一种基于维基百科知识库的中文文本分类方法研究1

苏小康[1,2]　何婷婷[1,2]　涂新辉[1,2]　何金卓[1,2]

1(华中师范大学 计算机科学与技术系，湖北 武汉 430079)

2(国家语言资源监测与研究中心网络媒体分中心，湖北 武汉 430079)

E-mail:Xiaokang86@gmail.com　tthe@mail.ccnu.edu.cn　tuxinhui@gmail.com　hejinzhuo@gmail.com

摘　要：传统的文本表示方法是基于词条的向量表示方法(Bag of Words or BOW)，文本信息中的每一个词条都被表示成该向量中的一个维度。尽管这样的表示方法简单而且常用，但是却难免会有一些限制，因为文本之间存在着复杂的潜在的联系，而且这些潜在的联系很难用词条向量表示出来。因此在文本表示中插入一些背景信息用以提高文本分类模型的精确度是很必要的。该文通过搜集维基百科全书信息作为背景知识来扩充文本信息从而达到克服传统向量表示方法(BOW)的一些缺点，实验证明该方法可以提高文本分类的精确度。

关键词：文本分类，信息扩充，维基百科

A Study of Chinese Text Classification Method Based on Wikipedia

SU Xiao-kang[1,2]　HE Ting-ting[1,2]　TU Xin-hui[1,2]　HE Jin-Zhuo[1,2]

1(Department of Computer Science, Huazhong Normal University, Wuhan, 430079;)

2(Monitor and Research Center for National Language Resource Network Multimedia Sub-branch Center, Wuhan, 430079)

E-mail:Xiaokang86@gmail.com　tthe@mail.ccnu.edu.cn　tuxinhui@gmail.com　hejinzhuo@gmail.com

Abstract: The traditional document representation is a word-based vector (Bag of Words, or BOW), where each term of the document is associated with a dimension of the vector. Although simple and commonly used, this representation has several limitations because of the complex relation between documents which BOW method can hardly deal with. To embed background information in order to enhance the accuracy of classification model is essential. In this paper, we proposed a approach of gathering the information of Wikipedia as background knowledge to enrich the information of documents, the method overcome the shortages of the traditional method(BOW).The model is proved that by use of Wikipedia to enrich the document information can improve the accuracy of document classification.

Key Words: Text Classification; Information Enrichment; WikiPedia

1、引言

随着互联网的发展，网络媒体的形式不断增多，网络上文本数量不断递增，面对海量的文本信息，人们想找到自己感兴趣的信息，文本分类技术因此产生并发展至今。如今，传统的文本分类技术的精确度已经不能满足人们的要求。因为文本信息存在着复杂的语义关系。一般来说，传统的文本分类技术都是把文本表示成一个基于词条的向量（Bag of Words or BOW）。这种传统的

[1]**项目资助：**国家自然科学基金(60773167)；国家十一五科技支撑计划课题“网络文化安全预警技术研究”(2006BAK11B03)；973 国家重点基础研究发展计划(2007CB310804)；教育部/国家外国专家局高等学校学科创新引智计划(B07042)；

文本表示方法把本来应该存在语义信息的文本分割成了一个个的词条，显然会存在一定的弊端。缺点大致上有以下几点：第一，BOW 方法没法识别同义词，比如，“好像”和“似乎”。第二，一词多义的情况。比如，“乔丹”既可以表示一个 NBA 球星，也可以表示一个体育运动品牌。第三，多个词条组成的词组被切分为多个词条，比如“机器学习”就会被分词系统分成“机器”和“学习”两个词条，那么原来应该有的语义信息就不存在了。这些缺点是传统的文本分类的技术很难突破的地方。因此，要解决这些存在的问题我们必须借助现有的资源对文本信息扩充，本文利用维基百科全书对文本中的词条都进行背景知识收集从而达到扩充文本信息的目的，实验结果说明，经过扩充文本信息以后文本分类精确度确实得到了提高。

本文接下来的结构将分为以下五部分：第二部分，相关研究工作；第三部分，维基百科全书介绍；第四部分，本文所述模型的结构以及采用的关键技术；第五部分，实验过程及结果分析；第六部分，下一步的研究工作。

2、相关工作

目前，在中文文本分类领域里，常用的模型有最小距离分类器，KNN分类器，Naive Bayes分类器和支持向量机分类器（SVM）等。文献[6]的作者提出了一种基于潜在语义分析和直推式谱图算法的文本分类方法来提高分类精度，该文认为传统的文本表示方法没有考虑到文本之间潜在的语义关系，该文利用奇异值分解技术对高维度的词条-文档矩阵进行处理，在潜在的语义空间子结构中重新表示文本，从而达到对文本分类精度的提高，实验证明该方法提高了分类的精度。虽然该文所提出的方法一定程度上实现了文本信息的扩充，但是对于那些更深层的信息比如文本想表达但是没有直接出现的信息是不能扩充出来的。

文献[7]的作者采用经典的SOM算法以及它的改进算法TGSOM把具有类概率的词条聚集成一组成为新的特征项。在此基础上计算文本聚合特征项权重，利用VSM表示文本，最后利用SPRNT决策树算法进行分类模型训练。该文通过词条聚合解决了原始文档向量可能出现的“维度灾难”的问题，极大的提高了分类性能，但是该文并没有考虑到对潜在的背景知识进行扩充的问题。

文献[8]的作者认为传统的向量空间模型在类特征向量和待判断特征向量维度相差很大时效果不好，然而在现实语料中往往又无法避免出现这种情况。因此该文提出了一种新的想法，基于HowNet对待判断文本进行一个词汇扩展，通过词汇扩展使得待判断文本尽量接近于它应该属于的那个完备的类特征向量，从而提高分类精度。但是该方法是通过关键词匹配进行简单的词汇扩充，并没有摆脱传统的BOW思想。

在英文文本分类领域，文献[1]的作者提出利用WordNet同义词集对文本词汇进行扩充处理，从而达到提高分类精度的目的。该方法类似于文献[9]所提出的基于HowNet进行中文词汇进行信息扩充的方法，同样的，它也并没有摆脱传统的BOW思想。

由于WordNet并不能充分的表示出文本所蕴含的丰富的语义信息，文献[2][3][4]的作者提出基于维基百科对英文文本信息进行扩充。

3、维基百科全书

维基百科全书是世界上最大的多语种的开放式的百科全书，与传统的百科全书最大的不同之处在于，它是面向互联网的，开放式的百科全书。它的基本组成单元叫“概念”，每个概念都由一篇文章来解释。

维基百科全书除了包含概念解释页以外，还包含了重定向页和消除歧义页。重定向页的存在原因和举例请参见表一。消歧页的存在是为了解释一词多义的情况，它让用户可以选择同名，不同义的条目进行跳转从而得到想要的信息。

表1　重定向页面原因和举例说明

原因：	用法说明：
简称	奥运会重定向到奥林匹克运动会
常见错别字	六合塔重定向到六和塔
不同的译名	圣弗朗西斯科重定向到旧金山
字母大小写	APM 重定向到 apm
别名和同义词	粤重定向到广东
其他语言	DNA 重定向到脱氧核糖核酸
相关文字	水解反应重定向到水解

概念与概念之间还存在着相互链接关系。图 1 举例说明了这一情况。（箭头方向表示概念之间跳转方向）

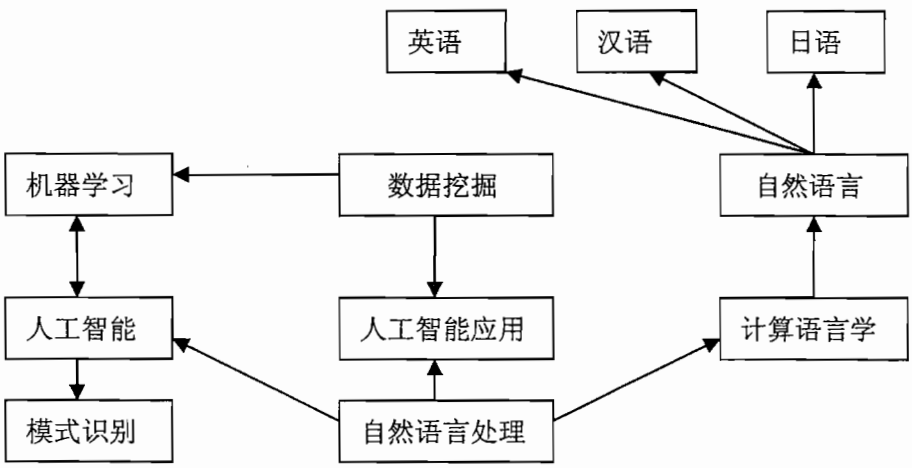

图1　维基百科概念之间链接关系举例

4、 基于维基百科全书进行文本扩充的文本分类模型

在这一节中，我们首先概要的描述本文所提出的模型的基本思想，然后对模型中所采用的关键技术作重点分析。

4.1 模型的基本思想

我们通过对传统的基于词条的分类模型进行仔细研究发现，传统的分类模型都是通过计算词条与词条之间的相似度得来的。词条与词条之间并不具备任何语义关系，因此传统的分类模型有很大的限制。本文所提出的模型不同于传统的分类模型，除了计算词条和词条相似度，我们还计算概念与概念之间的相似度，概念是在把维基百科作为背景知识，由词条扩充而来的。图 2 是本文所提出的分类模型的流程图。

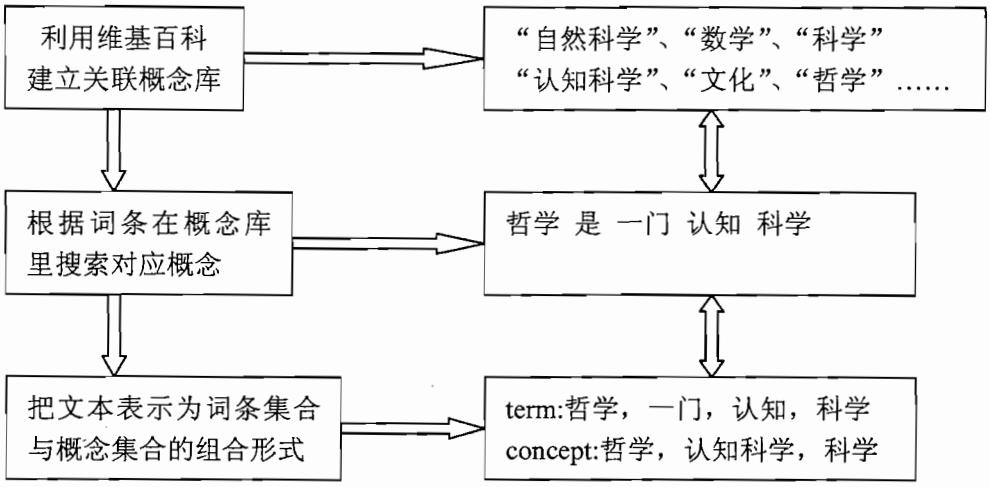

图2　基于维基百科知识库的文本分类模型流程图

4.2 维基百科关联概念库的建立

“概念”是维基百科全书的基本组成单元，每个概念被一篇文章解释，每篇文章中还包含有许多其他链接，每个链接都可以跳转到其他文章，但是在维基百科页面中，并不是所有的链接都是和这个页面所描述的概念息息相关的，比如在维基百科里面搜索词条“机器学习”，返回页面包含了许多链接，其中在介绍机器学习的应用时，有一个链接到“证券市场”。显然，这个概念和“机器学习”并不是紧密相关的。

因此，我们在建立概念库的时候，采用了一种严格匹配的策略，首先我们认为该概念与它所属的类别是直接关联的，比如“自然语言处理”属于3个类别：“人工智能”，“人工智能应用”，“计算语言学”，那么我们认为，“自然语言处理”与这3个概念是直接关联的。再次，如果这是一个重定向页面，比如“证券市场”被重定向到“股市”，我们也认为这两个概念是直接关联的。这样的严格的匹配机制虽然会丢失维基百科一些有用的链接，但是却大大的提高了效率，并且，后面的实验证明，这样做对文本分类的精确度有一个明显的提高。

4.3 对待判断文本和训练语料的扩充

对于一个经过预处理操作的待判断文本，我们可以把每个词条作为输入对维基百科相关概念库进行检索。这里，我们为了避免扩充后的文本出现“维度灾难”，我们采用严格匹配策略，即：这个词条作为一个概念在相关概念库里单独出现了，我们就检索出这一系列相关的概念，否则，我们不予以扩充。这样，我们认为，扩充出来的这些概念是和这个文本关联非常紧密的。比如对应一个文本片段“哲学是一门认知科学” 将会被扩充出“哲学”，“认知”，“认知科学”，“科学”这些概念出来。这些概念被加入到文本向量表示中用以扩充文本信息，一个被扩充后的文本，应该包含词条和概念。

因为本文接下来采用的相似度计算公式是改进以后的贝叶斯公式，因此训练语料的扩充和待判断文本的扩充有一点不一样的是要记录下扩充出的概念的tf（词频）值。把训练语料里面的每个词条作为输入对维基百科概念库进行检索得到一系列概念，如果根据该输入词条匹配到n个相关联的概念，那我们把这些概念的权重分别设置为tf/n（tf为该输入词条的tf值）。这样的设置是为了避免因为扩充出的概念总tf值太大影响到分类效果，当然这样的设置存在一定的不足，因为概念与概念之间的联系在程度上存在着不一样，考虑到效率问题，我们假设关联出的概念和

输入概念之间的联系程度是一样的，实验证明这样做是有一定的可行性的。

4.4 利用改进的朴素贝叶斯概率公式计算相似度

用数学公式表示贝叶斯分类模型：

$$P(Ci \mid d) = \frac{P(Ci)P(d \mid Ci)}{P(d)} = \frac{P(Ci)P((w1, w2...wn) \mid Ci)}{P(d)} \tag{1}$$

其中：Ci 表示某个训练语料集，d 表示某个测试文档，w1,w2…wn 表示测试文档的词条集合。

在朴素贝叶斯假设下，所有的词条之间都是相互独立的，那么上述公式(1)可以简化为：

$$P(Ci \mid d) = \frac{P(Ci)}{P(d)} P(w1 \mid Ci)P(w2 \mid Ci) \cdots P(wn \mid Ci) \tag{2}$$

因为本文所提出的模型是对文本进行概念扩充以后再进行计算的，测试文本和训练语料除了包含有词条以外还包含由词条扩充而来的概念，传统的朴素贝叶斯公式不再适用。所以我们提出了新的公式用来计算本模型中的贝叶斯概率：

$$P(Ci' \mid d') = \frac{P(Ci')P(d' \mid Ci')}{P(d')} = \frac{P(Ci')P((w1, w2...wn, c1, c2...cm) \mid Ci')}{P(d')} \tag{3}$$

$$P(ci \mid Ci') = \frac{\text{概念}ci\text{在训练语料中的概念集里出现的次数} + 1}{Ci'\text{训练语料的总概念数} + d\text{的概念数}} \tag{4}$$

其中：d'表示扩充以后的测试文档，Ci'表示扩充以后的训练语料集合，w1,w2…wn 表示测试文档的词条集合，c1,c2…cm 表示由该测试文档的词条集合扩充而来的概念集合。

和传统的贝叶斯概率模型相比，我们不单纯计算词条在训练语料集的共现概率，我们还考虑了由词条扩充出来的概念信息。分别计算词条的共现概率 P1 和概念的共现概率 P2。把两种概率进行线性组合：$P = \lambda P1 + (1 - \lambda)P2$

λ 参数的值由扩充文本的程度决定，即：$\lambda = \dfrac{\text{词条数}}{\text{词条数} + \text{概念数}}$

当测试文本没有关联到任何维基百科的概念时，那么我们并不对他进行扩充，λ 参数的值为 1，该公式退化为朴素贝叶斯概率公式。

5、实验过程及结果分析

5.1 实验语料说明

从维基百科网站下载获取中文版本 XML corpus，它包含现有的所有概念解释 XML 文档集合以及概念名、概念 ID、文档 ID、子概念 ID 之间相互关系映射文件。

训练语料选用的是 SougoC 中文文本分类标准文档集，包含（财经，IT，健康，体育，旅游，教育，招聘，文化，军事）九类共计 17 910 篇。其中，每类文本均为 1 990 篇。

测试语料从复旦大学文本分类测试语料中选取了相关的 4 个类别（财经，教育，体育，文化）

共计 2 950 篇，其中财经类 1 601 篇，教育类 61 篇，体育类 1 254 篇，文化类 34 篇。

5.2 实验步骤

为了进行对比，我们对应同样的训练语料和测试语料分别用了朴素贝叶斯文本分类模型和本文所提出的新的模型进行实验。

1）对维基百科 XML corpus 建立一个索引库，设置一个域存放这个概念关联的一系列概念。

2）对训练语料进行预处理操作，得到训练结果，检索索引库，得到相关概念。并对这些概念分别设置权重。进行扩充，生成新的训练结果。

3）对每个测试文本进行预处理操作得到词条集合，检索索引库，得到相关概念。进行扩充，生成新的文本。

4）根据改进以后的贝叶斯概率公式计算出扩充后的测试文本与扩充后的训练语料之间的贝叶斯概率值。并根据概率值对该文本做出分类决定。

5.3 实验结果及分析

根据上述实验步骤，分别从复旦大学文本分类测试语料中选取了财经，教育，体育，文化四类共计 2 950 篇进行两组实验，实验结果说明经过文本扩充以后的分类模型对分类精度有明显的提高。表 2 为实验结果：

表2　实验结果

	朴素贝叶斯分类模型	经过文本信息扩充的分类模型
财经（1601）	0.74690	0.83710
教育（61）	0.73770	0.85721
体育（1254）	0.75917	0.85624
文化（34）	0.70588	0.82630
合计（2950）	0.75145	0.84553

上述结果显示，朴素贝叶斯分类模型的精确度在 0.75 左右，本文所提出的模型分类精度在 0.85 左右。朴素贝叶斯分类模型是严格依赖于训练语料的一种概率模型，SogouC 文本分类标准语料是采集自网络上的新闻类文章，而复旦大学文本测试语料采集自报刊杂志等出版物。由于它们在文字表达上存在一定的差异性，所以朴素贝叶斯分类模型的精度没有达到应有的水平，而经过文本信息扩充的分类模型在精度上有一个明显的提高，说明本文所提出的方法是有效的。

6、总结与展望

因为文本信息之间存在着复杂的语义关系，传统的文本分类模型（BOW）的分类精度正处于一个瓶颈阶段。本文提出了一种新的方法，基于维基百科知识库对文本信息进行背景知识的扩充，本文接着还提出了一组改进过的贝叶斯概率公式，最后，我们对相同的数据分别进行了两组实验，实验结果证明本文所提出的方法是有效的。

但是，出于效率的考虑，本文所提出的方法在构建维基百科概念库的时候采用了一种严格的匹配策略，这种策略虽然实现起来效率很高，但是不可否认对于概念库的规模是有很大的缩减的。本文对训练语料扩充出的概念的权重设置也是基于一种严格意义上的假定做出的。本文所提出的模型是用改进的朴素贝叶斯概率公式进行相似度计算，虽然朴素贝叶斯分类器在文本分类中一直有着很好的效果，但是我们把扩充出的相关概念当做了词条来进行贝叶斯概率的计算，这样显然并没有最大限度的利用维基百科全书概念之间丰富的内在联系。

因此，我们未来的研究任务主要放在以下两个方面：第一，维基百科相关概念库的建设，提出一种新的能在效率和质量上兼备的算法。第二，考虑从以下三个方面去计算概念之间的相似度，基于概念解释文本内容[5]，基于概念所链接的概念集合[5]，基于概念与概念之间的距离[5]。

参 考 文 献

[1] M.de Buenaga Rodriguez,J.M.GHidalgo,and B.Diaz-Agudo.Using WordNet to complement training information in text categorization. In Recent Advances in Natural Language Processing II,volune 189.John Benjamins.2000.

[2] E.Gabrilovich and S.Markovitch.Feature generation for text categorization using world knowledge.In International Joint Conference on Artificial Intelligence,Edinburgh,Scotland,2005.

[3] E.Gabrilovich and S.Markovitch.Overcoming the brittleness bottleneck using wikipedia:enhancing text categorization with encyclopedic knowledge.In National Conference on Artificial Intelligence (AAAI),Boston,Massachusetts,2006.

[4] E.Gabrilovich and S.Markovitch. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis.In IJACT'07

[5] P.Wang,J.Hu,H.-J.Zeng,L.Chen,and Z.Chen.Improving text classification by using encyclopedia knowledge.In International Conference on Data Mining,pages 332-341,Omaha,NE,2007.IEEE.

[6] 戴新宇,田宝明,周俊生,陈家骏.一种基于潜在语义分析和直推式谱图算法的文本分类方法 LSASGT.电子学报,2008 年 8 月:第 8 期

[7] 蒋宗礼,徐学可,李帅.文本分类中基于词条聚合的特征抽取.哈尔滨工程大学学报,2008 年 11 月:第 11 期

[8] 孙宏纲,陆余良,刘金红,龚笔宏.基于 HowNet 的 VSM 模型扩展在文本分类中的应用研究.中文信息学报,2007 年 11 月:第 21 卷第 6 期.