新颖的无监督特征选择方法

朱颢东[1,2,3]，李红婵[1]，钟 勇[2,3]

(1. 郑州轻工业学院计算机与通信工程学院 郑州 450002; 2. 中国科学院成都计算机应用研究所 成都 610041;
3. 中国科学院研究生院 北京 石景山区 100039)

【摘要】针对有监督特征选择方法因为需要类信息而无法应用于文本聚类的问题，提出了一种新的无监督特征选择方法：结合文档频和K-Means的特征选择方法。该方法首先使用文档频进行无监督特征初选，然后再通过在不同K-Means聚类结果上使用有监督特征选择方法来实现无监督特征选择。实验表明该方法不仅能够成功地选择出最为重要的一小部分特征，而且还能提高聚类质量。

关 键 词 分类; 聚类算法; 文档频; 特征选择; K-Means
中图分类号 TP301.6 **文献标识码** A doi:10.3969/j.issn.1001-0548.2010.03.019

New Unsupervised Feature Selection Method

ZHU Hao-dong[1,2,3], LI Hong-chan[1], and ZHONG Yong[2,3]

(1.School of Computer and Communication Engineering, Zhengzhou University of Light Industry Zhengzhou 450002;
2. Chengdu Institute of Computer Application, Chinese Academy of Sciences Chengdu 610041;
3. The Graduate School of the Chinese Academy of Sciences Shijingshan Beijing 100039)

Abstract Due to unavailability of class label information, supervised feature selection methods can not be applied to text clustering. In this case, a new unsupervised feature selection method combined Document Frequency with K-Means is proposed. The method firstly employs document frequency to select initial unsupervised features, and then brings into unsupervised feature selection by means of mainly performing effective supervised feature selection methods on different K-Means clustering results. Experimental results show that the new method can not only successfully select out the best small part of features, but also can significantly improve clustering performance.

Key words classification; clustering algorithms; document frequency; feature selection; K-Means

有监督特征选择方法在文本分类中得到了广泛的应用，例如IG和CHI两种高效的有监督特征选择方法能移走文本中多达98%的单词而且还能不降低文本分类的性能[1]。文献[1]系统地研究了用于文本聚类的无监督特征选择方法，通过对文档频数、单词权、单词熵等多种无监督特征方法进行对比分析，发现这些无监督特征选择方法在不降低聚类性能的前提下通常只能移走90%左右的单词，如果再移走更多的单词，文本聚类的性能就会急剧下降。因此，用于文本聚类的无监督特征选择仍是一个亟待进一步研究的问题。

为此，本文提出了一个基于文档频和K-Means的无监督特征选择方法，该方法使用分类领域的有监督特征选择方法解决文本聚类领域的无监督特征选择问题。实验结果表明该方法不但能得到使用有监督特征选择所得到的聚类结果，而且还优于任何一种无监督特征选择所得到的聚类结果。

1 几种无监督特征选择算法

在文本聚类中最为常用的几种无监督特征选择方法有文档频、单词权、单词熵和单词贡献度[2]，下面对它们做一下简单介绍，具体请参考文献[2]。

1.1 文档频(DF)

文档频(document frequency, DF)是最易理解的一种无监督特征选择方法。某个词的文档频是在整个文本集中出现该词的文本数。文档频的理论前提是：词在某个类中出现次数过多或过少对问题是无贡献的，删除这些单词对聚类的结果不但没有负面影响，而且还可能会提高聚类结果，尤其是在那些稀单词恰好是噪声单词的情况[1]。

收稿日期：2008 – 09 – 17；修回日期：2009 – 03 – 26
基金项目：四川省科技计划项目(2008GZ0003)；四川省科技攻关项目(07GG006-019)
作者简介：朱颢东(1980 –)，男，博士，主要从事软件过程技术与方法、文本挖掘和智能信息处理方面的研究.

文档频的优点在于它速度十分快，其时间复杂度$O(n)$同文本数成线性关系，因此非常适合于海量文本数据集的特征选择[3]。

1.2 单词权(TS)

文献[4]提出的单词权(term strength, TS)被用于删除对文本检索没有贡献的单词。该方法的主要思想是：单词在相关的文本中出现的频率越高，在不相关的文本中出现的频率越低，该词的重要性越大。

在单词权方法执行过程中，由于要计算所有文本对之间的相似度，因此，该算法的时间复杂度较高，最低为$O(n^2)$。不过，单词权不依赖于类信息，是一种无监督的特征选择算法，所以能用于文本聚类。

1.3 单词熵(EN)

文献[5]提出的单词熵(entropy-based feature ranking，EN)是专门用于聚类问题的特征选择方法。该方法的基本思想是：不同的单词对数据的结构影响不同，单词重要性越大对数据的结构影响也就越大。该方法的缺点在于其时间复杂度太高，为$O(m\times n^2)$，作用于海量数据时性能十分低。

1.4 单词贡献度(TC)

文献[6]提出的单词贡献度，其基本思想是：单词对整个文本数据集相似性的贡献程度越大其重要性也就越大，即有：

$$TC(t)=\sum_{i\neq j} f(t,d_i)\times f(t,d_j) \qquad (1)$$

式中 $f(t,d)$ 表示单词t在文档d中的权重，该权重通常由该单词词频的对数乘以该单词的文档频，并进行归一化处理而获得[7]。文档频的使用增强了文档频低的单词的贡献，同时削弱了文档频较高的单词的贡献。通过进一步分析式(1)发现：单词的文档频越高其被累加的次数也就越多，从而平衡了单词权重和单词文档频之间的矛盾，那些只出现在1个文本中和出现在所有文本中的单词的贡献度都将为0，而那些相对出现较多且具有较高权重的单词的贡献度较大。

2 无监督特征选择方法思想

有监督特征选择在文本分类中得到了较为成功的应用，特别是IG和CHI两种有监督特征选择方法能移走文本中多达98%的单词而且还能略微提高文本分类的性能[8-9]。但是，有监督特征选择却因为依赖类信息而很少地被应用于文本聚类[10]，即使把它们应用于文本聚类，它们也只能在不降低文本聚类性能的前提下最多移走90%左右的单词，并且随着移走更多的单词文本聚类的性能会急剧降低[3]。在这种情况下，就很自然地产生了一个疑问：如果将那些在文本分类中被成功应用的有监督特征选择方法用于文本聚类，是否也能较大程度地提高文本聚类的性能？文献[6]做了一组较为理想的实验。在实验中事先查看了文档的类别信息，然后再使用IG和CHI两种有监督的特征选择方法进行特征选择。实验结果表明该两种有监督特征选择方法不但能够移走高达98%的单词而且还能持续提高文本聚类的性能。

有监督特征选择无法直接应用于文本聚类，因为该类方法需要依赖于文档的类别信息，而文档的类别信息正是文本聚类所要解决的问题。这似乎是个矛盾，但是这种矛盾传却达了一个新的启发信息：能否在聚类的结果上使用现存的有监督特征选择方法，然后再在无监督特征选择的基础上进行重新聚类?本文试图证明该问题。

文献[2]表明：文档频在删除90%单词时，它的性能与IG和CHI的性能相当，效率十分高。为此，本文在初始聚类特征选择时选用该方法。由于在众多聚类算法中，K-Means算法过程简单、易理解，因此本文使用该算法聚类。对于将要使用的有监督特征选择方法，本文选择降维效果较好的IG方法。

根据上面的选择，本文进行了一次实验，其过程为：首先利用文档频数在准备聚类的文档集上进行特征选择，然后再利用K-Means算法进行多次不同的聚类，紧接着再在每一个聚类结果上使用IG进行特征选择并再次进行聚类。该过程可以循环多次，直至达到满意的聚类效果。从实验结果发现再次聚类的结果几乎完全好于原始的聚类结果，而且当原始聚类的结果越好，聚类结果也就越接近理想的分类结果，在其基础上使用有监督特征选择方法选择出来的特征也就越接近于理想情况下使用相同方法所选择出来的特征，进而再次聚类的结果也就越好。

然而，在实际执行的过程中面临很大的问题：(1) 对待聚类的海量文档来说，具体聚类成多少个类别很难设定；(2) 各类别文档在大多数情况下分布是极不均衡的，有的类别文档数可能很大，而有的类别文档数又可能很小，甚至小到可以作为噪声处理。在该情况下，每个聚类结果都具有很大的不确定性，它们可能把一个大类分割成很多小类，也很可能把很多小类聚合成一个大类，而目前的各种方法又很难确定哪一个聚类结果更接近实际的分类。因此在单次聚类结果上进行的特征选择也具有不确定性，

选出的特征集质量高就会较大提高聚类性能，反之就会降低聚类性能。

为了解决该两个问题，文献[11]突破K-Means算法的限制，提出了一种较好的方法，该方法通过合并在不同初始条件下产生的多个不同K-Means聚类结果得到最后的聚类结果，不但能够有效地消除单次K-Means聚类结果的随机性，而且还能够发现任意形状的簇。受该方法的启发，发现K-Means算法所得到的解是一个局部最优解，它能从不同角度刻画数据的分布规律，不同的解之间多多少少都存在一种互补和加强的关系，因此，通过合并在不同K-Means聚类结果上所选择的特征集，可以很好地消除在单次聚类结果上进行特征选择的随机性。因此提出了一种新的用于聚类的无监督特征选择方法——结合文档频和K-Means的特征选择方法。该方法首先使用文档频进行特征初选，然后再通过指定不同的K值和初始点获得不同的K-Means聚类结果，紧接着再在不同聚类结果上使用IG获得特征子集，并把各特征子集合并获得最终的特征集，最后对所得的特征集进行微调以突出那些类区分能力较大的特征，并把调整后的特征子集作为最终的特征集。

之所以在最后加一个特征调整模块，主要是因为不同的词对分类的贡献是不同的，例如名词的分类能力就比助词的强。对某些文档向量进行调整时须突出分类重要词，调整的方法为：按一定比例降低非重要词所占比重。调整方法的特点为：(1) 突出分类重要词和文档本质意义；(2) 调整方法简单易行，只要扫描一遍文档词向量即可。

3 无监督特征选择方法描述

结合文档频和K-Means的无监督特征选择方法如下。

输入：N个待聚类的文档，最小文档频MIN_DF，最大文档频MAX_DF，最大聚类个数MAX_K，最小聚类个数MIN_K，聚类次数M，特征选择方法IG。

输出：一个集合P，该数组集合为最终所选择的特征子集，初始P为空；

步骤1：对整个文档集进行自动分词；

步骤2：根据停用词表去除停用词，称为第一次特征过滤，并统计词频和单词的文档频；

步骤3：移除DF值大于MAX_DF和低于MIN_DF的词，进行第二次特征过滤得到特征集Q；

步骤4：For(j=1;j<=M;j++)

(1) 随机地从[MIN_K,MAX_K]中选择一个数作为k；

(2) 从特征集Q中随机选择k个初始中心点，并依次进行k-Means聚类，得到类集C；

(3) 以类集C为基础，利用特征选择方法IG进行特征选择，得到特征子集H；

(4) P=P+H；

Endfor；

步骤5：对特征集P进行微调，以突出哪些贡献较大的特征词，然后输出调整后的特征集P。

从方法流程上看，该方法相当于一个适用于聚类领域的特征选择系统框架，在步骤的3中可以用其它任何有监督特征选择方法代替IG方法。当然步骤4的(1)和步骤4的(2)也可以用其他聚类方法代替，之所以用K-Means是因为该聚类方法简单易理解。总的来说本文提供的是一种适用于文本聚类领域的特征提取框架。

4 实验验证

实验数据选择复旦大学计算机信息与技术系国际数据库中心自然语言处理小组构建的中文文本分类语料库。分词时采用的是中科院计算所的开源项目“汉语语法分析系统ICTCLAS”系统。之所以选择有类别信息的语料库，是为了聚类后有个对比。分别使用文档频(可视为M=0)和本节提出的方法并使用k-Means算法进行聚类，其中本文方法进行了16次实验仿真，也即参数M分别为0到15。其实验结果如图1所示。

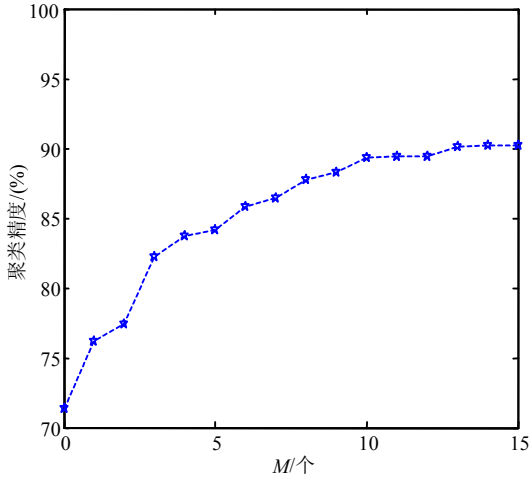

图1 聚类精度随M增加的变化趋势

从图1可以看出随着M的增加聚类精度增加，并且M达到一定次数时聚类精度就几乎不会再增加了，符合实际。因为聚类精度不可能随着M的无限增大而增大，毕竟所采用的聚类算法也有自身的不足。需要说明的是，图1并不是用于说明聚类算法的

精度大小的，而是用于证明本文所提出的方法思想，即聚类算法的精度是随着M的增加而增加，并最终趋于一个平衡状态的。

5 结 束 语

本文把分类领域的有监督特征选择方法用于聚类领域，提出了一种结合文档频和K-Means的无监督特征选择方法。该方法通过在不同聚类结果上使用有监督特征选择的思想，将那些在文本分类中非常有效但无法直接应用于文本聚类的有监督的特征选择方法成功地应用于文本聚类中，极大地降低了文本向量空间的维度，显著地提高了文本聚类的性能。实验表明该方法不仅能够成功地选择出较优秀的特征子集，而且还能显著提高聚类性能，从而在一定程度上解决了文本聚类中的无监督特征选择问题，该方法在文本聚类中有一定的实用价值。

参 考 文 献

[1] YANG Y, PEDERSEN I O. A comparative study on feature selection in text categorization[C]//Proc of International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1997: 412-420.

[2] 刘 涛, 吴功宜, 陈 正. 一种高效的用于文本聚类的无监督特征选择算法[J]. 计算机研究与发展, 2005, 21(3): 381-386.
LIU Tao, WU Gong-yi, CHENG Zheng. An effective unsupervised feature selection method for text clustering[J]. Journal of Computer Research and Development, 2005, 21(3): 381-386.

[3] JIANG Ning, GONG Xu-jun, SHI Zhong-zhi. Text clustering in high-dimension feature space[J]. Journal of Computer Engineer and Application, 2002, 21(2): 63-67.

[4] WILBUR J W, SIROTKIN K. The automatic identification of stop words[J]. Journal of Information Science, 1992, 18(1): 45-55.

[5] MANORANJAN D, LIU Huan. Feature selection for clustering[C]//Proc of Knowledge Discovery and Data Mining. Heidelberg: Springer Berlin, 2000:110-121.

[6] LIU Tao, LIU Sheng-ping. An evaluation on feature selection for text clustering[C]//Proc of International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2003:53-58.

[7] SALTON G. Automatic text processing: the transformation, analysis and retrieval of information by computer[C]//Proc of The ICML89. Pennsylvania: Addison-Wesley, 1989, 32(2): 11-17.

[8] XU Yan. A formal study of feature selection in text categorization[J]. American Journal of Communication and Computer, 2009, 6(4): 32-41.

[9] DESTRERO A, MOSCI S, MOL C D. Feature selection for high-dimensional data[J]. Computational Management Science, 2009, 6(1): 25-40

[10] JENNIFER G D, BRODLEY C E. Feature subset selection and order identification for unsupervised learning[C]//Proc of International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2000: 88-97.

[11] FRED A, JAIN A K. Evidence accumulation clustering based on K-means algorithm[C]//Proc of SSPR/SPR. Heidelberg: Springer Berlin, 2002: 31-37.

编 辑 蒋 晓