Application of An Improved DBSCAN Algorithm in Web Text Mining

Xie Ping[1st],Zhang Lin[2nd],Wang Ying[3rd],Li Qinqian[4th]
School of Control and Computer Engineering Academy
North China Electric Power University
Beijing, China
zhanglin200809@163.com

Abstract: **This paper studies the characteristics and key technology of Web text mining, and puts forward an improved DBSCAN density clustering algorithm.The algorithm combines the characteristics of hierarchical clustering effectively, it can confirmed class center well,and make the neighborhood parameter r self-adapt to the data sets with different density. To the data sets with different density, it can adjust parameters according to the dense degree. Simulation experiment results verify the proposed algorithm can improve the accuracy in the Web text mining.**

Keywords-Web text mining; DBSCAN; hierarchical clustering; self-adapt

I. INTRODUCTION

The Internet store huge amounts of data with the rapid development of computer technology and Internet technology. But people can not take good use of the useful information hidden in the large amounts of data resources because of the dynamic and amorphousness of the Web document .Therefore, how to obtain information and knowledge quickly and efficiently from the Web has become a hot topic .Tang jing has put foward DFSSM Web text mining system model and Web text clustering algorithm TLDFSSM, which realized the clustering of different groups users'characteristics[1].As we know most information on internet show as the text form,and these data are unstructured or semi-structured. Web text clustering [2] as an important part of data mining, can find hidden category information in Web text collection effectively. Yang Lili etc put forward a clustering integration algorithm [3] based on SEAM , which can determin the clustering number automatically. At present, there are a lot of research on text clustering algorithm,such as the algorithm based on hierarchy and density etc..But these algorithms only as single clustering algorithms, perform not as good as we wish.Now the development of integrated clustering technology is not enough and the work on mixing the clustering techniques can be done. This article focuses on the clustering algorithm in the Web text mining applications.

II. WEB TEXT MINING

Web text mining is one kind of the Web data mining [4],which digs the information that users are interested in,and finds the potential knowledge.

A. Basic process of Web text mining

The basic process of text mining generally includes Web text collection, Web text pretreatment, Web text representation, feature extraction, Web text mining, mining results evaluation, knowledge model acquirement.The Web text mining is a complex process. As shown in figure 1.

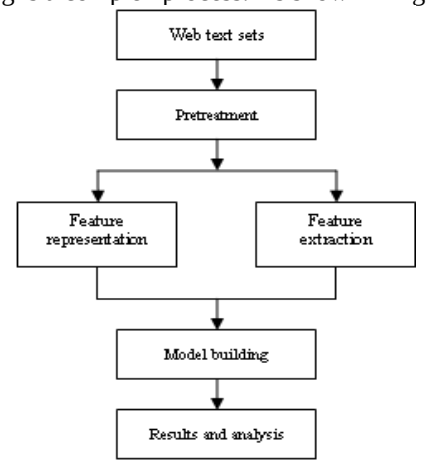

Figure 1 process of Web text clustering

B. Key technologies of Web text mining [5]

1） Web text collection and preprocessing

First of all, certain topic pages can be collected automatically by using the program,then Web text sets as the analytical base of following work are saved.To improve the quality of data and information,we need to do something in the Web text preprocessing, such as delete tabs, clear images, etc.,which can achieve the goal of data reduction and redundancy eliminating.

2） Feature representation and feature extraction

Before clustering analysis, the Web text format must be converted into data format, in order to facilitate the processing of clustering algorithm. Web text features is represented by a certain item (e.g., entry) to represent unstructured or semi-structured text information. Generally people use probability model,vector space model, Boolean model, etc for feature representation.

The process of text feature extraction is Removes some key characteristic vector witch are not too much useful and keep the key words witch can represent the text content. Its essence is mapping the high-dimensional data to low

dimensional space through transformation.After using vector space model to represent the characteristics, The dimensions of the vector space is very big,so the key of extraction is to finding the right map.

3） Web text clustering

Web text clustering refers to divide the original text collection into several clusters through analysis of Web text set. It requires that the quantity of text similarity as large as possible, and the similarity between the clusters as small as possible. Text clustering does not classify based on the predefined class table,but the given document collection.It gathers the relevant Web text into a class on the basis of the key vector.So the Web text clustering is a machine learning process without guidance.

III. CLUSTERING ALGORITHM

Clustering algorithm can mark off several classes by analysis the distribution characteristics of the data in database,and can also be used as a important part of data mining algorithm . The Web text data is different from traditional data objects. We can improve some common clustering algorithm, and apply directly on the Web text mining. Text clustering methods mainly include: dividing method, hierarchical method and method based on density, etc. Different methods have their own advantages and disadvantages of [6].

In this article,the improved algorithm involves the thought of the centeral point of K-MEANS algorithm(a classification method). First,we choose *k* objects randomly as *k* centeral point of initial partition. For other objects, calculate the distances between those *k* centeral point and them,and classify them into the nearest division. Then centeral points of each new class are calculated and classify again until the criterion function is convergence.

In this paper, we cluster Web text collection through the analysis of *r* neighborhood of each object ,which is based on the density of DBSCAN clustering algorithm, .if there are more than *n* objects in the *r* neighborhood of the point *p*, a new cluster with *p* as the center can be got. Iterate until any change of the clusters'shape is no longer happen. In this process, the clusters which have more relevance of density are merged autmatically.

DBSCAN algorithm can discover clusters with any shape, And can filter the isolated data (better performance in noise treatment). But before using this clustering method, the parameters *r* and n should be given in advance. The accuracy of the parameters depends on experience. Different texts have different characteristics of density.When the density of text is uniformity,We can use genetic algorithm and distance sorting to estimate the value *r* [7-8]. *r* and *n* can also be determined through analysis the statistical features of data to adaptation[9]. When density is nonuniformity, the effect of the method based on density clustering is bad, which is not applicable for high-dimensional data.We can determine threshold with nonuniform density through the equivalent rules between density clustering and grid clustering[10].

IV. IMPROVED CLUSTERING ALGORITHM

Among clustering algorithm about the Web text mining , K-MEANS algorithm based on the classification method and DBSCAN algorithm based on density method are widely used. The accuracy of K-MEANS algorithm which has low time complexity, and implement easily is not high because of the algroithm's strong dependence,which causes that the effect of K-MEANS is not ideal. DBSCAN algorithm with better cluster effect has good processing to the "noise" and can find arbitrary shape clusters in a given collection of objects,but the given parameter is difficult to determine in advance, only by experience, which influences the efficiency of the algorithm greatly.

This paper combines the advantages of both K-MEANS and DBSCAN. The improved algorithm not only can make class center determined preferable, but also can make parameters adjusted self-adapting to the data sets of different density which enhances the accuracy and efficiency of the algorithm.

A. Related definition and description

Here are some of the basic definition of the improved clustering algorithm

- Definition 1 *r* neighborhood of the point *p*: spherical area with *r* as the radius centered on any point *p* in the space.
- Definition 2 *k* average distance of point *p*: the average of distances of *k* points closest to the point p as aved *(p, k)*.
- Definition 3 *m* edge point set of *r* neighborhood of the point *p*: the point set which includes points that have greater distance than the *dm* from the point *p* in *r* neighborhood of the point *p*, as edge *(p, r, m)*. *dm* is the first m point distance from the nearest point *p* .
- Definition 4 absolute degree of outliers of point *p*: when *r* is the first *k* closest distance from point *p* and *m = k*,

$$\mathbf{stray}(\mathbf{p},\mathbf{k})=\frac{1}{\mathrm{m}}\sum_{q_i}\frac{\mathbf{aved}(q_i,\mathbf{k})}{\mathbf{aved}(\mathbf{p},\mathbf{k})},\ q_i\in \mathrm{edge(p,r,m)}. \qquad (1)$$

stray*(p,k)* called outlier degree of point *p*, absolute degree of outliers of point *p*

$$\mathrm{disstray}(\mathrm{p},\mathrm{k})=|\mathrm{stray}(\mathrm{p},\mathrm{k})-1|^{\alpha}\quad(\alpha=1,2,3) \qquad (2)$$

By the formula, you can see that the absolute degree of outliers is a positive number. The smaller the value of number is, the point has the higher similarity with the class.

B. Choice of parameter r

Every element of a set is in area with different density when the density is nonuniform. Here, the concept of density equals to the number of the points in *r* neighborhood of the point.

First of all, we use *k* average distance of point *p* as a measure of density. Make sure that *k* average distance of every point is calculated one and get all points' k average distance. Then ,several classes can be obtained by DBSCAN clustering with all average distances have

obtained . A largest distance d of point *o* which has the largest *k* average distance in every class can be got among distances of point *o* from its *k* closest points. It is reasonable to regard the value d as the radius parameter *r* of the class. Finally, we obtain several different radius parameter *r* using this method, ranking all the radius parameters with ascending order and coming into a one-dimensional set *U*.

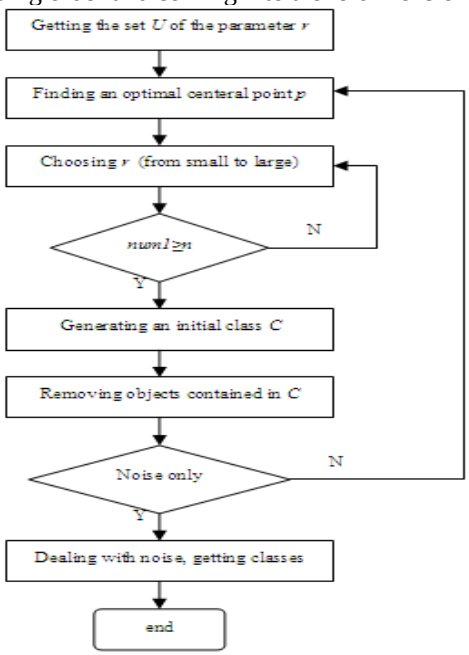

Figure 2 Fflow diagram of improved DBSCAN algorithm

C. *Choice of central point*

If we add density clustering algorithm to the concept of centeral point belonging to the classification method,the effect of clustering can be greatly improved,which can reduce the blindness of clustering, and improve the clustering efficiency. Initially given centeral point is random. Centeral point can be more reasonable through iteration and update with more points into the range of r neighborhood.

In this paper,we take a arbitrary point *p* from the data set as the centeral point and get the minimum value u1 from the set *U* as the radius *r*.In *r* neighborhood of the point *p*, get the number of object points *num1* and disstray*(p,k)*. Then, recalculate the centeral point *q* based on *r* neighborhood of the point *p*,At last, obtain the number of object points *num2* and disstray *(q, k)* within the neighborhood of *q*. If *num2* <*num1* and disstray *(q, k)* < disstray *(p, k)*, then *p* = *q*, repeat the above process to continue to find the centeral point of the optimal at this time; otherwise the p is the optimal centeral point.

D. *Improved DBSCAN algorithm*

(1) After finding the optimal centeral point *p*, compare the number of object points in *r* neighborhood of the point *p num1* with *n* .(*n* is set well in advance and remains unchanged).If *num1*<*n*, then choose a larger radius from *U*, until meeting *num1*≥*n*;

(2) An initial class *C* can be generated and remove data objects contained in class *C* from the data set;

(3) Repeat 3.1 and 3.2 and get a original class set, until remain the noise of data set at last. According to evaluation function, the noise point can be allocated to the closest class based on the distance from center of each initial class.

Flow diagram of improved DBSCAN algorithm is shown in figure 2:

V. SIMULATION EXPERIMENTS AND ANALYSIS

Test environment: Java programming language, Pentium Dual – Core 2.0 processor, 2 G memory, 250 G hard disk.

Web document collection: selecting 3460 news documents, including 680 news of economical theme, 1150 news of sport themes, 1630 news of entertainment theme.

With K-means algorithm, DBSCAN algorithm and the improved density clustering algorithm for independent testing, record and analyse the clustering results. Table 1, table 2 and table 3 for the test results.

TABLE I. CLUSTERING RESULTS OF K-MEANS ALGORITHM

Actual / Clustering	*Economy 680*	*Sport 1150*	*Entertainment 1630*
Economy	523	79	107
Sport	93	928	201
Entertainment	64	143	1322
Accuracy	76.9%	80.7%	81.1%

TABLE II. CLUSTERING RESULTS OF DBSCAN ALGORITHM

Actual / Clustering	*Economy 680*	*Sport 1150*	*Entertainment 1630*
Economy	565	50	90
Sport	50	965	169
Entertainment	65	135	1371
Accuracy	83.1%	83.9%	84.1%

TABLE III. CLUSTERING RESULTS OF IMPROVED DBSCAN ALGORITHM

Actual / Clustering	*Economy 680*	*Sport 1150*	*Entertainment 1630*
Economy	601	26	66
Sport	23	1049	102
Entertainment	56	75	1462
Accuracy	88.4%	91.2%	89.7%

VI. CONCLUSION

From the result of the above mentioned three algorithm, the effect of the K -MEANS algorithm clustering(the average accuracy is 79.6%) is inferior to DBSCAN algorithm(the average accuracy is 83.7%). As we can see, the accuracy of DBSCAN algorithm is slightly higher than K -MEANS algorithm,because the value of *k* is hard to determine. However, the improved algorithm clustering is obviously superior than the former two with the average accuracy is 89.8%. Inquiring into the reasons, using *k*

neighboring distance as the clustering radius plays a great role. And the method contributes to the noise removal. At last, we can draw the conclusion when dealing with high dimensions data types, the improved clustering algorithm can obtain better effect.

REFERENCES

[1] Tang J.Web Text Mining System and the Research of Clustering Algorithm.Telecommunication Construction,2004.(2):24-28.

[2] Li B.Reserch of Clustering Algorithm in Web Mining.Nanjing:Nanjing University of Posts and Telecommunications,2010.

[3] Yang L L.The Research of the Clustering Ensembles Based on SEAM Algorithm and it's Application on Test.Beijing:Beijing Jiaotng University,2009.

[4] Cao C C,Kang YH.Research in Web Data Mining.Modern Electronic Techinque,2007,(4):92-97

[5] Jin C X.Research and Application in Web Text Mining.Microcomputer Applications,2009.(7):54-56.

[6] Wei L.Comparison of Clustering Algorithms in Data Mining. Computer Knowledge and Technology,2007,(21):637-639.

[7] Lin C Y, Chang C C,Lin C C. Fundamental Informaticae,2005,68(4):315-331.

[8] Yue S H, Li P, Guo J D, et al.J Zhejiang University Science,2005,A6(1):71-78.

[9] Xia LN,Jing JW.SA-DBSCAN:A self-adaptive density-based clustering algorithm.Journal of the Gradate School of the Chinese Academy of Sciences,2009,26(4):530-538.

[10] Zhou S,ZhouA,Cao J.A Data-partitioning-based DBSCAN Algorithm.Journal of Computer Research & Development,2000,37(10):1153-1159.