情报学报 ISSN 1000-0135
第29卷第3期428-432，2010年6月

JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol.29 No.3，428-432 June 2010

doi:10.3772/j.issn.1000-0135.2010.03.007

基于粒子群优化的文档聚类算法[1)]

魏建香[1,2] 孙越泓[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.南京师范大学数学科学学院，南京 210097）

摘要 为了解决文献自动分类问题，提出了一种基于粒子群优化算法（PSO）的文档聚类算法并根据各种参数的变化策略进行了分析与比较。由于粒子运动的范围受到粒子最大速度 V_{max} 的影响，本文通过改变 V_{max} 的变化类型进行仿真比较，当 V_{max} 为凹函数，PSO算法具有较好的收敛性。同时，对惯性权重和学习系数进行了研究，提出了相应的变化策略：惯性权重线性递减，自身认知系数线性递增而社会认知系数线性递减。给出了PSO聚类算法的详细步骤，并根据各种变化策略进行了仿真分析，取得了较好的聚类效果。与标准的遗传算法（GA）相比，本文提出的PSO聚类算法具有更好的收敛效果。

关键词 文档聚类 最大速度 粒子群 遗传算法 参数优化

A Document Clustering Algorithm Using Particle Swarm Optimization

Wei Jianxiang[1,2], Sun Yuehong[3] and Su Xinning[1]

(1. *Department of Information Management, Nanjing University, Nanjing* 210093;
2. *Department of Information Science, Nanjing College for Population Programme Management, Nanjing* 210042;
3. *School of mathematics Science, Nanjing Normal University, Nanjing* 210097)

Abstract In order to implement documents automatic classification, a clustering algorithm based on PSO is proposed. As the scope of the particles is limited by the maximal velocity (V_{max}), this paper puts forward five type functions for V_{max}. Though simulation, the PSO algorithm may achieve better convergence when the type of V_{max} is a concave function. By researching inertia weight and learn coefficient of PSO, we design corresponding change strategies. The steps of the PSO algorithm are given in detail. The experiment result shows that PSO can obtain better performance than GA.

Keywords document clustering, maximal velocity, PSO, GA, parameter optimize

1 引 言

在信息化时代，从文献数据库检索所需的文献成为人们学习和研究的一个重要的途径和手段。如何充分利用日益庞大的文献数据库，为用户提供更好的文献资源服务，是图书情报学科重要的研究课题。利用聚类算法将文献进行自动地分类是解决此类问题的重要方法，聚类是人工智能和数据挖掘研究的重要领域，它的基本思想是利用相似性尺度来衡量事物之间的亲疏程度，在没有先验知识的情况下来实现自动的分类。聚类的方法都是根据研究对

收稿日期：2009年1月9日

作者简介：**魏建香**，男，1971年生，南京大学信息管理系博士生，南京人口管理干部学院信息科学系副教授，从事人工智能、数据挖掘研究。E-mail:jxwei@foxmail.com。**孙越泓**，女，1972年生，博士研究生，南京师范大学数学科学学院副教授，从事人工智能、图像处理研究。**苏新宁**，男，1955年生，南京大学信息管理系教授，博士生导师，从事信息处理与检索、知识管理、引文分析等。

1) 国家社科基金青年自选项目（09CTQ022）；江苏省“六大人才高峰”第六批资助项目（09-E-016）；教育部人文社会科学重点研究基地2008年度重大项目（08JJD870225）。

象本身的属性来构造模糊矩阵，在此基础上根据一定的隶属度来确定其分类关系。同样，文献聚类是通过计算文献之间的相似度，依据一定的策略，将相似度高的文献聚为一类。文献聚类的意义在于通过对文献数据库的挖掘，能发现潜在的、隐性的学科知识，包括学科之间的交叉关系、研究热点和学科增长点，能够帮助研究者掌握学科知识地图，为学术研究提供决策支持服务。

近几年，许多学者提出了文档聚类算法[1~3]，并且取得了较好的聚类效果。聚类的核心问题是确定最优的聚类中心，因此可以通过最优化方法来解决这一问题。模拟退火算法、遗传算法、禁忌搜索、神经网络优化算法、蚁群算法等最优化算法都可以应用于这一领域[4~6]。粒子群优化算法(PSO)作为一种新的基于迭代的最优化方法，是由美国的 Eberhart 和 Kennedy 于 1995 年提出的[7~9]。最初的想法是模拟鸟类觅食的过程来寻求最优解，它一经提出，就受到了业界的广泛关注。目前，PSO 优化算法经过各种改进并应于相当广阔的领域[10~16]，而将 PSO 算法应用于文档聚类的研究并不多，Cui Xiao-hui 等在文献[17]中提出了一种 PSO 和 K-means 的混合文档聚类算法。对于 PSO 算法本身，许多学者都对其中参数的设置进行了深入的研究和探讨[18~20]，主要是对惯性权重 ω、学习因子参数 c_1 和 c_2 的改进较多，但很少有文献对粒子运动的最大速度 V_{max} 进行研究。众所周知，鸟类在觅食空间中的活动范围 V_{max} 的限制，为了能找到食物，鸟类在开始时必须在一个全局的范围内进行搜索，当鸟类越来越接近食物时，它们就应该在一个很小的局部进行搜索。受此启迪，本文提出了一种基于最大速度 V_{max} 变化策略的 PSO 的文档聚类算法，并与传统的遗传聚类算法进行仿真比较，结果表明 PSO 具有更好的聚类性能。

2 算法比较

PSO 是一种基于群体智能理论的全局优化方法，通过群体粒子间的合作与竞争产生的群体智能指导全局搜索。与 PSO 算法类似，遗传算法也是一种基于迭代的优化工具，它起源于对生物系统所进行的计算机模拟研究，由美国密西根大学的 Holland 教授于 20 世纪 60 年代创造出的一种基于生物和进化机制的适合复杂系统优化计算的自适应概率优化算法。与遗传算法相比，粒子群算法具有以下特点：

(1)粒子群算法保留了基于种群的全局搜索策略，采用一种速度-位移的框架模型，操作简单、易于理解，避免了遗传算法中复杂的进化过程；

(2)粒子群算法采用实数编码，直接在问题领域进行求解，而遗传算法一般采用二进制编码，需要对结果进行解码；

(3)遗传算法适用于规模在 50 ~ 200 范围的种群，而粒子群算法对于种群的大小不十分敏感，种群数目下降时，对性能的影响不大，一般取 30 ~ 100 就足够了；

(4)粒子群算法需要自定义的参数比遗传算法少，易于实现，且每个粒子通过自身经验与群体经验进行更新，具有学习的能力；

(5)遗传算法通过基因继承实现信息共享，而粒子群算法则利用群体最优解 *gbest* 供粒子共享，信息是单向流动的。

3 问题分析

3.1 聚类数学模型

在文献聚类问题中，一般都是根据向量空间模型(VSM)思想，将文献聚类的样本空间表示成 $X = \{x_1, x_2, \cdots, x_n\}$，其中样本 $x_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ 为 m 维特征空间 R^m 中的一个点，现在要找到这样一个划分 $C = \{C_1, C_2, \cdots, C_k\}$，使得：

$$X = \bigcup_{i=1}^{k} C_i, C_i \neq \Phi, i = 1,2,\cdots,k, \text{ 且 } C_i \cap C_j = \Phi, j = 1,2,\cdots,k, \text{且 } i \neq j,$$

并且满足类内之和

$$E = \sum_{j=1}^{k} \sum_{x_i \in c_j} \| x_i - x_j^* \| \tag{1}$$

的值最小，x_j^* 表示类 C_j 的中心，x_i 表示划分在类 C_j 中的文献。

3.2 PSO 算法

本文选用 PSO 智能优化算法来确定最优的聚类中心。在 PSO 求解最优化问题时，通常将所求问题的解设计为搜索空间中一个粒子，每个粒子由三部分组成：当前位置 x、飞行速度 v 和粒子的适应度 f 组成，表示为 $P(x, v, f)$。在迭代的过程中，粒子通过更新两个“极值”来更新自己：一个是粒子本身所找到的最优解，我们称之为粒子的自身认知能力，记为 *pbest*；另一个是整个粒子群目前所找到的最优解，我们称之为粒子的社会认知能力，记为 *gbest*。在找到两个最优解以后，每一个粒子通过以下公式更新自己的速度和位置：

$$v_i^{(t+1)} = \omega v_i^{(t)} + c_1 r_1 (pbest_i^{(t)} - x_i^{(t)}) + c_2 r_2 (gbest^{(t)} - x_i^{(t)}) \quad (2)$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \quad (3)$$

其中，$pbest_i^{(t)}$ 为第 i 个粒子到第 t 代为止搜索到的历史最优解，$gbest^{(t)}$ 为整个粒子群到目前为止搜索到的最优解，V_i 是第 i 个粒子当前飞行速度，c_1 为自身认知系数，c_2 为社会认知系数，r_1、r_2 是[0,1]之间的随机数。

根据以上分析，我们用 k 个类中心点组成的集合来表示一个粒子，通过粒子群的不断运动，所求的最优解就是聚类的最优的类中心的集合。

4 算法设计

4.1 个体编码

个体用矩阵 $A = (a_1, a_2, \cdots, a_k)^T \subset R^{k \times m}$ 表示，作为一组 k 个类中心点。矩阵的每个行向量 $a_i = (a_{i1}, a_{i2}, \cdots, a_{im})$ 中的任意元素采用浮点数编码。

4.2 目标函数设计

个体适应度函数定义为：

$$f = \frac{100}{1 + E} \quad (4)$$

E 的定义见公式(1)。由于 E 表示各类别的类内距离，根据“类内距离尽可能小”的聚类目标，当 E 越小，个体的适应度越好。

4.3 算法描述

Step 1：种群初始化：

设种群规模为 M，n 篇文档分为 k 个类别。在 n 篇文档中随机地选取 k 篇文档的特征向量作为一个粒子的初始位置，如此反复 M 次，形成整个初始粒子群。

Step 2：更新速度：

首次运行时必须初始化每个粒子的运动速度，速度被限制在 $(-V_{max}, V_{max})$ 之间；否则，按照公式(2)更新粒子的速度，粒子的最大速度被限定在一定的范围内，如果更新后的速度大于 V_{max}，则取 V_{max}；如果小于 $-V_{max}$，则取 $-V_{max}$。

Step 3：更新位置：

根据公式(3)来更新粒子每一维的位置。由于每一维的位置值被限制在(0,1)之间，对于逃逸出问题空间的粒子的将被重新招回，将其位置赋给一个(0,1)之间的随机数。

Step 4：聚类操作：

由于每一个粒子代表了一组聚类中心点，根据“最近邻原则”将每一篇文献聚集到距离最近的类别中。

Step 5：更新 *pbest* 和 *gbest*：

根据公式(4)计算每个粒子适应度，求出每个粒子的最优解 *pbest*，并计算整个粒子群的最优解 *gbest*。对于每一个粒子，如果它目前的 *pbest* 比历史所经历的 *pbest* 好，则更新 *pbest*，并且粒子返回原来的位置；对于整个粒子群，如果目前的 *gbest* 比历史所经历的 *gbest* 好，则更新 *gbest*；

Step 6：终止操作：

如果已经满足迭代次数或 *gbest* 平均值已经不再发生变化，则算法终止；否则转 Step 2。

5 算法仿真与分析

为了测试 PSO 算法的可行性和有效性，我们从 2005 年 *CSSCI* 文献数据中选择 207 篇文献作为测试样本，包括文献学 70 篇、情报学 51 篇、图书馆学 86 篇。样本数据中，共有关键字 989 个，互异关键字 629 个，通过数据降维处理（如去掉出现频次为 1 的关键字；将含义相同而表述不同关键字用一个关键字代替等操作），剩余关键字为 112 个，降低了运算的数据维度。在实验时，设定公式(2)中的惯性权重为 $\omega = 0.9$ 且 $c_1 = c_2 = 2$，对 PSO 运动的最大速度 V_{max} 分情况进行了仿真，并与遗传算法在收敛性能方面作了比较。

5.1 参数设置及仿真

在高维空间中，粒子的运动轨迹受到最大速度 V_{max} 的影响。在粒子运动的初期，为了能使粒子在一个全局范围内进行搜索，因此 V_{max} 的值应该设置得大一些，当粒子运动一定的代数以后，逐渐靠近最优解时，粒子的速度应该控制在一个较小的范围内，以保证粒子的局部开采能力。由于递减函数可分为线性递减和非线性递减，而非线性递减函数又分为凹函数和凸函数两种。基于以上的分析，我们对粒子运动的最大速度 V_{max} 的值分五种类型进行比较，具体设置见表 1（表格中的 T_{max} 和 T 分别表示最大的迭代次数和当前迭代次数）：

表 1　V_{max}变化策略及变化函数

变化策略	V_{max}的变化函数
V_{max}不变化	$V_{max} = 1$
V_{max}线性变化	$V_{max} = (T_{max} - T)/T_{max}$
V_{max}非线性变化 1	$V_{max} = V_{max} \times \left(\frac{T_{max} - T}{T_{max}}\right)$
V_{max}非线性变化 2	$V_{max} = V_{max} \times (e-1)^{-\frac{T}{10}}$
V_{max}非线性变化 3	$V_{max} = 1 - \left(\frac{T}{100}\right)^2$

五种类型的 V_{max}变化趋势见图 1。

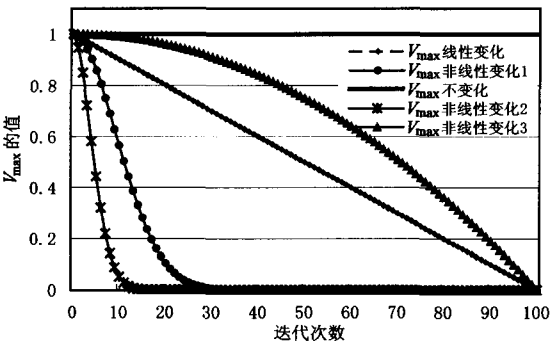

图 1　五种类型的 V_{max}的变化趋势

经过 30 次实验仿真，得到的平均目标收敛情况见图 2。

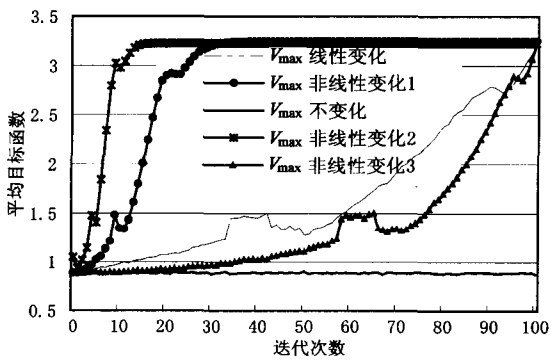

图 2　平均目标函数收敛情况对比

从图 2 可以看出：在粒子的运动过程中 V_{max}保持不变的情况下，算法在 100 代内不收敛；当 V_{max}线性变化时，平均目标函数收敛很慢且比 V_{max}非线性变化 3(为凸函数时)要快一些；当 V_{max}选择非线性变化 1 和非线性变化 2 时，此时 V_{max}变化函数为凹函数，收敛效果非常明显。V_{max}非线性变化 1 在第 30 代开始收敛，而 V_{max}非线性变化 2 在第 20 代就开始收敛。根据 PSO 算法的思想，前期的粒子需要在一个较大的范围内进行探测，因此 V_{max}在粒子运动初期的取值要大一些，而在运动后期，粒子需要一个较小的局部进行开采，以发现最优值，因此在一个较长的时间，V_{max}缓慢地变化到最小值，这种思想正好与凹函数的变化情况一致。因此，选择凹函数为 V_{max}的变化类型可以使 PSO 获得更好的收敛性。

上面的实验都是在惯性权重和学习系数确定的情况下进行的，这些参数通常是按照一定的经验值进行设置的。惯性权重 ω 一般都在[0.1,0.9]之间取值，文献[20]通过大量实验证明，如果 ω 随算法迭代的进行而线性减小，将显著改善算法的收敛性能，这是因为当 ω 从大到小变化时，粒子搜索的范围可以从一个较大的空间逐渐变化到很小的区域，这正好符合 PSO 的基本思想。因此 ω 的变化策略可以用公式 $\omega = \omega_{max} - T * \frac{\omega_{max} - \omega_{min}}{T_{max}}$ 表示。对于学习系数 c_1、c_2 的取值通常为 $c_1 + c_2 = 4$。由于粒子在运动的初期，粒子本身的经验不足因此需要更多地向群体学习，也就是说 c_1 的取值比 c_2 要小；当粒子运动到一定的阶段时，粒子本身积累了一定的经验后，自我学习的能力加强，因此 c_1 的取值比 c_2 要大。学习系数的变化策略为：在 $c_1 + c_2 = 4$ 时，c_1 线性递增而 c_2 线性递减。根据上述惯性权重和学习系数变化策略，我们进行了多次仿真实验，算法取得了明显的改进。

5.2　PSO 与 GA 的比较

为了进一步验证 PSO 算法的性能，我们将 PSO 算法与遗传算法进行了仿真比较。遗传算法的种群规模为 100，交叉概率 $p_c = 0.9$，变异概率 $p_m = 0.15$，最大进化代数 $T_{max} = 100$，误差因子 $\varepsilon = 0.0001$。PSO 算法的种群规模为 100，权重系数 $\omega = 0.9$，$c_1 = 2$，$c_2 = 2$，最大迭代代数 $T_{max} = 100$，误差因子 $\varepsilon = 0.0001$，V_{max}取上述非线性变化 2。仿真结果见图 3。

从图 3 中可以看出，PSO 算法在 20 代就开始收敛，而遗传算法收敛需要 56 代，说明 PSO 算法在收敛速度上比遗传算法快。

6　结　论

本文提出了一种基于 PSO 优化的文档聚类算

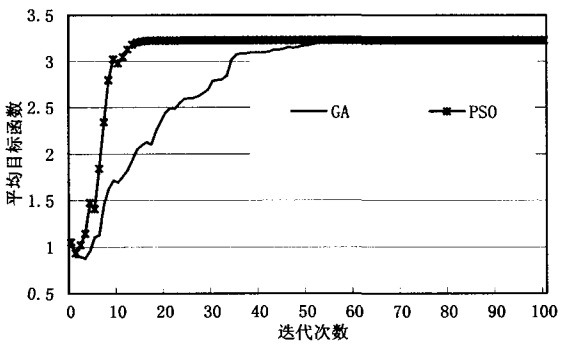

图3 PSO与GA仿真结果比较

法，并根据粒子运动的最大速度 V_{max} 的五种变化策略进行了仿真比较，结论是：当 V_{max} 的变化为凹函数时，PSO算法具有较好的收敛性。对惯性权重和学习系数的设置也进行了分析，提出了相应的变化策略。与遗传算法进行了仿真比较，结果表明PSO具有更好的收敛效果。由于本文提出的基于最大速度变化为凹函数策略的PSO算法具有更快的收敛速度和较好的收敛效果，因此可以更好地实际应用于文档自动分类系统。

参 考 文 献

[1] 宋江春，沈钧毅.一个基于双向近邻技术的多层文档聚类算法[J].情报学报，2006，25(4)：488-492.

[2] 吴景岚，朱文兴.基于K中心点的文档聚类算法[J].兰州大学学报(自然科学版)，2005，41(5)：88-91.

[3] 林春燕，朱东华.科学文献的模糊聚类算法[J].计算机应用，2004，24(11)：66-70.

[4] 白曦，吕晓枫，孙吉贵.融合模拟退火的遗传算法在文档聚类中的应用[J].计算机工程与应用，2006，42(23)：144-148.

[5] 张云，冯博琴，麻首强，等.蚁群-遗传融合的文本聚类算法[J].西安交通大学学报，2007，41(10)：1146-1150.

[6] 雷景生，伍庆清，王平.一种基于混合神经网络的Web文档聚类算法[J].计算机工程，2005，31(1)：12-14.

[7] Kennedy J，Eberhart R C. Particle swarm optimization[C]. In：IEEE International Conference on Neural Networks. Perth，Piscataway，NJ，Australia：IEEE Service Center，1995，Ⅳ：1942-1948.

[8] Kennedy J，Eberhart R C. Swarm Intelligence[M]. San Francisco，California：Morgan Kaufmann Publishers，2001.

[9] Eberhart R C，Kennedy J. A New Optimizer Using Particle Swarm Theory[C]// Proc. Sixth International Symposium on Micro Machine and Human Science，Nagoya，Japan，2001：39-43.

[10] Feng H M，Chen C Y，Ye F. Adaptive Hyper-Fuzzy Partition Particle Swarm Optimization Clustering Algorithm [J]. Cybernetics and Systems：An International Journal，2006，37(5)：463-479.

[11] 赫然，王永吉，王青，等.一种改进的自适应逃逸微粒群算法及实验分析[J].软件学报，2005，16(12)：2036-2044.

[12] 吕振肃，侯志荣.自适应变异的粒子群优化算法[J].电子学报，2004，32(3)：416-420.

[13] 吕艳萍，李绍滋，陈水利，等.自适应扩散混合变异机制微粒群算法[J].软件学报，2007，18(11)：2740-2751.

[14] Omran M，Engelbrecht A P. Particle Swarm Optimization Method for Image Clustering[J]. International Journal of Pattern Recognition and Artificial Intelligence，2005，19(3)：297-321.

[15] Zhang W，Liu Y T. Adaptive Particle Swarm Optimization for Reactive Power and Voltage Control in Power Systems [M]. Lecture Note in Computer Science，Springer Berlin，2005：449-452.

[16] Marco Mussetta，Stefano Selleri，Paola Pirinoli，et al. Improved Particle Swarm Optimization Algorithms for Electromagnetic optimization[J]. Journal of Intelligent and Fuzzy Systems，2008，19(1)：75-84.

[17] Cui X，Potok T E. Document Clustering Analysis Based on Hybrid PSO + K-means Algorithm[J]. Journal of Computer Sciences，vol. Special Issue，2005：185-191.

[18] 王俊伟，汪定伟.粒子群算法中惯性权重的实验与分析[J].系统工程学报，2005，20(2)：194-198.

[19] 胡建秀，曾建潮.微粒群算法中惯性权重的调整策略[J].计算机工程，2007，33(11)：193-195.

[20] 彭宇，彭喜元，刘兆庆.微粒群算法参数效能的统计分析[J].电子学报，2004，32(2)：209-213.

(责任编辑 马 兰)