基于小生境混合遗传算法的文本特征词聚类研究*)

戴文华[1,2] 何婷婷[1] 焦翠珍[2]

(华中师范大学计算机科学系 武汉 430079)[1] (咸宁学院计算机系 咸宁 437000)[2]

摘 要 提出一种基于小生境混合遗传算法的文本特征词聚类方法。该方法首先采用贝叶斯语义模型对语料库进行统计分析,并以K-L距离度量特征词间的距离,然后将小生境遗传算法与K-Means算法相结合,对文本特征词进行聚类,为文本特征词聚类提供了较高的效率和精确度。实验表明该方法是一种高效可行的文本特征词聚类方法。

关键词 小生境,遗传算法,K-L距离,K-means聚类,特征词聚类

Research of Text Feature Words Clustering Based on Niche Hybrid Genetic Algorithm

DAI Wen-Hua[1,2] HE Ting-Ting[1] JIAO Cui-Zhen[2]

(Department of Computer Science, Central China Normal University, Wuhan 4300790)[1]

(Department of Computer, Xianning College, Xianning 437005)[2]

Abstract Combined with the global optimization ability of Niche Genetic Algorithm and the efficiency of K-Means Algorithm, a new Text Feature Words Clustering method based on Niche Hybrid Genetic Algorithm is proposed. This method first uses the Bayesian Semantics Model to carry out statistical analysis on the corpus. K-L distance is used for measuring distances between feature words. Using this method, we can provide a higher efficiency and precision for Text Feature Words Clustering. Experiments indicate that Niche Hybrid Genetic Algorithm is an effective and feasible method for Text Feature Words Clustering.

Keywords Niche, Genetic algorithm, K-L distance, K-means clustering, Feature words clustering

1 引言

特征词聚类(Feature Words Clustering, FWC)是一种通过聚类的方式对文本特征词集合进行处理,将文本特征词划分为若干簇,并使得同一簇中的特征词具有尽可能大的相似度,而簇间特征词保持尽可能小的相似度。在自然语言处理领域,特征抽取、文本分类、歧义消解和语义关联分析等都要用到特征词聚类。正是由于其重要性,受到了国内外学者的广泛关注。

在特征词聚类的过程中,一般根据特征词之间的互信息进行词间相关度计算,然后根据词间相关度对特征词进行聚类。但是这种相关度计算并没有考虑到训练语料集中的文本分类信息。为了充分利用训练语料中的文本分类信息,提高特征词聚类的精确率,我们提出一种基于小生境混合遗传算法的文本特征词聚类方法。该方法采取贝叶斯语义模型对语料库进行统计分析,并以K-L距离度量特征词间的分布距离,然后将小生境遗传算法与K-means聚类算法相关结合,对特征词进行聚类。

2 贝叶斯语义模型

假设训练语料集包含 N 个文本 $D=\{D_1,D_2,\cdots,D_N\}$,这些文本分属于 M 个文本类别变量 $C=\{C_1,C_2,\cdots,C_M\}$,训练语料集共有 L 个文本特征词 $W=\{W_1,W_2,\cdots,W_L\}$。

根据假设,文本类别 C_j 出现的概率[1]满足公式(1)。

$$P(C_j)=\frac{\sum_{i=1}^{N}P(C_j|D_i)}{N} \tag{1}$$

特征词 W_t 出现在类别 C_j 中的概率满足公式(2)。

$$P(W_t|C_j)=\frac{1+\sum_{i=1}^{N}F(W_t,D_i)P(C_j|D_i)}{L+\sum_{s=1}^{L}\sum_{i=1}^{N}F(W_s,D_i)P(C_j|D_i)} \tag{2}$$

其中 $F(W_t,D_i)$ 表示特征词 W_t 在文本 D_i 中出现的次数,$P(C_j|D_i)$ 为文本 D_i 属于类别 C_j 的概率,当文本 D_i 属于类别 C_j 时,$P(C_j|D_i)=1$,当文本 D_i 不属于类别 C_j 时,$P(C_j|D_i)=0$。

特征词 W_t 出现时文本属于类别 C_j 的概率分布满足公式(3)。

$$P(C_j|W_t)=\frac{P(C_j)P(W_t|C_j)}{\sum_{k=1}^{M}P(W_t|C_k)P(C_k)} \tag{3}$$

如果将两个分布相似的特征词 W_s、W_t 组合成一个新的概念 $W_s\vee W_t$,则 $W_s\vee W_t$ 出现时属于类别 C_j 的概率分布满足公式(4)。

$$P(C_j|W_s\vee W_t)=\frac{P(W_s)}{P(W_s)+P(W_t)}P(C_j|W_s)+\frac{P(w_t)}{P(W_s)+P(W_t)}P(C_j|W_t) \tag{4}$$

3 特征词相似性度量

根据信息论原理,使用K-L距离,可以有效地判断两个

*)本文受咸宁学院科研重点项目(No. KZ0637);国家自然科学基金(No. 60442005, No. 60673040);国家社会科学基金(No. 06BYY029);教育部科学技术研究重点项目(No. 105117)基金资助。**戴文华** 硕士,副教授,主要研究领域为自然语言处理,数据库与数据挖掘;**何婷婷** 博士,教授,博士生导师,主要研究领域为自然语言处理、数据库与数据挖掘。

分布之间的距离。

上述贝叶斯语义模型中的两个概率分布 $P(C_j|W_s)$、$P(C_j|W_t)$之间的 K-L 距离可表示为公式(5)。

$$D(P(C_j|W_s)\parallel P(C_j|W_t))=\sum_{k=1}^{M}P(C_k|W_s)\log\frac{P(C_k|W_s)}{P(C_k|W_t)} \tag{5}$$

如果采用加权平均法计算 K-L 距离的平均值，则 $P(C_j|W_s)$、$P(C_j|W_t)$之间的平均 K-L 距离可表示为公式(6)。

$$\begin{aligned}&AvgDisKL(P(C_j|W_s),P(C_j|W_t))\\&=\frac{P(W_s)}{P(W_s)+P(W_t)}D(P(C_j|W_s)\parallel P(C_j|W_s\vee W_t))\\&+\frac{P(W_t)}{P(W_s)+P(W_t)}D(P(C_j|W_s)\parallel P(C_j|W_s\vee W_t))\end{aligned} \tag{6}$$

平均 K-L 距离 $AvgDisKL(P(C_j|W_s),P(C_j|W_t))$越小，表示两个概率分布 $P(C_j|W_s)$、$P(C_j|W_t)$越逼近，特征词 W_s、W_t 针对类别 C_j 相似度越大。

在特征词聚类中，我们以特征词间的相似性作为聚类划分的依据。特征词 W_s、W_t 间的非相似度 $NONSIM(W_s,W_t)$可用公式(7)表示。

$$NONSIM(W_s,W_t)=\sum_{j=1}^{M}\frac{P(C_j)}{\sum_{k=1}^{M}P(C_k)}AvgDisKL(P(C_j|W_s),P(C_j|W_t)) \tag{7}$$

4 基于小生境混合遗传算法的文本特征词聚类方法

典型的聚类方法有多种，其中 K-means 算法[2]是一种简单、高效的聚类算法。由于 K-means 算法在聚类中心的计算过程中采用了启发式方法，因而有效地降低了算法复杂度，提高了运算速度。也正是因为同样的原因，使得该算法对初始聚类中心的选择较为敏感，易于陷入局部最优解。

遗传算法(Genetic Algorithm, GA)是美国 Michigan 大学的 J. H. Holland 于上世纪 60 年代提出的一种模拟自然界生物进化机制的随机化搜索算法，适用于处理传统搜索方法难于解决的复杂约束优化问题。但是经典遗传算法中进化初期的超常个体会使得种群过早收敛到局部最优解。为了保证种群中个体的多样性，避免算法的早熟，人们将生境机制运用到遗传算法中，形成了小生境遗传算法。通过生境机制，后代中适应度超过其父辈的个体才能替代父辈个体，能有效地控制种群中相似个体的数量。

小生境遗传算法虽然具有较强的全局搜索能力，但是其局部搜索能力较低。为此，我们提出一种基于小生境混合遗传算法的文本特征词聚类方法[3,4]。该方法结合 K-means 算法的高效性和局部搜索能力，以及小生境遗传算法的全局搜索能力，为文本特征词的聚类提供了较高的效率和精确度。

4.1 算法描述

设种群规模为 G，聚类数为 K，最大进化代数为 G_{max}，算法描述如下：

(1)对训练语料集进行特征词抽取，得到文本特征词集 $W=\{W_1,W_2,\cdots,W_L\}$；

(2)利用前文中的公式计算文本特征词的概率分布 $P(C_j|W_s)$；

(3)根据文本特征词的概率分布，产生初始种群；

(4)对种群中所有个体，以其对应的 K 个特征词作为 K-means 算法的 K 个聚类中心；

(5)对种群中所有个体，根据聚类中心，按照最邻近法则将特征词集进行划分；

(6)对种群中所有个体，根据特征词的划分，调整各聚类中心；

(7)以选择概率 P_s 对所有个体进行选择操作；以交叉概率 P_c 对所选个体进行交叉配对；以变异概率 P_m 对个体进行变异操作；

(8)计算子代与父代个体的适应度，如果子代个体适应度大于其父代，则以子代个体替换其父代个体；

(9)判断遗传算法是否达到停止标准，如果达到，则转(10)，否则转(4)；

(10)选择适应度最大的个体，以其对应的 K 个特征词作为 K 个聚类中心，按照最邻近法则将特征词集进行划分，得到最终聚类结果。

使用小生境混合遗传算法进行文本特征词聚类时，必须考虑到在算法的实现过程中，编码方案、种群的初始化、适应度函数、遗传算子和停止标准等都是影响算法效率的非常关键的因素。下面将就这些问题进行讨论。

4.2 编码方案

采用浮点编码，可以克服二进制编码计算量大的缺陷，同时取消了编码和解码的过程，可相对缩短求解时间，在进化时表现出较好的搜索性能。同时，由于基于贝叶斯语义模型的特征词聚类为实数域求解问题，鉴于以上特点，在文本特征词聚类问题的求解中，一般采用浮点编码方案。

在小生境混合遗传聚类算法中，每条染色体由 K 个聚类中心组成，每个聚类中心对应一个特征词在训练语料中的概率分布。由于每个特征词针对 N 个文本类别具有 N 个概率分布，因此每条染色体是长度为 $K*N$ 的浮点码串。

4.3 种群初始化

针对 K-means 聚类算法的特点，经过优化的聚类中心的中点必定在所有样本点中心附近，因此在种群初始化的时候，我们并不采用随机数生成算法(随机产生浮点数作为染色体基因位)，而是通过随机选择特征词的分布概率作为聚类中心染色体基因。这样不仅避免了随机数生成法必须人为确定随机数上下限的缺点，同时由于限定了染色体基因的范围，使得遗传速度更为加快。

在聚类中心确定的情况下，聚类划分采用最邻近法则决定。具体规则如下：

若 W_i，j 满足公式(8)，则 W_i 属于第 j 类。

$$NONSIM(W_i,Z_j)=\min_{k=1,2,\cdots,K}(NONSIM(W_i,Z_k)) \tag{8}$$

其中 W_i 为特征词集中的第 i 个特征词，Z_j 为第 j 个聚类中心对应的特征词，K 为聚类数。

新的聚类中心采用公式(9)进行计算。

$$z_{ij}=P(C_i|Z_j)=\frac{\sum_{k=1}^{N_i}P(W_k)P(C_j|W_k)}{\sum_{k=1}^{N_i}P(W_k)}\quad(i=1,2,\cdots,K\quad j=1,2,\cdots,M) \tag{9}$$

其中 z_{ij} 表示第 i 号中心对应的特征词 Z_i 针对文本类别 C_j 的概率分布，N_i 为特征词类别 FC_i 中特征词个数，W_k 为属于特征词类别 FC_i 的特征词。

4.4 适应度函数

在使用特征词非相似性进行文本间的非相似性度量时，我们将遗传算法的适应度函数定义如下：

$$Fitness(Ind)=\frac{1}{1+\sum_{j=1}^{K}\sum_{i=1}^{N_j}NONSIM(W_i,Z_j)} \tag{10}$$

(下转第 223 页)

实验可以看出，时间是话题的重要特征，其应该在话题检测研究得到更加充分的应用。

参 考 文 献

1 The 2003 Topic Detection and Tracking (TDT2003) Task Definition and Evaluation Plan. Available at: http://www.nist.gov/speech/tests/tdt/tdt2003/evalplan.htm, April 2003

2 Allan J, Carbonell J. Topic Detection and Tracking Pilot Study: Final Report. In: Proceeding of the DARPA Broadcast News Transcriptions and Understanding Workshop, February, 1998. 194～218

3 Yang Yiming, Carbonell J, Brown R, Pierce T, Archibald B T, Liu Xin. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems Special Issue on Applications of Intelligent Information Retrieval, 1999, 14(4): 32～43

4 Yang Yiming, Ault T, Pierce T, Lattimer C. Improving text categorization methods for event tracking. In: Proc. ACM SIGIR, 2000. 65～72

5 Lavrenko V, Allan J, DeGuzman E, LaFlamme D, Pollard V, Thomas S. Relevance models for topic detection and tracking. In: Proc. Human Language Technology Conference (HLT), 2002

6 Allan J, Papka R, Lavrenko V. On-line New Event Detection and Tracking. In: Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Australia, August 1998

7 Allan J, Lavrenko V, Jin H. First story detection in TDT is hard. In: Proc. 9th Conference on Information Knowledge Management CIKM, McClean, VA USA, 2000. 374～381

(上接第 203 页)

公式的含义是：计算各特征词类别 FC_j 中的特征词 W_i 到该类别中心对应的特征词 Z_j 的距离，并求这些距离之和，得到各类的适应度。所有类的适应度之和加 1 并求倒数，得到染色体 Ind 的适应度。

4.5 遗传算子

选择策略对算法性能的影响起到举足轻重的作用。在种群进化过程中，采取精英保留策略，首先保留遗传过程中的精英个体。然后采用轮盘赌法，由适应度函数对应的概率分布确定把当前群体中的第 i 个体 Ind_i 按选择概率 $Ps(Ind_i)$ 抽出，并进行交叉和变异，以提高群体的平均适应度。$Ps(Ind_i)$ 由公式(8)进行计算。

$$Ps(Ind_i)=\frac{Fitness(Ind_i)}{\sum_{j=1}^{G}Fitness(Ind_j)} \tag{11}$$

其中 G 为群体大小，$Fitness(Ind_i)$ 为第 i 号个体的适应度。

遗传交叉算子有多种，可以选择单点交叉和多点交叉，这里我们根据实际情况选择单点交叉。交叉方法为：从当前群体中按轮盘赌法选择两条染色体，随机选取交叉点位置，将两条染色体从交叉点处分为左右两半段，按概率 P_c 依次将两条染色体的右半段互换并重新连接，得到两条新的染色体。

在遗传算法中，变异算子以一个较小的概率 P_m 随机地改变染色体上的某些位串。变异算子使用的几率不大，没有必要对染色体上的每个基因位都考察是否变异，只需随机地考察其中的某些位。变异过程为：随机地确定基因的变异位置，以事先确定的变异概率 P_m 对这些位置的基因进行变异。

与种群初始化方法相似，在进行染色体变异时，我们同样根据 K-means 算法的特点，随机选取特征词的分布概率作为变异点基因值。

4.6 算法停止标准

在实际系统中，我们采取如下停止标准：

(1)固定最大遗传代数 $GNUM$。当算法进行 $GNUM$ 代遗传后停止。最大遗传代数 $GNUM$ 依赖于模型复杂度；

(2)根据遗传收敛的程度，我们给出另一种停止标准，即群体的平均适应度连续多代遗传后仍无明显变化，遗传算法停止。

5 实验结果与分析

针对本文提出的基于混合遗传算法的文本特征词聚类问题，我们进行了两项实验。实验中各参数：种群规模 $G=100$，最大进化代数 $GNUM=100$ 代，交叉概率 $P_c=0.86$，变异概率 $P_m=0.02$，精英个体数 $Elite=4$，聚类数 $K=10$。

实验一(算法稳定性比较)：分别采用 K-means 聚类算法、小生境遗传算法 NGA、小生境混合遗传算法 NHGA(K-means+NGA)对 10 篇测试文档中 635 个特征词进行聚类，每种算法测试 50 次。测试结果见表 1。

从实验结果可知 K-means 算法是一种快速的聚类算法，但算法稳定性较差，这正是 K-means 算法对初始聚类中心依赖性的体现；NGA 算法在聚类稳定性上优于 K-means 算法，但是遗传进化代数较大，运算时间相对较长；NHGA 是一种高效而稳定的遗传算法，这正是由于 NHGA 算法结合 NGA 算法的稳定性和 K-means 算法的高效性所获得的优势。

表 1 算法稳定性比较

	K-means	NGA	NHGA
平均进化代数/迭代次数	28	39	15
获得最优解的次数	36	44	49

实验二(聚类准确率比较)：我们从国家语委现代汉语语料库的 10 种类型文档中各抽取 3000 篇文档作为训练语料，再从训练语料中每类文档抽取 10 篇作为测试语料。测试语料经过特征抽取后得到 2156 个特征词，使用训练语料对这 2156 个特征词进行统计。将这些统计数据作为聚类的原始数据，分别采用上述三种算法进行文本特征词聚类。聚类结果见表 2。

表 2 聚类准确率比较

	K-means	NGA	NHGA
聚类准确率	74%	79%	91%

从实验结果可以看出，该实验与实验一中的结果较为一致。实验表明 K-means 算法聚类准确度较低，NHGA 则是一种准确度较高的聚类算法。

结论 本文通过给出一种将小生境遗传算法和 K-means 算法相结合的小生境混合遗传算法，克服了传统 K-means 算法对初始聚类中心的依赖，充分发挥了小生境遗传算法的全局优化能力和 K-means 算法的快速局部寻优能力，有效地均衡了算法对聚类空间的探索和开发能力，实验证明该算法是一种高效可行的文本特征词聚类方法。

由于 K-means 算法必须人为确定 K 值，本文所提出的 NHGA 算法同样存在这种问题，下一步我们将进一步分析并使用遗传算法对 K 值的选择进行优化处理。

参 考 文 献

1 姜宁，史忠植. 文本聚类中的贝叶斯后验模型选择方法[J]. 计算机研究与发展，2002，39(5)：580～587

2 Clausi D A. K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation [J]. Pattern Recognition, 2002, 35: 1959～1972

3 焦翠珍，戴文华. 基于混合并行遗传算法的多目标约束优化技术研究[J]. 沈阳农业大学学报，2006，37(1)：125～127

4 Glover F, Kelley J, Laguna M. Genetic algorithms and tabu search: a hybrids for optimization [J]. Computers & Operations Research, 1995, 22(1): 111～134