第45卷　第3期
2009年5月

南京大学学报(自然科学)
JOURNAL OF NANJING UNIVERSITY
(NATURAL SCIENCES)

Vol. 45, No. 3
May, 2009

Design and simulation of a document clustering algorithm based on genetic algorithm*

Wei Jian-Xiang[1,2**], *Liu Huai*[3], *Su Xin-Ning*[1]

(1. Department of Information Management, Nanjing University, 210096, China)
(2. Department of Information Science, Nanjing College for Population Programme Management, 210042, China)
(3. School of Electrical and Electronic Engineering, Nanjing Normal University, 210042, China)

Abstract: Among various document algorithms, K-means is a classical one. However it is a greedy algorithm, which is sensitive to the choice of cluster center and is much easier to result in local optimization. As genetic algorithm (GA) is a global convergence algorithm and the best cluster center can be found easily, a new dynamic document clustering method based on GA is presented in this paper. Reviewing all kinds of traditional document clustering methods, the partial similarity of keywords was not taken into account, so the document similar matrix is a sparse matrix. To some extent, the accuracy of document similarity is influenced. In this paper, some new formulas are given which are improved based on the traditional method. The formulas take the partial similarity of keywords into account, thus improving the accuracy of the calculation of similarity. In this algorithm, the single individual is presented by a matrix which consists of K cluster centers. All individuals are encoded by floating-point numbers. The reciprocal of the sum of mean square deviation of intra-class distance plus one is adopted as the fitness function. The smaller the fitness function, the littler probability that the individual can be selected to enter the next generation. The optimal cluster center is finally found by the following iteration process: selection, crossover, mutation and so on. The simulation results show that the accuracy of this classification can reach over 98 percent and the algorithm is superior to K-means in performance. Thus, the algorithm of this paper is an effective method of document clustering.

Key words: document clustering, genetic algorithm, similarity, cluster center

基于遗传算法的文档聚类算法的设计与仿真

魏建香[1,2],刘　怀[3],苏新宁[1]

(1.南京大学信息管理系,南京,210093)
(2.南京人口管理干部学院信息科学系,南京,210042)
(3.南京师范大学电气与自动化工程学院,南京,210042)

摘　要: 在各种聚类算法中,K-means是一种基于划分的经典算法.但是由于K-means方法对于初始中心点的选择非常敏感,有可能导致聚类结果收敛于局部,本文提出了一种基于遗传算法来对类中心点进行全局寻优的文档

* Foundation Item: National Natural Science Foundation of China(10771076)
Received Date: 2008－11－11

** Corresponding Author, E-mail: jxwei@foxmail.com

聚类算法. 在传统相似度计算的方法中，文档相似矩阵为绝大部分元素为 0 的稀疏矩阵，忽略了关键字之间的部分相似性，影响了文档之间的相似度. 为此，本文改变了传统相似度计算的方法，通过关键字之间的部分相似度，设计出更加精确的文档相似度计算公式. 在遗传算法的设计中，将 K 个类中心点组成的矩阵作为初始个体，采用浮点数进行编码；适应度函数采用所有类内距离的均方差之和加 1 的倒数表示，当类内均方差之和越小，则个体的适应度越大，被选择进入下一代的概率也越大. 通过选择、交叉和变异等步骤对聚类的中心点进行反复迭代寻优，最终找到最优的类中心点. 通过实验仿真，K-means 收敛速度快，聚类的平均目标函数大于 genetic algorithm (GA)且正确率明显小于 GA. 本文提出的 GA 算法的分类正确率能达到 98%以上，与传统的 K-means 方法相比，聚类的准确性更高，说明本文提出的算法是一种行之有效的文档聚类方法.

关键词： 文档聚类，遗传算法，相似度，类中心
中图分类号： TP 18

Clustering analysis is an important research field of artificial intelligence and data mining. Its basic idea is to use characters to measure the degree of similar relationship among objects and to achieve automatic classification in the absence of prior knowledge. All the clustering approaches are to construct the fuzzy matrix in accordance with their own attributes, and then on this basis, to determine their classification relations according to the degree of affinity. Similarly, document clustering is to put the document of high similarity together by computing the similarities among documents and certain strategy. Document clustering is very meaningful. Through data mining of document database, we can identify much potential and hidden knowledge, such as the cross-relationship between subjects, the focus of researches and academic growth point. Such knowledge can not only help the researchers to master the subject knowledge map, but also provide the academic researches with decision-making services. Among all the methods of document clustering, the reference[1] presents an approach named dividing classification, which is based on the similarity of documents, but actually the impacts of clustering haven't been discussed effectively. In the reference[2], the author uses the genetic algorithm to optimize the value of K of K-means without knowing the number of category. The reference[3] is a clustering method based upon the similarity of document keywords. The reference[4] brings forward a new algorithm based on the structure of GML files. However, these methods did not take the cluster center into consideration. It is worth mentioning that the kernel issue of clustering is to find the best cluster centers, the choice of which has a direct influence on the clustering. At present, the most widely used method is K-means based on the objective function. However, its objective function exist the local minimum[5] and it is a greedy algorithm, so it's much easier to result in local optimization. What's more, such a method is extremely sensitive to the choice of cluster center. Many scholars made plenty of improvement on K-means[6~10]. For example, P. S. Bradley and others proposed a method of selecting a number of subsets randomly from the data and repeating carrying out K-means to get the initial cluster center, but such a center is probably the suboptimal result.

Genetic algorithm (GA) is a search algorithm developed from the biologically natural selection and evolution mechanism. Because of its abilities of self-adaptation and self-organization, it is widely used to solve some complicated optimization problems and it has nothing to do with the question itself. It is simple, common, robust, general-purpose and suitable for the parallel processing. GA performs operations on the group which is made up of individuals, and then makes it possible for individuals to exchange information. In this way, individuals can be evolved from generation to generation and approach the optimal solution step by step. As mentioned above, many algorithms are sensitive to the initial center, while GA can overcome such a shortage. Many works use it in document clustering[2,11]. So we can use the genetic algorithm to find out the optimal cluster center in the global situation. In this paper, we adopt genetic algorithm to find out the optimal cluster center in the global situation. The main contents of this paper include the improvement on the traditional similarity calculation, the detail design of genetic algorithm, the simulation of the clustering algorithm and comparisons with the traditional K-means method. The main contents of this paper include the improvement on the traditional similarity calculation, the detail design of genetic algorithm, the simulation of the clustering algorithm and comparisons with the traditional K-means method.

1 Correlative definitions

As we all know, keywords can directly reflect the characteristics of documents and the themes of subjects, so the similarity of documents can be computed by keywords. In order to classify the documents, we should give a definition to the matrix of documents' similarity which is computed by Euclidean distance. The definition is constituted on the basis of the following assumptions: the total number of documents is n; the sum of all various keywords is m; keywords collection $W=\{w_1, w_2, \cdots, w_m\}$ and the number of category is k.

Definition 1 document similar matrix ($n\times n$) is defined as:

$$\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & & S_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ S_{n1} & S_{n2} & & S_{nn} \end{bmatrix} \tag{1}$$

Among all documents, the similarity of any two documents i and j is S_{ij}, and the similarities of all the documents make up of document similar matrix. The matrix is a symmetric matrix whose diagonal value is 0. Where $S_{ij}=\sqrt{\sum_{k=1}^{m}(Q_{ik}-Q_{jk})^2}$, Q is the vector of document-keyword matrix in Definition 2. Because there are m keywords in n documents, we must compute the component of every document in m-dimension space, and then construct the document-keyword matrix.

Definition 2 document-keyword matrix ($n\times m$) is defined as:

$$\begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1m} \\ Q_{21} & Q_{22} & & Q_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ Q_{n1} & Q_{n2} & & Q_{nm} \end{bmatrix} \tag{2}$$

$Q_{ij}(i=1,2\cdots,n;j=1,2\cdots,m)$ represents the jth component of the document i in m-dimension space. It needs to be attained through the similarity of keywords, so we have to give a definition of keywords similarity.

Definition 3 supposing there are two keywords w_i and w_j, their characters in length are l_i and l_j, the maximum length of character strings which are continuously the same is l, so the similarity of these two keywords can be defined as follows:

$$T(w_i, w_j)=\begin{cases} \dfrac{l}{l_i+l_j-l} & \text{while } l>=4 \\ 0 & \text{while } l<4 \end{cases} \tag{3}$$

Obviously, $T(w_i, w_j)\in[0,1]$. This formula takes the partial similarity of keywords into account, thus improving the accuracy of the calculation of similarity. For example, there are two Chinese character keywords: "public library" and "digital library". In many methods of clustering documents, the similarity of the two keywords is defined as 0 (That is, they are completely different). To some extent, the accuracy of similarity is influenced. When we use Equ 3, the result is 0.428 6, which is more accurate than that of the traditional method.

In Definition 2, Q_{ij} can be computed by similarities of all keywords in documents i and w_j. As there commonly are 3～5 keywords in the document i, we define Q_{ij} as the maximum value of similarity of all the keywords in document i and keyword w_j.

Definition 4 Component of document-keyword matrix is defined as follows:

$$Q_{ij} = \max(\{T(w_i, w_j) \mid w_i \text{ is any keyword in the document.}\}) \quad (4)$$

$Q_{ij} \in [0,1]$. $T(w_i, w_j)$ is calculated by Equ 3. In the traditional clustering method, the value of Q_{ij} is usually 0 or 1. In other words, the document-keyword matrix is a sparse matrix. While in Definition 4, the situation has been changed. The value of Q_{ij} is no longer always 0 or 1, but ranges from 0 to 1. Equ 4 makes more accurate similarity.

Definition 5 The aim of clustering is "The intra-class distance is as short as possible, while the inter-class distance is on the contrary". To achieve this aim, we usually adopt three ways to design objective function. One is to use the sum of mean square deviation of intra-class distance. Another one is to use the sum of mean square deviation of inter-class distance. The third one is the combination of the above two ways. In this paper, we choose the first way. The clustering objective function is defined as:

$$E = \sum_{j=1}^{k} \sum_{x_i \in c_j} (x_i - x_j^*)^2 / n_j \quad (5)$$

Where $x_j{}^*$ represents the center of cluster c_j, n_j is the amount of documents in cluster c_j. It represents the sum of mean square deviation of intra-class distance and indicates the discrete degree among all objects. It is clear that the smaller the value of E is, the better the clustering effect is.

2 Design of clustering algorithm

Genetic algorithm is a way to search the optimization in high-dimension space based on an idea of natural selection and evolution which has the capacity of global optimization. On the assumption that the number of category is k, this paper uses GA to search the best cluster center. Steps of this algorithm are given as follows:

(1) Encoding: adopting floating-point code. Individual is represented by the matrix $A = (a_1, a_2, \cdots, a_k)^T \subset R^{k\times m}$ which consists of k cluster centers. Each component a_i presents cluster center and every element of a_i is encoded by floating-point number.

(2) Group initialization: supposing the amount of initialized group is M, matrix collection $X = (A_1, A_2, \cdots, A_n)$ represents the group collection. Elements of every matrix are composed of k × m random real numbers in the range from 0 to 1.

(3) Design of fitness function: firstly we take every component of each individual as the cluster center, and then we will compute the similarities among all documents and cluster centers according to Equs 1－4. Then, according to the minimum distance principle, we gather the documents to the most similar category. Thus, k cluster is formed. Finally, we compute the sum of mean square deviation of all intra-class distance by Equ 5. We can de-

fine individual fitness function as:

$$f = \frac{1}{1+E} \tag{6}$$

From the formula, we can see that the smaller the value of E is, the better the individual fitness is.

(4) Selection: it means to pick up some fine individuals from the current group and determine which individual can enter the next generation. We use a strategy that is the combination of choiceness and sorting. At first, we size down the individuals in terms of fitness function and the former h individuals enter the next generation directly. Then we compute the fitness of the remaining individuals in sequential order by the following formula [12]:

$$P(C) = \left[b + (a-b)\frac{(M - \mathrm{Rank}(C))}{M-h-1} \right] / (M - h) \tag{7}$$

Where M is the size of groups, Rank(C) is the serial number after the sorting of individual, and Rank$(C) \in \{h+1, h+2, \cdots, M\}$, $a+b=2$ and $a \in \{1,5,2\}$. By the roulette strategy, we choose $M-h$ individuals, cross and mutate them, and then generate $M-h$ new individuals. Consequently, it is easier for us to retain the optimal individuals and adjust the worst ones, thus enhancing individual's ability of fitness and guaranteeing a certain selection pressure.

(5) Crossover: choose two individuals randomly, cross them and generate a symmetrical and random number r between 0 and 1. If $r < p_c$, perform the cross, and then generate the new individuals A' and B' by the following formula:

$$\begin{aligned} A' &= rA + (1-r)B \\ B' &= rB + (1-r)A \end{aligned} \tag{8}$$

(6) Mutation: generate a random number r between 0 and 1, if $r < p_m$, perform the mutation. We adopt non-symmetrical mutation algorithm[13]. For an individual A, if a_i is chosen to be mutated, the corresponding component of a_i is changed as follows:

$$a_{ij}{}' = \begin{cases} a_{ij} + \Delta(t, a_{ij}^{\max} - a_{ij}) \ \mathrm{rand}(0,1) = 0 \\ a_{ij} - \Delta(t, a_{ij} - a_{ij}^{\min}) \ \mathrm{rand}(0,1) = 1 \end{cases}$$

$$j = 1,\ 2 \cdots m \tag{9}$$

Where, $\Delta(t,y) = yr\left(1 - \frac{t}{T}\right)^b$, $a_{ij}^{\max}$ and $a_{ij}^{\min}$ are the maximum and the minimum of elements in the row vector. T is the maximum iteration times and t is the current one. Usually $b=2$, and it determines the non-symmetrical system parameter. $(t, y) \in [0, y]$, so the probability that (t, y) is equal to 0 approximately increases with the growth of T. Such a characteristic enables the algorithm to search the global situation equably at the beginning and become convergence in the local.

(7) Termination: if the error of the fitness function of individuals between the new generation and the previous one is less than the given error parameter ε or the iteration times has reached the maximum T, the algorithm will end, or else turns to the step (4).

3 Simulation of clustering algorithm

In order to test the feasibility and effectiveness of the algorithm, 600 documents are chosen from CSSCI (Chinese social science citation index) of year 2005 as the test samples, which include Philology 190, Intelligence Science 140 and Library Science 270. In genetic algorithm, we set the size of group $n=30$; the probability of crossover $p_c = 0.9$; the probability of mutation $p_m = 0.15$, the maximum iteration times $T = 100$; error parameter $\varepsilon = 0.0001$. After running K-means and GA ten times, we make a comparison with each other from the following three aspects.

(1) Accuracy of classification

Table 1 Accuracy rates of classifications

Method	Philology	Intelligence science	Library science
K-means	94. 7%	92. 9%	96. 3%
GA	98. 3%	99. 2%	98. 7%

(2) Clustering objective function

Average value of objective function is an important factor to evaluate whether a clustering method is superior or not. According to Equ 5, the comparative results are as follows:

Table 2 Objective function of two algorithms

Method	Objective function	The Iteration time of convergence
K-means	10. 46	15
GA	10. 25	73

(3)Convergence of algorithms

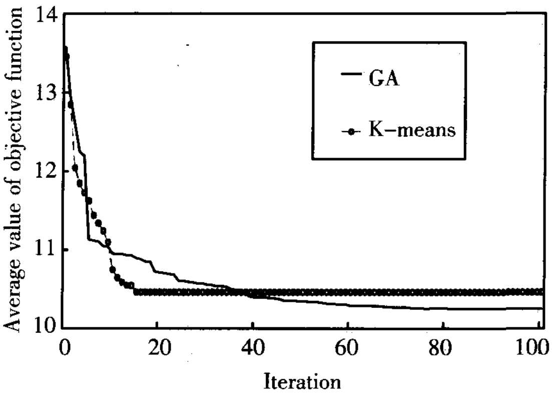

Fig. 1 Comparison of the convergences between GA and *K*-means

From Table 1, we can see that the rate of correctness of the GA can reach over 98%. In Figure 1, the *K*-means algorithm converges quickly but prematurely. As shown in Figure 1, the *K*-means's average value of objective function is sharply reduced from 13. 45 to 10. 46 within 15 iterations and fixed at 10. 46, while GA's average value of objective function is converged from 13. 57 to 10. 25 within 73 iterations. The reduction of the average value of objective function in GA is not as sharp as in *K*-means and becomes smooth after 73 iterations. Table 2 shows that GA can generate the higher clustering compact result than *K*-means.

In the process of clustering simulation, there will be a situation that the individual number of a certain category is 0. That is, the objective function *E* tends to infinity and results in the overflow of the value of *E*. During simulating, we make a correction on the value of *E* by adding 100 to *E* (*E* values from 10 to 20 commonly). In this way, we not only avoid the overflow of data, but also take it into account that after the increase of the value of *E*, it is more impossible for this individual to be chosen to enter the next generation, thus avoiding repeating the bad situation in the new generation.

4 Conclusions

This paper brings forward a document clustering algorithm based on GA to search the best cluster center in the global situation. Through the simulation, we know that GA can generate the better result than *K*-means. On the similarity, many improvements have been made on the basis of the traditional method and the accuracy of computations has been enhanced. But there are many problems which need to be solved, such as: (1) For realistic problems we can easily get hundreds of unique keywords, so each individual is a matrix of several hundreds of real numbers. And it is known that the size of an individual that GA need to evolve satisfactory solutions grows exponentially with the length of the representation. So, we must search a way to reduce the dimension of clustering space so that the algorithm can be applied to large dataset. (2) In Definition 3, we give a constraint that the longest continuous common subsequence is shorter than that in Definition 4, maybe it is a better fit only for Chinese characters rather than for the other languages. We have to do much research on it.

K-means can converge quickly and suit large datasets but its clustering quality depends on the initial seeds. GA can get the global solution but it needs large numbers of iterations and computation. Therefore, we may use GA to find the better seeds for *K*-means at first. Combining GA with *K*-means for document clustering is also the next step in our research.

References

[1] Yang J L. Classification based on the similarity of document set. Journal of the China Society for Scientific and Technical Information, 1999, (S1):87～89. (杨建林. 基于文献集相似度的分类方法. 情报学报, 1999,(S1):87～89).

[2] Casillas A, González de Lena M T, Martínez R. Document clustering into an unknown number of clusters using a genetic algorithm. International Conference on Text Speech and Dialogue, 2003,43～49.

[3] Lin C Y, Zhu D H. Fuzzy clustering for the literature of science and technology. Computer applications, 2004(11): 66～70. (林春燕, 朱东华. 科学文献的模糊聚类算法. 计算机应用, 2004(11):66～70).

[4] Miao J X, Ji G L. Clu-GML: An algorithm for clustering geography markup language documents by structure. Journal of Nanjing University (Natural Sciences), 2008,44(2):188～194. (苗建新, 吉根林. GML 文档结构聚类算法 Clu-GML. 南京大学学报(自然科学), 2008,44(2):188～194).

[5] Selim S Z, Ismail M A. *K*-means-type algorithms: a generalized convergence theorem characterization of local optimality. IEEE Transactions Pattern Analysis and Machine Intelligence, 1984, 6(1):81～87.

[6] Bradley P S, Fayyad U M. Refining initial points for *K*-means clustering. Advance in Knowledge Discovery and Data Mining. Cambridge: MIT Press, 1996.

[7] Raymond T N, Han J W. Efficient and effective clustering methods for spatial data mining. Proceeding of the 20th VLDB Conference Santiago, Chile, 1994, 144～155.

[8] Suo H G, Wang Y W. An improved *k*-means algorithm for document clustering. Journal of Shandong University (Natural Science), 2008, 43 (1):60～64. (索红光, 王玉伟. 一种用于文本聚类的改进 *k*-means 算法. 山东大学学报(理学版), 2008,43(1):60～64).

[9] Shi Z. Efficient online spherical *K*-means iustering. Proceedings of the 2005 IEEE International Joint Conference on Neural Networks. Montreal, IEEE Press, 2005,3180～3185.

[10] Cao F Y, Liang J Y, Jiang G. Initial cluster centers choice algorithm for *K*-means based on neighborhood model. Computer Science, 2008, 35(11):181～184. (曹付元, 梁吉业, 姜 广. 基于邻域模型的 *K*-means 初始聚类中心选择算法. 计算机科学, 2008,35(11):181～184).

[11] Gareth J, Alexander M R, Chawchat S, *et al*. Non-hierarchic document clustering using a genetic algorithm. Information Research, 1995, 1(1).

[12] Zhang T, Zhang H, Wang Z C. Float encoding genetic algorithm and its application. Journal of Harbin Institute of Technology, 2000, 32(4): 59～61. (张 彤, 张 华, 王子才. 浮点数编码的遗传算法及其应用. 哈尔滨工业大学学报, 2000,32(4):59～61).

[13] Wang D G, Liu Y X, Li S. Hybrid genetic algorithm for solving a class of nonlinear programming problems. Journal of Shanghai Jiaotong University, 2003(12):1953～1956. (王登刚, 刘迎曦, 李 守. 求解一类非线性规划问题的混合遗传算法. 上海交通大学学报, 2003(12):1953～1956).