一种改进的特征选择方法在文本分类系统中的应用

李长虹, 李堂秋
(厦门大学软件学院, 福建 厦门 361005)

摘要: 在介绍文本分类的背景及传统基于向量空间模型特征选择不足之处的同时, 提出了不同特征选择方法相结合的文本分类模型. 该模型首先对文本进行分析, 把文本表示成向量空间的形式. 文本在经过预处理后, 按一定规则提取关键词, 关键词的提取中增加了对名词短语的识别. 特征选择的方法上, 结合了文档频数和互信息量, 并对他们进行了改进. 实验结果表明, 使用新方法进行分类所得到的分类精度得到了一定的提高.

关键词: 文本分类; 特征选择; 文档频数; 互信息量
中图分类号: TP 302 **文献标识码**: A **文章编号**: 0438-0479(2005)Sup-0239-04

随着网络的迅猛发展, 网络上的信息量迅猛增加. 如何对这些在线文档进行有效的组织和高效的管理, 成为人们迫切需要解决的问题. 文本分类也成为解决该问题的一项关键技术. 文本分类是把一个自然语言文本根据其主题归入到某一预先定义好的某一个分类体系中的一类或几类的过程. 文本自动分类就是使用计算机根据一定的分类规则实现文本的自动归类的过程. 目前, 对于文本分类所采用的技术主要有, Naive Bayes, K-nearest neighbor, support vector machines, boosting 等[1,2]. 本文中, 我们详细描述了一个基于向量空间模型的文本分类系统的研究与设计技术. 对分类的粒度选择, 在基于词的文本分类的基础上增加了对名词短语的识别. 对特征选择的方法, 结合了文档频数和互信息量, 并对他们进行了改进, 这样加强了对低频词和类间分布差异相近的特征的处理, 解决了文本分类中一些特征处理不足之处.

1 系统设计

图 1 为本文所采用的文本分类系统结构图, 主要由训练模块和分类模块两个模块组成. 训练文档和测试文档需要相同的预处理和特征提取方法, 只有这样, 通过训练集学习获得的分类规则才能用于测试文档进行分类. 训练模块对训练文档进行预处理、特征选择和提取、参数训练, 生成分类规则. 分类模块用训练得到的分类规则, 通过分类器对测试文档进行分类.

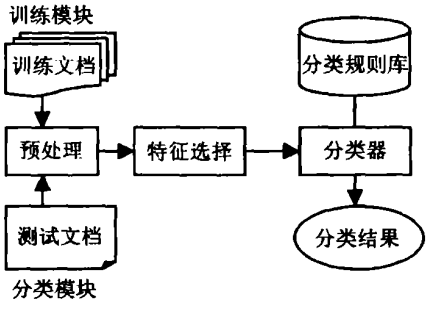

图 1 文本分类系统结构
Fig. 1 Structure of text categorization system

1.1 文本预处理

文本预处理的目的是把文本整理成一些规范的格式, 便于接下来的步骤操作, 主要包括以下的一些部分:

(1) 名词短语的抽取

在文本分类中, 通常将文档中出现的词作为一个单元, 独立的抽取出来, 作为分类的特征. 但是由于单独的词并不总是表达意思的原子单位, 造成了一些分类信息的丢失, 导致了分类精度的降低. 例如需要对 artificial intelligence, machine tools, education 3 个主题进行文本分类, 那么对于词组 machine learning, 如果将 machine 和 learning 作为一个整体从文档中抽取出来, 比 machine 和 learning 分别单独的抽取更能表达文档的主题. 所以在提取单个词作为文档分类特征的基础上, 把名词短语抽取出来加入到系统的特征中, 会在一定程度上改善分类的效果. 我们对名词短语的定义如下: $NP = \{A, N\}^* N$, 例如, 对 the quick

收稿日期: 2005-03-04
作者简介: 李长虹(1976-), 男, 硕士研究生.

brown fox jumps over the lazy dog 这个句子,按照定义,brown fox,quick brown fox 和 lazy dog 都将作为名词短语抽取出来. 把名词词组作为特征的问题是会在系统中潜在的引入大量的冗余特征,并且大量低频名词短语的出现有可能为分类引入噪声. 如上述的 brown fox, quick brown fox 就会为分类引入冗余特征. 通过借助于已知的名词短语词典,可以将这些名词短语过滤掉. 如果在文档中出现 machine learning, machine language, interest rate, monetary interest 等名词短语,同时他们也在名词短语词典中存在,对于这样的名词短语就可以从文档中提取出来,作为表述一个完整意思的单元. 考虑到名词词组和单个词在文本分类中重要性的不同,对名词词组权重的计算需要考虑词组的长度、名词词组中各个词的权重以及在文本中的分布等多个因素.

(2) 词干还原

在系统中,通过使用 stemming 的方法,完成名词复数的去除、动词时态的转换、动词第三人称转换等工作. 即把名词的复数形式恢复为其原来的形式,把动词在各种时态的形式和在第三人称下的形式恢复至其原来的形式.

(3) 禁用词去除

对大量的介词、冠词、连词、代词(如 in、the、and、this、very)等无具体含义的虚词词汇,另外还有一些比较常见的实词,如 take, let 等,这些词在几乎所有的文本中都有很高的出现频率,对文本没有区分作用,还会干扰关键词所起的作用,降低了分类系统的处理效率与准确率,应予滤除.

1.2 特征选择

寻找一种有效的特征选择方法,降低特征空间的维数,提高分类的精度和效率,是文本分类需要面对的重要问题[3,4]. 目前常用的特征选择方法主要有文档频数 DF、互信息 MI[5]、信息增益 IG 和 x^2 统计量[6] 等. 在本系统中,我们结合文档频数和互信息量两种方法实现对特征的选择. 文档频数 DF 是最简单的特征选择方法,其值为训练文档集合中该词出现的文档数. 文档频数针对所出现的低频词特征进行过滤,稀有单词要么不含有有用信息,要么太少,不足以对分类产生影响. 文档频数的不足之处是某些低频词可能在某一类文档中并不稀有,而且包含重要的分类信息. 我们将在文档频数的基础上改进,以过滤掉真正对文档分类不起作用的低频无用信息. 具体步骤如下:

步骤 1 统计词条 W_i 在 $D_k(k=1,2,\cdots,m)$(其中 D_k 为包含词 W_i 的文档)中出现的频数 $\mathrm{freq}^{wi}D_k$(不为零).

步骤 2 计算词 W_i 的 $\max(\mathrm{freq}^{wi}D_k)$ 与 $\min(\mathrm{freq}^{wi}D_k)$ 的比值 T_{W_i}.

步骤3 对于如下的一个三元组$(W_i, P(W_i)T_{W_i}, DF_{W_i})$, $P(W_i)$ 为词条 W_i 在训练文档集中出现的频率, DF_{W_i} 为词条 W_i 的文档频数, 根据 $(1+P(W_i)T_{W_i})DF_{W_i}$ 对 W_i 进行评分,然后对评分值进行排序,这样就可以按照预先设定的阈值,将某一文档数以下的词过滤掉.

词汇和类别的互信息量:

$$\mathrm{MI}(W_i, C_j) = \log\left(\frac{p(W_i \mid C_j)}{p(W_i)}\right),$$

其中的 W_i 是词汇, C_j 是类别, $p(W_i)$ 表示词 W_i 在所有文本中出现的概率, $p(W_i \mid C_j)$ 表示词 W_i 在类别为 C_j 的文本中出现的概率. 互信息量越大,表明词汇和类别之间的共现概率也越大. 互信息量的主要不足之处是没有考虑关键词在文档集合中出现的频率,可能造成了不同频率词互信息量的相近. 例如对于特征 A 和 B, A 特征在所有文档中出现的概率为 0.8, 其在某一类 C_j 中出现的概率为 0.2, B 特征在所有文档中出现的概率为 0.2, 其在 C_j 中出现的概率为 0.05, 那么根据互信息量的计算,特征 A、B 对 C_j 的贡献是一样的,但是实际上很显然,特征 A 和特征 B 相比,对 C_j 是更重要一些,因此,我们对互信息量的计算作如下的改变:

$$\mathrm{MI}(W_i, C_j) = p(W_i \mid C_j)\log\left(\frac{p(W_i \mid C_i)}{p(W_i)}\right),$$

这样,特征 A 和 B 在计算了互信息量后,对 C_j 的重要性就有了差异. 通过互信息量的处理,过滤掉了在各类间概率分布比较均匀,分布差异不大的特征,保留了大量出现在某一类中,而在其他类中出现次数比较少的特征和在有限几个类中出现,而在其他的类中出现的情况相对比较少的特征. 经过文本预处理和特征选择两个阶段的处理后,文档被表示成为一个特征的序列. 由于不同的特征在分类时,对类别的重要性是有差异的,因此需要对关键词进行权重处理. 我们采用最为常用的 TF-IDF 方法[7], 词频 TF 为关键词在文档中出现的次数,逆文本频率 $IDF = \log|D|/DF(W_i)$, $|D|$ 为训练集合中文档总数, $DF(W_i)$ 为出现 W_i 的文档数. TF-IDF 方法提升了低频词在分类中的重要性. 但是同时,也可能提高与分类无关的噪声词的词频差异,导致分类精度的降低. 由于我们在特征选择时,通过文档频数的处理,已经过滤掉了不含有用信息的低频词对分类的影响,所以对低频词权重的加强,并不会导致噪声的增加而淹没有用信息,该权重处理方法是可行的. 对名词词组 $NP_i = \{a_1^{(i)}, a_2^{(i)}, \cdots, a_k^{(i)}\}$, a_1 为 NP_i 的第一个成分, k 为词组的长度,对 TF_{P_i} 乘以一个权值

$k(1+\frac{TF_{d_1^{(i)}}+TF_{d_2^{(i)}}+\cdots+TF_{d_k^{(i)}}}{TF_{a_1}+TF_{a_2}+\cdots+TF_{a_k}})$, TF_{a_1} 为 a_1 在文档中出现的次数, $TF_{a_1^{(i)}}$ 为 a_1 在 NP_i 中出现的次数, 该权值的调整考虑了词组的长度, 以及组成词组的各个词在文档中的分布对词组权值的影响. 对每一个特征对应的向量进行归一化处理, 降低高频词和低频词对分类效果的影响.

1.3 训练算法[8,9]

训练算法是分类系统的核心部分, 目前存在多种基于向量空间模型的训练算法, 例如, 支持向量机算法、神经网络方法、K 个最近邻居方法(K Nearest Neighbor) 和贝叶斯方法等, 本系统中采用 KNN 算法对文本进行分类. 该算法的基本思路是在给定新文本后, 考虑在训练文本集中与该新文本距离最近(最相似) 的 k 篇文本, 根据这 k 篇文本所属的类别判定新文本所属的类别. 具体的算法步骤为:

步骤 1　根据特征项集合重新描述训练文本向量.

步骤 2　在新文本到达后, 根据特征词分词新文本, 确定新文本的向量表示.

步骤 3　在训练文本集中选出与新文本最相似的 k 个文本, 计算公式为:

$$\mathrm{Sim}(d_i,d_j)=\sum_{k=1}^{M}W_{ik}\times W_{jk}\Big/\sqrt{\sum_{k=1}^{M}W_{ik}^2\sum_{k=1}^{M}W_{jk}^2}.$$

步骤 4　在新文本的 k 个邻居中, 依次计算每类的权重, 计算公式如下:

$$p(\vec{x},C_j)=\sum_{\vec{d}_i\in KNN}\mathrm{Sim}(\vec{x},\vec{d}_i)y(\vec{d}_i,C_j),$$

其中, $\vec{x}$ 为新文本的特征向量, $\mathrm{Sim}(\vec{x},\vec{d}_i)$ 为相似度计算公式, 与上一步骤的计算公式相同, 而 $y(\vec{d}_i,C_j)$ 为类别属性函数, 即, 如果 $\vec{d}_i$ 属于类 C_j, 那么函数值为 1, 否则为 0.

步骤 5　比较类的权重, 将文本分到权重最大的那个类别中.

2 实验及分析

实验中, 我们使用 Rainbow 文本分类系统作为实验平台, 系统所带的新闻语料库由 20 000 篇新闻组短文章组成, 预先分为 20 个类别, 每类 1 000 篇文章. 通过扫描文档集合, 总共得到 37 344 个不同的词条. 过滤掉禁用词之后, 剩余 36 820 个不同的词条. 通过名词短语的抽取, 得到 37 245 个不同的词条. 实验过程中, 从每类中选取 900 篇作为训练样本, 另外 100 篇作为测试样本. 分类效果的评价采用系统自带的评价算法: 平均精度. 平均精度的计算是在设定训练样本数以及重复测试若干次实验的基础上计算其平均的正确率.

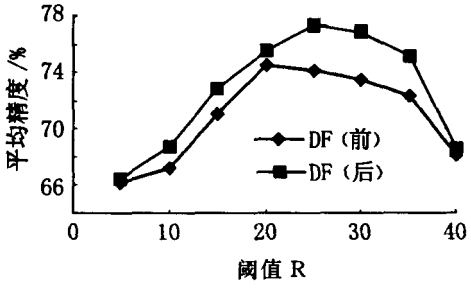

图 2　DF 阈值的选择与平均精度的关系

Fig. 2　Relationship of DF threshold and average precision

从图 2 中, 可以看到, 随着阈值的提高, 经过 DF 处理之后, 分类精度都得到了提高, 没有经过改进的文档频数在阈值为 20 时, 分类精度最高, 达到了74.5%, 而文档频数在经过改进之后, 在阈值为 25 时, 分类精度最高, 为 77.3%. 说明在滤掉对分类不起作用的低频无用信息对提高精度是有帮助的. 但是随着阈值的提高, 分类的精度都出现下降, 在阈值为 40 时, 分类精度都接近 68%. 可能是由于大量的表达类别信息的特征都被过滤掉了, 造成了精度的降低.

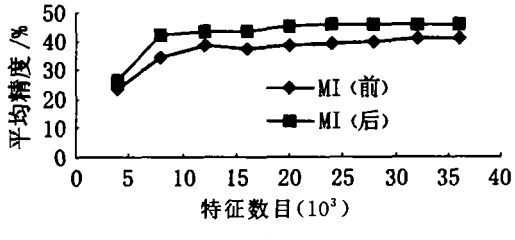

图 3　特征数目与平均精度的关系

Fig. 3　Relationship of feature numbers and average precision

从图 3 中, 可以看到, 特征数在 10 000 之前, 分类精度得到了明显的提高, 在特征数超过了 15 000 之后, 分类精度趋于平稳. 互信息量作为特征选择方法, 分类精度是很低的, 考虑特征在文档中出现的频率之前和之后的互信息量, 分类精度分别为 41%, 45%. 精度提高的原因是, 我们保留了某些高频词, 而降低了对低频词的倚重.

图 4 比较了 3 种特征选择方法对平均精度的影响. 由于在互信息量的计算过程中, 利用到该词条的文档频数, 因此组合的特征抽取方法并不会增加额外的计算量. 从实验数据可以得知, 结合了改进的文档频数

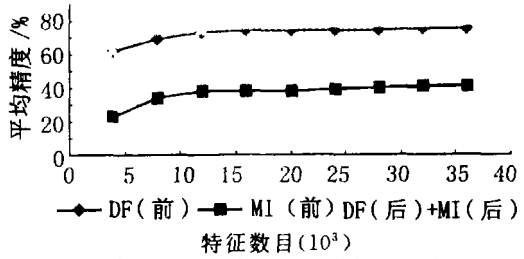

图 4 3 种特征选择方法与平均精度的关系

Fig. 4 Relationship of three feature selection methods and average precision

和互信息量的分类精度明显高于使用互信息量和文档频数的分类精度. 这是由于一方面过滤了可能为分类引入噪音的稀有单词, 另一方面, 去除了特征分布相对均匀的某些单词的结果.

3 结束语

随着对网络信息处理需求的增加, 自动文本分类将会越来越多的受到人们的重视. 本文讨论了文本分类系统实现中的若干关键步骤. 对文本预处理、特征选择、分类方法进行了详细的介绍. 我们针对文档频数和互信息量在特征选择上的不足, 对低频词和类间分布差异不大的特征的处理作出了改进, 提高了系统的分类精度.

在系统实现过程中, 对特征权重的衡量方法还存在一些不足之处, 对名词词组的处理还需改善. 在今后的工作中, 将进一步改善特征选择方法, 并在多种分类算法尝试, 比较其性能.

参考文献:

[1] 王国胜, 钟义信. 支持向量机的若干新进展[J]. 电子学报, 2001, 29(10): 1 397– 1 400.

[2] 杨岳湘, 田艳芳, 王韶红. 基于模糊聚类和 Naive Bayes 方法的文本分类器[J]. 计算机工程与科学, 2002, 24(5): 18 – 21.

[3] 陆玉昌, 鲁明羽, 李凡. 向量空间法中单词权重函数的分析与构造[J]. 计算机研究与发展, 2002, 39(10): 1 205– 1 210.

[4] 陈治平, 林亚平, 彭雅. 基于最小类差异的无关信息预处理算法[J]. 电子学报, 2003, 39(10): 1 750– 1 753.

[5] Kenneth Ward Church, Patric K Hanks. Words association norms, mutual information and lexicography[A]. Proceedings of the 27th Annual Meeting[C]. Vancouver: Association for Computational Linguistics, 1989. 76– 83.

[6] Dunning T E. Accurate methods for the statistics of surprise and coincidence[J]. Computational Linguistics, 1993, 19(1): 61– 74.

[7] Salton G. Introduction to Modern Information Retrieval [M]. New York: Mc Graw Hill Book Company, 1983.

[8] 庞剑锋, 卜东波, 白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001, 18(9): 23– 26.

[9] 周水庚, 关佶红, 胡运发, 等. 一个无需词典支持和切词处理的中文文本分类系统[J]. 计算机研究与发展, 2001, 38(7): 839– 844.

Application of Improved Feature Selection in Text Categorization System

LI Chang-hong, LI Tang-qiu

(School of Software, Xiamen University, Xiamen 361005, China)

Abstract: A new text categorization system using composite feature selection method was designed. Texts are described in the form of Vector Space Model through model analysis. After pretreatment for texts, keywords which include noun phrase were extracted on the basis of rules from texts. On feature selection, document frequency was combined with mutual information, and performance was improved. Experiments show that the precision of text categorization is improved through using new method.

Key words: text categorization; feature selection; document frequency; mutual information