International Journal of u- and e- Service, Science and Technology
Vol.9, No. 2 (2016), pp.343-352
http://dx.doi.org/10.14257/ijunesst.2016.9.2.33

Research on Data Mining Algorithm based on Business Cloud Platform for Mobile Internet

Fenghua Liu

Shandong Women's University
wsh1234wl@126.com

Abstract

Mobile Internet is a mainstream access and communication technology, due to access to Internet anytime and anywhere, the business will varied, and bring mass data, but the data processing has different characteristics, the delay and energy consumption are also different. Therefore, it is necessary to apply to different data mining methods in the cloud platform, so as to adapt to different business applications, this paper proposes an improved Apriori algorithm, theory and simulation can prove that the method is effective.

***Keywords**:Mobile Internet, Cloud Computing,Data Prediction,Data Mining*

1. Introduction

Due to wireless communication havethe characteristics that can communicate at anytime and anywhere. Its application has become more and more widely, and the application of wireless communications equipment also more and more, data transmission has become exponential growth; with the progress of the society, computer technology experienced fifty or sixty years of development, and played a great role in promoting the progress of human society. When the arrival of the new millennium, the era of information is coming, it brings massive information integrate in our daily life. The Internet has brought together the function of traditional media such as cable television, radio and paper reading and broke their monopoly. In recent years, the potential of the mobile Internet market in the world is increasingly being found, people living in contact with each other, calls, exchange of always generated a lot of information. How to reuse the information generated, and even discover the effective information, the technology of data mining is becoming more and more important.Enterprises through of communication process derived from a variety of information mining and analysis, it can comprehensive understand users' interest, habits, and the degree of active business and business in the future development trend, etc. in order to facilitate scientific and optimize the allocation of network resources, and improve the development of the network information resources and use efficiency, further turn it into productivity. And then making scientific inference, for the company develop new products or to provide auxiliary help for senior decision making. Cloud computing is a recently emerging hot professional IT vocabulary, but the total cloud computing is integrated distributed processing and computing, parallel computing and virtualization technology. It can enable the storage resource to be expanded after the virtual, and create a platform and have the distributed processing, parallelization, and grid computing function.It avoids the single use of these technologies very well, resulting in low hardware usage, expand or reduce the resources cannot dynamically, data placement problem. At the same time, the virtual technology can be copied, mobile, maintain easily, and can optimize the management of large clusters, and provide security for running process. Through the use of these advantages, cloud computing related functions in the mobile business, logistics, retail, geological exploration and other fields to play a huge role. With the intelligent mobile phone, 4G network came into being, customers can according to personal preferences in any time

ISSN: 2005-4246 IJUNESST

and place randomly use mobile data services, mobile operators, the development of a variety of mobile communication business, such as: mobile Internet, mobile phone is used to watch video, mobile email, download music and others, making information flow suddenly increased exponentially, but this is just improve a source of value-added business profit point.When the cloud computing was proposed, it overcome the data mining period to lower the cost of access to huge data calculation problem, full use of Cloud Calculation fast, the movement of the access to low cost storage and computing power. Research personnel want to work out a mining algorithmwhich dynamic, and has high expansion, with the capacity to store as a unit for low cost calculation, to overcome the shortcoming of conventional technology, ultimately reduce the operation input can improve efficiency. Cloud computing continues to move forward, and attracted the attention of many industries, the relevant enterprise and research institutions will combine cloud computing and existing data mining algorithm to do research and application.

Cloud computing often confusedwith the traditional grid computing(distributed computing, a virtual super computer that consisting with a group of loosely coupled computer sets, is used to perform a large task), utility computing (a way of a package and billingof IT resources), autonomic computing (a computer system which has the function of self-management). In fact, the deployment of cloud computing depend on computer cluster, and it also absorbed the features of autonomic computing and utility computing, but it is different from system organization, purpose and working wayof grid composition, Cloud computing is the product of the integration of traditional computer technology such as grid computing, distributed computing, parallel computing, utility computing, network storage, virtualization, load balancing and so on.

With the trend that the various types of data all over the worldwhich producesevery day is growing, database, statistics, parallel calculation, machine learning, neural network, pattern recognition, data visualization, information retrieval, image and signal processing techniques have laid the foundationfor data mining, which are applied in data miningof all aspects. In the 1980s,data mining technology began to gradually develop. Data mining technology is developing rapidly due to currently in the possession of the world's great amount of data and social of these data resources transformation for the huge demand of information and knowledge resources. The application of data mining in all walks of life, in other words, as long as the local data has been stored,data mining will be used.

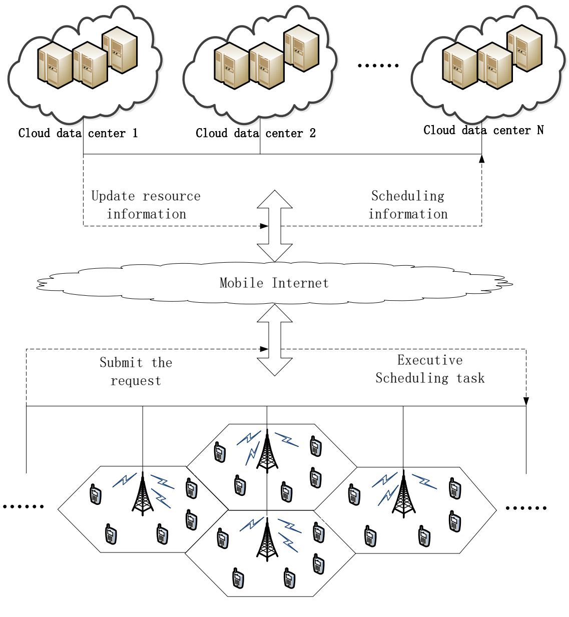


Figure 1. Mobile Internet Cloud Platform

In the large and disorderly data, it is valuable and valuable for enterprises to extract valuable and implicit information scientifically and effectively. For example, operators can provide users with a variety of business, it will take full account of the inherent properties of the businesswhen classify the business. Recorded after the business open, active, cancellation of records to carry out data mining, we can find some users will also ordered several business, we can real-time bound together into the promotion, so as to improve the business volumein business promotion, full permission to disinter the business value for the local. After the business order analysis, we can clearly know what business is the most popular, the most value-added, in order to help senior decided to develop new business and how to develop the old business. It can realize the accurate and fast sales target product function which can realize the regular retail business pattern, this profit through the excavation knowledge to make the corresponding sales plan. This shows that the data mining of the cloud environment for the company operations or research has practical significance.

2 Related Works

2.1 Research Status of Cloud Computing

[6] applied a continuous time Lagrange dual scheduling principle to dispatch data resources, its optimization objective function is minimizing transmission energy consumption. The energy efficiency and delay transmission structure of online perception are studied by [5].[7]considered the energy efficiency based control algorithm based on the time-varying channel conditions. However, the conditions of the design of the algorithm are too ideal, considering that a user transfers data.

Subsequently, the paper [8]provided the service for the N users in a base station at a base station in N channels, and an adaptive channel was proposed. The paper [9] proposed the method of adaptive carrier aggregation based on adaptive carrier aggregation to change the energy efficiency of data download and upload. It can be seen, the existing research is still based on physical transport layer and application layer scheduling combining way, the focus remains at the physical layer, application layer business are mainly concentrated in upload and download data transmission, which is obviously does not meet the requirements of mobile Internet. The computational quantity of Internet is studied by [10], and the calculation of the dynamic distribution is accord with the method of cloud computing. In order to improve the efficiency of the distribution, [11] is used to calculate the load distribution of the workload. This method simplifies the computational complexity of the allocation algorithm, but the effect of its allocation is not as good as other computational complexity of the algorithm. Based on the [12], the resource allocation method is put forward, which can allocate resources quickly, and the computational complexity is lower. However, these algorithms cannot be applied to the mobile Internet, because of the low computing power of its individual nodes, and the communication capability among each node is lower than that of the Internet. Therefore, we need to find a scheduling algorithm for mobile Internet.Through studying the load estimation model of [12], a new VM dynamic supply strategy is proposed based on the distributed online collection algorithm.To improve cloud computing efficiency based on MapReduce,[13]proposed a computing resource auction mechanism, this mechanism can appropriate computational complexity at the expense of obtaining higher efficiency. However, these two methods are needed to predict the work load. Therefore, [8] proposed a does not need to predict the work load of VM dynamic scheduling mechanism, the mechanism only according to the current user of computing resources on demand through the auction to configure and to provide users with VM.When using the auction mechanism to the biggest drawback of resource scheduling is to calculate the complexity is high, in order to reduce the computational complexity, the literature [9] according to the real auction mechanism proposed a based on strategies that combine auction of VM dynamic allocation and supply the strategy is according to the user request VM real bid value for judgment, the user only real bid to ensure the maximum of benefit, the advantage of this strategy lies in the judgment of the auction and does not require the previous auction information, can effectively reduce the computational complexity. However, the above methods are based on data center to design, and did not consider the wireless bandwidth constraints, although it can be used for mobile cloud computing scenarios, but also facing some problems.For example: when the data center to the super busy area of a large number of users distributing VM may cause by small bandwidth limit makes the system unable to timely data service delivery to the user's situation happened, not only affects the service quality of mobile cloud computing, and also affect the cloud provider's benefit.

2.2 Research Status of Data Mining

After data mining decades of data on the development of data mining has evolved into a cross discipline, including statistics, machine learning, database, pattern recognition, intelligent, parallel computing [7] and subjects of mutual penetration.From the original regularity data to the present messy, huge data, our research object is getting wider and wider. In the face of the growing data, we cannot meet the requirements of the single terminal mining. So we put forward the parallel computing, which is the effective way to extract and analyze the recessive information quickly. It can be imagine if we can divide the original data into several blocks, and it will be processed in parallel on a number of CPU or multiple servers, and the speed will be greatly improved.Common methods of parallelization are: MPI, MapReduce and OpenMP. OpenMP, a boot compiled method, mainly for common use of memory resources for multi task parallel program design, shared memory, visible OpenMP is a stand-alone operationat most of the time. MPI is a parallel language, or said at least not exactly, it is a library, MPI to unify the parallel operation of each sub structure, data transmission, rely on the machine mutual information and communication to achieve. The ultimate goal of the specification is to serve the inter task communication even though its content is huge. MapReduce can be separated to describe, map and reduce, which is a specified map function, from a health value of another group in the middle of the health value of the MapReduce framework intermediate health value is handed to the reduce function, it will have the same numerical keys to simplification into smaller values.Future data mining has the trend of the development of large data volume, complex structure and practical application significance. Data volume is big, in some e-commerce platform is particularly evident, the amount of the general assembly to influence the way and the efficiency of the excavation, which is difficult to improve the analysis. In the development of unstructured data, the most important is how to change from unstructured to structure, and there is no structure and relationto structure and relation. It includes the technology of text mining, natural language processing and so on. Analysis of the socialization mainly refers to the analysis of the media and network information, to a variety of views on a variety of platforms of the attitude of service product, which can analyze the user's social circle, circle of preferences, inside the circle which is guide which is being a leader, findingthe key point, which is conducive to the promotion of products.

3. Proposed Scheme

There is the relation between the data in the database, therefore, it needs to screen them; according to the characteristics of data mining algorithm and the topological structure of the mobile Internet, improved the traditional association rules, forming new algorithm which can fast convergence, this algorithm can integrated a variety of needs, according to the needs of various between association of size and needs of the sequence, and it can quickly get the data association. In the association rule, the most classic is the Apriori algorithm, it was improved to makeit to adaptmobile Internet.

3.1 Apriori Association Rules

There is a rule or rule that can be used in the database among the data association.The value of a variable or between a property contains a rule-related nature that can be called an association. The purpose of the analysis is to explore the association knowledge within the database. Simple correlation, in chronological order, according to the causal relationship between the composition of the association. The association rule of the data in the original database is not the only one, so the rules generated by the correlation analysis are also credible. In the course of business operations, especially in places such as supermarkets, watching high goods shelves, it has a superb collection of beautiful things. The general customer may not thought of which the relationship, why there is such a place, even a small part of the customer there will be some questions, it feel not very convenient.In fact, the supermarket's merchandise placed in a certain extent, the reference to the customer to buy goods preferences, convenience, price and other factors. The purpose of this process is for a large number of buyers choice commodity information of a data mining process, mining is to find association rules of goods and commodities, the association between customer preferences,the association between convenience, the association between price， between the connection of goods. This is an association rule that finds out patterns of customer buying behavior, such as purchasing a toothbrush which will affect the purchase of what goods. The corresponding enterprise will put the shelf position of these commodities, and stock management to take the corresponding consideration. Therefore, the value of the commodity related information is increasingly valued by enterprises.

$I = \{I_1, I_2, ..., I_n\}$isthe set of eachdatabase option,$I_1, I_2, ..., I_n$isthe database option, Dis theset of database transactions,it contains the transaction T, which is anitemset, andT contained in I.Any database options are composed of an item set A, which is the total collection of the item sets of all the options, $A \subseteq T$ which is a subset of the total itemsets.

Confidence: also credibility and accuracy. T this event occurs in the event of the transaction R, the probability that the event Q also occurred. Expression is

$$PC = P(\mathrm{Q}|R) \qquad (1)$$

The effect of confidence, credibility and accuracy is the intensity of the rules.For the degree of supportcontains the event P and event Q,which is the ratio that contains events $\mathrm{P} \cup \mathrm{Q}$in all events,

$$PS = P(R \cup Q) \qquad (2)$$

Apriori algorithm is a correlation algorithm, which has the basic characteristics and the range of the impact is very wide. It describes the potential relationships among the data items, and it belongs to the single dimension, single layer, Boolean association rules in classification. The Apriori algorithm uses the sequential search to search the frequent sets of cyclic patterns. The basic problem of association rule is two:

1) finding all the frequent itemsets. All the itemsets should reach "frequent", and the frequency of these itemsets should satisfy the minimum support degree, which is related to the performance of the whole association rules.

2) according to obtained the frequent itemsets, when they are greater than or equal to the minimum confidence threshold, a strong association rules will be generated.

The specific method of the algorithm as follows:

First, scanning target item set, in the situation which meeting the minimum support and confidence threshold, finding frequent 1-itemset,denoted as T_1 ,Then based on

T_1,satisfying the minimum support degree and the confidence level of the new specifiedto generateT_2, that is the frequent 2-itemset,and then on the basis of T_2 to search T_3, the frequent 3-itemset; a search forT_4, T_5, T_6 ... , until cannot find the more frequent sets,which is the frequent K-itemset. Each layer needs to scan the entire database when you findT_K .

Apriori algorithm has three important properties:

Aprioriproperty1, a frequentitemset can launch its all non-empty subsets equally frequent. Apriori property 2,If the itemsetT is not frequent, then any superset of T is not frequent. Apriori property 3: for a certain item of the K in the, if the frequent (K-1) -item subset of the item is less than K, the item is not a frequent item for frequent K- entries. Though Apriori algorithm in a certain extent by its nature of compression of the frequent itemsets of size and improve the mining performance, but it still exists some disadvantages: first, large-scale database itself there is a large number of frequent sets. After the algorithm will produce a large number of candidates of the frequent pattern; secondly, the number of scan data in the operation of the algorithm more and more repeated, a waste of time. After connecting stage algorithm to after pruning, using the Apriori property, reducing the size of data set, it will improve the efficiency of mining.But once more than that amount, and then setting support threshold low, the efficiency of the algorithm will be significantly reduced, the main reason is: first, when computing support, Apriori algorithm using sequential search, scanning the database one time, candidate itemsets are searched. If the candidate set is too large, or the number of candidate sets is too large, there will be a lot of pressure when the input and output, which will reduce the efficiency of computing. Second, the number of CPU resources is consumed in the process of dealing with the candidate itemsets. In short, in the face amount of information with the time change exponentially increasing case, the original Apriori algorithm scanning the database many times spend a lot of time and CPU resources, and it also restricts the efficiency of the Apriori algorithm raise. Scientific researchers are increasingly attached to this, some of the improved method came into being.

3.2 Improved Apriori Association Rules

Apriori algorithm of scanning transaction database and pattern matching calculating the set of candidate support, improved algorithm using SQL queries against the relational tables to calculate the candidate set of support degree, frequent n-itemsets and their support is stored in a relational table $\text{tb\_item\_n}(\text{item 1}, \text{item 2}, \text{L} \ \text{item n})$, first calculation rules to enhance the degree, generating association rules, and then determine credibility is set greater than the minimum confidence threshold.

In the mining process, using SQL statements to query the relationship table, first calculated frequent itemsets which meet the minimum support count, then, calculating correlation rules, finding all positive correlation rules, deleting negative association rules and has nothing to do with the rule. Finally, calculating rules of credibility generated all meet the minimum support degree and the credibility of positive association rules. The algorithm makes use of the redundancy rule of correlation pruning, and the computation of the degree of upgrading greatly reduces the space complexity by using SQL query table.

3.2.1 Determining the Frequent l- itemset

Determine the flow chart of the frequent l-itemsetas shown in Figure 2.

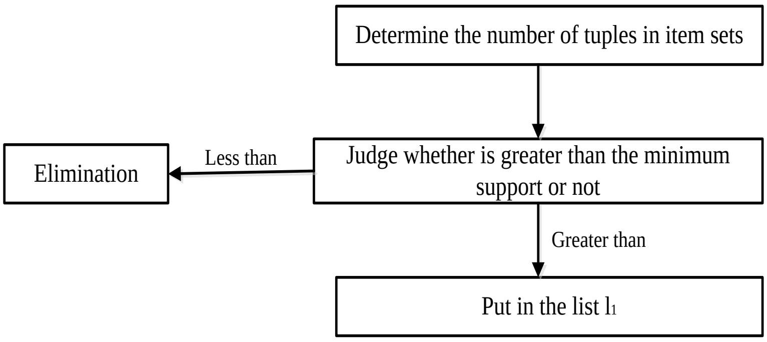

Determine the number of tuples in item sets
Judge whether is greater than the minimum support or not
Less than
Elimination
Greater than
Put in the list l_1

Figure 2. Determine the Frequent l-itemsetflow Chart

3.2 Determine the frequent n-itemset
Determine the flow chart of the frequentn-item set as shown in Figure 3.

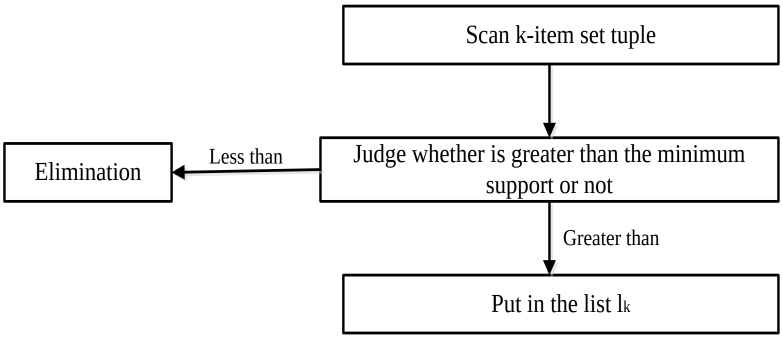


Figure 3. Determine the Frequent n-item Set Flow Chart

First calculating ascension of the rules , when the positive correlation between the event A and B , it is concluded that the rule $A \Rightarrow B$, when the negative correlation the event A and B , it is concluded that the rule $A \Rightarrow -B$, when it is independent between the event A and B , no rules are generated. The flow chart of the generating rule is shown in Figure 4.

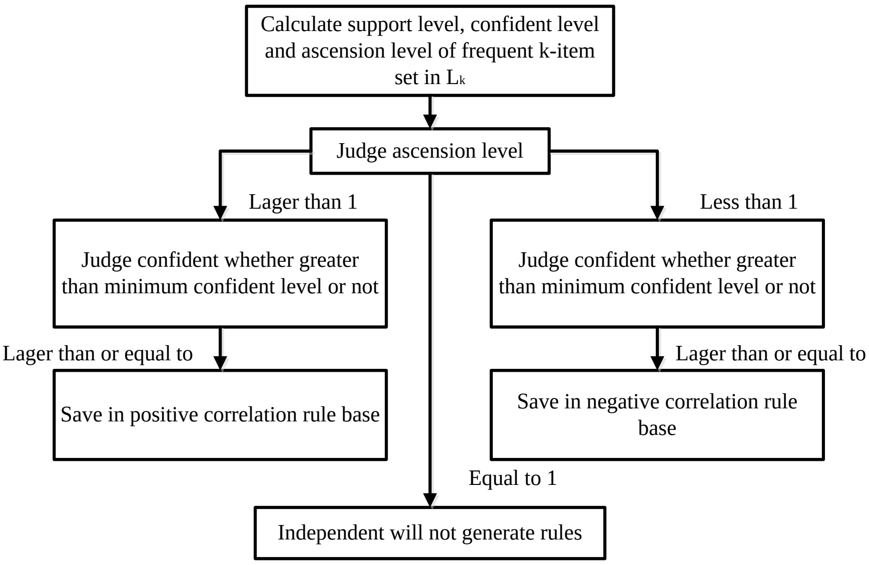

Calculate support level, confident level and ascension level of frequent k-item set in L_k
Judge ascension level
Lager than 1
Less than 1
Judge confident whether greater than minimum confident level or not
Judge confident whether greater than minimum confident level or not
Lager than or equal to
Lager than or equal to
Save in positive correlation rule base
Save in negative correlation rule base
Equal to 1
Independent will not generate rules

Figure 4. Generating Rule Flow Chart

The procedure of generate rules as follows:

```
for(k=2;$l_k \neq \phi$;k++)
for frequent k-item set every item set P in $L_k$
{
/*calculate rules e $\Rightarrow$ p-e confidence level,support level,ascension level*/
    Confidence=$\frac{p.count}{e.count}$;
    Lift=$\frac{confidence}{\frac{(p-e).count}{n}}$;
    Support=$\frac{p.count}{n}$;
    if Lift>1 then
    {
    if confidence min_conf} then
    R_S=R_S$\cup${e $\Rightarrow$ p-e};
    }
    Else if Lift<1 then
    {
    confidence=1-$\frac{p.count}{n}$;
    support=$\frac{e.count-p.count}{n}$;
    Lift=$\frac{confidence}{1-\frac{(p-e).count}{n}}$;
    if confidence min_conf then
    R_S=R_S$\cup${e $\Rightarrow$ p-e};
    }
    else
    e and (p-e) are independent， no rules are generated;
    }
```

4. Experiment Results and Analysis

As the calculation and simulation of the cloud computing model for the mobile Internet, Therefore, it is necessary to build a cloud simulation platform, the cloud model platform used cloud computing simulation of general platform CloudSim and the distributed parallel calculation based on the development of and by means of the platform can be through the resources of the computer simulation data storage and transmission, but lacking of topological changes of the link, according to the actual situation, the this has been modified. This modification based on the topology of the graph changes the transmission of data and transmission time. The simulation environment includes the computer configuration environment. The computer simulation environment is shown in Table 1

Table 1. VM Configuration

	处理器	内存	硬盘
VM_1	1× 2 GHz	4 GHz	500 GB
VM_2	2× 2 GHz	8 GHz	1 TB
VM_3	4× 2 GHz	16 GHz	2 TB
VM_4	8× 2 GHz	32 GHz	4 TB

The simulation of the journey is as shown in the figure below, according to virtual task and scheduling to achieve the model, written for the development of the core algorithm of the scheduling interval according to the different simulation environment, need to set up separately.

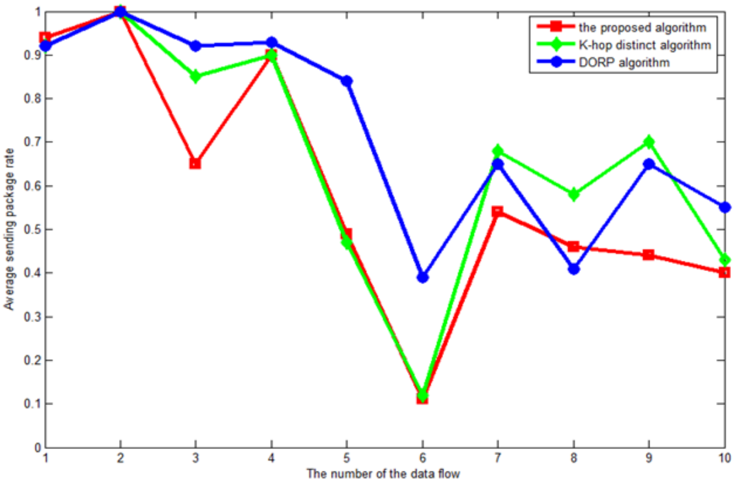


Figure 5. The Resource Utilization Rate Comparison

From the resource utilization rate of figure 5, under resources utilization conditions in data flow increased, the proposed algorithm significantly lower than the other schemes, we can see that the proposed scheme has advantage in resource mobilization, which data mining used APRORI algorithm, it can be more effective allocation of resources, and improve resource utilization efficiency.

5. Conclusions

According to the characteristics of mobile Internet data structure, in cloud services platform of database under the conditions, it proposed data mining algorithms which adapt to the characteristics of mobile Internet topology, this algorithm based on the Aprori algorithm, the mining method has been improved. The experimental results show that the proposed algorithm can effectively improve the utilization efficiency of system resources, and data mining effects is fast.

References

[1] Z. C. Huang, P. P. K. Chan and W. W. Y. Ng, “Content-based image retrieval using color moment and Gabor texture feature”, Machine Learning and Cybernetics (ICMLC), International Conference, IEEE, **(2010)**.

[2] M. A. Herráez, F. J. Ferri and S. M. Picot, “A hybrid multi-objective optimization algorithm for content based image retrieval”, Applied Soft Computing, vol. 13, no. 11, **(2013)**, pp. 4358-4369.

[3] G. G. Wan and Z. Liu, “Content-based information retrieval and digital libraries”, Information Technology and Libraries, vol. 27, no. 1, **(2013)**, pp. 41-47.

[4] N. D. Thang, T. Rasheed and Y. K. Lee, “Content-based facial image retrieval using constrained independent component analysis”, Information Sciences, vol. 181, no. 15, **(2011)**, pp. 3162-3174.

[5] P. Järventausta, S. Repo and A. Rautiainen, “Smart grid power system control in distributed generation environment”, Annual Reviews in Control, vol. 34, no. 2, **(2010)**, pp. 277-286.

[6] M. E. ElAlami, “A new matching strategy for content based image retrieval system”, Applied Soft Computing, vol. 14, **(2014)**, pp. 407-418.

[7] N. Amoda and R. K. Kulkarni, “Efficient Image Retrieval using Region Based Image Retrieval”, International Journal of Applied Information Systems (IJAIS)–ISSN, **(2013)**, pp. 2249-0868.

[8] D. Liu, X. S. Hua and H. J. Zhang, “Content-based tag processing for internet social images”, Multimedia Tools and Applications, vol. 51. no. 2, **(2011)**, pp. 723-738.

[9] P. P. K. Chan, Z. C. Huang and W. W. Y. Ng, “Dynamic hierarchical semantic network based image retrieval using relevance feedback”, Machine Learning and Cybernetics (ICMLC), International Conference, IEEE, **(2011)**.

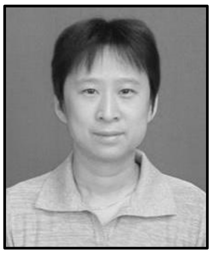
[10] T. Furuya and R. Ohbuchi, “Visual Saliency Weighting and Cross-Domain Manifold Ranking for Sketch-Based Image Retrieval”, MultiMedia Modeling. Springer International Publishing, **(2014)**.

[11] A. Arampatzis, K. Zagoris and S. A. Chatzichristofis, “Dynamic two-stage image retrieval from large multimedia databases”, Information Processing & Management, vol. 49, no. 1, **(2013)**, pp. 274-285.

[12] R. Jin, G. Yang and G. Agrawal, “Shared memory parallelization of data mining algorithms: Techniques, programming interface, and performance”, Knowledge and Data Engineering, IEEE Transactions, vol. 17, no. 1, **(2005)**, pp. 71-89.

[13] J. Ekanayake and G. Fox, “High performance parallel computing with clouds and cloud technologies”, Cloud Computing. Springer Berlin Heidelberg, **(2010)**, pp. 20-38.

Author



Fenghua Liu, He received her B.S degree in computer science and M.E degree incomputer software and theory from Shandong University,China in1997 and 2006 respectively. She is currently researching on Data Mining, Cloud Computing and Software Engineering.