基于混合并行遗传聚类的文本特征抽取方法研究*)

戴文华[1] 焦翠珍[1] 何婷婷[2]

(咸宁学院计算机系 咸宁 437100)[1] (华中师范大学计算机科学系 武汉 430079)[2]

摘 要 同义词和近义词现象以及强关联语义信息加大了文本向量的特征维数,对文本分类的效率和精度都会带来极大影响。为了有效降低文本向量的特征维数,提出一种基于混合并行遗传聚类的文本特征抽取方法。该方法首先使用K-means聚类算法进行特征词粗粒度聚类,然后采用混合并行遗传算法对各类特征词进行细粒度聚类,最后对各聚类中的特征词进行分析并压缩,得到最终能反映文本类别特征和语义信息的文本特征词集合。实验证明,该方法是一种有效的文本特征抽取方法,能切实提高文本分类的效率和精度。

关键词 并行遗传算法,K-means聚类,特征抽取,文本特征词

Research on Text Feature Extraction Based on Hybrid Parallel Genetic Algorithm

DAI Wen-hua[1] JIAO Cui-zhen[1] HE Ting-ting[2]

(Department of Computer, Xanning College, Xianning 437100, China)[1]

(Department of Computer Science, Huazhong Normal University, Wuhan 430079, China)[2]

Abstract Issues of synonymy and strong relational semantic information increase the feature dimension of text vector, which embarrasses the efficiency and precision of text classification. In order to decrease the feature dimension of text vector, a method of text feature extraction based on hybrid parallel genetic clustering algorithm was proposed in this paper. Firstly, K-means algorithm is used to perform thick-granularity clustering for feature words. Successively, hybrid parallel genetic algorithm is used to perform thin-granularity clustering for feature words. Finally, Feature words in each cluster are analyzed and compressed to form feature word set which reflects the feature of text classes and semantic information. The experiments validate our method for text feature extraction is effective.

Keywords Parallel genetic algorithm, K-means clustering, Feature extraction, Text feature words

1 引言

文本分类中的同义词和近义词现象,对文本分类的效率和精确度都会带来较大影响。同义词和近义词的出现会增加文本向量空间的维数,大量消耗文本分类的机器时间。同时,同义词和近义词的出现,造成文本特征词的权值统计误差,给文本分类带来潜在的影响。

如果在文本分类前能设法挖掘出文本特征词间的潜在语义信息,将文本中的同义词和近义词消解,必将给文本分类问题带来很大利益。

除此之外,在文档集合中,有很多这样的词:它们之间具有较强的语义相关性,而且对文本的类别分布具有相同或相似的分布特征。如果我们在文本分类前能发现这些特征词,并将它们进行语义重构,这样也会大大地缩减特征词的维数,减少文本分类的计算量。

传统的特征抽取方法多采用潜在语义索引来发现文本特征词间的语义关联,从而进行同义词和近义词消解以及潜在语义信息的发现。然而,潜在语义索引除了会涉及大规模的矩阵运算外,最终形成的语义信息也是难以理解的。

为了解决潜在语义索引中出现的问题,并有效降低文本向量的特征维数,我们提出一种基于混合并行遗传聚类的文本特征抽取方法。该方法的总体思路是:首先使用K-means聚类算法进行特征词粗粒度聚类,然后采用混合并行遗传算法对各类特征词进行细粒度聚类,最后对各聚类中的特征词进行分析并压缩,得到最终能反映文本类别特征和语义信息的文本特征词集合。

相对于文本聚类来说,文本分类中的特征词抽取过程有了类别信息的帮助。如果能充分利用这些类别信息,必将对特征词抽取带来有效的帮助,给文本分类准确率和召回率都带来极大改善。因此在文本特征词的向量表示中,我们将类别信息加入其中,采用贝叶斯语义模型,形成富含类别信息的特征词向量。

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:

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

特征词 W_t 出现在类别 C_j 中的概率为

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

*)国家自然科学基金(No. 60442005, No. 60673040)、国家社会科学基金(No. 06BYY029)、教育部重点研究项目(No. 105117)、湖北省教育厅科研重点项目(No. D200728002)。**戴文华** 副教授,硕士。

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，则有 $P(C_j|D_i)=1$；当文本 D_i 不属于类别 C_j 时，则 $P(C_j|D_i)=0$。

特征词 W_t 出现时，文本属于类别 C_j 的概率分布为

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

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，文本属于类别 C_j 的概率分布为

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

3 文本特征词的表示及相似性度量

文本特征词按式(5)进行描述：

$$V(W)=\{P(C_1|W),P(C_2|W),\cdots,P(C_M|W)\} \tag{5}$$

根据信息论原理，使用 *K-L* 距离[2]，可以有效地判断两个分布之间的距离。

上述贝叶斯语义模型中的两个特征词 W_s，W_t 相对文本类别的概率分布之间的 *K-L* 距离可表示为

$$D(P(C|W_s)\,\|\,P(C|W_t))=\sum_{k=1}^{M}P(C_k|W_s)\log\frac{P(C_k|W_s)}{P(C_k|W_t)} \tag{6}$$

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

$$AvgDisKL(P(C|W_s),P(C|W_t))=\frac{P(W_s)}{P(W_s)+P(W_t)}D(P(C|W_s)\,\|\,P(C|W_s \vee W_t))+\frac{P(W_t)}{P(W_s)+P(W_t)}D(P(C|W_s)\,\|\,P(C|W_s \vee W_t)) \tag{7}$$

平均 *K-L* 距离 $AvgDisKL(P(C|W_s),P(C|W_t))$ 越小，表示两个概率分布 $P(C|W_s)$ 和 $P(C|W_t)$ 越逼近，特征词 W_s 和 W_t 的相似度越大。

在特征词聚类中，我们以特征词间的相似度作为聚类划分的依据。特征词 W_s，W_t 间的非相似度 $NONSIM(W_s,W_t)$ 可表示为

$$NONSIM(W_s,W_t)=AvgDisKL(P(C|W_s),P(C|W_t)) \tag{8}$$

4 基于混合并行遗传聚类的文本特征抽取(HPGA-ClustFE)方法

特征抽取的主要目的就是要发现文本中的同义词、近义词以及语义相关的特征词，并对它们进行特征重构，从而缩减特征词维数，减少文本分类的计算量。由于同义词、近义词以及语义相关的特征词在文本中具有较强的相似性，我们考虑应用聚类方法对特征词进行潜在语义的挖掘。面对大量的特征词，必须对它们进行精确聚类，才能发现它们之间的语义关系。为了精确有效地对特征词进行聚类，我们将采用一种基于混合并行遗传算法的聚类方法。

4.1 基于混合并行遗传算法的聚类方法

并行遗传算法(Parallel Genetic Algorithm，PGA)[5-7]是一种适用于复杂约束优化问题的多种群并行进化的遗传算法，该算法能有效克服标准遗传算法(Genetic Algorithm，GA)的早熟收敛问题，具有较强的全局搜索能力。K-Means 算法是一种快速高效的聚类方法，具有较强的局部搜索能力，然而该算法对初始聚类中心的选择较为敏感，同时聚类数 *K* 的选择也较为困难。

通过分析上述两种算法，考虑使用并行遗传算法对 K-Means 算法的初始聚类中心和聚类数进行动态优化，从而得到一种高精度聚类方法——基于混合并行遗传算法的聚类(HPGAClust)方法。具体模型如图 1 所示(以两种群并行遗传为例)。

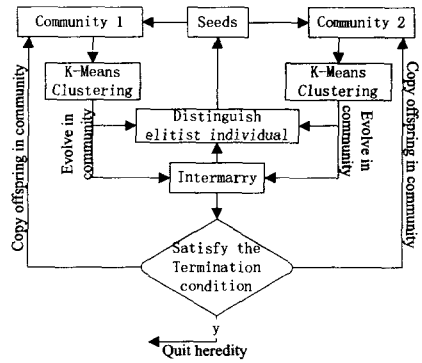

图 1 基于混合并行遗传算法的聚类算法

该算法采用基于“联姻”策略的并行遗传算法模仿人类的联姻策略，尽可能防止具有相同基因结构的个体进行交配，以避免算法的早熟。算法以 $M(M\geq 2)$ 个子种群并行进化。当不同种群之间个体满足联姻条件时，这些个体将两两联姻，并将联姻后代中的精英个体复制到相关的源种群。在遗传过程中，为了保留优良基因，我们还采用精英个体保留策略，将联姻后代和源种群中的精英个体进行比较，保留优者，作为种子参与下一代遗传。

并行遗传算法中所使用的染色体编码方案、适应度函数和遗传算子如下。

(1)可变长染色体编码方案

在聚类问题中，由于聚类中心数难以确定，只能凭经验设置。这种凭经验确定的聚类中心数会对聚类结果产生偏差，因此我们采用并行遗传算法以动态方式来确定聚类中心数，相应的染色体采用可变长染色体编码方案。染色体基因由初始聚类中心对应的特征词在特征词集合中的编号表示。

(2)适应度函数

由于染色体采用可变长编码，因此聚类中心的个数并不固定，其适应度函数定义如下：

$$Fit(Ind)=\frac{1}{1+\sum_{j=1}^{Len(Ind)}\sum_{X_i\in C_j}NONSIM(X_i,Z_j)} \tag{9}$$

其中 $Len(Ind)$ 为个体 *Ind* 的染色体长度。(9)式的含义是：计算各类中的特征词与该类中心的非相似度，并求这些非相似度之和，得到各类的适应度。所有类的适应度之和加上 1 并求倒数，得到染色体 *Ind* 的适应度。

(3)插入删除交叉算子

针对可变长染色体编码，我们特意设计了插入删除交叉算子，以适应遗传进化过程中染色体长度的变化。

插入删除交叉算子的主要思想是：将一个染色体的一段基因删除，并将这段基因插入另一个染色体的某一位置，删除被插入染色体中的重复基因。通过该操作后，被删除基因段的染色体将变短，而被插入基因段的染色体变长。

4.2 特征词的粗聚类

在文本特征抽取过程中，如果直接采用前面所提到的基于混合并行遗传算法的聚类方法来对特征词进行精确聚类，将会花费大量的机器时间。我们考虑：如果首先对这些特征词用快速的 K-Means 算法进行粗略聚类，将它们划分为多个大的类别，然后用基于混合并行遗传算法的聚类方法对这些大类进行精确聚类，这样必定会节省大量的机器时间，而且能得到较好的结果。

特征词的粗聚类步骤可描述如下：

① 按式(3)计算原始特征词集合中的特征词对文本类别的分布概率 $P(C_j|W_i)$，形成特征词的向量表示；

其中 C_j 为第 j 个文本类别，$j=1,2,\cdots,M$，M 为文本类别个数；W_i 为第 i 个特征词，$i=1,2,\cdots,N$，N 为特征词个数。

② 根据经验值，设置特征词聚类数目为 $K=\lceil\sqrt{N}\rceil$；

③ 以特征词作为样本，用 K-Means 算法将特征词集聚为 K 类。

为了保证粗聚类结果的相对精确，可以将粗聚类过程进行多次，取最优的一次作为最终结果。

4.3 特征词的精聚类

特征词的精聚类就是为了使聚类结果更加精确，而采用一些精度较高的聚类算法，对经过粗聚类的各类特征词进行聚类，使得聚类结构变得更为精细。

在本项研究中，我们采用前面所提出的基于混合并行遗传算法的聚类方法对特征词进行精聚类。

4.4 特征重构

特征重构就是找到分布相同或相似的特征词，并将它们合并，形成一个新的能反映文本类别特征和语义信息的文本特征词。显然，采用这种处理方法所得到的语义信息要比潜在语义索引方式得到的语义信息直观。

经过精聚类得到的各个子类中的所有特征词，在特征分布上必定存在很大的相似性。通过这些特征词的分布状况分析，可以对它们进行压缩和整理，形成新的特征词集。具体步骤如下：

① 在各子类中找到那些对一个或多个分类具有相同或相似分布的特征词，将它们合并，形成一个新的特征词。

例如，设有 10 个文本类别，图 2 中的特征词 W_1 和 W_2 对于类别 C_2，C_4 和 C_7 的分布几乎相同，我们可将它们进行合并。而对于特征词 W_3，它对于类别 C_2，C_4 和 C_7 的分布与特征词 W_1 和 W_2 的分布完全不同，因此不能将它与特征词 W_1 和 W_2 合并。

② 在各子类中找到那些对大多数类别都具有相同或相似分布的特征词，将它们从特征词集合中删除。这是由于这些词对文本的分类已经失去了意义，不能给文本分类带来可用信息，它们在特征词集合中的出现只会影响分类的速度。

例如，设有 10 个文本类别，图 3 中的特征词 W_1 和 W_2 对于类别 C_1，C_3，C_4、C_6，C_7 和 C_8 的分布几乎相同，则将它们从特征词集合中删除。

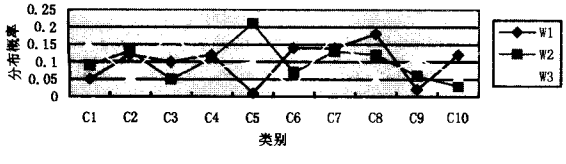

图 2 待合并特征词的分布情况

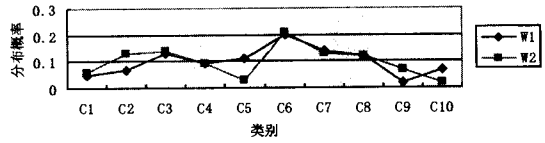

图 3 待删除特征词的分布情况

通过上面的处理，就能删除那些对文本分类毫无贡献的特征词，同时能消除特征词中的同义词和近义词，使新的特征词集合能更有效地反映文本分类的特征。

由多个特征词重构的特征词在进行特征词的权值计算时要注意：新特征词的词频应以原始特征词的词频均值进行统计。

5 实验及结果分析

为了验证本文所提出的基于混合并行遗传聚类的文本特征抽取(HPGAClustFE)方法的实际性能，我们在国家语委现代汉语语料库的 5 类文档集中各随机抽取不重复的 20 篇文本，共 100 篇文本，形成一个文档集。采用同样的方式共形成 5 个文档集。针对上述 5 个文档集，我们进行了两项实验。实验中各参数为：种群大小 Gsize=100，最大进化代数 Gnum=100 代，交叉概率 Pc=0.86，变异概率 Pm=0.02，精英个体数 Elite=4。

(1)实验 1 HPGAClustFE 方法特征抽取效果测试

为了验证 HPGAClustFE 方法的特征抽取性能，首先将 5 个文档集合中的文本经过预处理，得到原始特征词集，然后采用 HPGAClustFE 方法对原始特征词集进行特征抽取。实验结果见表 1。

表 1 基于混合并行遗传聚类的文本特征抽取结果

文档集编号	1	2	3	4	5
原始特征词数	3516	3237	3426	3329	3257
结果特征词数	2539	2321	2528	2420	2332
压缩比	27.8%	28.3%	26.5%	27.3%	28.4%
文本平均相似度	93.8%	92.9%	94.1%	93.5%	92.2%

说明：文本平均相似度等于每篇文本使用原始特征词表示与使用结果特征词表示的相似度之和除以总文本数。

从表 1 可以看出，本文提出的文本特征抽取方法是一种可行的方案，压缩比和精确度都达到了比较满意的程度。

(2)实验 2 特征抽取对文本分类算法性能的影响测试

仍然采用与实验 1 中相同的 5 个文档集分别作为训练集，然后采用相同的采集方式在国家语委现代汉语语料库中提取第 6 个文档集作为测试集。使用 KNN 算法对 5 个训练集进行训练，并用第 6 个文档集进行测试。以宏平均准确率 MP、宏平均召回率 MR 和宏平均 F1 值 MF1 进行对比。MP，MR 和 MF1 的计算公式如下：

$$MP=\frac{1}{n}\sum_{i=1}^{n}P_i \tag{10}$$

$$MR=\frac{1}{n}\sum_{i=1}^{n}R_i \tag{11}$$

$$MF1=\frac{2\times MP\times MR}{MP+MR} \tag{12}$$

其中 P_i 为第 i 类的准确率，R_i 为第 i 类的召回率，n 为训练集分类数。

针对 5 个训练文档，测试文档的分类结果见表 2。

表2 特征抽取对文本分类算法性能的影响情况

测试值 / 文档集编号	MP		MR		MF1	
	特征抽取前	特征抽取后	特征抽取前	特征抽取后	特征抽取前	特征抽取后
1	0.64	0.74	0.63	0.73	0.63	0.73
2	0.67	0.77	0.61	0.72	0.64	0.74
3	0.65	0.71	0.66	0.75	0.65	0.73
4	0.68	0.76	0.65	0.73	0.66	0.74
5	0.65	0.72	0.61	0.77	0.63	0.74

从表2可以看出，经过特征抽取后，文本分类的MP，MR和MF1值被普遍提高了将近8个百分点，说明对特征词的抽取能为分类文本带来较高的精确度，在文本分类中应该引起足够重视。

结束语 本文提出一种基于混合并行遗传聚类的文本特征抽取方法，通过特征词的粗聚类、特征词的精聚类和特征重构，有效地消除了文本中的同义词和近义词现象以及强关联语义信息，降低了文本特征维数，为文本分类精度和效率的提高提供了有效的帮助。实验证明，该方法是一种有效的文本特征抽取方法。

本文主要研究了针对文本分类问题的特征抽取方法，但是对于通过特征词聚类挖掘出来的特征词之间的语义关联没有加以分析。下一步的工作将对特征词聚类挖掘出的语义关系进行分析，并以可视化方式描述特征词间的语义关联。

参考文献

[1] Mühlenbein H. Evolution in time and space-the parallel Genetic Algorithm [M]. Rawlins, Foundations of Genetic Algorithms. Morgan Kaufmann, 1991

[2] Liu Juan, Iba H. Selecting informative genes with parallel GA in tissue classification[J]. Genome Informatics, 2001, 3(12): 14-23

[3] Pettey C B, et al. A Parallel Genetic Algorithm// Proc. of the Second ICGA. 1987: 155-161

[4] Tanese R. Parallel genetic algorithm for a hypercube// Proc. of the second ICGA. 1987: 177-183

[5] Glover F. Future Paths for Integer Programming and Links to Artificial Intelligence [J]. Computers and Operations Research, 1986, 13: 533-549

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

[7] Lee L J. Similarity-based approaches to natural language processing. Ph. D. Thesis. Harvard University, 1997

(上接第174页)

Info(*Leak*)。在统一的描述方法下，通过扩定义新的属性及其关系，可以描述多种安全属性，如：公平属性。因此，本方法具有很强的扩展性和广泛的适用性。总体上来说，本方法优于其他方法，如表2所示。

表2 各种方法比较

方法	准确性	简洁性	可扩展行	适用性
属性统一描述方法	高	较高	高	高
类BAN逻辑	差	高	差	差
CSP方法	高	中	中	中
串空间	高	差	差	中
SPI演算	较高	高	中	中

结束语 安全属性的形式化描述是证明协议安全性的关键问题之一。不同的形式化分析方法对于安全属性有不同的形式化描述方法。然而，它们局限于具体的分析方法和少数的安全属性，不具备普遍性，从而影响了分析方法的使用范围和有效性。本文在迹方法的基础上，提出了一种统一的安全属性形式化描述方法。将安全属性抽象成属性动作及其匹配关系，在协议分析时，通过确定具体的属性动作和匹配关系，可以准确且一致地形式化描述协议的某个安全属性，使得基于迹的分析方法可以有更广泛的适用范围，分析更多类型的安全协议。本文还在这个方法下，具体分析了安全协议的认证、保密和公平性属性的形式化表达。通过比较分析，该方法与其他方法相比，具有准确、简洁和扩展性强的特点，在总体上优于其他方法。

参考文献

[1] Burrows M, Abadi M, Needham R. A logic of authentication. Technical Report 39, Digital Systems Research Center, 1989

[2] Thayer FJ, Herzog JC, Guttman JD. Strand spaces: Proving security protocols correct [J]. Journal of Computer Security, 1999, 7(2/3): 191-230

[3] Lowe G. Breaking and fixing the Needham-Schroeder public-key protocol using FDR. Software-Concepts and Tools, 1996, 17: 93-102

[4] Abadi M, Gordon A D. A calculus for cryptographic protocols: The spi calculus. Information and Computation, 1999, 148(1): 1-70

[5] Boreale M. Symbolic trace analysis of cryptographic protocols// Proceedings of ICAL P01. volume 2076. LNCS 2076. Springer Verlag, 2001: 667-281

[6] Abadi M, Blanchet B. Analyzing security protocols with secrecy types and logic programs. Journal of the ACM, 2005, 52(1): 102-146

[7] Kremer S, Ryan M D. Analysis of an Electronic Voting Protocol in the Applied Pi Calculus// Proceedings of the European Symposium on Programming (ESOP'05), Lecture Notes in Computer Science Series. Springer Verlag, 2005

[8] Dolev, Yao D. On the security of public key protocols. IEEE Transactions on Information Theory, 1983, 29(2): 198-208

[9] Focardi R, Gorrieri R. A Classification of Security Properties. Journal of Computer Security, 1995, 3(1): 5-33

[10] Abadi M. Security protocols and their properties. In Foundations of Secure Computation, volume 175 of NATO Science Series: Computer & Systems Sciences. IOS Press, 2000: 39-60

[11] Kremer S, Markowitch O, Zhou J. An intensive survey of fair non-repudiation protocols. Computer Communications, 2002, 25(17): 1606-1621

[12] 卿斯汉. 电子商务协议中的可信第三方角色[J]. 软件学报, 2003, 14(11), 1936-1943