一种基于隐含子类信息的粗糙集中文文本分类方法*)

金凯民　苗夺谦　段其国
(同济大学计算机科学与技术系　上海 200092)

摘　要　中文文本分类是中文信息检索和Web挖掘等领域的研究热点。现有的一些分类方法在特征选择阶段存在不足,忽略了隐含的子类信息。本文提出了一种提升隐含子类的关键词权值的方法,从而可以发现有价值的子类信息,进而使用粗糙集构建分类器。实验结果表明这种方法在不增加待约简词汇数量的情况下有效地提高了文本分类的查全率。

关键词　中文信息处理,粗糙集,文本分类,向量空间模型

A Chinese Text Classification Method Using Implied Sub-class Information and Rough Set

JIN Kai-Min　MIAO Duo-Qian　DUAN Qi-Guo
(Dept. of Computer Science and Technology, Tongji University, Shanghai 200092)

Abstract　Chinese Text Classification is a hot area of Information Retrieval and Web Mining. Existing methods have some defect in the phase of Feature Selection. They ignore the hidden sub-class information. This paper suggests a method to enhance the weight of key words of hidden sub-classes, so that we can discover valuable information of sub-classes, then we use Rough Set to construct classifier. The result of the experiment indicates that this method can effectively improve the recall of text classification, without increase the amount of words need reduction.

Keywords　Chinese information processing, Rough set, Text classification, VSM

近年来随着网络和信息技术的发展,电子文档的数目增长迅速。信息检索等技术已经成为辅助人们高效、准确地查阅信息的重要工具。文本分类作为信息检索和文本处理的一个重要分支,也是信息检索和信息智能处理的基础。在分类工具中,很多学者对文本分类器做了深入研究[1~3]。目前主流的分类器有Rocchio[1]、KNN[4]、Naïve Bays[5]、最大熵模型[6]等等。粗糙集作为一种简单有效的分类工具,其在文本分类中的应用,也受到了越来越多的关注[7,8]。使用粗糙集可以进行特征约简,从而降低特征向量的维数,获得简洁的分类规则。当样本数和属性个数增大的时候,由于计算复杂度增加,通常的约简方法很难胜任,因而只能按照一定权值排列后截取前面一部分属性,再对这些属性进行约简。

1　分类器

Rocchio是一种传统的信息检索方法,该方法为一个类别构造一个核心向量,用待分类文本与该核心向量比较,取相似度最大的类别。此方法适用于样本呈正态分布的情况,而当样本分布不那么规则时,分类效果不佳。究其原因,是因为单纯的一个特征向量和相应的相似度公式很难体现样本的分布。

相比Rocchio,KNN方法则是另一个极端。KNN将待分类样本与训练集中每个样本比较,取最相似的 k 个样本所在的类别。显然这种方法可以胜任样本任意一种分布情况,但也引入了KNN分类方法的一般缺点:当 k 较小时,存在偶然误差,而 k 大时计算复杂度增大[4]。对KNN的改进有很多。胡荣等人将人脸识别中用于特征脸技术的NFL(nearest feature line)算法引入到KNN中,提出一种kNFL(k nearest feature line)算法用于文本分类[9],取得了较好的效果。

粗糙集理论是Z. Pawlak在上世纪80年代初提出的一种新的处理模糊和不确定信息的分类工具。在文本分类中,将所有的文本集合作为论域,每个文本作为论域中的对象,特征词的集合作为属性,离散化后的词语权重作为属性的取值。可以看出用粗糙集作为分类器时属性集合是很大的,最大可能和汉语中的词汇量相当,而粗糙集的属性数量以在数千个为宜。一般采用截取权重较大词汇的方法,以控制算法复杂度。粗糙集文本分类的效果可参见卢娇丽等人的工作[7]。为了增加粗糙集分类器的鲁棒性,降低约简算法复杂度,并向大规模文本推广,山西大学李钝等人提出一种将聚类和粗糙集分类相结合的方法[8],各个子样本集以聚类中心为代表,在此基础上构造分类器。

图1　文本类别中的子类示意图

Rocchio和KNN这两种方法都没有考虑文本类别的内部结构,将类别中心取1个和 N 个两种极端,未能很好地体现和利用样本的特性。从而引出了先聚类后分类的方法。新的kNFL方法在找出一定数量的FL(feature line)的基础上

*)基金项目:本文得到2006年博士学科点专项科研基金(20060247039)和国家自然科学基金(60475019)的资助。金凯民　硕士研究生。苗夺谦　博士,教授,博士生导师。段其国　博士研究生。

应用KNN,降低了由于样本噪声引起的偶然误差和算法复杂度,一定程度上有了改进。kNFL算法中FL类似于各个子类别的中心,kNFL算法达到了类似先聚类后分类的效果。我们认为,一个文本类别往往包含着许多子类,如图1所示,因此应当考虑子类别中心的作用。

无论是先聚类还是先算出特征线,都旨在利用样本的空间分布特征,以增加分类器的鲁棒性,并尽可能降低算法复杂度。但聚类和特征线计算又增加了处理时间。本文从文档向量模型中的词汇权值入手,试图提升代表小规模子类主要特征的词汇权值,提高这些词的重要性并保证其能在截取后的词汇属性集中保留下来,以提高文本分类的查全率。这种方法避免了聚类或特征线方法繁重的计算任务,而又能体现类别的分布结构信息。在权值处理后,再使用粗糙集进行特征的约简和分类器的构造。

2 文本的形式化表示

2.1 文档向量表示

目前在信息处理中,文本有向量空间模型(VSM)、语义网络、框架模型等表示方法。其中的向量空间模型简单明了,得到了广泛的应用。对文档集 $D=\{d_1,d_2,\cdots,d_n\}$,从中取出 m 个特征词 $T=\{t_1,t_2,\cdots,t_m\}$,而选取的标准是根据各个特征词的权重,选出排列靠前的一部分特征词。增加了权重的文档向量模型可表示为$(w_1,w_2,\cdots,w_m)$。

2.2 汉语分词

分词是中文信息处理的基础,也是和其它语言比如英语文本信息的主要区别之一。一般采用正向或逆向的最大长度匹配方法,将词典中所能找到的最大长度的词作为当前分词,再用隐马尔可夫模型进行词性标注。

我们利用北京大学汉语分词词库进行汉语分词和词性标注,该词库中已还有大量专有名词。我们只将文本中的名词性词汇作为样本特征词汇,以减小后期用粗糙集进行分类的词汇数。

2.3 权重计算

在文档向量空间模型中,目前有很多种权重函数来计算关键字在文档向量中的权值,如布尔权重函数、TF-IDF权重函数、ITC权重函数、Okapi权重函数等。鉴于Okapi权重函数的优点[10],我们选取它为权重函数并对其进行改进。Okapi权重函数的公式如下:

$$a_{ik}=\frac{tf}{0.5+1.5\dfrac{dl}{avg\_dl}}\log\left(\frac{N-df+0.5}{df+0.5}\right)$$

其中 a_{ik} 为词汇 i 在文本 k 中的权重;tf 为在训练文本 k 中词语 i 出现的次数;df 为训练集中出现词语 i 的文本数;dl 为文本长度;avg_dl 为训练集所有文本长度的平均值;N 为训练集中文本的数目。

由公式可以看出,当 df 大到一定程度时,权重计算结果会是负值。直观地看,那些在很多文章中出现的词汇,比如"人"、"国家"等在每个类别中都大量出现。作为干扰项,这些权重为负值的词汇将在权重的后处理中首先被除去。词汇的权值代表了它们对分类的重要性。在很多情况下,由于初始词汇集的庞大的数量,需要在原始词汇集按一定规则(通常是按权值由大到小排列)截取最重要的一部分词汇,用于后续的分类。Okapi权重函数主要考虑词频因素,对它的改进有不少,山西大学卢娇丽等人[7]将词汇在文章中的位置以及在类别中的位置引入权重计算。给出现在标题、摘要中的词汇和集中出现在同一类中的词汇以较高权重。但即便这样,由于部分子类关键词在截取后的词汇集中丢失,即使在封闭测试中也存在一定程度的误识。

2.4 权重计算的改进

无论是Okapi还是TFIDF权重函数,都基于一个假设,即:稀有词比常用词包含更新的信息。在极端情况下,只在一篇文章中出现的词有很高的权值。这样整个训练集受单个样本的影响波动很大,同时会使一些更有用的关键词被埋没。

正如上文中所述,同一类别中的样本往往是聚族分布,比如"历史"类别的文档中,三国史、唐史、明史这些子类分别聚团,提取这些子类的特征更有意义。我们为了突出父类中隐含子类的关键字,对那些符合"子类"标准的文本的关键词在权值上予以一定的提升。将集中出现于某一父类中,同现词数量较多的那些样本视为子类。其提升函数delt为:

$$delt=\begin{cases}s-\dfrac{s}{a^2}(dfc*\log_2 wfc-a)^2 & 0<dfc<2a\\ 0 & \text{其他}\end{cases}$$

$$a'_{ik}=a_{ik}(1+delt)$$

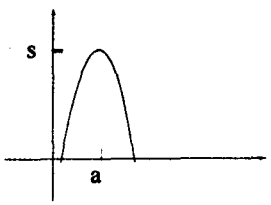

图2 权值提升函数示意图

其中 dfc 为出现词汇 i 的文档数与该文档所属类的总文本数的比值;wfc 为出现词汇 i 的文档中,除 i 外其他共同出现的词汇数,一般来说拥有多个相同词汇的几个文档,其属于某一潜在子类的可能性较大。对其取以2为底的对数后,只有出现2个以上共同词汇的被保留了下来;a 取一个较小的值以控制某个未知子类的规模;s 控制对子类关键词的提升程度。整个提升函数 dfc 类似于一个中值滤波器,那些频率过低或过高的词汇都可以无变形地通过该滤波器,而那些可能是某个潜在子类关键词的词汇被提升了起来。同时山峰形状的二次函数能够在一定程度上缓解因不重要词汇的提升而带来的噪音。当 s 取0或 a 取0时,提升函数 $delt$ 恒为0,$a'_{ik}=a_{ik}$,a'就退化为普通的Okapi权重函数。

该权重提升函数与所采用的权重计算函数无关。可适用于Okapi和TFIDF等函数。旨在将聚类等思想提前到权重计算阶段,以提高后期处理的准确性,降低算法复杂度。

我们用此公式对一篇题为"河南新蔡县人口计生系统党风行风建设受表彰"的国内新闻进行潜在子类关键词提升,结果如下(s取3,a取0.03):

表1 新闻文本的tf、df、delt及权重值计算结果示例表

关键词	tf值	Df值	Delt值	权重
新蔡县	3	5	1.3	14.51
河南省	4	10	0	5.34
计划生育	1	4	2.4	9.26
人口	4	12	0	2.91
党风	1	4	2.4	10.14
建设	1	12	0	1.12
……				

从上表可以看出，尽管一些关键词的 tf 值(训练文本 k 中词语 i 出现的次数)和 df 值(训练集中出现词语 i 的文本数)都很低，如“计划生育”、“党风”，这些词在 Okapi 权重函数中不会取得较高的分值，但它们不为 0 的 $delt$ 值可以给它们一倍左右的权值提升。同时对于“河南省”等本身权值较高的词汇，其 $delt$ 值为 0，提升滤波器对它们没有影响。

2.5 词汇权重离散化

在离散化之前，先将权重小于等于 0 的项(干扰项)滤去。这里采用简单的等距离离散化方法，将权值离散化为 0 到 10 共 11 个级别。

此时上表的结果如下：

表 2 新闻文本的 tf、df、delt 及离散化后权重值计算结果示例表

关键词	tf 值	Df 值	Delt 值	离散化后的权重
新蔡县	3	5	1.3	5
河南省	4	10	0	2
计划生育	1	4	2.4	5
人口	4	12	0	1
党风	1	4	2.4	7
建设	1	12	0	0
……				

3 利用粗糙集理论获取规则

粗糙集理论作为一种新的机器学习方法[11]，已经在很多方面进行了应用。其中文本分类是粗糙集的几个经典应用领域之一。我们采用粗糙集作为文本分类工具。在本次实验中，选取 100 篇国际新闻，100 篇国内新闻。各留 100 篇做测试。在前期分词、权重过滤之后，尚有 3216 个特征词汇，我们选取权重排名前 1000 个特征词汇，在此基础上进行属性约简，得到分类规则。

在 ROSETTA 粗糙集平台[12]上进行知识约简获取规则，在 200 篇测试集上做开放测试，结果如下：

表 2 新闻文本的 tf、df、delt 及离散化后权重值计算结果示例表

参数选取	S=3，a=0.03		S=0，a=0(传统方法)	
查全率	国际新闻	91.9%	国际新闻	87.2%
	国内新闻	94.3%	国内新闻	91.6%

由上表可见，改进后算法提高了开放测试的查全率。

总结和展望 通过滤波器器的方法将隐藏未知子类的特征词汇赋予较高的权值，使我们能够充分利用文本分类中类别的结构信息，提高查全率。在我们提升小规模子类特征的时候，虽然确实能把小规模子类的一些重要特征提升起来，但同时也提升了一些并不重要的特征，使噪声加大，因而如何选择合适的参数使结果最佳将是我们进一步的研究工作。

参 考 文 献

1 Sabstiani F. Machine learning in automated text categorization [J]. ACM Computing Surveys, 2002, 34:1～47
2 Lewis D R. A comparison of two learning algorithms for text categorization [A]. In: Symposium on Document Analysis and IR,[C]. Las Vegas: University of Nevada Press, 1994. 81～93
3 Yang Y, LIU X. A re-examination of text categorization methods [A]. In: Proc. of the 22nd Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval[C]. Berkeley: ACM Press, 1999. 42～49
4 杨丽华，戴齐. KNN 文本分类算法研究. 微计算机信息，2006，22
5 Mccallum A, Nigam K. A comparison of event models for naive Bayes text classification [A]. AAAI-98 Workshop on Learning for Text Categorization [C], Madison: AAAI Press, 1998. 22～28
6 Nigam K, Lafferty J, Mccalum A. Using maximum entropy for text classification [A]. IJCAI-99 Workshop on Machine Learning for Information Filtering [C], Stockholm: IJCAI Press, 1999. 35～42
7 卢娇丽，郑家恒. 基于粗糙集的文本分类方法研究. 中文信息学报，2005，19(2)
8 李钝，梁吉业. 利用聚类和粗糙集进行文本分类研究. 计算机工程与应用，2003，07-0186-03
9 胡荣，罗庆云. kNN 算法在文本分类中的改进. 南华大学学报(自然科学版)，2005，19(3)
10 Robers S E, Walker S. Okapi/Keenhow at TREC8 [A]. In: E. M. Yoorhees, D. K. Harmann, eds. Proceedinging of Eighth Text Retrieval Conference(TREC-8)[C], Gaithersburg, 2000
11 苗夺谦. Rough Set 理论及其在机器学习中的应用研究[D]. 北京：中国科学院自动化研究所，1997
12 http:// rosetta. lcb. uu. se/

(上接第 125 页)

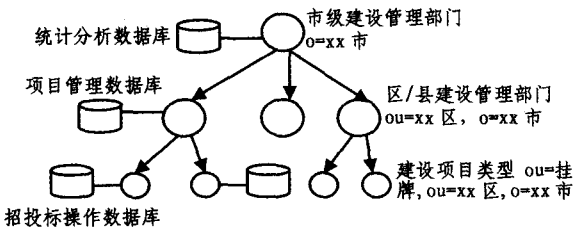

图 5 建设项目招投标管理数据框架

目录数据分布通过网络将市级管理部门与区/县管理部门连接在一起，区县部门通过招投标管理系统进行业务操作和关系数据库更新，包括项目公共、资格审查、评分规则、评标计算和结果公示。不同类型项目(招标、挂牌、拍卖)的评分规则和标准、专家分组和数量是不同的。

通过应用系统的功能开发，在混合数据框架下的每个目录节点(市/区/项目类型)，管理部门能够进行逐级管理和汇总分析，体现了本文提出的混合数据框架的可行性和优势。

结论 由于业务复杂程度的加大，以及各种专业化业务处理系统的增多，企业和单位进行信息系统规划和资源整合的难度很大。本文提出一个利用目录和关系数据结构优势的混合数据框架，可以既保持业务处理系统灵活性和独立性，又能够进行组织架构及其变化的管理，使数据做到有效的统一、分布和共享，对信息系统建设和整合中的数据结构化程度，数据组织和处理效率有一定的特色和优势。

参 考 文 献

1 王芳，张顺达，等. 基于 LDAP 的对象存储系统元数据的组织与管理. 计算机工程与科学，2007(3)
2 薛宏智，王俊，等. 基于 LDAP 的企业访问控制系统设计与实现. 计算机与信息技术，2006(Z1)
3 李明，连乔，等. SQL 标准符合性测试的框架. 计算机工程与应用，2003(20)
4 王建芳，阎保平，等. 目录的数据管理模型的研究与实现. 计算机工程，2007(10)
5 罗艳兵，蔡鸿明. 应用企业数据模型解决信息孤岛的研究. 新西部(下半月) 2007(1)
6 RFC2253. Lightweight Directory Access Protocol (v3)
7 ISO/IEC 9075:1992, Database Language SQL