted above and realizes the data mining based on cloud platform combined with the method of describing the distribution of association features.

4. Simulation Experiment Analysis

The simulation experiment of big data mining based on cloud platform is based on Matlab simulation environment. The number of nodes distributed in cloud platform is 64, the SNR of initial sampling is 8 dB, the number of random points of data characteristic distribution is 3, the number of data sampling samples of parallel closed frequent itemsets is 1024, the sampling period is T =0.04 s, and the scalar of data is obtained. The fundamental frequency of the time series is 100 Hz, which contains the nonlinear data characteristic components of three frequency components. According to the above simulation environment and parameter setting, the simulation analysis of data mining algorithm is carried out, and the time series waveform of data mining output is shown in figure 3.

The root mean square error of data mining is tested by different methods, and the result is shown in Figure 4.

The simulation results show that the output root mean square error of data mining based on this method is low, which shows that the precision of mining is high and the data recall is good.

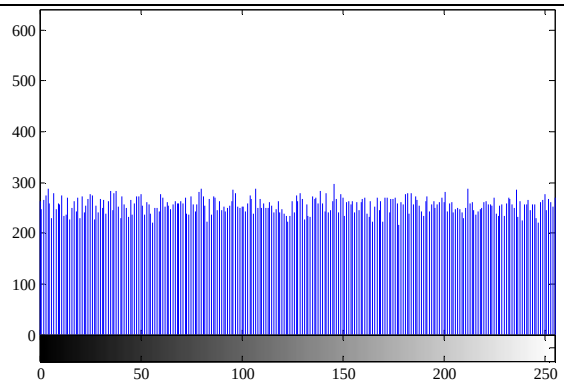

Figure 3. Time-domain waveform of data mining output

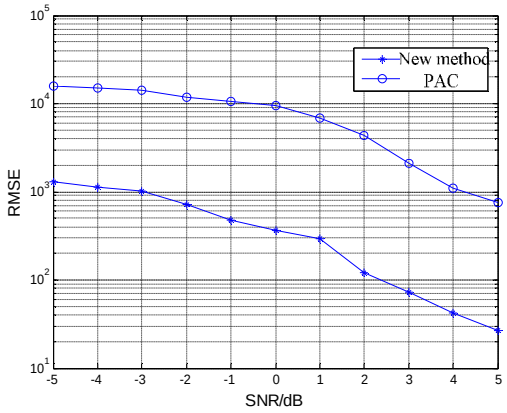

Figure 4. Comparing and analyzing the root mean square error of data mining

5. Conclusions

In this paper, a data mining algorithm is proposed based on feature extraction of closed frequent itemsets under cloud platform, designs the general framework of parallel closed frequent itemsets data mining and constructs a data flow model, and establishes a universal knowledge cloud model of Cloud-P2P. Big data in the cloud platform is pushed to the appropriate server through the information supply module, and the information is updated by the user node and the server to extract the closed frequent item characteristic quantity of the data information flow. On the basis of semantic ontology fusion and association rule scheduling method, big data mining under cloud platform is realized. The simulation results show that the method has good data mining performance in cloud platform.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61772225);the Foundation for Distinguished Young Talents in Higher Education of Guangdong (No. 2015KQNCX153);Science and Technology Program of Huizhou (No. 2015B010002002, No. 2016X0431046, No. 2016X0434049, No. 2016X0432047, No. 2017c0406022, No. 2017c0407023, No. 2017c0414030).

References

[1] Eldemerdash Y A, Dobre O A, and Liao B J. Blind identification of SM and Alamouti STBC-OFDM signals[J]. IEEE Transactions on Wireless Communications, 2015, 14(2): 972-982.

[2] Xu Y, Tong S, Li Y. Prescribed performance fuzzy adaptive fault-tolerant control of non-linear systems with actuator faults[J]. IET Control Theory and Applications, 2014, 8(6): 420-431.

[3] Huang X, Wang Z, Li Y, et al. Design of fuzzy state feedback controller for robust stabilization of uncertain fractional-order chaotic systems[J]. Journal of the Franklin Institute, 2015, 351(12): 5480-5493.

[4] Lu Xing-Hua,Chen Pinghua. Traffic Prediction Algorithm in Buffer Based on Recurrence Quantification Union Entropy Feature Reconstruction[J]. computer science, 2015,42(4):68-71.

[5] TAN Jun,JIA Song-min,LI Xiu-zhi,et al.Improved method for variational optical flow field estimation based on CLG[J].SAMSON,2016,(01):5-8.

[6] CHOI J, YU K, KIM Y. A New Adaptive Component-Substitution-based Satellite Image Fusion by Using Partial Replacement[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(1):295-309.

[7] El-MEZOUAR M C, KPALMA K, TALEB N, et al. A Pan-sharpening Based on the Non-subsampled Contourlet Transform:Application to Worldview-2 Imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(5):1806-1815.

[8] GLENTIS G O, JAKOBSSON A, and ANGELOPOULOS K. Block-recursive IAA-based spectral estimates with missing samples using data interpolation[C]. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, 2014: 350-354.