中 文 信 息 学 报
第 13 卷 第 6 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 6

中文文献的层次分类方法*

战学刚　林鸿飞　姚天顺

东北大学计算机科学与工程系　沈阳　110006

摘要　现有的分类系统通常忽略类别体系的层次结构,在对文献进行分类时,往往很难区分类别相近的文献属于哪一类。本文基于向量空间模型,提出根据类别体系的层次结构,自顶向下,逐层分类的方法。其目的是提高分类精度;并根据概念词典,将同义词或下位概念映射到单一的概念词上,由这些概念词构成一个规模很小的特征集,以缩小特征向量空间的维数,从而减少分类系统的计算量。此外,通过对类别层次体系的分析,压缩特征向量,从另一方面减少分类系统的计算量。

关键词　文献分类　向量空间模型　类别层次结构

Hierarchical Method for Chinese Document Classification

Zhan Xuegang　Lin Hongfei　Yao Tianshun

Department of Computer Science, Northeastern University　Shenyang　110006

Email: ics@mail.neu.edu.cn

Abstract　Existing statistical document classification systems often ignore the hierarchical structure of the pre-defined topics. This makes it difficult to identify which category a document belongs to when the possible categories are somewhat similar. In this article, we propose a top-down classification method according to the hierarchical structure of topics. The purpose is to improve precision and reduce computation of classification systems. Through a concept dictionary (thesaurus), we map the synonyms or lower-level concepts in a document to a small set of concept words that are used as terms. This reduces the computational complexity from another aspect by reducing the dimension of the vector space.

Keywords　Document classification　Vector space model　Topic category hierarchy

一、引言

文献分类就是将大量的自然语言文献归结到一个(或多个)预定义的文献类别中。近年来,随着文本信息的不断增多,人们对大规模文本信息自动处理也提出了更高要求。有效的信

* 本文于 1999 年 3 月 22 日收到

息检索需要有良好的索引和文献内容概括。文献分类便是解决这类问题的一种手段。文献分类一般是通过统计方法或知识工程方法来实现的[1]。知识工程方法需要编制大量的推理规则,因此其开发费用相当昂贵。这种方法的一个例子是卡内基集团为路透社开发的 Construe 系统[2]。该系统的开发工作量达 10 个人年。相比之下,统计方法由于其相对简单的机制,为大多数实用文献分类系统所采用[3]。

在基于统计的各种分类方法中,应用最广的是向量空间法(VSM)和 Bayes 方法。其它的统计方法,大都是这两种方法的变形或改进(如 kNN 方法)。它们的共同特点是[1]:

- 忽略文献的语言学结构
- 把文献类和文献都作为特征项的集合对待
- 利用加权特征项构成向量作为文献或文献类的表示
- 根据词频信息计算权值。

统计分类方法的基本假设是文章的内容与其中的词汇有着必然的联系。因此,许多分类系统都直接用词(主要是名词)或词组作为特征项,并将文献的向量表示与文献类的向量表示逐个比较,以确定文献的类别。这就要求系统在此之前确定文献类的特征向量,即用手工标注的文献集作为训练文献,求出各个类别的特征向量。然而,对于较大规模的训练文献集,往往有数百个类别和成千上万的特征项,系统的计算量极大。而且系统的性能对训练集的依赖性非常大。此外,由于这些系统通常忽略类别体系的层次结构,而将各个文献类作为单一的实体看待,这就加剧了系统对训练集的依赖性,系统的分类精度也受到影响。

Schutze[4]等人的实验表明,特征筛选是处理这些问题的有效方法。我们可以将训练集中那些对于主题类别不具区分能力或区分能力很小的词汇从特征项中删除。而 Koller[5]等人则进一步证实,将特征项从 1600 条缩减到 600 条,竟使系统的分类精度明显提高。即使这样,系统的计算量仍然很大,系统稳定性也受到限制。

本文就上述的问题,基于向量空间模型,提出两点解决方案:

1. 根据类别体系的层次结构,自顶向下,逐层分类。其目的是提高分类精度,降低计算量,并在一定程度上减少对训练集的依赖性。

2. 根据概念词典,将同义词或下位概念映射到单一的概念词上,由这些概念词构成一个规模很小的特征集。这样可以大大降低特征向量空间的维数,减少分类系统的计算量。

二、层次分类方法

如上所述,基于统计的分类方法都是从文献中提取词汇信息,并以特征向量的形式来表示文献。

当两个特征向量很相近,但其对应的文献又属于不同类别时,系统区分它们的能力严重依赖于训练文献集。而事实上,这时起区分作用的往往是为数极少的一些特征项。在这种情况下,如果我们考虑类别体系的层次结构,则对应的文献又往往分别属于某一主题类的两个子类。此外,如果不考虑类别体系的层次结构,则对于存在相邻层次关系的类别(如图 1 中的农业和粮食生产),分类精度很难保证。

当主题类别差异很大,且类别数目很小时,不需很多特征项便可将它们区分开来。例如,当主题类别只有“计算机”和“农业”时,很容易区分一篇关于打印机的文章是属于哪一类的。

基于上述原因,我们考察类别体系的层次结构。

2.1　类别层次

当我们根据类别体系列出其层次结构时，便得到了类似于图 1 所示的森林状层次结构图。其中每一节点的子节点代表其子类。

显然，每一节点的子节点数都只占全部节点数的一小部分（第一层的节点数可能略多些，但就整体而言，所占比例仍然很小。）。

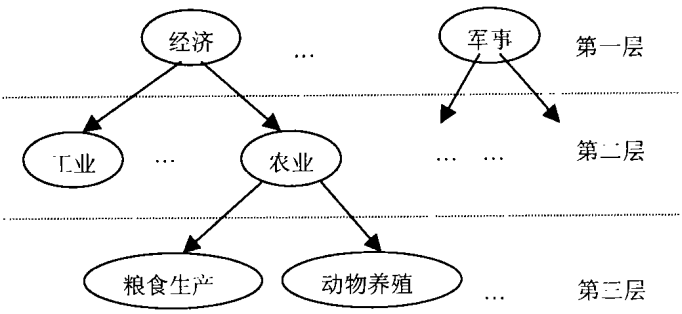

图 1　类别体系的层次结构

如果我们根据这种层次结构，自顶向下地逐层分类，则可以提高分类精度，降低计算量。这是因为在第一层上，主题类别差异很大，而且类别数目较小，这样，区分一篇文章属于哪一类便比较容易。而在此后的进一步细分过程中，分类路径是沿第一层的某个节点的各个子节点逐层向下。每一次细分，都是在很小的类别范围内进行的，所以，分类精度易于保证。从这一过程中，我们还可以看出，文献的特征向量并不是与所有的类别特征向量进行比较，而是只与部分类别特征向量进行比较，这就大大减少了比较次数。

此外，如前文所述，两个特征向量很相近，但其对应的文献又属于不同类别，或主题类别差异很大时，只需很少的特征项便可区分文献的类别。因此，我们从第一层的各个类别向量以及以下各个类别的子类别向量中抽取出那些区分能力较强的特征项，并将原特征向量中的其它分量置为零，构成新的特征向量，这样便突出了这些类别特征向量的区分能力，并在一定程度上减少对训练集的依赖性，提高分类系统的稳定性。而由于这些新的类别特征向量中非零分量很少，我们可以用

（下标值，特征值，下一项）

形式的链表或类似的数组结构来表示特征向量，这种变换的实际作用，是把原来特征向量中的其它分量置零。但当非零分量很少时，利用上述表示方法，可以进一步减少分类时的计算量。

为了叙述方便，我们称这种新的特征向量为压缩的特征向量。

从图 1 中可以看出，在类别体系的层次结构各个子树中，层次较低的节点的各个子节点间的相似度要高于层次较高的节点的各个子节点间的相似度。相应地，其对应的特征向量中区分能力较强的特征项的数目也较少，即对应的压缩的特征向量中特征项数目较少。

以下我们介绍特征项和特征向量的表示方法，然后给出类别特征向量的构造方法。

2.2　概念层次

考察一篇文章所描述的内容，主要是看其中表达被描述对象的名词或名词短语，以确定被描述对象。尽管文章的谴词造句随作者的风格和习惯而不同，但描述同一主题的文章所涉及的概念是相似的。

同一概念可以用多个同义词或短语来描述。因此，可以将文章中的名词或名词短语映射到它所描述的概念上，用一个词来表示，我们称之为概念词，并用概念词作为特征项来构成特

征向量。

此外，根据类别体系，一些个体概念，完全可以由其对应的群体概念来表示。如“猪”、“牛”、“羊”等可由“牲畜”来表示。表示组成关系的概念也可以由其上位概念代替，如“硬盘”、“显示卡”等可用“计算机”代替。

这样可以大大减少特征项的数目，从而降低分类时的计算量。

2.3 特征向量的压缩方法

我们用所有的概念词构成一个概念词表。假设词表中有 N 条概念词，那么用于描述文献和类别的特征向量便有 N 个分量。

对于文献，其对应的特征向量 $V=(w_1, w_2, \cdots, w_N)$；其中，

$$w_k = f_k * idf_k$$

式中 f_k 是第 k 个特征项在文献中出现的频率；

idf_k 是第 k 个特征项在训练集中的反比文献频率，$idf_k=[\log_2 n-\log_2 d_k]+1$；

n 是训练集的文献总数，d_k 是训练集中含第 k 个特征项的文献数。

对于文献类，其对应的特征向量 $V=(w1, w2, \cdots, wN)$；其中，

w_k 是训练集中属于该文献类及其子类的所有文献的特征向量第 k 个分量的和。

在实际应用中，可以将各个向量规范化，使得向量的模为 1。方法是将各个分量除以

$$\sqrt{\sum_{i=1}^{N} W_i^2}$$

下面我们给出各文献类特征向量的构造和压缩方法：

1.［构造所有文献的特征向量］ 利用分词程序对训练集中的文献进行逐篇扫描。在此过程中，将有关的名词或名词短语映射到概念词表并进行计数；记录每篇文献的类别，并为它构造一个频率向量（$f_1, f_2, \cdots, f_N$）；其中 f_k 是第 k 个概念词在该文献中出现的次数。这一过程结束后，我们可以根据所有的频率向量和文献总数 n，计算出各个特征项在训练集中的反比文献频率 idf_k。将各个频率向量的分量乘以相应的 idf，便得到了训练集中所有文献的特征向量 $V_i=(w_{i1}, w_{i2}, \cdots, w_{iN})$。

2.［构造类别特征向量］ 根据类别体系的层次结构，逐类将属于该类及其子类的所有文献的特征向量相加，便得到了各个文献类的原始特征向量。为了后续处理方便，我们将各个向量规范化，使得向量的模为 1。

3.［压缩类别特征向量］ 用每一节点的所有直接子节点的特征向量（作为行）构成一个矩阵，删除所有全为零的列和所有全非零的列，然后逐列计算出各个列非零元素的乘积，并依各列零元素数目（降序）和乘积的大小（升序）将各列排序（同时记录下各列的原始位置），构成一个新的矩阵。逐行从新矩阵的各列中自左向右选取 n_i 个非零元素及其原始位置构成压缩的特征向量。其中 n_i 是根据各节点所在层次 i 预先确定的分量个数（阈值）。根据前面的讨论，$n_1 \geqslant n_2 \geqslant n_3 \cdots$。

通过训练集确定了各个类别特征向量后，系统便可以进行文献分类工作。分类算法的基本步骤如下：

1.［构造文献特征向量］ 利用分词程序对文献进行逐篇扫描。在此过程中，将有关的名词或名词短语映射到概念词表并进行计数；并为它构造一个频率向量（$f_1, f_2, \cdots, f_N$）；将各个频率向量的分量乘以相应的 idf，得到该文献的特征向量 $V=(w_1, w_2, \cdots, w_N)$。

2.［首层分类］ 利用夹角余弦公式，计算 D 与类别层次结构中第一层各节点的类别特征

向量间的相似度，确定 D 所属的类别。

3. [下层分类] 继续利用夹角余弦公式，计算 D 与当前类别的各直接子节点的相似度，确定 D 所属的子类别。直至当前节点无子节点或无法区分所属的子类。当前节点即是文献 D 所属的类别。

由于我们采用了压缩特征向量来表示各个类别，相似度的计算须利用类似于稀疏矩阵算法来进行(向量是矩阵的特殊形式)。其具体算法请参阅有关的数据结构文献。欧氏空间的夹角余弦公式如下：

设 $V_1=(w_{11}, w_{12}, \cdots, w_{1N})$，$V_2=(w_{21}, w_{22}, \cdots, w_{2N})$，则

$$Sim(V_1, V_2)=\cos\theta=\frac{\sum_{k=1}^{N} W_{1k} * W_{2k}}{\sqrt{(\sum_{k=1}^{N} W_{1k}^2) * \sum_{k=1}^{N} W_{2k}^2)}}$$

此外，我们也可以根据上述的算法框架，利用其它的 VSM 变体(如 kNN)进行分类。

三、实验结果

我们按照上述方法，构造了一个实验系统，以考察这种方法的有效性。实验系统所使用的分词词典含 72 000 个词条，其中 2 ~4 字的名词词条为 21931 条。我们删去使用频率常用词汇和冷僻词汇，从中选出 3 600 个名词用于分类系统，根据《中国分类主题词表》(华艺出版社，1994)进一步将这些名词归结到 920 个概念词。针对新闻语料，并基于《中国分类主题词表》，将分类体系确定为 108 个类别，其层次结构为 3 层。我们把已标注的 4 000 篇《人民日报》和新加坡《联合早报》的新闻语料划分成两部分，800 篇作为训练语料，3 200 篇用作测试(10MB)。通过训练语料，我们删除了概念词表中未被用到或极少用到的词汇，只保留了 660 项词条。

评价与测试文献自动分类算法两个重要指标：查全率(recall)、查准率(precision)和平均查准率(average precision)，这与文献检索系统评估方法类似。查全率是指通过分类算法被正确分类的文献占未分类之前属于该类的文献的百分比，查准率是指通过分类算法被正确分类的文献占被分类为该类的文献的百分比。通过设定不同的阈值，我们可以得到查全率在 0%、10%、20%、…、100%处的查准率，并计算出平均查准率作为分类系统的性能指标。有关性能评估的详细描述，见参考文献[6]。

针对不同的参数设置，实验结果如表 1 所示。

在实验 5 中，未压缩类别特征向量；而在实验 6 中，则直接用我们选出的名词词条作特征项。从表 1 中可以看出，实验 4 的平均查准率最高，而实验 6 的平均查准率最低。实验 3 和 4 的效果均优于实验 5。这说明适当地删掉一些区分度较小的特征项，能够起到消除噪声的作用，从而提高分类系统的性能。实验 6 则说明了，直接以词汇作为特征项，其分类效果不如以概念词作为特征项。而且其分类时间大约是其它实验的两倍(在 Pentium 166，32M 内存的 PC 上)。

表 1　3200 篇文献的分类结果

实验	词条数 N	n1	n2	n3	平均查准率
1	660	32	24	16	78.43%
2	660	64	32	16	79.83%
3	660	128	64	32	81.16%
4	660	256	128	64	81.44%
5	660	660	660	660	80.67%
6	3600	3600	3600	3600	74.75%

四、结论

本文基于向量空间模型，提出了根据类别体系的层次结构，自顶向下，逐层分类的方法。并利用小特征集，以减少系统在分类时的计算量。

尽管同一系统的性能，可能随特征集和训练语料的选择而稍有变化，但针对相同的训练语料和测试语料，我们的实验结果说明，自顶向下的逐层分类方法可以提高系统的分类精度并减少分类时的计算量。此方法可操作性强，适用面较广。稍做调整后，该设计思想也适用于其它基于统计的分类方法。

需要进一步解决的问题之一是如何根据预定义的类别体系和已有的词典确定用于分类的特征词典。特征词典的构造无论对于分类系统还是检索系统都是非常重要的。在我们的实验中所使用的特征词典是基于汉英机译系统的汉英词典[7]（其中包含语义码，可据此确定同义词和概念从属关系）手工构造的。目前，我们正在进行计算机辅助特征词典构造方面的研究。我们曾尝试过利用汉语分析器对文献标题进行语法分析[8]，并通过概念词典进行确定性映射来进行分类。实验表明，当文献类别差异较大时该方法效果很好，但当文献类别差异较小时，其分类效果很难令人满意。这也是我们提出利用类别体系的层次关系来提高分类精度的原因之一。

参 考 文 献

[1] Yiming Yang. An Evaluation of Statistical Approach to Text Categorization, http://www.cs.cmu.edu/~yiming

[2] Kennech W Church, Lisa F Rau. Commercial Applications of Natural Language Processing, Comm. of ACM, Nov. 1995, 38(11)

[3] 吴立德等. 大规模中文文本处理. 上海：复旦大学出版社，1997

[4] Schutze H, Hull D, Pedersen J. A Comparison of Selective Bayesian Network Classifiers. In: ICML—96, 1996

[5] Koller D, Sahami M. Toward Optimal Feature Selection. In: Proceedings of ICML—96, 1996

[6] Salton G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, Pennsylvania, 1989

[7] 姚天顺等. 自然语言理解. 北京：清华大学出版社，1995

[8] 战学刚，姚天顺. 基于汉语分析的中文分类方法. 见：1998 中文信息处理国际会议论文集. 北京：清华大学出版社，1998