A New Text Clustering Method Based on KSEP

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 algorithm is a classical division algorithm, and can solve the problem of isolated points, But it often converges to local optimum. This article presents a improved social evolutionary programming(K-medoids Social Evolutionary Programming,KSEP). The algorithm is the k-medoids algorithm as the main cognitive reasoning algorithm. and Improved to learning of Paradigm、Optimal paradigm strengthening and attenuation and Cognitive agent betrayal of paradigm. This algorithm will increase the diversity of species group and enhance the optimization capability of social evolutionary programming, thus improve the accuracy of clustering and the capacity of acquiring isolated points.**

***Index Terms*—Text clustering, K-medoids algorithm, social evolutionary programmi ng**

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.

Social evolutionary programming (SEP) is an algorithm based on paradigm conversion into global search algorithm[1-2], has been used to solve the problem of clustering[3-4], But not solved the problem of isolated points. This article presents a improved social evolutionary programming(K-medoids Social Evolutiona -ry Programming,KSEP).The algorithm K-medoids algorithm is as cognitive subject's cognitive reasoning algorithm; raised awareness of the main new paradigm in the study of clustering in the way; propose a new paradigm of the optimal formula to strengthen and decay. This algorithm will increase the diversity of species group and enhance the optimization capability of social evolutionary programming, thus improve the accuracy of clustering and the capacity of acquiring isolated points.

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. Conventionally, text mining is based on the whole vector space model(VSM)and its similarity measurement -s 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 computational complexity grows exponentially. To overcome this difficulty, a nonEuclidean space similarity measurement 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 measure -ment not only breaks out the Euclidean space constraint, but also contains the structural information of documents. Therefore, a novel clustering algorithm, named weight affinity propagation(WAP)，is developed by combining the new similarity measurement and AP. In addit -ion, as a benchmark dataset, Reuters-21578 is used to test the proposed algorithm. Experimental results show that the proposed method is superior to the

Manuscript received September 30,2011; revised November 20,2011;accepted November 26,2011.

doi:10.4304/jsw.7.6.1421-1425

classical k-means, traditional SOFM and affinit -y propagation with classic similarity measurement [7].

PENG Jing et al proposed a novel text clustering 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 metho -d is better than traditional algorithms[8].

In addition, Hamerly G[9],WagstaffK[10],Tao Li[11] ,G. Forestier [13],Wen Zhang[14],Linghui Gong[15] and Argyris Kalogeratos[16]Were also proposed the method of text clustering.However, these methods are not effectively solve the problem of isolated points. So,This article presents a improved social evolutionary programming (K-medoids Social Evolutionary Programming,KSEP). 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

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

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

IV. TEXT CLUSTERING METHOD BASED ON KSEP

A. K-medoids-based body of cognitive reasoning algorithm

In order to enhance the ability of algorithms to find outliers, this algorithm will be k-medoids algorithm as the individual's cognitive.

K-medoids algorithm operating principle is shown below: 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[12]. The function is as follows:

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

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.

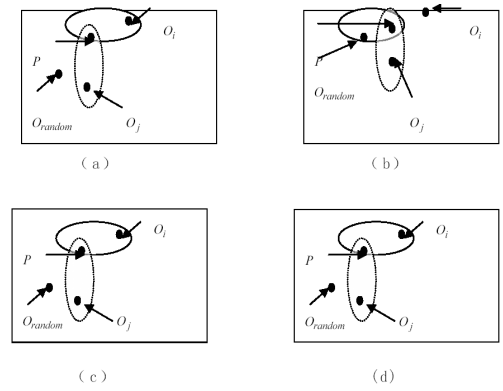

Figure1.k-medoids algorithm clustering process figure

K-medoids clustering algorithm follows four main processing as Fig1.

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} .

B. Evolved self-optimization process of paradigm-based learning and updating

A good paradigm is a good viable solution record. Here, F stands for paradigm, M for the number of paradigms, $F[i]$ for *NO.i* paradigm ($i = 1,2 \cdots M$). M number of paradigms are arranged according to object function value $f(F[i])$ in ascending sequence, as shown below: $f(F[1]) \le f(F[2]) \le \cdots f(F[M-1]) \le f(F[M])$. We can obtain series of individuals through application of K-means algorithm. Once a new individual $F[l]$ is obtained, it is inserted in the proper position of M number of paradigms arranged in object function value ascending manner if its object function is smaller than a certain object function value already having individuals, i.e. for $j \in (1, M)$,if $f(F[j-1]) < f(F[l]) < f(F[j])$, then $F[j] = F[l]$, $F[j+1] = F[j], \cdots, F[M] = F[M-1]$. Such as, in the entire evolving process, M number of paradigms are constantly in a dynamic updating status.

C. Learning paradigm form of cognitive agent in cluster

A new paradigm produced in $NO.k$ generation cognitive agent should refer to $NO.(k-1)$ generation paradigm. In the k generations, all sorted in the selected paradigm after 1 / 3 of the paradigm(Because the function value in accordance with small to large order, after the 1 / 3 of the paradigm is the paradigm of the best part of all). If there are M paradigm, that is, select the M / 3 (rounded up) a paradigm. In this M / 3 个 paradigm paradigm in a randomly selected. In this paradigm, the category $h(h \in (1, c), c$ is the number of clusters)of the number of datum $i(i \in (1, n)$, n is the numberof data to be clustered) to undergo can be explicitly displayed. It is believable that some data near clustering center p (p is the mean value of category h) under category h embody better cognitive behavior of the agents so that the heritage of such behavior should proceed from $NO.k$ individual, i.e. each categories under category h reserves certain data (to the preset ratio, e.g. ratio=0.45 as assigned in this article and taken round numbers upward). The reserved data are still of c number of categories and the rest will be allocated to these categories according to the similarity (Euclidean distance is used in this article) to the clustering center (means of all data under the category) as to complete heritage and generate a new paradigm.

D. Optimal paradigm strengthening and attenuation

To strengthen SEP local self-optimizing ability, learning probability p_1 of "currently most optimal paradigm" $F[1]$ may be artificially enlarged. Mean-while, p_1 value should also be attenuated step by step in order to prevent entire social population convergence toward most optimal paradigm to lead to reduction of global self-optimization capability. The specifics are illustrated as follows:

If a new "currently most optimal paradigm" $F[1]$generated in k generation, in the process of $k+1$ generation clusters, the probability of learning paradigm $F[1]$ is designated as p_1 , $p_1 \in (0,1)$, and the probability $p_i (i = 2,3, \cdots, M)$ learned by other paradigms as

$$p_i = [1 / f(F[i])](1 - p_1) / \sum_{i=2}^{M} 1 / f(F[i]) \quad (3)$$

In general, the more the algorithm is more close to the evolution of post-its optimal solution, In order to not destroy the optimal solution as much as possible,the learning probability p_1 of Optimal paradigm, In the period (such as the first half of the cycle) and given its relatively small value, Later re-assigned a higher value. This is to keep a good paradigm, but also can increase the diversity of population.

In the process of each generation clusters between $k+2$ generation and $k+t$ generation(the supposition is renewed once more in $k+t$ generation of "currently most optimal paradigm"), the probability p_1 of paradigm $F[1]$ learned by other paradigms in turn as:

$$p_1^{k+i}=\begin{cases} p_1^k\times(100-\mu_1^{(i-1)})/100 & \text{if} \quad u\le\dfrac{(t-2)}{3} \\ p_1^k\times(100-\mu_2^{(i-1)})/100 & \text{if} \quad \dfrac{(t-2)}{3}<u<\dfrac{2(t-2)}{3} \\ p_1^k\times(100-\mu_3^{(i-1)})/100 & \text{if} \quad u\ge\dfrac{2(t-2)}{3} \end{cases} \tag{4}$$

In which, $i\in 2,3,\cdots,t)$, parameter μ_1,μ_2,μ_3 controls attenuation rate, its right shoulder mark $(i-1)$ is the times of power. The less the μ_1,μ_2,μ_3, the slower the attenuation. In general, $\mu_1\in(2.5,3),\mu_2\in(1.5,2.5),\mu_3\in(1,1.5)$ Other "paradigm" genetic rate $p_i,i\in(2,3,\cdots,M)$ will stil luse Eq. (3) for computation.

E. Cognitive agent betrayal of paradigm

①. Assume cognitive agent mutation probability threshold α, which is used to determine whether a certain cognitive agent bears the nature of betrayal, whereas behavioral mutation probability threshold β is used to determine on which time or times of specific behavior in the entire process the cognitive agent inclining to betray fall into betrayal.

②. Prior to cognition of each cognitive agent, a random number is given by an evenly distributed generator. If it is not greater than mutation threshold α, it is considered that it does not have betrayal nature and its behavioral process rigorously follows "cognitive agent learning paradigm" form to complete genetic process as mentioned above; otherwise, this agent has the nature of betrayal and is continued instep ③.

③. If the cognitive agent is identified bearing betrayal nature, a random number is assigned by the evenly distributed generator. If the random number is not greater than behavioral mutation rate β, the behavior does not belong to betrayal behavior and follow existing paradigm genetic form as described in cognitive agent learning paradigm; otherwise, this agent has the nature of betrayal and chaotic mutation operator will be applied to produce a new individual.

The value of α and β,in the early to give its larger value, given its relativ -ely small in the latter part of the value.

V. 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). respectively. 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 KSEP algorithm are used for clustering analysis.

A. Experiment 1

First, k-medoids algorithm is used for clustering analysis. The results are shown in Table 1.

TABLE 1 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, KSEP algorithm is used for clustering analysis. The results are shown in Table 2.

TABLE 2 RESULTS FROM GA-K ALGORITHM

	IE	CE	MRI	M	SE	CAN
Wrong articles	9	11	8	7	9	0
Correct articles	91	89	92	93	91	5
Percentage of correct ones	91	89	92	93	91	100
Time(second)	1893					

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.

VI. SUMMARY

Text clustering is widely used in real world and an important subject for data mining. K-medoids algorithm is a more classical clustering algorithm, but its accuracy is lower. This paper embeds k-medoids algorithm into Social Evolutionary Programming, and Improved to learning of Paradigm、Optimal paradigm strengthening and attenuation and Cognitive agent betrayal of paradigm. This algorithm will increase the diversity of species group and enhance the optimization capability of social evolutionary programming, thus improve the accuracy of clustering and the capacity of acquiring isolated points.

ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Foundation of China (Grant No.70971077), Shandong Province Doctoral Foundation (2008BS0 1028), Natural Science Foundation of Shandong Province (Grant No.ZR 2009 HQ005, ZR2009HM008).

REFERENCES

[1] Yu Yixin, Zhang Hongpeng. A social cognition model applied to general combination optimization problem. Proceedings of the first international conference on machine learning and cybernetics, November4-5,2002 Beijing China,1208~1213.

[2] Sebastien Picault, Anne Collinot, Designing Social Cognition Models for Multi-Agent Systems through Simulating Primate Societies,Proceedings of ICMAS98(3rd International Conference on Multi-Agent Systems),1998,238~245.

[3] HAO Zhangang, Building Text Knowledge Map for P -roduct Development based on CSEP Method ,2009 International Conference on Computer Network and Multimedia Technology，2009, 12 : 1081-1085.

[4] HAO Zhangang,YANG Jianhua, Building Knowledge M -ap for Product Development based on GAKME Method. The Second International Workshop on Education Technology and Computer 2010,3:696-699.

[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 // Pm -ceedalgs of the 17th Annual Conference on 1\eural hlfamatiou Pmcessalg Svstmls(NIPS).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 lt -eration[A]. In proceedings of the 12th ACM international Conference on Multimedia[C]. New York ACM Publisher 2004 364- 367.

[12] Zhu Ming, Data Ming,HeFei:China Science and Tec -hnology University Press, 2002,129-164.

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

[14] 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.

[15] 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.

[16] 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 at Shandong Institute of Business and Technology in YanTai of Shandong province.