Personalized Overseas Chinese Education Model Based on Map-Reduce Model of Cloud Computing

Zhehuang Huang,Jianxin Huang
Huaqiao University, Quanzhou ,China

Abstract—**The rapid updates of the resources and media in the big data age provide new opportunities for oversea Chinese education. It is an urgent task to effectively use the big data to boost the development of oversea Chinese education. However, very few studies are conducted in this area. Map-Reduce is a programming model of cloud computing used for the parallel computing of the large-scale data sets and this model enables programmers to run their own programs in the distributed system. In this paper we proposed a personalized overseas Chinese education model based on Map-Reduce mechanism , which can analyze the behavioral habits and personal preferences of users from a large pool of Chinese educational resources. In this way, the customer needs can be accurately grasped and their favorite resources are recommended from huge amounts of resources. The proposed model has a good application prospect for overseas Chinese education .**

Index Terms—**overseas Chinese education, cloud computing, Map-Reduce ,online learning, student modeling.**

I. INTRODUCTION

As the political pluralism, cultural diversity and economic globalization are developing rapidly around the world in recent years, the comprehensive national power of China is also improving and the Chinese rejuvenation keeps moving ahead. Meanwhile, the popularization of Chinese culture also flourishes and new opportunities are presented for the oversea Chinese education. Phenomena such as "craze for Chinese language" and "craze for Chinese culture" successively come up in various countries of the world, and Confucius Institutes and Confucius Classrooms have been established in many regions in recent years[1,2]. It shows that a broader development space can be explored in the cause of Chinese language and art education. Oversea Chinese education and the international promotion of Chinese have made great contributions to the cause of teaching Chinese as a foreign language. However, the teaching resources of Chinese education are relatively dispersed and weak abroad, and the educational method of Chinese education is too simple because of the history and international environment. These factors limit the development of the Chinese education, especially the overseas Chinese education. Therefore, it is of great significance to conduct studies on the reception psychology, reception rules and learning motivation of the oversea Chinese learners who are different in learning stages, groups and backgrounds[3,4]. The results of studies

should become an inexhaustible driving force for the sustainable development of oversea Chinese education.

With the mobile internet, intelligent terminal, cloud computing and the technologies on internet of things continuously develop in recent years, the data presents an explosive growth and the picture scroll of the big data age is unfolding. It has become a main task to make full use of the big data technology to obtain its inherent huge value in the big data age. Currently, the big data processing has been widely applied in the fields such as mobile internet [5], medical data management [6], biological recognition [7], X-Learning [8],logistic storage [9] and other areas. In the big data age, it is a key and difficult point to accurately and conveniently retrieve the resources that satisfies the users' individual demands from a large pool of Chinese educational resources.

Cloud computing [10-12]is an emerging business calculation model that receives the wide attention of the public. Hadoop[13,14] is a cloud computing platform to develop and parallel process the large-scale data more easily and this platform mainly features strong dilatation, low cost, high efficiency and good reliability etc. Hadoop platform contains two parts: Hadoop distributed file system and MapReduce computing model. MapReduce[15] is a distributed programming model with great efficiency and also a realization of processing and generating the large-scale data set. MapReduce was initially designed by Google lab as a distributed parallel computing framework, and then the Hadoop distributed computing platform was used to realize its open-resource implementations by the Apache organization.

In this paper, we presents an intelligent recommendation model for oversea Chinese education resource based on cloud computing. This paper is organized as follows. Firstly, a resource recommendation model for oversea Chinese education is presented and some key algorithm of our proposed model is implemented. Secondly, we implement the model in the cloud computing environment . Finally some discuss and conclusions are given on the model.

II. RESOURCE RECOMMENDATION MODEL FOR OVERSEA CHINESE EDUCATION

In this section , we proposeda resource recommendation modelfor oversea Chinese education. Firstly, the students' interest information, Browse history and emotional information is used asmulti knowledge source to construct student's information vector space; secondly, Fuzzy c-Means algorithm is usedto implementthe clustering of

students; Finally, according to the preference of the most adjacent three neighbors, the Chinese education information is recommended. The proposed model is shown in Figure 1.
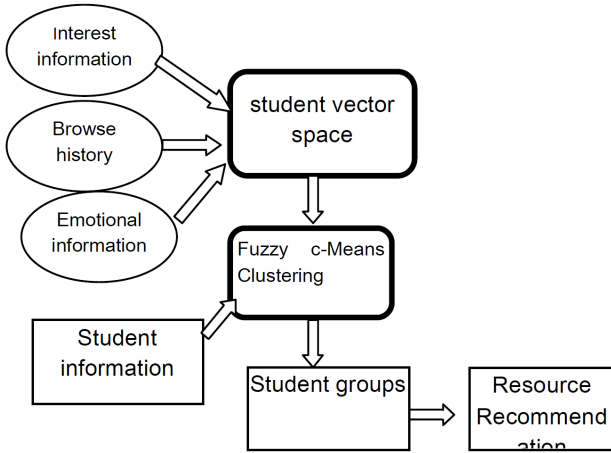

Figure 1. Resource recommendation modelfor oversea Chinese education

A. Student vector space model

Vector space model (VSM) was proposed by Salton et al in 1970s, and was successfully applied to the famous SMART text retrieval system.

By the vector space model, the text data is converted into structured data which can be processed by computer. The similarity between two documents is transformed into the similarity between two vectors. The most common method of similarity measurement in text processing is the cosine distance.

The student modeling will employ vector space approach, which is used to construct a vector $W = (w_1, w_2, \text{L}, w_n)$ based onthe user's interest in the keyword, assuming another user vector is $V = (w_1, w_2, \text{L}, w_n)$,Then the matching degree of the two vector is:

$$sim(V,W) = \frac{\sum_{i=1}^{n} v_i \times w_i}{\sqrt{\sum_{i=1}^{n} v_i^2} \times \sqrt{\sum_{i=1}^{n} w_i^2}}.$$

$sim(V,W)$ shows the matching degree of these two directions.

B. StudentClustering

Cluster is one of the important research topics in data mining. Cluster is a process of constituting the multiple categories or clusters composed of similar objects by using the set of physical or abstract objects. Many implementation algorithms of clusters are available, but the Fuzzy c-Means algorithm [16,17]is a fuzzy clustering algorithm which is widely applied and has a sound theory in the clustering algorithms based on the target functions. Many objects are medium in nature and have the properties of "both A and both B", namely the fuzziness. In the

fuzzy clustering analysis, an object can be placed in different categories and the subjection degrees ranging from 0 to 1 are used to express the degree of each object belong to each category, namely establishing the uncertainty description of the object against the category. Given that many students have extensive needs, the fuzzy clustering technology is used to cluster students, which better reflects the multiple interests of many students. In addition, the better clustering results can be obtained in the clustering analysis and the real world can be reflected more objectively.

The clustering process of students based on FCM clustering algorithm are as follows:

(1) Input the number of clusters s and student vector information.

(2)Obtain student information vectors $X = \{x_1, x_2, \text{L}, x_n\}$, where n is the number of users, x_i is a student vector;

(3)Initial user membership matrix $U = \{u_1, u_2, \text{L}, u_n\}$, Where $u_j = (u_{1j}, u_{2j}, \text{L}, u_{sj})$, u_{ij} is the membership degree of a student j belong toa class i.

(4) Obtain clustering center $c_i, i = 1, 2, \text{L}, s$. Where m is weighted index.

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}$$

(5) Calculatedcost function.

$$J(U, c_1, c_2, \text{L}, c_s) = \sum_{i=1}^{s} \sum_{j=1}^{n} u_{ij}^m sim(c_i, x_j).$$

(6) Calculate the new user membership matrix, if it meet the iteration stop conditions, then turn to steps (7), otherwise the return (4)▢

$$u_{ij} = \frac{1}{\sum_{k=1}^{s} \left(\frac{sim(c_k, x_j)}{sim(c_i, x_j)} \right)^{\frac{2}{m-1}}}$$

(7) Output the membership degree and clustering center of users.

The flow chart of the student clustering algorithm is shown in Figure 2.

In selecting the initial clustering number, the system first selects those users with a high percentage of page visits and a low mutual similarity as the center of the initial clustering according to the characteristics of resources and students. Subsequently, the membership matrix of initial users is determined according to the center of clustering, whichcan effectively reduce the iteration times and quickly converge to the global optimum.
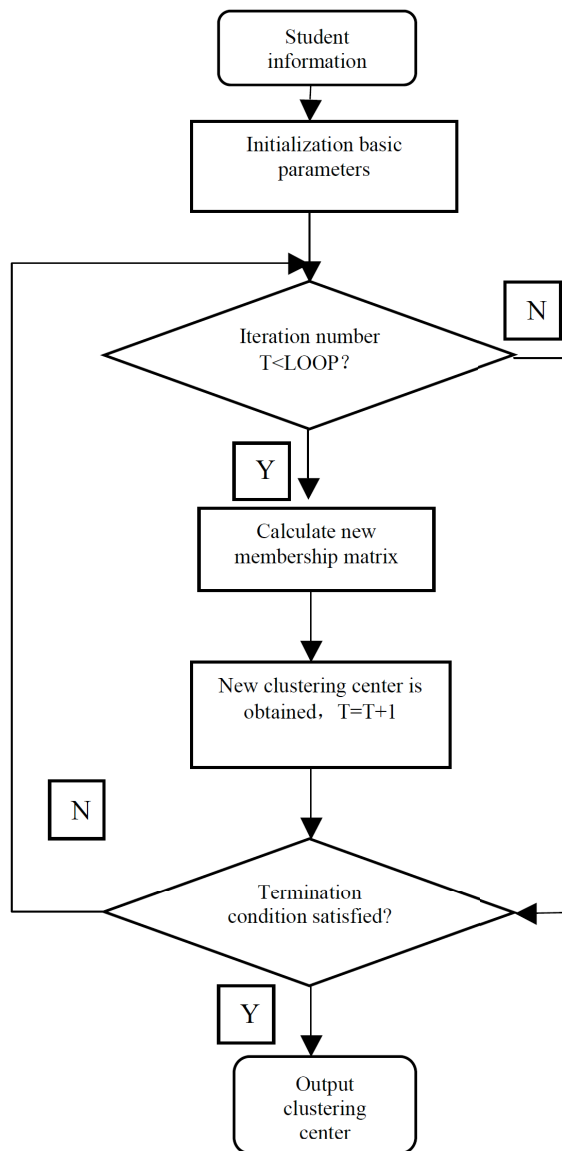
```
┌─────────────────┐
│     Student     │
│   information   │
└─────────────────┘
         │
         ▼
┌─────────────────┐
│ Initialization  │
│ basic parameters│
└─────────────────┘
         │
         ▼
    Iteration number        N
      T<LOOP?
         │
         │ Y
         ▼
┌─────────────────┐
│  Calculate new  │
│ membership matrix│
└─────────────────┘
         │
         ▼
┌─────────────────┐
│ New clustering  │
│ center is       │
│ obtained，T=T+1 │
└─────────────────┘
         │
    N    ▼
   Termination
condition satisfied?
         │
         │ Y
         ▼
┌─────────────────┐
│     Output      │
│   clustering    │
│     center      │
└─────────────────┘
```

Figure 2.   The flow chart ofstudent clustering

C. The recommendation of Chinese educational resources

After the target students are determined, the recommended resources can be generated and revealed to the target users according to the recent neighboring resource access. This system obtains the resources that these students are interested in and recommend these categories to these students by analyzing the latest three neighboring browsing information.

III.   THE OVERSEA CHINESE EDUCATION MODEL BASED ON MAP-REDUCE MECHANISM

With the continuous increase in the number of visiting students, the oversea Chinese education recommendation system meets its bottlenecks. The stand-alone system cannot afford the high expenditures, so it is necessary to realize the recommendation system in the distributed parallel system. This section provides the model of the Chinese language and art education based on the Map-Reduce mechanism of cloud computing.

A. Map-Reduce mechanism ofCloud computing

Hadoop is a framework for the distributed processing of the big data. This framework can divide the application programs into many small work units and these units can be put at any cluster node for execution. Hadoop distributed file system(HDFS) is fault tolerant and is used in the cheap hardware. Moreover, HDFS also provides the high throughput for visiting the data of the application program and it is suitable for those application programs with large data sets. Map-Reduce is a programming model used for the parallel computing of the large-scale data sets and this model enables programmers to run their own programs in the distributed system, which are realized by Map function and Reduce function respectively. Map function is used for mapping a group of key values into a new group of key values, while Reduce function is used to aggregate the results. In this way, the users can conveniently develop and run the application programs that process huge amounts of data on Hadoop. The calculation flow chart of Map-Reduce mechanism is shown in Figure 3.

B. Parallel implementation of the model

With the increasing amount of data, the realization of the Chinese education recommendation model is limited by the time complexity and the space complexity. This problem may be solved by applying the parallel processing technology in the students clustering and the efficient parallel algorithm is designed for improving the algorithm's performance of processing huge amounts of data.

This paper realizes the parallel Chinese education recommendation model of in the Hadoop cloud computing environment. One task is decomposed into multiple tasks and each task is sent to each nodefor calculation. Subsequently, the result of the multitasking is summarized and then the final analysis result is reached. In the distributed system, the machine clustering can be regarded as the pool of hardware resources for decomposing the parallel tasks. The decomposed task is sent to each idle machine resources for processing and it can greatly improve the computational efficiency. Meanwhile, the independence of the resource undoubtedly provides the best design guarantee for the extension of the computing cluster. The parallel model is shown in Figure 4. Mapfunction clusters the task block available, while Reduce function combines the clustering results of the task block.

The steps of the parallelmodel based on Map-Reduce mechanism of cloud computing is:

1. Student information stored in Hadoop Distributed File System.
2. Construct vector space model based on the students' interest, interest and emotion information.
3. The student information block is sent to each node for clustering, each node completes the following steps.
   - Map functionis used to complete the fuzzy C means algorithm, andthe students information are clustered;
   - Return the clustering results to the master.
4. In the Reduce stage, the clustering arrays are combined into the original output files according to their respective positions in the inputfiles.
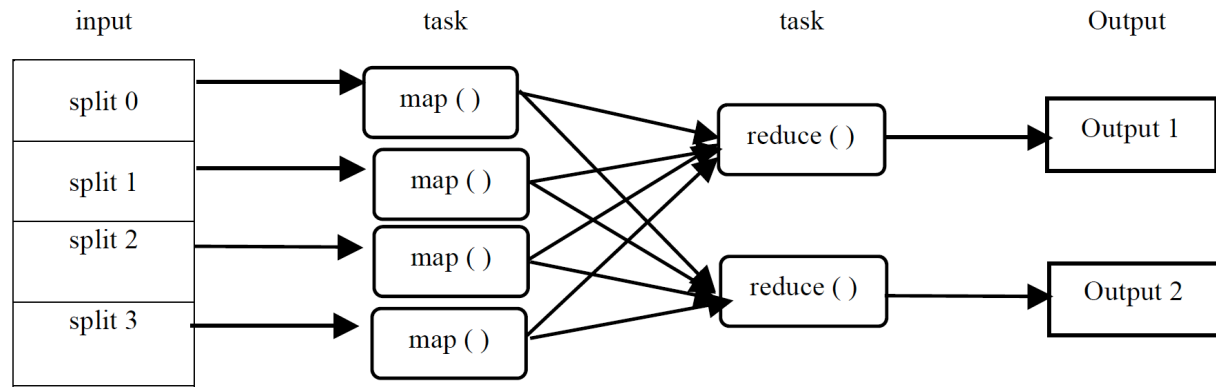5. Recommend educational resources to the students.
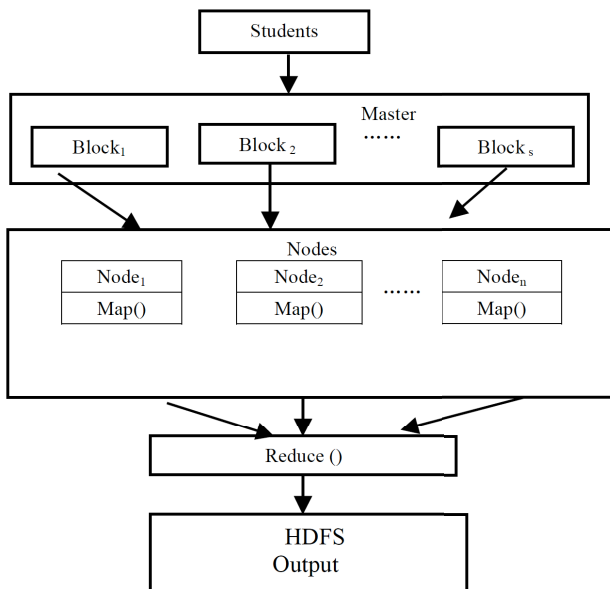
Figure 3.   MapReduce computing framework

Figure 4.   The parallel implement of oversea Chinese education model based on Map-Reduce mechanism

Figure 5.   The speedup of parallel model

IV.   EXPERIMENT AND ANALYSIS

In this section, we implement the parallel oversea Chinese education model on Hadoop cloud computing platform. There are totally three computers. One is the master node supplying service for NameNode and Resource Manager. The other two computers are used as slave nodes to supply service for computing tasks. Figure 5 shown the speedup of our system.

From figure 5, we can find out that the proposed model can achieve a higher speedup in the Hadoop cloud computing environment, and reduce the execution time . When the number of students is relatively small , the increase of speedup is not obvious. The main reason isthe constant reading, writing and data transmission are involved which consumed most of spare time. With the increase of the number of students, the speedup is also gradually increased.

V.   CONCLUSION AND PROSPECT

In this paper, we present a personalized overseas Chinese education model based on Map-Reduce model of cloud computing. This is the first effort to introduce Map-Reduce model to overseas Chinese education . We can get the following conclusions:
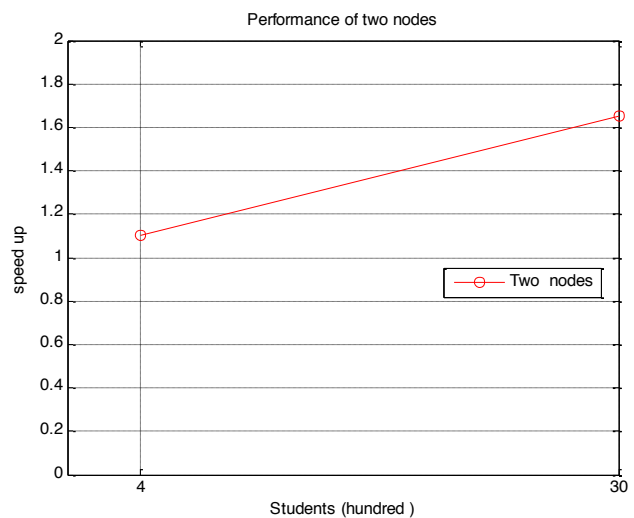
1. This proposed model can be used to recommend the Chinese education resources according to the students' individualized needs, and it has a good application prospect for overseas Chinese education .

2. This Map-reduce model can reduce the execution time and meet the needs of the student's real-time access.

The clustering algorithm has important influence on the performance of the model. In the future work, we will introduce more effective clustering and classification algorithm.

REFERENCES

[1]   D Starr. "Chinese Language Education in Europe: the Confucius Institutes". European Journal of Education,vol 44, no 1, 2009, pp 65-82. http://dx.doi.org/10.1111/j.1465-3435.2008.01371.x

[2]   Yang, R. "Soft Power and Higher Education: An Examination of China's Confucius Institutes Globalisation", Societies and Education,vol 8, no 2, 2010, pp 235-245. http://dx.doi.org/10.1080/14767721003779746

[3]   Gusev M, Ristov S, Velkoski G, et al. "E-learning and benchmarking platform for parallel and distributed computing". International .Journal of Emerging Technologies in Learning, vol.9, no.2, 2014,pp.17-21. http://dx.doi.org/10.3991/ijet.v9i2.3215

[4]   Vijayan Sugumaran,Veda C. "Storey,Ontologies for conceptual modeling:their creation,use,and management". Data & Knowledge

Engineering, vol.42, no.3, 2002, pp.251-271. http://dx.doi.org/10.1016/S0169-023X(02)00048-4

[5] Zhang H Z, Zhang Z Y, Dai H Y. "Gossip-based information spreading in mobile networks". IEEE Trans Wirel Commun,2013, 12, pp.5918–5928. http://dx.doi.org/10.1109/TWC.2013.100113.130619

[6] Yu H, Wang D. "Research and implementation of massive health care data management and analysis based on hadoop". IEEE, 2012, pp.514-517. http://dx.doi.org/10.1109/iccis.2012.225

[7] Kamath U, Compton J, Islamaj-Dogan R, et al. "An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice site prediction". IEEE/ACM Trans Comput Biol Bioinform, 2012, 9,pp.1387-1398

[8] Xiangsheng K. "Big Data X-Learning "Resources Integration and Processing in Cloud Environments". International Journal of Emerging Technologies in Learning, vol.9, no.5, 2014, pp.22-26. http://dx.doi.org/10.3991/ijet.v9i5.3783

[9] Morrison D A. "Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis". Wiley Interdiscip Rev Data Mining Knowl Discov, 2014, 4,pp. 296–312. http://dx.doi.org/10.1002/widm.1130

[10] Iosup A, Ostermann S, Yigitbasi M N, et al. "Performance analysis of cloud computing services for many-tasks scientific computing". Parallel and Distributed Systems, IEEE Transactions on, 2011, 22(6): 931-945. http://dx.doi.org/10.1109/TPDS.2011.66

[11] Whaiduzzaman M, Sookhak M, Gani A, et al. "A survey on vehicular cloud computing". Journal of Network and Computer Applications, 2014, 40,pp.325-344. http://dx.doi.org/10.1016/j.jnca.2013.08.004

[12] Al-Zoube M, Abou El-Seoud S, Wyne M F. "Cloud computing based e-learning system". International Journal of Distance Education Technologies, vol.8, no.2, 2010, pp.58-71. http://dx.doi.org/10.4018/jdet.2010040105

[13] Wang C C. "A sharable platform for e-learning based on cloud computing architecture". International Journal of Cloud Computing, vol.2, no.1, 2013, pp.89-99. http://dx.doi.org/10.1504/IJCC.2013.050957

[14] Moise D, Shestakov D, Gudmundsson G, et al. "Indexing and searching 100M images with map-reduce". Acm International Conference on Multimedia Retrieval, 2013,pp.17-24. http://dx.doi.org/10.1145/2461466.2461470

[15] Vasudevan A, Swetha M, Hyba H, et al. "Map-Reduce Based High Performance Clustering On Large Scale Dataset Using Parallel Data Processing". Data Mining and Knowledge Engineering, vol.6,no.5,2014, pp.181-185.

[16] Duan L S, Liu P Y, Xie F F. "Sample weighted fuzzy C-means algorithm based on simulated annealing". Computer Engineering & Design,vol.34,no.6, 2013,pp.2004-2008.

[17] Ji Z, Liu J, Cao G, et al. "Robust spatially constrained fuzzy c-means algorithm for brain MR image segmentation". Pattern Recognition, vol.47, no.7, 2014, pp. 2454–2466. http://dx.doi.org/10.1016/j.patcog.2014.01.017

AUTHORS

Zhehuang Huang: Corresponding author. He is a lecturer in the School of Mathematics Sciences, Huaqiao University, 362021, Quanzhou, China. His research interest covers E-learning, intelligent algorithm, machine learning and data fusion.

Jianxin Huang: He is a Professor in School of Mathematics Sciences, Huaqiao University, 362021, Quanzhou, China. His research interest covers E-learning and machine learning.