

基于自适应混沌粒子群的 Web 搜索结果聚类研究

童亚拉

(湖北工业大学 理学院, 湖北 武汉 430068)

摘 要: 提出了基于自适应混沌粒子群的 Web 搜索结果模糊 C-均值算法, 用粒子群算法代替模糊 C-均值算法梯度下降的迭代过程, 同时引入自适应的平衡搜索策略加快算法收敛和提高去噪能力, 在增强局部搜索能力的同时引导粒子群跳出局部极值点. 这样不仅在一定程度上解决了网页文档不确定性的问题, 而且获得快速、稳定的聚类效果.

关键词: Web 搜索结果聚类; 混沌粒子群; 模糊 C-均值算法; 自适应策略

中图分类号: TP301

文献标识码: A

文章编号: 1000-7180(2010)01-0173-04

Research of Web Search Results Clustering Based on Adaptive Chaos Particle Swarm Optimization

TONG Ya-la

(School of Science, Hubei University of Technology, Wuhan 430068, China)

Abstract: A fuzzy C-means clustering algorithm based on adaptive chaotic particle swarm optimization (ACPSO) is proposed in the thesis. On one hand, interactive procedure based on FCM is replaced by that of PSO; on the other hand, a balanced adaptive search strategy is embedded so as to accelerate algorithm convergence, improve the capacity of de-noising, enforce local search ability, and escape from local optimization. It will not only solve indeterminacy of Web document in parts, but also obtain stable clustering results quickly.

Key words: web search results clustering; chaotic PSO; fuzzy C-means algorithm; adaptive strategy

1 引言

在垂直搜索引擎中, 搜索结果排序问题是一个非常突出的问题. Web 挖掘中的聚类技术是一个有效的方法, 它有助于搜索结果文档形成簇, 使搜索结果以类目的形式呈现, 减少信息的总量, 有利于用户找到所需信息, 提供个性化服务^[1].

Web 搜索结果聚类是一个优化问题, 通过迭代运算使系统的目标函数达到一个极小值^[2]. 目前许多聚类分析方法都以启发式算法为基础, 求解效率虽高, 但对初始参数敏感、而且容易陷入局部最优, 难以保证准确性和一致性^[3]. 解决有效途径之一就是引入随机搜索过程, 如遗传算法、禁忌搜索或模拟

退火等. 同样, 粒子群算法也是一种非常有效的随机搜索机制, 而且这种存在多代理结构的算法非常易于并行处理, 收敛速度也很快. 文献[4]使用粒子群算法对自组织特征映射网络(self-organizing feature map, SOM)进行训练代替 SOM 的启发式训练方法, 用于聚类得到较优的识别效果; 文献[5]将混沌粒子群算法与模糊 C-均值算法相结合作聚类分析, 并指导丙烯酸酯反应器参数的调整, 效果显著; 文献[6]采用基于混合粒子群算法 + K 均值算法的文档聚类分析, 避免了 K 均值算法对初值敏感、陷入局部最优的缺陷. 这表明基于粒子群的聚类算法能有效处理例外, 有较好的抗噪声能力. 因此, 文中在对 Web 搜索结果进行聚类分析时, 将经典的聚类方

法如模糊 C-均值聚类(Fuzzy C-mean, FCM)算法与自适应混沌粒子群算法(Adaptive Chaos Particle Swarm Optimization, ACP SO)相结合,既发挥经典聚类算法的优势,也有效利用粒子群算法的寻优能力,使搜索结果的类目形成过程实际上就是不断对聚类结果用粒子群算法优化的过程,利用改进的粒子群算法全局寻优、快速收敛的优点,对聚类准则函数进行优化处理,得到近似最优解,从而对搜索结果进行较准确地划分。

2 模糊聚类与混沌粒子群算法

2.1 模糊聚类

FCM 可用模糊矩阵 $m = (m_{ij})$ 将数据集 X 分为 c 类, m_{ij} 为第 i 个样本属于第 j 类的隶属度,簇中心集合为 $C = (c_1, c_2, \dots, c_K)$, m_{ij} 满足如下条件:

$$m_{ij} \in [0, 1], i = 1, 2, \dots, N; j = 1, 2, \dots, K$$

$$\sum_{j=1}^K m_{ij} = 1, \sum_{i=1}^N m_{ij} > 1$$

模糊聚类在数学上可表示为如下目标函数求极值问题:

$$\min J(m, C) = \sum_{i=1}^N \sum_{j=1}^K m_{ij}^m \|x_i - c_j\|^2 \quad (1)$$

$$\text{s.t. } \sum_{j=1}^K m_{ij} = 1, m_{ij} \in [0, 1] \quad (2)$$

其中, $m > 1$ 为模糊化程度常数,用来控制 m 的模糊程度,目前 m 的选择一般取 $1 < m < 5$ 。

模糊聚类就是求使目标函数最小的隶属度矩阵 $U = [m_{ij}]_{n \times c}$ 及类中心 $C = (c_1, c_2, \dots, c_K)$ 。

$$m_{ij} = \frac{\sum_{k=1}^k \|X_j - c_j\|^{2/(m-1)}}{\|X_j - c_k\|^{2/(m-1)}} \quad (3)$$

$$m_{ij} = \begin{cases} 0 & (\forall i \in H_i) \\ 1 & (\forall i \in I_i) \end{cases} \quad (4)$$

$$c_j = \frac{\sum_{i=1}^N m_{ij}^m x_i}{\sum_{i=1}^N m_{ij}^m} \quad (5)$$

FCM 聚类算法迭代过程如下:

- (1) 随机初始化 $C^{(0)}$, 根据式(3) 计算 $m^{(0)}$;
- (2) 计算聚类中心. 给定 $C^{(k)}$, 根据式(5), 计算 $m^{(k)}$;
- (3) 重新计算隶属度. 给定 $m^{(k)}$, 根据式(3) 和式(4), 计算 $C^{(k)}$;
- (4) 若 $\max |m_{ij}^k - m_{ij}^{k-1}| < e$, 则停止迭代; 否则, 令 $k = k + 1$, 转(2)。

FCM 算法采用梯度下降法对优化目标函数, 因而对初值和噪声数据敏感; 聚类效果受到初始聚类中心影响很大; 由于随机选定初始聚类中心, 结果容易陷入局部最优。

2.2 混沌粒子群算法

鉴于 PSO 算法本身不具备遍历特性, 引入混沌机制, 使每个粒子不仅在个体极值与群体极值的引领下调整自身的位置, 同时还作混沌运动, 以达到尽量遍历整个搜索区域的目的, 从中获取有用信息, 并用以逐代改进个体极值与群体极值, 流程如下:

(1) 初始化相关参数。

(2) 混沌初始化粒子位置和速度。

① 随机产生一个分量值在 $0 \sim 1$ 间的 n 维向量 $x_1 = (x_{11}, x_{12}, \dots, x_{1n})$, 根据 $X_{k+1} = mX_k(1 - X_k)$, 得到 N 个向量 $x_1, x_2, \dots, x_N \sqrt{b^2 - 4ac}$;

② 根据 $cx_i^{(k+1)} = 4cx_i^{(k)}(1.0 - cx_i^{(k)})$, $i = 1, 2, \dots$ 将 x_i 的各分量载波到对应变量的取值区间;

③ 计算 t 时刻粒子群的适应值 p_i^t , 并从 N 个初始群体中选择性能较好的 M 个解作为初始解, 随机产生 M 个初始速度;

(3) 比较当前粒子适应度与个体极值和全局极值, 如果性能优, 则更新。

(4) 更新粒子的速度和位置。

(5) 混沌优化最优位置 $gBest = P_g = (p_{g1}, p_{g2}, \dots, p_{gD})$. 将 $P_{gi}(i = 1, 2, \dots, D)$ 映射到 Logistic 方程 $X_{k+1} = mX_k(1 - X_k)$ 的定义域 $[0, 1]$, $x_i = (p_{gi} - a_i)/(b_i - a_i)$, 然后用 Logistic 方程迭代产生混沌变量序列 $x_i^{(m)}(m = 1, 2, \dots)$, 再把混沌变量序列逆映射 $p_{gi}^{(m)} = a_i + (b_i - a_i)x_i^{(m)}$ 返回到原解空间, 得 $p_g^{(m)} = (p_{g1}^{(m)}, p_{g2}^{(m)}, \dots, p_{gD}^{(m)})$; 在原解空间对混沌变量经历的每一个可行解 $p_g^{(m)}$ 计算其适应值, 得到性能最好的可行解 p^* 。

(6) 用 p^* 取代当前群体中任意一个粒子的位置。

(7) 若满足停止条件, 则搜索停止, 输出全局最优位置, 否则返回(3)。

3 自适应混沌粒子群模糊 C-均值聚类

3.1 算法分析

由于 Web 文档存在多样性和大量性的特点, 一个给定网页可能属于多个类, 聚类时应既考虑类与类之间排斥的特征, 又要考虑其间的联系. 目前大多数 Web 聚类方法很少考虑类与类间存在一定交叉

概念.因为FCM算法是一种聚类不定归属的方法,用隶属度函数表示每个网页属于类别的程度,能合理解决文本多样性的问题.FCM算法是基于梯度下降的局部搜索算法,如果初始簇中心选在局部最优附近,会陷入局部极小值,而且要想得到好的聚类效果,需要对不同的初始解多次试验,既费时,效率又太低,因此将FCM算法和PSO算法相结合,利用改进的粒子群算法全局寻优、快速收敛的特点,对聚类准则函数进行优化,得到一个近似最优解,从而对搜索结果进行较准确地划分.

3.2 编码规则

由于被优化问题的解是寻找最优聚类中心,所以采用基于聚类中心的编码方式,即每个粒子的位置是由 k 个聚类中心组成, k 个聚类中心顺序排列构成一个粒子,粒子的长度是 $D = k * s$, s 是聚类中心的维数,一个粒子代表文档集的一个可能解,每次迭代时粒子在矢量空间根据自己的经验和所有粒子的经验调整聚类中心的位置.

3.3 适应度函数的确定

由于模糊C-均值算法是基于数值数据的,可以计算聚类中心,并定义目标函数,通过不断修改聚类中心最终使目标函数稳定.一个粒子代表一种聚类中心的选取,适应度的高低表明在这种聚类中心选取下聚类效果的优劣,所以对每个个体的评价即为对样本空间的评价,据此,对粒子群算法全局极值和个体极值的选取只需根据适应度函数到每个聚类中心距离即可.

$$J(m, C) = \sum_{i=1}^N \sum_{j=1}^K m_{ij}^m \|x_i - c_j\|^2 \quad (6)$$

$$f(x_i) = \frac{1}{J(m, C)} + 1 \quad (7)$$

式中, $\|x_i - c_j\| = d_{ij}$ 表示文档 j 和聚类中心 C_i 间的欧氏距离.中止条件为满足最大代数或 $J(m, A)$ 小于规定的 e 或模糊矩阵不再改变.

3.4 自适应策略与类标记

混沌优化的引入提高了算法寻优能力,同时也加重了算法的运算负担,基于模拟退火机制的选择策略,可以在运算量增加不多的情况下有效提高算法寻优能力.运行初期粒子群依靠自身跳出局部极值点,模拟退火概率较高.随着迭代次数增多,粒子群的多样性逐渐降低,粒子群跳出局部极值点的能力越来越差,发生早熟收敛的可能性越来越高,此时模拟退火概率应较小,从而克服早熟收敛现象.因此文中设计的模拟退火概率为

$$P_r = \begin{cases} 1 & s_1^2 > s_2^2 \\ \exp[-\frac{s_2^2 - s_1^2}{T_k}] & s_1^2 \leq s_2^2 \end{cases} \quad (8)$$

式中, s_1^2, s_2^2 为当代与前代粒子群的多样性度量,文中选用粒子群群体适应度方差 s^2 作为多样性测度^[7], s^2 越小,粒子群趋于收敛; T_k 为退火温度, $T_k = T_{k-1} \lambda$, λ 是小于1的一个常量,文中 $\lambda = 0.93$.

聚类算法中类标记是很重要的,文中采用文献[8]中的算法标记类别.聚类完毕后,理论上得到 K 个聚类原形 $(C_1, C_2, \dots, C_i, \dots, C_k)$,及各样本对各聚类原形的隶属度 m_{ij} ,根据 m_{ij} 和模式集 P 中的文本归到 K 个类中.

3.5 算法流程

如上所述,基于ACPSO的Web搜索结果模糊聚类算法流程如下:

输入:文本集合, s 个Web文档的特征向量

输出:合理的特征向量族、孤立点族和分类文档

Step 1:利用向量空间模型将待聚类Web文档表示成模式集 P .

Step 2:初始化.给定粒子群规模 n ,迭代次数 $t = 0$,最大迭代次数 n_{max} ,起止温度 T, T_0 和退火速度 λ ,模糊指数 m 的值,混沌扰动区域半径 r ,学习因子 C_1, C_2 ,惯性权重 w_{max}, w_{min} 随机选择 c 个文档矢量作为聚类中心矢量 $C^{(0)} = (c_1, c_2, \dots, c_k)$,并组成一个粒子;初始化该粒子的速度,反复执行该步直到粒子的数量符合要求 n ,共 $n \times c$ 个粒子;计算每个粒子的 $pBest$ 和 $gBest$.

Step 3:FCM算法.根据式(3)、(5)计算新的聚类中心和隶属度 m_{ij} ,组成相应粒子.

Step 4:ACPSO算法.

① 按式(7)计算每个粒子的适应度值,更新 $pBest, gBest, s_1^2, s_2^2$ 和 T_k .

② 如果 $gBest$ 未进化,则计算模拟退火概率 P_r ,若 $P_r > \text{rand}[0, 1]$,则转Step 7.

③ 混沌扰动.

Step 5:计算粒子的隶属度矩阵及聚类中心.

Step 6:判断是否达到最大次数,如达到,停止迭代,输出最佳解,否则转Step 3.

Step 7:以Step 6的结果,执行标类算法.

Step 8:以Step 7的结果,执行文本归类法.

Step 9:处理没有归入的文档到other类.

4 仿真实验

为了算法验证有效性,文中通过常用搜索引擎

采集不同主题的数据集进行实验。

4.1 数据集

在 Yahoo 上根据不同的主题进行 40 次搜索, 下载每次搜索到的前 20 个文档构成 40 个文档类, 每个类包含 800 个文档; 用人工方法剔除无关文档, 分成 40 个类, 以此作为分类准确性的基准; 然后分别采用文中算法、K-均值算法和 STC 算法对此文档集聚类。因为数据是随机采集的真实 Web 文档, 所以具有一定的复杂性: 文档大小差别悬殊, 预处理后文档的字节数在 152B~32KB 不等; 文档内容复杂, 许多文档包含多个主题。

4.2 实验结果及分析

采用查准率和查全率来度量文本检索质量。如预期一样, 文中方法的准确率最高。由于文中的算法允许类间重叠, 平均聚类精度高于 K-均值算法, 而 STC 算法充分利用了文本的特性, 将短语运用在聚类划分和聚类描述中, 除聚类划分较好外, 对聚类描述也比较符合实际。K-均值算法可用来处理各种数据集的聚类问题, 计算复杂度较低、速度快, 但算法受输入参数影响很大, 性能不稳定, 结果不是很理想。三种聚类方法的准确率比较如图 1 所示。

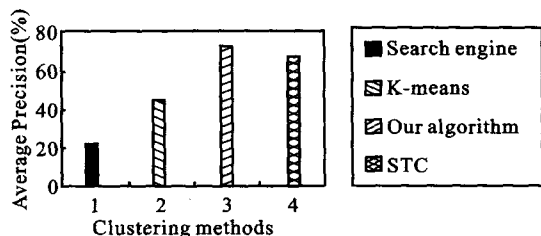


图1 三种聚类方法准确率

用 Yahoo 在 Internet 上搜索前面 40 个主题的相关文档, 下载每个主题的前 10 个文档, 并以 5 的增幅逐步递增到前 45 个文档, 形成文档数依次为 500、750、1000、1250、1500、1750、2000 和 2250 的 8 个文档集; 然后用文中算法和 K-均值算法、STC 算法对文档集聚类, 计算平均聚类时间。结果表明: 随着文档数的增多, 三种算法的平均执行时间增加, 但是文中算法平均执行时间的增幅较小, 增长趋势较为缓慢, 说明文中算法的效率较高。三种聚类方法的执行效率比较如图 2 所示。

5 结束语

文中将 ACPSO 算法和 FCM 算法相结合, 进行

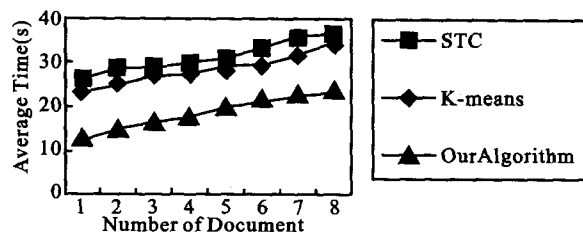


图2 三种聚类方法执行效率

搜索结果聚类分析: 一方面用 FCM 算法解决 Web 文档不确定性问题, 另一方面用 ACPSO 算法优化聚类结果、粒子群算法代替 FCM 基于梯度的迭代过程, 同时引入自适应平衡搜索策略加快算法收敛, 跳出局部最优, 获得快速、稳定的聚类效果, 避免了 FCM 选取聚类个数和聚类中心的随机性。本方法具有简单高效, 具有高查全率和查准率的特点, 实现简单, 便于在垂直搜索引擎中应用。

参考文献:

- [1] 吴远红, 徐妙君, 范银琛. 元搜索引擎中基于 OWA 算子的查询结果合成研究[J]. 微电子学与计算机, 2007, 24, (9): 22-24.
- [2] 徐科, 崔志明, 郑冬冬. 元搜索引擎中基于用户兴趣的查询结果合成研究[J]. 微电子学与计算机, 2004, 23 (7): 200-202.
- [3] 张选平, 马琮, 蒋宇, 等. 一种基于概念抽取的相关词推荐模型[J]. 微电子学与计算机, 2006, 23(5): 163-169.
- [4] 唐贤伦, 仇国庆, 李银国, 等. 基于粒子群优化和 SOM 网络的聚类算法研究[J]. 华中科技大学学报: 自然科学版, 2007, 35(5): 31-33.
- [5] 吕强, 俞金寿. 基于粒子群优化的模糊 c-均值聚类算法——在丙烯酸酯反应器参数优化上的应用[J]. 计算机工程与应用, 2005(22): 211-214.
- [6] Cui X, Potok T E. Document clustering analysis based on hybrid PSO + K-means algorithm[J]. Journal of Computer Sciences, 2005(Special Issue): 27-33.
- [7] Carisle A, Dozier G. Tracking changing extrema with adaptive particle swarm optimizer[C]// World Automation Congress. Spain, Seville, 2002, 13: 265-270.
- [8] 黄钢石, 陆建江, 张亚非. 球型模糊 c 均值算法在中文文本聚类中的应用[J]. 系统仿真学报, 2004, 16(3): 516-518.

作者简介:

童亚拉 女, 博士, 副教授. 研究方向为智能计算、数据挖掘。