Parallel MapReduce for Clustering of Residential Customers Energy Behavior

WANG Zheng[1, a], YANG Yongbiao[2,b], SONG Jie[2,c], JIANG Ling[1,d], YU Jiancheng[1,e] and WANG Xudong[1,f]

[1] State Grid Tianjing Electric Power Company, Tianjing 300010, China

[2] State Grid NARI technology Co.,Ltd., Nanjing211106, China

[a]wangzhengncepu@163.com, [b]yangyongbiao@sgepri.sgcc.com.cn, [c]song-jie@sgepri.sgcc.com.cn, [d]yujiancheng@126.com, [e]ling.jiang@tj.sgcc.com.cn, [f]xudong.wang@tj.sgcc.com.cn

Keywords: Mapreduce, Fuzzy c-means clustering, Analysis of electric behavior, Big data.

Abstract. In allusion to the problem about electricity behavior analysis in the low efficiency of dealing with huge amounts of data, we puts forward the Fuzzy c-means clustering (Fuzzy c-means clustering, FCM) parallel algorithm based on Mapreduce technology. By decomposing the iterative process of FCM algorithm into two steps of Map and Reduce, it can effectively improve the efficiency of similarity computing between the data objects and the clustering centers. On this basis, the four characteristics of resident electrical data are clustering analyzed by using the proposed FCM parallel algorithm. The experimental results show that the proposed algorithm can improve the efficiency of mass data clustering analysis and also proves the feasibility of the model.

Introduction

With the increase of the power consumption and the personalized requirements of the service, electric power enterprises provide customers with electric energy products as well as undertake professional guidance on the use of electricity to improve power utilization efficiency and utilization level of the task [1]. The satisfaction of these requirements depends on the data acquisition and data analysis techniques.

In recent years, there have been some experts and scholars research on data analysis. The paper [2] briefly introduces the basic principle and related technologies of the distributed and parallel retrieval technology Hadoop_MapReduce. A clustering method that classifies customers in the same contract codes into k groups using the k-means algorithm suggested in paper [3]. In the paper [4], the K-means clustering algorithm is realized by using Mapreduce model, but the analysis of the data is not implemented.

In this paper, we propose a parallel computing algorithm based on c-Means clustering (FCM) in allusion to the demand of the analysis of mass data in the field of intelligent power. This algorithm is implemented in the Mapreduce framework, and can be used to analyze the behavior of the user with the FCM algorithm, and improve the efficiency and feasibility of the analysis by parallel computing. The experimental results show that the proposed algorithm can accurately analyze the resident user's electrical data, and quickly and accurately determine the user's power mode.

Distributed computing architecture for electrical data analysis

Distributed batching calculation to the mass data is the key to improve the efficiency of clustering computing, and the theoretical foundation of batch computing framework is the Mapreduce computing framework of Google. Mapreduce will highly abstract the complex parallel computing process to two functions, Map and Reduce, and can run on a large scale computing cluster. Using the Mapreduce framework, large scale computing tasks can be decomposed into many small sub tasks by Map process. Since the sub task is decoupled from each other, it can be processed in parallel. The results that Map output will be merged through the Reduce function and then generate the final results.

The Hadoop platform is the representative of the Mapreduce's open source implementation, and now it is widely used by Internet companies for large-scale data analysis.

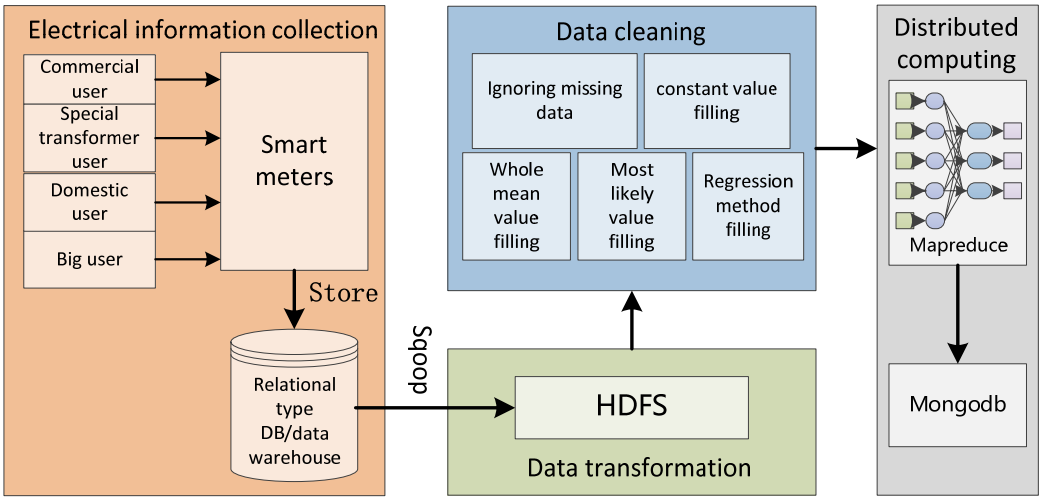

Fig. 1 Distributed computing architecture for electrical data analysis

Fig.1shows that the distributed architecture based on Mapreduce is mainly divided into foursteps: data collection, data transformation,data cleaning and distributed computing. Among them, the electric information collections rely on two-way smart meters、transform of the collection terminal and so on to achieve the collection of data, and stored in the electrical information collection systemas a basis of the electricbehavior analysis.Inthe distributed computing phase, the distributed FCM clustering algorithmbased on Mapreduce which is proposed in the paper used to complete the clustering of electrical behavior data. The clustering centers obtained from it can be used to describe the feature of electrical user group, and the fuzzy membership degree of each participating clustering data objects about different clustering, can be used to judge the cluster which electrical users belonging to.After the completion of the clustering process, resultthat is in form of key value pairs can be stored in non-relational database(NoSQL), such as Mongodb,so it will be convenient to query and do further data mining based on clustering results.

Analysis of the characteristics of the power consumption based on FCM algorithm

FCM clustering algorithm

In this paper, we use the fuzzy C-means (C-means Fuzzy, FCM) clustering algorithm to analyze the power behavior. FCM clustering algorithm based on objective function is suitable for processing large amounts of data, and the algorithm is simple, so it is easy to be realized on the computer. It is suitable for dividing the complex data set based on time series, which is in agreement with the characteristics of the data. The core idea of FCM algorithm is solved by J_m (U, P) min "{J" _m solution minimum value (U, P)}, thus obtaining the optimal partition matrix and clustercenter matrix. For the classification of object model space contains n members set X={x_1, x_2, x_n}. U can be expressed as:

$$U = [\mu_{ik}] = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \vdots & \mu_{ik} & \vdots \\ \mu_{c1} & \cdots & \mu_{cn} \end{bmatrix}, \text{ among them, } 1 \le i \le c, 1 \le k \le n \quad (1)$$

Among them,$\mu_{ik} = \mu_{x_i}(x_k)$ represents the subordinate relationship between the sample x_k and the subset $X_i(1 \le i \le c)$. For FCM the range of μ_{ik} is [0, 1]. Namely, the subordinate relationship between each sample and the subset X_i can be represented by a real number of 0~1, and $P = \{p_i, 1 \le i \le c\}$ said the cluster center matrix class I subset of X_i.Optimization objectives can be expressed as:

$$J_m(U,P) = \sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2 \quad 1 \le m < \infty \quad (2)$$

Among them, *m* is the smoothing factor, and the *m* control mode is the degree of sharing among the classes. The bigger the *m* is, the more fuzzy clustering results are obtained. In general, in order to control the clustering results not to be too vague, the *m* value is set to 2. d_{ik}indicates the distance

between the sample k to the i cluster center p_i can be represented by different types of model. Using Euclidean distance representation in this paper:

$$d_{ik} = \left[\sum_{j=1}^{s} \left| x_{ij} - x_{kj} \right|^2\right]^{\frac{1}{2}} \tag{3}$$

FCM algorithm updates the membership degree μ_{ik} and clustering center p_i. When the iterative convergence is obtained, the membership degree and the clustering center can be used to classify the data sets and to determine the relationship between the dataobjects and the classification. Through the iterative process in the control stop valve ε and the number of iterations of the b, we can solve the following formula:

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left[\frac{d_{ik}}{d_{jk}}\right]^{\frac{2}{m-1}}} \tag{4}$$

$$p_i = \frac{\sum_{k=1}^{n} (u_{ik})^m x_k}{\sum_{k=1}^{n} (u_{ik})^m} \tag{5}$$

Analysis of power consumption based on FCM

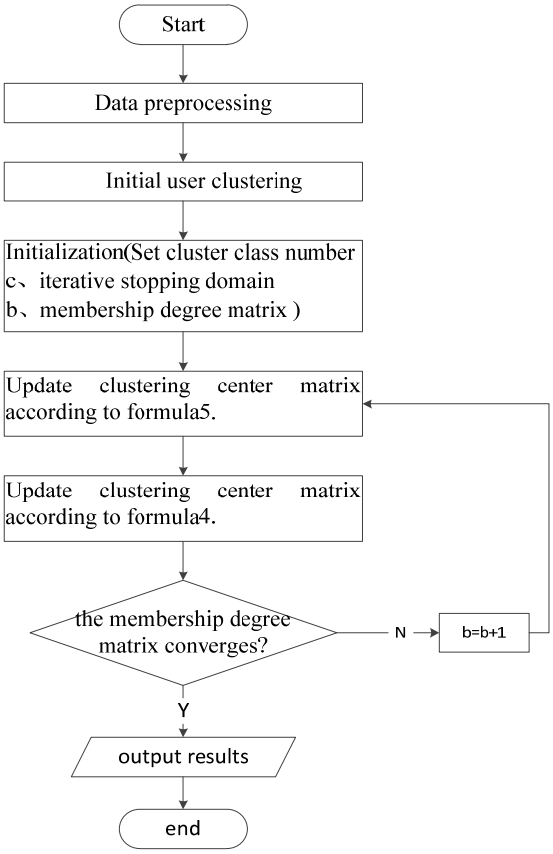

Fig.2 Theanalysis algorithm of electric behavior based on FCM

Residential users, large users of smart meter installed with PLC and wireless communication technology which is used to transmit the data of user with a certain frequency information acquisition system. We select four kinds of data objects as clustering analysis:

(1) x_{i1} power consumption: Daily total power;

(2) x_{i2} load rate: Average load / maximum load;

(3) x_{i3} peak electric coefficient: Peak time consumption/daily total power;

(4) x_{i4}valley electric coefficient: Valley power consumption/daily total power.

The acquisition frequency of the smart meter is a point every 15 minutes. So the daily collection is 96, and total daily electricity is the sum of the 96 data. The average load is /96. Total electricity consumption in peak time and valley power consumption are total electricity consumption in peak and valley time. So the x_k of each sample is a four-dimensional vector. FCM algorithm based on the use of electrical behavior analysis process as shown in Fig.2:

In data pre processing stage, the need for the missing data by difference algorithm was filled with, for beyond the threshold data were corrected. Next, the FCM clustering algorithm for initialization, including setting cluster category c, iteration stopping domain epsilon and iteration steps b=0 and the membership degree matrix U^0 . Next, according to the formula (4) and formula (5), the membership degree and clustering center will be updated during the iterative process unless the set stop condition$\left\|U^{(b)} - U^{(b+1)}\right\| < \varepsilon$come. At this point, the output of the clustering center is characterized by electricity behavior. And the degree of membership matrix determines the close degree of each sample and the behavior characteristics.

Clustering validity verification

The result of clustering analysis is closely related to the setting of data sample and parameter, it is unable to obtain the label information associated with data objectbecause clustering is an unsupervised learning process, so we need to introduce the clustering validity verification algorithm to evaluate the clustering results effectively. The contents of the evaluation include quantification of clustering's intra cluster compactness and the inter cluster separation. For the fuzzy clustering algorithm, the representative clustering validity verification method includes index Vxie of Xie-Beni[5].Vxie based on geometric structure, using the "compactness" and "separation" to measure the clustering quality of different division.For the FCM algorithm used in this paper, V_{xie}proves the cluster's effectivenessby obtaining the minimum value of thefollowing formula:

$$V_{xie}(U) = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(\mu_{ik})^m(d_{ik})^2}{n\cdot(\min\{c_i - c_j\})} \tag{6}$$

Distributed calculation method of FCM

In order to adapt to the MapReduce calculation model, we need to in parallel reform the electrical behavior analysis algorithm which were based on FCM algorithm. The FCM iteration is decomposed into two stages of Map and Reduce, on Map stage, a function can effects different data sets in the different date node. The Map output is a set of records in the form of <key, value> pairs stored on that node. After the end of map stage, the calculated model will be transmitted to the node which will undertake the work of Reduce, and dispose the key(like merge etc.) from the Map stage ,than output the final results in the form of<key, value>. Due to the Map and Reduce steps can run distributedly on multiple computers and highly abstract process of distributed computing, so the calculated model of MapReduce can analyze large-scale data (1TB) conveniently and effectively.

After studied the calculated model of MapReduce and combined with process of FCM algorithm, we find the similar computing that using formula(3) to calculate the distance of the sample to the current clustering center is the most frequent count. To clustering process of FCM which have n samples of the object in the classification of k, each iterations needs to do distance calculation about n*k time and each calculations needs to do variance calculation of s dimension's characteristics. According to the idea which can greatly improve the working efficiency of FCM, we put forward the FCM clustering algorithm based on MapReduce, the process is shown in Fig. 3.

(1) Copy the electric consumption data from relational databases (such as Oracle) to the distributed file system (HDFS), determine the clustering number *c* and stop domain *ε* according to the need of clustering;

(2) Based on the last clustering result, the initial clustering center can be determined and then transmit those data to the data nodes that participating in the distributed computing;

(3) Do some pretreatment to electric data and produce the key-value pairs of < user, profile >, in which the user is the user's unique identifier and profile contains the characteristics of the data object x_{i1}~x_{i4};

(4) All the key-value pairs of < user profile > is divided into several data subset, and transfer to the Map function. The Map function do membership degree calculation according to the formula (4) and store the result in the intermediate key-value pair of <*i*, μ_i> in which i is clustering number, and μ_i is membership degree about all objects in data subsets to the ith clustering;

(5) Transmit the calculation result of the Map function to Reduce nodes, Reduce task will merge the intermediate key-value according to the clustering number, then obtain a new clustering center by calculating formula (5);

(6) Repeat step (2) ~ (6), until the matrix of membership degree satisfies the stopping domain condition, the distributed FCM algorithm ends, export clustering results including the clustering number, the clustering center and the final membership degree of each clustering.

Through the six steps above, we can realize the distributed clustering analysis of the user's electrical information by using Mapreduce computing model on the Hadoop platform. And we can also obtain the classification of user groups through the final clustering center matrix and get the membership degree of every sample data to the cluster, so then determine its classification.

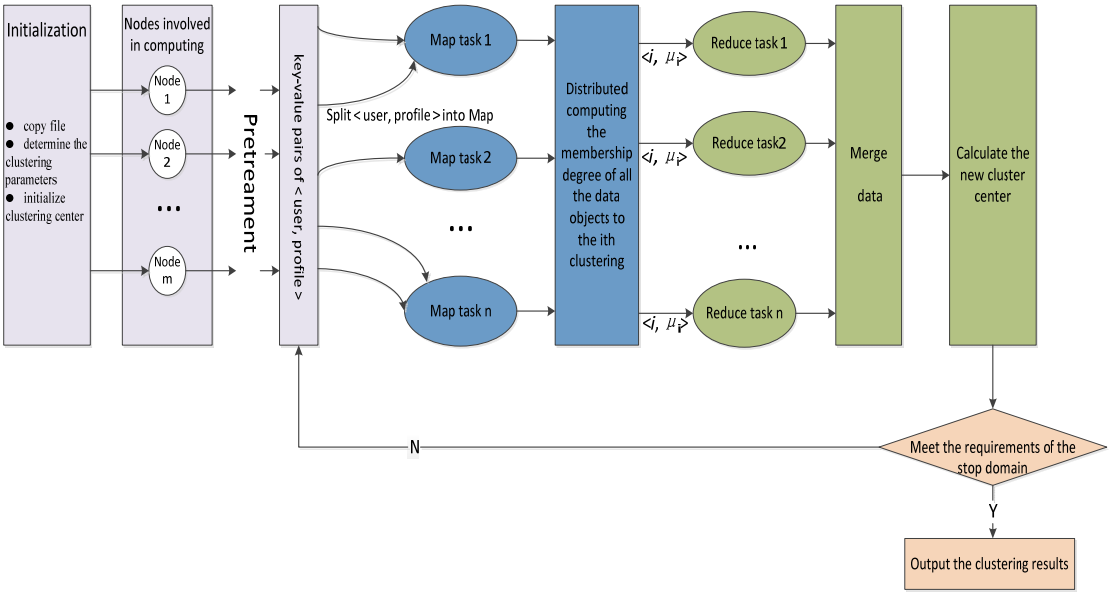

Fig.3 Parallel computation of FCM based on Mapreduce

Summary

In allusion to the application scenarios of user's electrical behavior clustering analysis in Sino-Singapore Tianjin Eco-City, we put forward the computing process of clustering analysis by using parallel computing technology and achieve the parallel design of Fuzzy c-means (FCM) clustering algorithm. The FCM parallel algorithm proposed in the paper can accurately complete the user classification of the Sino-Singapore Tianjin Eco-City, and mining the potential value of the vast amount of data so that it can provides a useful reference for the user to participate in the demand side response and to develop the optimal power strategy.

Acknowledgements

This work was supported by the Science and Technology Foundation of State Grid Corporation ((A research on the construction model of the internet thinking based smart grid innovative demonstration zone, SGTJDK00DWJS1500101).

References

[1] Hu xuehao. Smart Grid—A Development Trend of Future Power Grid[J]. Power System Technology, 2009(14): 1-5. (in Chinese)

[2] AiLing Duan, HaiFang Si. Research and Practice of Distributed Parallel Search Algorithm on Hadoop_MapReduce [S].Control Engineering and Communication Technology (ICCECT), 2012 International Conference. IEEE Conference Publications, 2012:105-108.

[3] Young-Il Kim,Jin-Ho ShinSong, Jae-Ju,Il-Kwan Yang. Customer clustering and TDLP (typical daily load profile) generation using the clustering algorithm [C]. Transmission & Distribution Conference & Exposition: Asia and Pacific, 2009. IEEE Conference Publications, 2009: 1-4.

[4] Prajesh P Anchalia． Improved MapReduce k-Means Clustering Algorithm with Combiner [C]． 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. Cambridge, IEEE, 2014: 12 – 17.

[5] Xie, X. L, Beni,G.. “A Validity Measure for Fuzzy Clustering”, IEEE Trans. on Pattern Analysis and machine Intelligence, Vol.13, No4, 1991.