A New Text Clustering Method Based on KGA

ZhanGang Hao
Shandong Institute of Business and Technology, Yantai ,China
Email:zghao2000@hotmail.com

***Abstract*—Text clustering is one of the key research areas in data mining. K-medoids is a classical partitioning algorithm, which can better solve the isolated point problem, but it often converges to local optimization. In this paper, we put forward a new genetic algorithm called KGA algorithm by putting k-medoids into the genetic algorithm, then we form a local Optimal Solution with multiple initial species group, strategy for crossover within a species group and crossover among species groups, using the mutation threshold to control mutation. This algorithm will increase the diversity of species group and enhance the optimization capability of genetic algorithm, thus improve the accuracy of clustering and the capacity of acquiring isolated points.**

***Index Terms*—Text clustering, K-medoids algorithm, genetic algorithm**

I. INTRODUCTION

Text clustering methods have been some. K-menas algorithm and k-medoids algorithm are efficient, able to effectively handle large text collection, but will generally converge to a local minimum, it is difficult to ensure that the global minimum.Some text clustering algorithm are proposed, For example, SKM algorithms, WAP algorithms and other algorithms[5-11]. Most algorithms can be more efficient to solve the problem text clustering. However, these algorithms find isolated points in the results in terms of weak.

Genetic algorithm is a random optimization algorithm based on natural selection and genetics. [1] It has been tried to use genetic algorithm to solve partitioning problem, [2-4] resulting in either undesirable outcome or failure to solve the sloe point problem. This paper presents a new genetic algorithm using K-medoids algorithm to optimize the population to the local population to evolve the optimal solution, which can improve the convergence speed; propose a new niche of the population based on the generation and evolutionary approach to improve the diversity of the population; propose a new cross-method; proposed variation of the threshold, to control the genetic variation of the algorithm to preserve the elite individual.

II. LITERATURE REVIEW

In the past few years, some people have been studied for text clustering. Xu Sen et al proposed Spectral clustering algorithms for document cluster ensemble problem . In this paper, two spectral clustering algorithms were brought into cument cluster ensemble problem. To make the algorithms extensible to large scale applications, the large scale matrix eigenvalue decomposition was avoided by solving the eigenvalue decomposition of two induced small matrixes, and thus computational complexity of the algorithms was effectively reduced. Experiments on real-world document sets show that the algebraic transformation method is feasible for it could effectively increase the efficiency of spectral algorithms; both of the proposed cluster ensemble spectral algo-rithms are more excellent and efficient than other common cluster ensemble techniques, and they provide a good way to solve document cluster ensemble problem[5].

DHILLON I S et al proposed SKM algorithms (sphe -rical K-means). It Has been proved to be a very efficient algorithms. However, SKM algorithm is gradient-based algorithm, the objective function with respect to the concept of vectors in R^d is not strictly concave function space. Therefore, different initial values will converge to different local minima, that algorithm is very unstable[6].

Guan Renchu et al proposed WAP(weight affinity propagation)algorithms. Abstract Affinity propagation (AP) is a newly developed and effective clustering algorithm. For its simplicity, general applicability, and good performance, AP has been used in many data mining research fields. In AP implementations, the similarity measurement plays an important role. Conve -ntionally, text mining is based on the whole vector space model(VSM)and its similarity measurements often fall into Euclidean space. By clustering texts in this way, the advantage is simple and easy to perform. However, when the data scale puffs up, the vector space will become high-dimensional and sparse. Then, the computat -ional complexity grows exponentially. To overcome this difficulty, a nonEuclidean space similarity measure -ment is proposed based on the definitions of similar feature set(sFS}，rejective feature set(RFS) and arbitral feature set(A F S).The new similarity measurement not only breaks out the Euclidean space constraint, but also contains the structural information of documents. Theref -ore, a novel clustering algorithm, named weight affinity propagation(WAP)，is developed by combining the new similarity measurement and AP. In addition, as a be -nchmark dataset, Reuters-21578 is used to test the proposed algorithm. Experimental results show that the proposed method is superior to the classical k-means, traditional SOFM and affinity propagati -on with classic similarity measurement[7].

PENG Jing et al proposed a novel text clustering

Manuscript received September 30,2011; revised November 17,2011;accepted November 23,2011.

doi:10.4304/jsw.7.5.1094-1098

algorithm based on Inner product space model of semantic. Abstract Due to lack considering the latent similarity information among words, the clustering result using exist clustering algorithms in processing text data, especially in processing short text data, is not ideal. Considering the text characteristic of high dimensions and sparse space, this paper proposes a novel text clustering algorithm based on semantic inner space model. The paper creates similarity method among Chinese concepts, words and text based on the definition of inner space at first, and then analyzes systematically the algorithm in theory. Through a two phrase processes, i. e. top-down"divide" phase and a bottom-up"merge" phase, it finishes the clustering of text data. The method has been applied into the data clustering of Chinese short documenu. Extensive experiments show that the method is better than traditional algorithms[8].

In addition, Hamerly G[9],WagstaffK[10],Tao Li[11] ,G. Forestier [15],Wen Zhang [16],Linghui Gong[17] and Argyris Kalogeratos[18] Were also proposed the method of text clustering.However, these methods are not effectively solve the problem of isolated points. So, we put forward a new genetic algorithm called KGA(k-medoids genetic algorithm,KGA) algorithm by putting k-medoids into the genetic algorithm. Compared with the k-menas algorithm, the KGA algorithm not only can better solve the problem of isolated points, and be able to find the global optimum. Compared with the K-medoids algorithm, isolated point of the search algorithm better, and be able to find the global optimum. With the new algorithm, the KGA algorithm can not only efficiently, but more good points to solve the problem in isolation.

III. CHARACTERISTIC DENOTATION OF TEXT

A Chinese Text Categorization model first makes Chinese text groups participle and vector, forming a characteristic group, followed by the extraction of a most optimum characteristic sub group from all characteristic groups using characteristic extraction algorithm according to characteristics evaluation function.

Chinese text transforms non-structural data to structural data by the treating the participles, using text vector space model. The basic idea of VSM can be explained in such a way, each article in the text group is denoted as a vector in a high dimensional space according to predefined vocabulary order. Word in predefined vocabulary order is viewed as the dimension of the vector space and the weight of the word is viewed as the value of the vector in a certain dimension of the high dimensional space, consequently, the article is denoted as a vector in a high dimensional space. The advantage of VSM is that it is simple, not demanding on semantic knowledge and easy for calculation.

This model defines text space as a vector space composed of orthogonal words vector. Each text d is denoted as a normalized characteristic vector V(d)=(t1,w1(d)； …ti,, wi (d);…;tn, wn (d)), ti is the characteristic word in text d;, wi (d) is the weight of ti in d, calling V(d) the vector space expression of text d, Wi (d)= $\psi(tf_i(d))$. ψ uses TF•IDF function, which has many formulas in actual application. The one used by this paper is:

$$w_i(d)=\frac{(\log(\mathrm{tf}_i)+1.0)\times\log(N\mid n_i)}{\sqrt{\sum_{i=1}^{l}[(\log(\mathrm{tf}_i)+1.0)\times\log(N\mid n_i)]^2}} \qquad (1)$$

In the formula, tf_i is the frequency of characteristic word ti in text d , N is the total text number in the text group, n_i is the number of texts in the text group that contain characteristic word t_i , l is the number of characteristic words in text d .

IV. BRIEF INTRODUCTION TO K-MEDOIDS ALGORITHM

The primary idea of the k-medoids algorithm is that it firstly needs to set a random representative object for each clustering to form k clustering of n data. Then according to the principle of minimum distance, other data will be distributed to corresponding clustering according to the distance from the representative objects. The old clustering representative object will be replaced with a new one if the replacement can improve the clustering quality. A cost function is used to evaluate if the clustering quality has been improved. The function is as follows:

$$\Delta E = E_2 - E_1 \qquad (2)$$

where ΔE denotes the change of mean square error; E_2 denotes the sum of mean square error after the old representative object is replaced with new one; E_1 denotes the sum of mean square error before the old representative object is replaced with new one. K-medoids clustering algorithm follows four main processing:

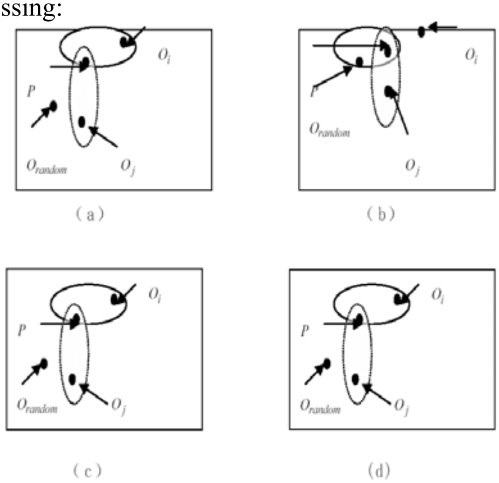

Figure1.k-medoids algorithm clustering process figure

If ΔE is a minus value, it means that the clustering quality is improved and the old representative object should be replaced with new one. Otherwise, the old one should be still used.

The procedure of the k-medoids algorithm is as follows:

(1) Choose stochastic k objects as the initial clustering representative objects from n data;

(2) Circulate steps from (3) to (5) until every clustering doesn't change;

(3) According to the distance (generally using Euclidean distance) between each datum and the corresponding clustering representative object and according to the minimal distance principle, distribute each datum to the corresponding clustering;

(4) Randomly choose a not representative object O_{random} and calculate the cost ΔE of changing with the stochastic representative object O_j chose;

(5) If ΔE is minus, replace O_j .with the O_{random} .

V. TEXT CLUSTERING METHOD BASED ON IMPROVED GENETIC ALGORITHM

Genetic algorithm is a good algorithm, but when doing text clustering, there are all kinds of problems. This paper presents an improved genetic algorithm will be able to improve the efficiency of clustering, but also can better solve the problem of isolated points.

A. Encoding

Suppose we divide dataset n into k subsets, then we may consider two denotation ways. The first one is that we can classify a certain text into a certain category, after all the texts have been classified, the denotations representing text file and the category it belongs to form a chromosome. For example, a chromosome is denoted as $r=\left\{r_{12},r_{23},\cdots,r_{ij}\cdots r_{nk}\right\}$, r_{ij} denotes I articles that belong to the jth category, $1\le i\le n$, $1\le j\le k$. The second way is to form the chromosome by the center of each clustering. For example, a chromosome is denoted as $r=\left(r_1,r_2,\cdots,r_k\right)$, r_i denotes the ith clustering center is r_i . In general, there are many text files as clustering, if the first way is used, the chromosome will be too long, resulting in increased difficulty for crossover and mutation. So the second way is adopted by this paper. There are many encoding of genetic algorithm[12-13], we use real-coded.

B.The formation of initial species group

Initial species group can be generated by random function to form an initial group matrix. However, this matrix is so random that the quality of chromosome of the whole species group can't be ensured. Therefore, this paper adopts k-medoids algorithm to optimize the species group generated randomly, resulting in a new species group matrix as the initial species group matrix of the genetic algorithm.

For example, we perform clustering VI category against 100 text files to generate initial species group with real number coding. The species group matrix generated by random function is shown as follows:

$$\begin{pmatrix} 1 & 3 & 5 & 26 & 40 & 50 \\ 2 & 41 & 23 & 4 & 65 & 76 \\ 3 & 7 & 90 & 84 & 32 & 67 \\ 82 & 45 & 70 & 9 & 8 & 15 \end{pmatrix}$$

This is a species group matrix comprising 4 rows and 6 columns, which denotes that 4 rows mean there are 4 individuals in this species group and 6 columns mean each chromosome contains 6 genes, which is the category number that need to be clustered. Optimizing this initial matrix by k-medoids algorithm result in the following species group matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} \end{pmatrix}$$

The matrix is the initial population matrix. To avoid the genetic algorithm into a local optimum, we must enhance the population diversity. This article draws ideas niche genetic algorithm, using the mechanism of niche crowding genetic technology, and thus to maintain the population diversity.

C. Fitness function

This paper uses mean square error as fitness function and defines

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}\left|p-m_i\right|^2 \qquad (3)$$

E is the mean square error sum of all data objects and corresponding clustering center, p is a point in the space representing objects and m is the mean value of clustering C_i . The fitness function used by this paper meets the major conditions required for designing fitness function.

D. Genetic operator

This paper uses roulette selection, the basic idea of which is that the probability of each individual is selected is positively proportional to its fitness. The specific operation is expressed as follows:

$$p(a_j)=\frac{f(a_j)}{\sum_{i=1}^{n}f(a_i)}, \qquad j=1,2,\cdots,n \qquad (4)$$

$p(a_j)$ denotes the probability the jth individual is selected, $f(a_j)$ denotes the value of fitness function of the jth individual and n denotes the number of total individuals.

In this paper, the genetic algorithm adopts multiple initial species group strategy, the crossings of which include crossover within a species group and crossover among species groups. For crossover within a species group, to retain sound gene fragment, this paper uses single point crossover, which represents a crossover point randomly set in an individual coding string, and then portion of the genes of pairing individuals is exchanged at this point. For crossing among species groups, for every 50 generations evolved, random pairing crossing among species groups are performed (for odd groups, the group left after all the other even groups have finished pairing go into the next cycle), single point crossing is also adopted by crossing among groups. Crossing rate normally takes 0.4-09[14].

In order to retain the diversity of species groups, mutation operators are needed. However, mutation might destroy valuable genes. Hence, this paper sets a mutation threshold ∂ . Before mutation, a random number should be produced. If this number is greater than ∂ , mutation happens; if this number is not greater than ∂ , then the genes are retained without mutation occurring. Crossing rate normally takes 0.001-0.1[14].

E. Criteria to stop the algorithm

The first one is to fix the maximum genetic algebra. The algorithm stops as the maximum algebra appears.The second one is according to the degree of convergence. The algorithm stops as the mean fitness of the species group undergoes no change after a few generations.

VI. EXPERIMENTAL ANALYSIS

This paper picks up 505 articles in 6 categories from CQVIP as experiment data. The first 5 categories contain 100 articles each and the last category contains 5, which form the isolated points. The first 500 articles for the experiment are sourced from http://dlib.cnki.net/kns50/. The 5 categories are industrial economy（IE）, cultural and economic (CE), Market Research and Information (MRI), Management(M), service economy(SE). respectiv -ely. The last category is current affair and news(CAN) sourced from http://www.baidu.com/. After having undertaken basic treatment and dimension reduction to these files, k-medoids algorithm and KGAalgorithm are used for clustering analysis.

A.Experiment 1

First, k-medoids algorithm is used for clustering analysis. The results are shown in Table1.

TABLE Ⅰ RESULTS FROM K-MEDOIDS ALGORITHM

	IE	CE	MRI	M	SE	CAN
Wrong articles	59	60	55	52	57	1
Correct articles	41	40	45	48	43	4
Percentage of correct ones	41	40	45	48	43	80
Time(second）	32.5					

As can be seen from the above experiments, K-medoids algorithm for text clustering, the time is very short, very efficient, but also better identify isolated points. However, clustering results are not satisfactory,, clustering accuracy is very low.

B. Experiment 2

Then, KGA algorithm is used for clustering analysis. The results are shown in Table 2.

TABLE Ⅱ RESULTS FROM KGA ALGORITHM

	IE	CE	MRI	M	SE	CAN
Wrong articles	12	13	8	9	9	0
Correct articles	89	87	94	91	88	5
Percentage of correct ones	89	88	90	91	90	100
Time(second）	12375					

As can be seen from the experiment 2, algorithms presented in this paper KGA increased with time despite the many, but clustering effect is very good. As can be seen from Table 2, significantly reduced the number of false papers, the correct number of articles increased significantly, but also to identify well isolated point.

VII. SUMMARY

Text clustering is widely used in real world and an important subject for data mining. Both k-medoids and genetic algorithms can be used to study it although each method has shortcomings. This paper embeds k-medoids algorithm into genetic algorithm, proposing new tactics for initial species group, crossover and mutation as well as a new algorithm KGA. This algorithm increases the diversity of species groups, enhances genetic algorithm's capability to search ideal targets and improves clustering accuracy and its capability to aquire isolated points.

ACKNOWLEDGEMENTS

This paper was supported in part by the National Natural Science Foundation of China （Grant No.70971077）,Shandong Province Doctoral Foundation (2008BS01028),Natural Science Foundation of Shandong Province (Grant No.ZR2009HQ005).

REFERENCES

[1] D.B.Fogel. An introduction to simulated evclutionary optimization[J], IEEE Trans.Neural Network,vcl.5,no.1, 1994,3~14.

[2] D.R.Jones and M. A. Beltramo.Solving partitioning problems with genetic algorithms[C], in Proc. 4th Int. Conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman, 1991,442~457.

[3] HE Ting-ting,DAI Wen-hua,JIAO Cui-zhen, Research of Text Clustering Based on Hybrid Parallel Genetic Algorithm[J], JOURNAL OF CHINESE INFORMATION PROCESSING,2007,21(4),55-60.

[4] QIN Xiao,YUAN Chang-an, Text clustering method based on genetic algorithm and SOM network[J],Computer Applications, 2008 ,28(3), 757 -760.

[5] XU Sen, LU Zhi-mao,GU Guo-chang, Spectral clustering algorithms for docu -ment cluster ensemble problem[J], Journal on Communications, 2010, Vol. 31 No.6,58-66.

[6] DHILLON I S, MODHA D S. Concept decompositions for large sparse text data using clustering[J]. Macliine Learning, 2001, 42(1-2):143-175.

[7] Guan Renchu，Pei Zhili，Shi Xiaohu,Yank Chen，and Liana Yanchun, Weight Affinity Propagation and Its Application to Text Clustering[J], Journal of Cor -mputer Research and DeveloprnenL, 2010,47(10), 1733- 1740.

[8] PENG Jing ,YANG Dons-Qin, TANG Shi-Wei, FU Yan, JIANG Han-Kui, A Novel Text Clustering Algorithm Based on Inner Product Space Model of Semantic[J], CHINESE JOURNAL OF COMPUTERS, 2007, 30(8), 1354-1362.

[9] Hamerly G, Elkan C.Learning the k in k-means// Pmcee -dalgs of the 17th Annual Conference on 1\eural hlfamatiou Pmcessalg Svstmls (NIP S).2003,281-289.

[10] WagstaffK, Cardie C,Rogers S, Schroedl S Constranied K-mearns clustering with background knowledge In Brodley CE, Danyluk AP, eds. Proc of the 18th Int 1 Conf on Machine Learning[M].William stow M organ Kauf m ann Publishers 2001.577-584.

[11] Tao Li Docunent clustering via Adaptive Suhspace lterat ion[A]. In proceedings of the 12th ACM international Conference on Multimedia[C]. New York ACM Publisher 2004 364- 367.

[12] J.N.Bhuyan, V.V.Raghavan,and V.K.Elayavalli,Genetic algorithm for clustering with an ordered representation,in Proc. 4th Int. Conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman, 1991,408~420.

[13] YUAN C, TANG C,WEN Y,etal, Convergence of genetic regression in data mining based on gene expression pmgranming and optimized solutions[J],International Journal of Computer and Application, 2006, 28(4): 359-366.

[14] WANG Xiao-ping, CAO Li-ming, Genetic algorithm: Theory, application and software realization, Xi’an: Xi’an Communication University Press, 2002.

[15] G. Forestier, P. Ganrski , C. Wemmert. Collaborative clustering with background knowledge [J]. Data & Knowledge Engineering,2010,69(02):211-228.

[16] Wen Zhang a, Taketoshi Yoshida b, Xijin Tang c, Qing Wanga， Text clustering using frequent itemsets[J], Knowledge-Based Systems,2010,23(5),379-388.

[17] Linghui Gong, Jianping Zeng , Shiyong Zhang，Text stream clustering algorithm based on adaptive feature selection[J], Expert Systems with Applications, 2011, 38 (3),1393-1399.

[18] Argyris Kalogeratos, Aristidis Likas，Document cluste -ring using synthetic cluster prototypes[J], Data & Knowledge Engineering, 2011,70(3), 284-306.

ZhanGang Hao 1976,3. Obtained from Tianjin University in 2006 PhD in Management. Research areas: text mining, knowledge management, evolutionary algorithms,

He is ASSOCIATE PROFESSOR Shandong Institute of Business and Technology in YanTai of Shandong province.