第41卷 第10期 西 安 交 通 大 学 学 报 Vol.41 №10
2007年10月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Oct. 2007

蚁群-遗传融合的文本聚类算法

张云，冯博琴，麻首强，刘连梦
（西安交通大学电子与信息工程学院，710049，西安）

摘要：针对蚁群算法容易出现停滞现象而不能对解空间进行全面搜索的问题，提出了一种蚁群-遗传融合的文本聚类算法.该算法将影响蚁群算法性能的4个参数作为遗传算法中的染色体进行编码，基于此又设计出相应的适应度函数以及选择交叉变异算子，通过多次迭代找出最优的参数组合，并将其应用到文本聚类问题上.经与经典的k均值聚类算法、基本的蚁群聚类算法的仿真比较，结果表明所提出算法的聚类效果更好，在3个测试集上的F度量值要比k均值聚类算法分别提高5.69%、48.60%、69.60%，所以更适合于处理较大规模的数据集.

关键词：蚁群算法；遗传算法；融合；文本聚类
中图分类号：TP391 **文献标识码**：A **文章编号**：0253-987X(2007)10-1146-05

Text Clustering Based on Fusion of Ant Colony and Genetic Algorithms

Zhang Yun, Feng Boqin, Ma Shouqiang, Liu Lianmeng
(School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: Focusing on the problem that the ant colony algorithm gets into stagnation easily and cannot fully search in solution space, a text clustering approach based on fusion of ant colony and genetic algorithm is proposed. The four parameters which influence the performance of ant colony algorithm are encoded as chromosomes, thereby the fitness function, selection, crossover and mutation operator are designed to find the combination of optimal parameters through a number of iterations, and then it is applied to text clustering. The simulation results show that compared with the classical k-means clustering and the basic ant colony clustering algorithm, the proposed algorithm has better performance and the value of F-measure is enhanced by 5.69%, 48.60% and 69.60% respectively in 3 test data sets. Therefore it is more suitable for processing larger datasets.

Keywords: ant colony algorithm; genetic algorithm; fusion; text clustering

在信息过载的今天，文本聚类问题已经成为信息处理领域中的一项重要研究课题[1-3].

蚁群算法(ACA)[4]是一个增强型学习系统，具有分布式的计算特性，也具有很强的鲁棒性[5]，但仍存在一些问题，如它的搜索时间较长，容易出现停滞现象，以及不能对解空间进行全面的搜索，而且在参数选择上缺乏理论的支持等.遗传算法(GA)是一种模拟自然界中生物遗传机制的随机搜索算法，该算法在搜索过程中不易陷入局部最优，即使在所定义的适应度函数非连续、不规则和伴有噪声的情况下，也能以极大的概率找到全局最优解，并易于与其他算法相结合，形成性能更好的问题求解方法.

本文针对蚁群算法存在的问题，结合遗传算法的优越性，提出了一种蚁群-遗传融合聚类算法(AGCA)，利用遗传算法求解组合优化问题的能力来确定蚁群算法中的多个参数的最优组合，并将蚁群-遗传融合算法应用到文本聚类问题之中.

收稿日期：2007-02-01. 作者简介：张云(1982～)，女，博士生；冯博琴(联系人)，男，教授，博士生导师. 基金项目：国家高技术研究发展计划资助项目(2006AA01Z210)；西安交通大学“教育振兴行动计划”资助项目.

1　聚类模型及评价函数

聚类是一个将数据集划分为若干簇的过程,并使得同一个簇内数据对象具有较高的相似度,而不同簇中的数据对象相似度较低.通常,利用对象间的距离可进行相似度度量.聚类模型的形式化描述如下.

给定数据集 $X=\{x_1,x_2,\cdots,x_n\}$,对于 $\forall i\in\{1,2,\cdots,n\}$, $x_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}$ 是 X 的一个对象,对于 $\forall l\in\{1,2,\cdots,p\}$, x_{il} 是 x_i 的一个属性.根据数据的内在特性,可将 X 聚类为 $C=\{C_1,C_2,\cdots,C_k\}$,其中 $\bigcup_{i=1}^{k}C_i=X$, $\forall i,j\in\{1,2,\cdots,k\}$, $C_i\neq\varnothing$, $C_j\neq\varnothing$,且 $C_i\wedge C_j=\varnothing(i\neq j)$. $K=\{X,C\}$ 称为一个聚类空间, C_i 成为聚类空间的第 i 类(簇).

聚类结果评价函数的选取非常重要,因为不同的评价函数将产生不同的聚类效果,在数据集中,常用所有对象与相应聚类中心的均方误差之和(SSE)作为评价函数,即

$$E_{\mathrm{SSE}}=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(d(x_i,c_j))^2 \tag{1}$$

式中: n_j 表示第 j 个聚类中的元素数; $d(x_i,c_j)$ 为数据 x_i 与聚类中心 c_j 的距离.SSE 越小,说明聚类效果越好.

在文本聚类的过程中,也可采用外部评价方法,即 F 度量值(F-Measure)法[6],它组合了信息检索中查准率与查全率的思想,是信息检索领域的一种系统性能测试指标.一个聚类 j 及与此相关的分类 i 的查准率与查全率定义为

$$P=N_{ij}/N_i;\quad R=N_{ij}/N_j \tag{2}$$

式中: N_{ij} 是在聚类 j 中分类 i 的数目; N_i 是分类 i 中所有的元素数; N_j 是聚类 j 中所有的元素数.分类 i 的 F-Measure 定义为

$$F(i)=\frac{2PR}{P+R} \tag{3}$$

则最终聚类结果的评价函数为

$$F_{\mathrm{Measure}}=\frac{\sum_i N_iF(i)}{\sum_i N_i} \tag{4}$$

F_{Measure} 值越高,说明聚类的效果越好.

2　蚁群-遗传融合算法

2.1　基于蚂蚁觅食原理的聚类算法

生物学家经过长期观察发现,蚂蚁在寻找食物的过程中都会在其路径上释放一种称作信息素的化学物质,其他蚂蚁能够感知到这种信息素.大量蚂蚁组成的蚁群集体行为表现出一种信息正反馈现象,蚁群算法就是在蚂蚁群体行为研究的基础上提出来的一种启发式算法.

在基于蚂蚁觅食原理的文本聚类过程中,将待聚类的数据视为具有不同属性的蚂蚁,聚类中心看作蚂蚁所要寻找的食物源,那么数据聚类过程就可以看作是蚂蚁寻找食物源的过程.在每个搜索周期,蚂蚁根据路径(到达聚类中心)上的信息量以及启发信息来计算转移概率,从而决定下一次的转移位置.在第 t 次聚类过程中,数据 x_i 到第 j 个聚类的状态转移概率

$$p_{ij}=\frac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}(t)]^{\beta}}{\sum_{l\in\{1,2,\cdots,k\}}[\tau_{il}(t)]^{\alpha}[\eta_{il}(t)]^{\beta}} \tag{5}$$

式中: α 为信息启发式因子,反映蚂蚁在运动过程中积累的信息量(即残留信息量 $\tau_{ij}(t)$)的相对重要程度; β 为期望启发式因子,反映蚂蚁在运动过程中启发信息(即期望值 $\eta_{ij}(t)$)的相对重要程度; $\tau_{ij}(t)$ 表示在第 t 次聚类中,数据 i 到第 j 个聚类中心的残留信息量,初始时刻各路径上的信息量相等,即 $\tau_{ij}(0)=\tau_0$; $\eta_{ij}(t)$ 为能见度函数,反映蚂蚁在运动过程中的一种先验性、确定性因素, $\eta_{ij}(t)=\frac{1}{d_{ij}}$,其中

$$d_{ij}=d(x_i,x_j)=\Big(\sum_{l=1}^{p}(x_{il}-x_{jl})^2\Big)^{1/2}$$

当所有蚂蚁完成一次聚类之后,聚类中心发生变化,并且每个数据到聚类中心的信息量按照下列规则进行调整,即

$$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\Delta\tau_{ij}(t)$$

$$\Delta\tau_{ij}(t)=\frac{Q}{d(x_i,c_j)} \tag{6}$$

式中: ρ 表示信息素的挥发程度; $1-\rho$ 表示信息素的残留程度; $\Delta\tau_{ij}(t)$ 表示本次循环中数据 i 到聚类 j 的信息素增量; Q 为一常量, Q 越大,蚂蚁在已经走过的路径上的信息素累积加快,在一定程度上影响算法的收敛速度.蚂蚁每一次在不同聚类中心之间的转移都将导致聚类中心发生变化,从而开始下一次的聚类过程,直至聚类结果稳定为止.

2.2　蚁群遗传聚类算法

在基于蚂蚁觅食原理的蚁群聚类算法中,需要初始化大量的参数,这些参数对算法的性能影响极大.如何确定最优参数组合使蚁群算法求解性能最佳,一直是一个极其复杂的组合优化问题,目前尚无完善的理论依据,大多是根据经验而定[4,7-8],文献

[9]给出了利用蚁群算法解决TSP问题时的参数取值范围：α取0～5，β取0～5，ρ取0.1～0.99，Q取10～10 000. 但是，该参数是否适用于其他问题，目前尚没有定论. 本文沿用上述参数取值范围，在文本聚类应用上进行了反复实验，发现聚类结果并不理想.

据此，本文提出了一种蚁群算法与遗传算法相融合的方法，利用遗传算法的组合优化能力，将影响蚁群算法性能的α、β、ρ、Q这4个参数作为遗传算法中的染色体进行编码，各个染色体的取值范围足够大，并对适应度函数、选择交叉变异算子进行设计，通过多次迭代，最终找出最优的参数组合，使得蚁群算法在解决具体问题的时候达到最优性能. 适应度函数的设定应该与所求解问题本身的特性结合起来，本文利用了文本聚类结果的评价函数*F*-Measure作为适应度函数，并根据蚁群聚类结果的好坏来判断当前个体的优劣.

蚁群-遗传聚类算法的伪代码形式化描述如图1所示.

```
步骤1：初始化参数设置，设置染色体的取值范围.
步骤2：生成初始种群.
    For i=1 to PopSize do
        generate population[i].chrom randomly
步骤3：利用蚁群算法求解适应度值.
For i=1 to PopSize do{//For every individual in population
   (α,β,ρ,Q)=population[i].chrom;
    For j=1 to k do//select the initial centers{C1,C2,…,Ck}
   centers[j]=dataset[j];
While(N<max N|| the center is not changed) {
    for every xi in X//for each data xi in dataset X
       for j=1 to k do
         compute Pij by equation(5)
       move xi to Cj, 其中 pij=max{Pij, 0, j≤k}
    for j=1 to k do//updata centers
       updata new centers[j]
     for j=1 to TotalNum do//updata trail
        compute τij(t+1) by equation(6)
}
population[i].fitness=fm, compute the fm by equation(4)
步骤4：遗传操作.
    对当代所有个体根据其适应度值进行排序，根据轮盘
    赌选择法进行新一代个体选择，然后对选择出来的个体
    进行单点交叉以及变异操作，得到新一代的种群.
步骤5：重复步骤3、步骤4，直至迭代次数达到预置值.
```

图1 蚁群-遗传聚类算法的伪代码形式化描述

3 仿真实验

3.1 数据集

选取的实验数据为国际通用的文本标准测试集Reuters-21578，它包含了路透社1987年的新闻，全部文档使用英语. 该测试集经路透社人工汇集和分类，共包含21 578篇文档，135类. 本文采集了多组测试文本集，每一组文本集的数量、主题等均不相同，各个文本集描述见表1. 文本采用最常用的向量空间模型$(w_1, w_2, \cdots, w_n)$，其中w_i为第i个特征项的权重，特征选择采用信息增益法，阈值设为0.5，从而得到的每一个文本集的特征个数分别为145、155、173. 权重计算采用最流行的LTC权重计算方法，聚类性能指标分别用F_{Measure}和SSE来衡量，测试文本集所用算法分别为k均值(k-means)聚类算法，基本蚁群聚类算法(ACA)以及蚁群-遗传融合的聚类算法(AGCA).

3.2 实验结果

3.2.1 聚类效果比较 将聚类个数K设置为数据集的类别数，蚁群聚类算法的参数设置如下：$\alpha=95$，$\beta=45$，$\rho=0.5$，$Q=400$，$\tau_0=20$，最大迭代次数maxN=30. 蚁群-遗传聚类算法参数设置为：种群大小PopSize=30，交叉概率$p_C=0.8$，变异概率$p_M=0.8$，最大代数$\text{maxN}_{\text{AGCA}}=50$. 为了验证各种算法对初始聚类中心的敏感度，可分2种情况选择初始聚类中心：Initc=1，选择较差的初始聚类中心，即选择的初始聚类中心均来自于同一个分类；Initc=2，选择较好的初始聚类中心，即从不同的分类中随机选择一个点作为初始聚类中心. 在3组测试集上的实验结果见表2、表3与图2、图3.

从实验结果中可以看出，AGCA、ACA的聚类效果均优于k-means聚类算法，尤其是AGCA. 当选择较差的初始聚类中心时，AGCA在3个测试集上的F_{Measure}值分别比k-means聚类算法提高5.69%、48.60%、69.60%，同时也说明AGCA比k-means聚类算法更适合处理规模较大的数据集. 另外，从不同的初始聚类中心选择方法得到的聚类结

表1 测试文本集描述

数据集	文本总数	类别数	主题
dataset1	300	4	gnp, gold, jobs, ship
dataset2	560	6	coffee, crude, grain, interest, money-supply, trade
dataset3	1 240	8	coffee, crude, gold, interest, money-supply, ship, sugar, gnp

表2　3种聚类算法在3个测试集上的测试结果

算法	$F_{Measure}$					
	dataset1		dataset2		dataset3	
	Initc1	Initc2	Initc1	Initc2	Initc1	Initc2
k-means	0.772 438	0.963 841	0.433 219	0.879 980	0.427 970	0.857 516
ACA	0.772 438	0.963 841	0.457 711	0.880 174	0.671 905	0.857 479
AGCA	0.816 356	0.963 841	0.643 776	0.906 593	0.725 849	0.865 642

表3　3种聚类算法在3个测试集上的均方误差和

算法	E_{SSE}					
	dataset1		dataset2		dataset3	
	Initc1	Initc2	Initc1	Initc2	Initc1	Initc2
k-means	118.632	105.434	239.560	184.583	436.445	321.072
ACA	118.632	105.434	290.777	184.621	370.970	321.070
AGCA	117.440	105.434	215.216	199.757	389.805	321.104

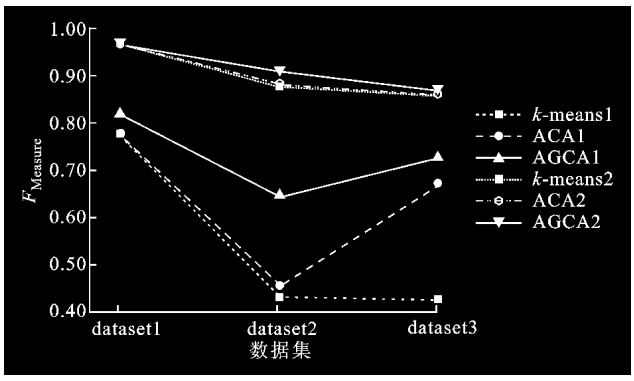

图2　不同聚类算法聚类结果

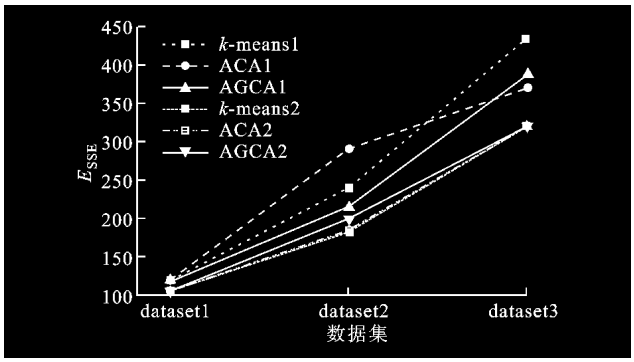

图3　不同聚类算法的均方误差和

果可以看到,ACA同*k*-means聚类算法一样,对初始聚类中心点非常敏感,如果初始聚类中心选择不当,很难得到良好的聚类效果,而融合了遗传算法的蚁群算法大大减轻了因初始聚类中心选择不当所产生的影响.

从图3中还可以看到,虽然*k*-means聚类算法的聚类目标可使得SSE最小,但是AGCA在保证了最优$F_{Measure}$的前提下,大多数情况获得的SSE值均比*k*-means的小,从而进一步说明了本文算法是一种有效的聚类算法.

需要指出的是,在算法效率方面,AGCA需要多次执行基本蚁群聚类过程,其效率要低于ACA、*k*-means算法,然而对比AGCA带来的优越性,例如减轻了对初始聚类中心的敏感度,能自动寻找蚁群算法对多个参数的全局最优组合,以及提高了聚类效果等等,这一缺点是可以接受的.

3.2.2　*不同的参数取值范围对蚁群算法性能的影响*　蚁群算法在解决经典的TSP问题时的经验参数取值范围是[9]:$0\leqslant\alpha\leqslant5$,$0\leqslant\beta\leqslant5$,$0.1\leqslant\rho\leqslant0.99$,$10\leqslant Q\leqslant10\ 000$.本文实验参数取值范围是:$0\leqslant\alpha\leqslant100$,$0\leqslant\beta\leqslant100$,$0\leqslant\rho\leqslant1$,$10\leqslant Q\leqslant1\ 000$.其他参数设置同3.2.1节.利用它们分别在3个测试集上进行测试,遗传算法在50代种群中发现的最优解对应的$F_{Measure}$如表4所示.

通过对比表4中数据可以看出,ACA在解决TSP问题上的经验取值范围并不适用于求解任意问题,如针对文本的聚类问题.与蚁群算法相比,特别是在求解不同的问题时,本文算法能更好地解决各个参数的组合取值问题.

表4 蚁群遗传算法在不同参数下的最优 $F_{Measure}$(Initc=1)

	$F_{Measure}$		
	dataset1	dataset2	dataset3
文献[9]取值范围	0.507 291	0.331 047	0.521 478
本文取值范围	0.816 356	0.643 776	0.725 849

4 结 论

遗传算法作为一种自适应全局优化概率搜索算法,在组合优化问题的求解中取得了良好的效果.蚁群算法作为一类模拟生物群体突现聚集行为的非经典算法,已成为近年来研究的热点.本文提出的蚁群-遗传融合聚类算法,将遗传算法融入到蚁群算法之中,利用遗传算法求解组合优化的能力来确定蚁群算法的各个参数的最优组合,并将其应用到文本聚类问题上,结果取得了较好的聚类效果.今后研究的重点是,进一步提高蚁群-遗传聚类算法的效率,并基于本文方法对蚁群算法的各个参数取值再做试探性的理论研究.

参考文献:

[1] 刘远超,王晓龙,徐志明,等. 文档聚类综述[J]. 中文信息学报,2006,20(3):55-62.
Liu Yuanchao, Wang Xiaolong, Xu Zhiming, et al. A survey of document clustering [J]. Journal of Chinese Information Processing, 2006, 20(3):55-62.

[2] Sasaki M, Shinnou H. Spam detection using text clustering [C]// International Conference on Cyberworlds. Los Alamitos, USA: IEEE Computer Society, 2005: 316-319.

[3] He Feng, Ding Xiaoqing. Combining text clustering and retrieval for corpus adaptation [C/OL]// Proceedings of SPIE. [2007-01-31]. http://spiedigitallibrary.api.org.

[4] Dorigo M, Blum C. Ant colony optimization theory: a survey [J]. Theoretical Computer Science, 2005, 344(2/3):243-278.

[5] Zhu Xingliang, Li Jianzhang. An ant colony system-based optimization scheme of data mining [C]// Proceedings of the 6th International Conference on Intelligent Systems Design and Applications. Los Alamitos, USA: IEEE Computer Society, 2006:400-403.

[6] van Rijsbergen C J. Information retrieval [M]. 2nd ed. London:Butterworths, 1979.

[7] 吴春明,陈治,姜明. 蚁群算法中系统初始化及系统参数的研究[J]. 电子学报,2006,34(8):1530-1533.
Wu Chunming, Chen Zhi, Jiang Ming. The research on initialization of ants system and configuration of parameters for different TSP problems in ant algorithm [J]. Acta Electronica Sinica, 2006, 34(8):1530-1533.

[8] 黄永青,梁昌勇,张祥德. 基于均匀设计的蚁群算法参数设定[J]. 控制与决策,2006,21(1):93-96.
Huang Yongqing, Liang Changyong, Zhang Xiangde. Parameter establishment of an ant system based on uniform design [J]. Control and Decision, 2006, 21(1):93-96.

[9] 段海滨. 蚁群算法原理及其应用[M]. 北京:科学出版社,2005.

(编辑 苗凌)