基于标题的中文新闻网页自动分类

钱爱兵[1]　江　岚[2]
[1](南京中医药大学经贸管理学院　南京 210046)
[2](南京大学信息管理系　南京 210093)

【摘要】借鉴 tf-idf 加权思想,利用新闻标题来做中文新闻网页自动分类的依据,构建基于标题的中文新闻自动分类方法,并设计多个实验对各种基于标题的中文新闻网页自动分类方法进行评测。实验结果表明,基于标题对中文新闻网页进行自动分类,可以大大缩短判断处理时间,节省存储空间,且准确率较高,特别是改进的类目加权法分类效果最好。

【关键词】词频/逆文档频率　新闻标题　中文新闻网页　自动分类
【分类号】TP391　G202

Automatic Classification Based on News Titles for Chinese News Web Pages

Qian Aibing[1]　Jiang Lan[2]
[1](School of Economy and Commercial Management, Nanjing University of Chinese Medicine, Nanjing 210046, China)
[2](Department of Information Management, Nanjing University, Nanjing 210093, China)

【Abstract】This paper describes automatic Chinese news Web pages classification by using news title based on tf-idf weighting scheme, and constructs correlation degree of news title which determines appropriate category for each news Web page. The performance of this proposed method is evaluated in terms of top one score, top two score, and top three score. The experimental evaluation demonstrates that improved tf-idf weighting scheme with categories provides high accuracy with the classification of Chinese news Web pages.

【Keywords】tf-idf　News title　Chinese news Web pages　Automatic classification

1　引　言

近年来,随着 Web 技术的飞速发展与普及,各种电子文本在数量和类别上不断累积,造成了有效管理与利用的难题,从而,电子文本分类的需求应运而生。作为电子文本信息重要组成部分的中文新闻网页也面临着同样的挑战,迫切需要通过标准化的分类加以规范,实现新闻行业之间、新闻行业和广大用户之间的新闻信息互换、存储、处理和共享。传统意义上的文本分类工作均由人工来完成,即根据文本管理者的需求与期望事先定义或选定类别,再由人工阅读文本,根据其主题大意给予适当类别标示。由此可以看出:人工分类的周期长、成本高且效率低下,难以适应中文新闻网页迅猛增长的实际情况,因此,实现自动分类是中文新闻网页分类工作的必由之路。标准化是自动化的基础和前提,但长期以来,中文新闻信息没有统一的分类标准,这一瓶颈严重制约了中文新闻信息自动分类技术的研究与开发。直到 2006 年 1 月 5 日,我国第一部中文新闻信息分类国家标准——《中文新闻信息分类与代码》正式颁布实施,才填补了中文新闻信息分类法的空白,也从此打破了中文新闻信息自动分类研究的僵局。

收稿日期:2008-07-02
收修改稿日期:2008-07-23

目前,国内外关于文本自动分类的研究,虽然所采用的分类算法不尽相同,包括:Naive Bayes[1,2]、SVM[3-5]、kNN[6]、Rocchio[7]等,但绝大部分都是依据文件内容进行分类,因此,需要对整篇文章的文本作相关的预处理,包括:分词、停用词过滤、关键词抽取等。如果处理的文本是网页,那么在分词之前还要进行正文抽取操作。处理过程相当麻烦和耗时,计算量也十分庞大,且占用大量存储空间,并且正文抽取、关键词抽取的质量也将直接影响自动分类的精度。邓茜[8]等人为中文新闻信息自动分类标引设计了一个总体框架,并初步实现,但从文中看不到具体的实验数据,因而,无法对分类效果进行评价。侯汉清[9]等人针对网页文本信息的基本特征进行网页自动标引和自动分类,标引和分类过程较为科学,但是网页格式的复杂多变决定了自动抽取网页文本信息特征本身带有极大的不确定性,从而直接影响分类的准确率。此外,他采用的核心分类标准是《中图法》,也难以适应新闻信息的分类。何琳[10]等人提出基于标引经验和机器学习相结合的多层自动分类方法,该方法的核心在于关键词的自动标引,主要解决学术论文的自动分类问题,而本文的分类对象是中文新闻网页。姜远[11]等人提出基于词频分类器集成的文本分类方法,该方法具有一般通用性,但是词频统计的范围是全文,运算代价相对较大。本文提出的分类方法,从本质上说也是一种基于词频的集成式学习方式的分类,不同之处在于词频统计的范围是标题而不是全文。

文件标题通常代表文章的中心和主旨,这一特点在新闻中体现的尤其明显。新闻撰稿人为了在短短的新闻标题中表达整篇新闻的中心思想和主题内容,达到吸引读者的目的,所使用的字和词都相当言简意赅,都有特别的意义,且都很重要。众所周知,某一特定文档内的高频词语以及该词语在整个文档集合中的低文档频率,可以产生出高权重的 tf - idf,因此,tf - idf 倾向于过滤掉常见的词语,保留重要的词语。同理可得,如果待分类新闻标题中的词语在某个新闻类目中的词频较高,但文档频率较低,则该种类型的词语具有较强的新闻类目特征,即新闻标题中一旦出现这样的词语,则该篇新闻属于特定新闻类目的概率就极大。

本文正是基于上述原因,借鉴 tf - idf 的思想,利用新闻标题来做中文新闻网页自动分类的依据,构建基于标题的中文新闻自动分类方法,并设计多个实验对各种基于标题的中文新闻网页自动分类方法进行评测,实验结果表明:用标题来做中文新闻网页分类可以大大缩短判断处理时间,且准确率较高,特别是改进的类目加权法分类效果最好。

2 新闻网页预处理

新闻标题经过分词后得到一组词汇和短语,而基于标题对新闻网页进行自动分类,必须统计这些词汇或短语在新闻类目中的词频以及文档频率。为了快速、高效地获取这些数据,成功实现新闻网页自动分类,需对其进行相关的预处理操作,包括:

(1)手工分类,创建新闻分类标注语料库;

(2)抽取新闻网页正文,形成“干净”的新闻文本;

(3)对正文文本进行分词,创建全文索引。

2.1 创建新闻分类标注语料库

此处的新闻分类标注语料库是由人工分好类的中文新闻网页构成(虽然人工分类难免会出现偏差,但本文所用的语料库由多名具有专业背景的人员进行加工,因而可将偏差降至最低),它是实现中文新闻网页自动分类的基础和前提。但是,现有的中文新闻网页语料库均不符合本文的实验要求。例如 Sogou 提供的文本分类语料库[12]只包含 10 个一级类目,没有采用《中文新闻信息分类与代码》,而且也没有将标题信息单独抽取出来。北大网络实验室提供的中文网页分类训练集——CCT2006[13]同样只有 8 个一级类目,并且全文数据嵌在 HTML 源码中。为了解决这一问题,笔者利用《江苏法院网络舆情采集与索引系统》共采集到分属于 195 个网站的 24 634 张与江苏法院相关的中文新闻网页,然后分别交由 8 名具备图书馆分类、编目背景知识的人员进行手工分类,其中有 2 人是从事图书馆编目工作达 6 年以上的专家,另外 6 人分别是情报学、图书馆学和档案学专业的硕士研究生。

考虑到这些新闻网页的内容均围绕法律和司法主题,笔者将《中文新闻信息分类与代码》一级类目——“法律、司法”下的 15 个二级类目[14]作为分类标准,进行自动分类实验。剔除所有无关网页和重复网页,最终得到 14 353 张有效网页,形成本文的新闻分类标注语料库。一级类目“法律、司法”之二级类目以及各类目的网页分布情况如表 1 所示。

表 1　实验类目及网页分布情况

序号	类别代号	类别名称	网页数量
1	02.01	法制建设	585
2	02.03	法律服务	274
3	02.05	知识产权保护	249
4	02.07	法律、法规、法令	1 290
5	02.08	国际法	93
6	02.11	司法	103
7	02.12	检察	50
8	02.13	审判	371
9	02.14	社会治安综合治理	163
10	02.15	社会公共安全	133
11	02.17	国家安全	17
12	02.21	犯罪、案件	7 406
13	02.23	刑罚、刑罚执行	359
14	02.98	法学研究	597
15	02.99	法律、司法其他	2 663

下文所述之各种自动分类方法和实验均基于该新闻分类标注语料库。

2.2　新闻网页正文抽取

由于中文新闻网页中存在大量噪音内容，例如，导航、广告、相关链接和版权声明等，因此，在进行自动分类之前，必须将其过滤，只保留正文文本，才能提高自动分类的准确率。通过对中文新闻网页进行相关的统计分析，笔者发现这样一个现象：网页正文文本结点的文本密度通常都要远远大于非正文部分，从而提出一种新的网页正文抽取方法——文本密度判别法，该方法的基本原理是：利用贝叶斯判别准则[15]求出文本结点的密度区分阈值，该阈值能够使得发生文本结点误判的平均损失达到最小，然后将各文本结点的文本密度与该密度区分阈值进行比较，大于密度区分阈值的结点就判定为正文文本结点，小于或等于密度区分阈值的结点则判定为非正文文本结点，最后将所有判定为正文文本结点的文本连接起来就是需要抽取的网页正文。

由于本文的重点是自动分类，因此，对如何利用文本密度判别法进行新闻网页正文抽取不再赘述。

2.3　正文文本分词及索引的创建

新闻网页经过正文抽取以后，形成相对“干净”的新闻文本，笔者使用 SharpICTCLAS 分词器[16]对其进行分词，分词词典采用《人民日报分词词表》[17,18]。在正文文本分词的基础上，还需为新闻网页建立全文索引，这就涉及到搜索引擎技术。目前，各商业公司均不愿意将自己的搜索技术公布于众，而搜索引擎又集成了数据库管理、信息检索、自然语言处理、机器学习等诸多技术，因此，要独立开发一个全文搜索引擎系统，其难度可想而知。

Lucene[19]作为一个由 Java 实现的成熟、自由、开源的信息检索软件包，具有清晰的整体架构，强大的索引、分析、过滤、搜索功能和高度的可扩展性，近几年来已经成为最受推崇和青睐的 Java 开源信息检索软件包。最为关键的是 Lucene 已得到 Apache 软件许可协议的授权，并且有多个用不同编程语言实现的版本。因此，笔者基于 C#版的 Lucene[20]为新闻分类标注语料库实现了一个轻量级的新闻全文搜索引擎，索引数据项包括：新闻标题、新闻全文、人工分类类目、新闻发布时间。下文所述的新闻全文搜索引擎如无特殊说明均指该搜索引擎。

3　基于标题的自动分类方法

基于标题的新闻网页自动分类方法是指利用新闻标题作为分类依据的自动分类方法。该方法的核心思想是：通过计算待分类新闻标题与新闻分类标注语料库中每个类目之间的关联度来判断新闻所属的类目，关联度值越大，该新闻属于相应类目的概率也越大，一般的做法是取关联度值最大的类目作为该篇新闻的正式类目。关联度计算公式如(1)式所示：

$$\mathrm{Score}(c,t) = tf_{c,t} \times idf_{c,t} \times w_t \qquad (1)$$

其中，Score(c,t)表示词语 t 相对于新闻类目 c 的关联度；$tf_{c,t}$表示在新闻类目 c 中词语 t 的词频；$idf_{c,t}$表示在新闻类目 c 中词语 t 的逆文档频率；w_t 表示词语 t 的加权因子。

为了获得新闻标题与新闻语料库中每个类目之间的关联度值，先对待分类的新闻标题进行分词和停用词过滤，然后将剩余的词语提交给新闻全文搜索引擎，按照新闻类目自动统计出与这些词相对应的 tf－idf 值，再利用此 tf－idf 值与加权因子的乘积作为自动分类的依据。

由公式(1)可以看出：寻找到加权因子 w_t 的相对较优的加权方式成为实现新闻网页自动分类的关键。本文根据加权因子计算公式的不同，相应地提出以下 4 种方法：词长加权法；简单类目加权法；经典类目加权法；改进的类目加权法。

3.1　词长加权法

词长加权法是指根据新闻标题中词语的不同长度

进行加权。

该方法的具体实现步骤是:先对待分类的新闻标题进行分词和停用词过滤,再将剩余的词提交给新闻全文搜索引擎,进而得到这些词语相对于各个新闻类目的 tf - idf 值,再将这些 tf - idf 值分别乘以对应的词长加权因子,全部加总后分数最高的新闻类目即为该新闻标题应被分类的新闻类目。

借鉴文献[21],笔者定义了词长加权因子的计算公式,如(2)式所示:

$$w_t = \begin{cases} 0 & x = 1 \\ \log_2 x & 2 \le x < 8 \\ 3 & x \ge 8 \end{cases} \quad (2)$$

其中,w_t 表示词 t 的词长加权因子,x 表示词 t 的词长,则关联度计算公式如(3)式所示:

$$Score(c,t) = tf_{c,t} \times idf_{c,t} \times w_t = \begin{cases} 0 & x = 1 \\ tf_{c,t} \times idf_{c,t} \times \log_2 x & 2 \le x < 8 \\ tf_{c,t} \times idf_{c,t} \times 3 & x \ge 8 \end{cases} \quad (3)$$

例如,在新闻语料库中有一篇新闻,原始分类属于“02.07 - 法律、法规、法令”,标题为《江苏常州法院推出措施限制不讲诚信的赖账者》。先对标题进行分词和停用词过滤,然后针对剩余的词语按照词长加权分类法进行处理,得到以下关联度表,如表 2 所示:

表 2 『江苏常州法院推出措施限制不讲诚信的赖账者』词长加权法之关联度表

待分类标题	江苏常州法院推出措施限制不讲诚信的赖账者								
原始分类	02.07 - 法律、法规、法令								
	江苏	常州	法院	推出	措施	限制	诚信	赖账者	合计
02.01 - 法制建设	2.67	5.97	1.29	18.28	5.04	14.63	26.59	0	74.47
02.03 - 法律服务	3.15	54.8	1.12	34.25	6.68	21.08	39.14	0	160.22
02.05 - 知识产权保护	2.77	9.96	1.38	8.03	4.15	16.6	49.8	0	92.69
02.07 - 法律、法规、法令	3.55	4.86	1.22	44.35	10.6	21.04	44.61	5 869.12	5 999.35
02.08 - 国际法	2.58	1.63	1.27	31	23.25	31	0	0	90.74
02.11 - 司法	3.96	10.3	1.11	51.5	3.32	17.17	34.33	0	121.69
02.12 - 检察	3.13	3.85	1.16	16.67	7.14	25	25	0	81.94
02.13 - 审判	2.99	5.71	1.13	26.5	5.89	16.86	41.22	0	100.3
02.14 - 社会治安综合治理	3.13	6.79	1.13	7.41	6.27	23.29	14.82	0	62.84
02.15 - 社会公共安全	3.69	16.63	1.24	66.5	5.78	26.6	44.33	0	164.78
02.17 - 国家安全	1.21	2.43	5.67	5.67	0	17	0	0	31.98
02.21 - 犯罪、案件	4.57	13.03	1.19	17.43	4.83	7.72	18.17	255.58	322.52
02.23 - 刑罚、刑罚执行	2.94	23.93	1.2	359	19.94	39.89	179.5	0	626.41
02.98 - 法学研究	3.73	7.11	1.16	28.43	3.9	5.15	20.59	0	70.07
02.99 - 法律、司法其他	5.17	115.78	1.39	9.48	6.19	8.48	19.02	0	165.52

将关联度按类目进行加总,最终以“02.07 - 法律、法规、法令”类目的关联度总分 5 999.35 为最高分,因此该标题在词长加权法中被分在“02.07 - 法律、法规、法令”类目下。词长加权法的封闭测试实验结果如 4.3 节所示。

经过深入分析后发现该方法存在一个明显的不足:它将新闻标题中词长相同的每个词语简单地视为一律平等,且乘以一样的加权因子,而实际上,这些词语都具有独特的意义,往往与特定的新闻类目相关,即只在特定的新闻类目中出现,因此,考虑到新闻类目对自动分类结果的影响,笔者又设计了以下三种基于新闻类目的自动分类方法。

3.2 简单类目加权方法

简单类目加权法是指在分类的过程中考虑新闻类目对自动分类结果的影响,但加权因子的计算方式比较简单。该方法的核心思想是:由于某些词语只会出现在几个特定的新闻类目中,因此,当这种类型的词语出现时,给予它较大的权值,即加权因子的值。加权因子值的大小由该词语所出现的新闻类目数来判定,计算公式如(4)式所示:

$$w_t = N / N_t \quad (4)$$

其中,N 表示总类目数,本文默认为 15,N_t 表示出现过词语 t 的新闻类目数,则对应的关联度计算公式如(5)式所示:

$$Score(c,t) = tf_{c,t} \times idf_{c,t} \times w_t = tf_{c,t} \times idf_{c,t} \times N/N_t \quad (5)$$

例如,在新闻语料库中有这样一篇新闻,原始分类属于『02.05 - 知识产权保护』,标题为『"稻香村"商标起纷争』。『稻香村』在全部 15 个新闻类目中,只在 6 个类目里出现过,则『稻香村』的加权因子为 15/6 = 2.5。利用词长加权法进行自动分类,该篇新闻被错分到『02.21 - 犯罪、案件』里,而利用简单类目加权法进行处理,得到以下关联度表,如表 3 所示:

表 3　『"稻香村"商标起纷争』简单类目加权法之关联度表

待分类标题	"稻香村"商标起纷争				
原始分类	02.05 - 知识产权保护				
	稻香村	商标	起	纷争	合计
02.01 - 法制建设	1 170	97.5	2.3	146.25	1 416.05
02.03 - 法律服务	0	91.33	1.85	34.25	127.43
02.05 - 知识产权保护	2 468.67	26.45	1.82	139.74	2 636.67
02.07 - 法律、法规、法令	215	21.86	1.85	75.88	314.6
02.08 - 国际法	0	23.25	2.74	0	25.99
02.11 - 司法	0	0	1.87	51.5	53.37
02.12 - 检察	0	0	1.61	0	1.61
02.13 - 审判	0	61.83	2.05	61.83	125.72
02.14 - 社会治安综合治理	0	81.5	1.55	163	246.05
02.15 - 社会公共安全	266	44.33	1.72	0	312.06
02.17 - 国家安全	0	0	2.125	0	2.125
02.21 - 犯罪、案件	124.5	1.5	1.73	22.64	150.37
02.23 - 刑罚、刑罚执行	0	44.88	2.94	0	47.82
02.98 - 法学研究	0	27.14	1.65	54.27	83.06
02.99 - 法律、司法其他	1 775.33	24.66	1.86	71.97	1 873.82

依照新闻类目,将各个词语的关联度值加总,发现『02.05 - 知识产权保护』的总分 2 636.67 为最高分,因此该篇新闻在简单类目加权法中被分在『02.05 - 知识产权保护』类目下。

关于简单类目加权法的封闭实验结果如 4.3 节所示。

3.3　经典类目加权法

简单类目加权法计算加权因子的思想与 idf 类似,仅计算方式不同,因此,笔者借鉴经典的 idf 加权思想又设计出 icf(inverse class frequency)加权模式,即逆类目频率。利用 icf 加权模式对新闻标题中的词语进行加权称为经典类目加权法。加权因子的计算公式如(6)式所示:

$$w_t = \log(N/N_t) \quad (6)$$

则关联度计算公式如(7)式所示:

$$score(c,t) = tf_{c,t} \times idf_{c,t} \times w_t = tf_{c,t} \times idf_{c,t} \times \log(N/N_t) \quad (7)$$

关于经典类目加权法的封闭实验结果如 4.3 节所示。

3.4　改进的类目加权法

由公式(4)、(6)可以看出:经典类目加权法和简单类目加权法在加权思想上是一致的,只是加权方式略有不同,因此,为了寻找到相对较优的加权因子,在此基础上增加另外两种类似的加权公式,如公式(8)、(9)所示:

$$w_t = (N/N_t)^2 \quad (8)$$

$$w_t = (N/N_t)^3 \quad (9)$$

根据导数的数学意义可知:导数就是瞬时变化率,也就是函数在某一点的变化率。对于一元函数来讲,就是自变量在某一点取得一个改变量时,函数将以多大的比例发生改变。因此,为了比较关于 N/N_t 的 4 种加权方式的加权幅度,将加权因子 w_t 看作因变量,N_t 看作自变量,分别对公式(4)、(6)、(8)、(9)求一阶导数,结果分别如公式(10)、(11)、(12)、(13)所示:

$$w_t' = -N/N_t^2 \quad (10)$$

$$w_t' = -1/(N_t \times \ln 10) \quad (11)$$

$$w_t' = -2N^2/N_t^3 \quad (12)$$

$$w_t' = -3N^3/N_t^4 \quad (13)$$

一阶导数的曲线图直接对应原函数的变化趋势大小,因此,根据公式(10)、(11)、(12)、(13)分别画出 w_t' 关于 N_t 的 4 条二维曲线。本文中,N_t 在 1 到 15 之间,相应地 w_t' 的值均较小,如果将四条曲线放在一张二维图上,则很难直观区分各自的变化趋势,如图 1 所示,因此,先对其进行两两比较,再获得全局顺序,如图 2、图 3 和图 4 所示。

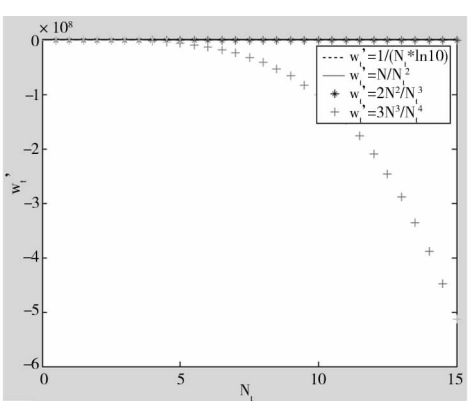

图 1　加权因子变化幅度对照图 1

由图 2、图 3 和图 4 可知 4 种加权因子的加权幅度由大到小的顺序分别为:$(N/N_t)^3 > (N/N_t)^2 > N/N_t$

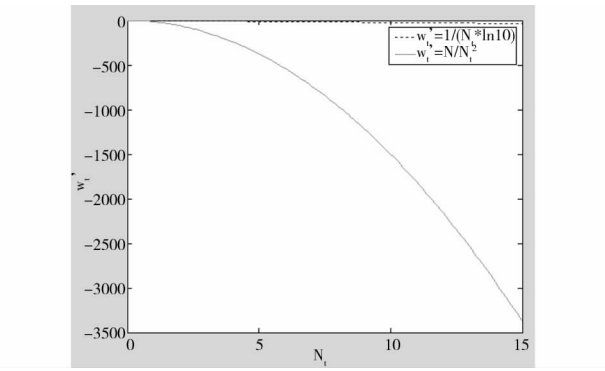

图 2　加权因子变化幅度对照图 2

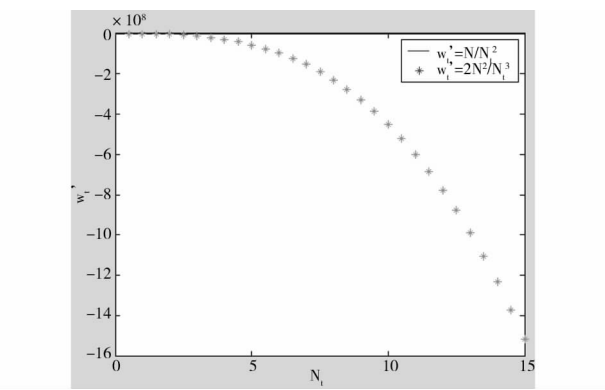

图 3　加权因子变化幅度对照图 3

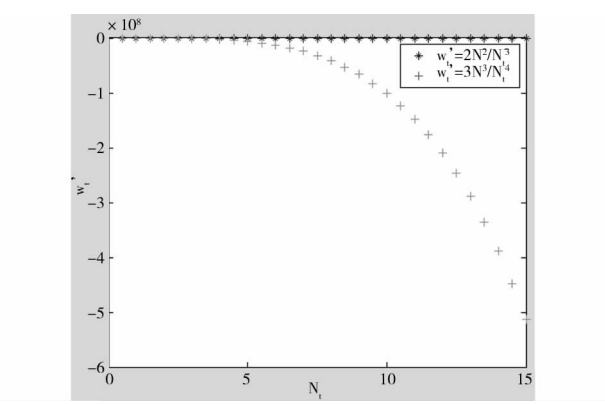

图 4　加权因子变化幅度对照图 4

$>\log(N/N_t)$。对应到实际含义就是：$\log(N/N_t)$ 会降低词语的特殊性，而 $(N/N_t)^2$ 和 $(N/N_t)^3$ 则会拉大特殊词语与常用词语之间的差距。

为了寻找加权幅度较为合适的加权因子，本文利用新闻语料库中的数据对其分别进行封闭实验，具体实验步骤参见 4.2 节，实验结果如表 4 所示。

表 4　加权因子 4 种计算方式的对照性实验结果

w_t	正确率(%)
N/N_t	72.04%
$\log(N/N_t)$	72.54%
$(N/N_t)^2$	73.87%
$(N/N_t)^3$	71.04%

由表 4 可以看出，由于 $\log(N/N_t)$ 降低了词的特殊性，所以造成分类效果不甚理想，而 $(N/N_t)^2$ 则拉大特殊词与常用词之间的差距，所以效果最好，但如果取到三次方 $(N/N_t)^3$，正确率反而会下降，因此，取 $(N/N_t)^2$ 作为改进的类目加权法的加权因子，关联度计算公式如(14)式所示：

$$Score(c,t) = tf_{c,t} \times idf_{c,t} \times w_t = tf_{c,t} \times idf_{c,t} \times (N/N_t)^2 \quad (14)$$

通常情况下封闭实验强调的是模型的拟合性能，因此实验的正确率都会很高，但由表 4 可以看出 4 组封闭实验的正确率并不理想，主要是由于新闻语料库在加工过程中存在两个方面的问题：

(1)分类类目存在不一致的情况，即当某篇新闻属于两个或两个以上类目时，为了统计的简便性，在语料库中被认为只属于一个类目；

(2)人工分类本身就是一项主观性较强的工作，谬误在所难免，从而导致正确率下降。

由于存在过拟合等问题，开放实验的正确率一般比封闭实验要低很多，所以本文在封闭实验准确率不理想的情况下就没有再进行开放实验，为了弥补这一缺陷，笔者进一步提出了前二高分正确率、前三高分正确率，具体计算公式参见 4.1 节。

4　实验结果与分析

为了对比词长加权法、简单类目加权法、经典类目加权法和改进的类目加权法的自动分类性能，用经过预处理的新闻语料库(参见 2.1 节)分别对其进行封闭实验。

4.1　性能评价指标

考虑到性能评价指标的简洁性、直观性和可操作性，笔者没有采用一般意义上的查准率、召回率和 F1 测度值，而是设计了第一高分、前二高分和前三高分共三个指标。

(1)第一高分是指将自动分类过程中关联度最高

的新闻类目视为自动分类的类目,若该类目与新闻语料库中原始定义的类目一致,则判定自动分类结果正确,反之则否。

(2)前二高分是指将自动分类中关联度最高的前两个新闻类目视为自动分类的类目,在这两个类目中只要有一个类目与新闻语料库中原始定义的类目一致,则判定自动分类结果正确,反之则否。

(3)前三高分是指将自动分类中关联度最高的前三个新闻类目视为自动分类的类目,在这三个类目中只要有一个类目与新闻语料库中原始定义的类目一致,则判定自动分类结果正确,反之则否。

相应地,第一高分正确率、前二高分正确率和前三高分正确率的计算公式如(15)、(16)和(17)式所示:

$$第一高分正确率=\frac{系统判别得分最高的类目为正确类目的网页数}{测试网页总数} \quad (15)$$

$$前二高分正确率=第一高分正确率+\frac{系统判别第二高分的类目为正确类目的网页数}{测试网页总数} \quad (16)$$

$$前三高分正确率=前二高分正确率+\frac{系统判别第三高分的类目为正确类目的网页数}{测试网页总数} \quad (17)$$

4.2　实验步骤

本文封闭实验的步骤如下:

(1)对新闻标题进行分词,并去除停用词;

(2)将步骤(1)产生的词汇或短语逐项提交给新闻搜索引擎(参见 2.3),并按照表 1 设定的类目计算关联度值并加总;

(3)对步骤(2)中产生的关联度值进行排序,依次统计第一高分、前二高分和前三高分的正确率;

(4)重复(1),遍历新闻语料库中的所有新闻。

4.3　实验结果

(1)词长加权法

利用经过预处理的新闻语料库对词长加权法进行封闭实验测试,实验结果如表 5 所示。

由表 5 可知:若将第一高分的新闻类目视为原始定义的类目,则正确率为 68.72%;若以前二高分的两个新闻类目来测定,正确率则有 81.74%;以前三高分的三个新闻类目来测定,正确率可达到 86.95%。

(2)简单类目加权法

利用经过预处理的新闻语料库对简单类目加权法进行封闭实验测试,实验结果如表 6 所示。

表 5　词长加权法实验结果数据

词长加权法	正确率
第一高分	68.72%
前二高分	81.74%
前三高分	86.95%

表 6　简单类目加权法实验结果数据

简单类目加权法	正确率
第一高分	72.04%
前二高分	84.76%
前三高分	90.65%

由表 6 可知:若将第一高分的新闻类目视为原始定义的类目,则正确率为 72.04%;若以前二高分的两个新闻类目来测定,正确率则有 84.76%;以前三高分的三个新闻类目来测定,正确率可达到 90.65%。

(3)经典类目加权法

利用经过预处理的新闻语料库对经典类目加权法进行封闭实验测试,实验结果如表 7 所示。

由表 7 可知:若将第一高分的新闻类目视为原始定义的类目,则正确率为 72.54%;若以前二高分的两个新闻类目来测定,正确率则有 86.26%;以前三高分的三个新闻类目来测定,正确率可达到 91.22%。

表 7　经典类目加权法实验结果数据

经典类目加权法	正确率
第一高分	72.54%
前二高分	86.26%
前三高分	91.22%

(4)改进的类目加权法

利用经过预处理的新闻语料库对改进的类目加权法来进行封闭实验测试,实验结果如表 8 所示:

表 8　改进的类目加权法实验结果数据

改进的类目加权法	正确率
第一高分	73.87%
前二高分	87.56%
前三高分	93.53%

由表 8 可知:若将第一高分的新闻类目视为原始定义的类目,则正确率为 73.87%;若以前二高分的两个新闻类目来测定,正确率则有 87.56%;以前三高分的三个新闻类目来测定,正确率可达到 93.53%。

根据表 5、表 6、表 7 和表 8 的数据可以看出:以第一高分作为判别标准,则改进的类目加权法自动分类

结果最好,就新闻语料库而言,第一高分的正确率可达78.87%。前二高分和前三高分的准确率数据同样说明了这一规律。

4.4 结果分析

由于本文重点阐述自动分类方法,因此,直接利用现有的开源资源——中科院分词插件对新闻标题进行分词,关于分词过程中的歧义词识别、新词识别等问题,以下的实验结果未作阐述。

通过对分类错误的新闻网页进行深入分析后发现,主要包括以下几个方面的情况:

(1)单从新闻标题来看,当新闻可属于两种或两种以上的新闻类目时会发生分类错误。

例如,在新闻语料库中有一篇标题为『"毒医生"一审被判死刑』的新闻,原始分类属于"02.13 - 审判",用改进的类目加权法进行处理,被分到"02.21 - 犯罪、案件"类目下,关联度数据统计结果如表9所示。如此分类不能说完全错误,因为从新闻标题本身来看,既可属于"02.13 - 审判"类目,也可属于"02.21 - 犯罪、案件"类目。

表9 『"毒医生"一审被判死刑』关联度表

待分类标题	"毒医生"一审被判死刑						
原始分类	02.13 - 审判						
	毒	医生	一审	被	判	死刑	合计
02.01 - 法制建设	83.57	24.38	11.7	1.87	9.29	18.28	149.09
02.03 - 法律服务	44.89	16.91	3.53	1.25	3.85	15.49	85.92
02.05 - 知识产权保护	49.8	0	3.23	1.32	4.7	0	59.05
02.07 - 法律、法规、法令	34.86	19.85	7.17	1.48	6.48	18.17	88.01
02.08 - 国际法	15.5	0	2.74	1.27	9.3	3	31.81
02.11 - 司法	17.17	103	6.44	1.61	5.72	17.17	151.1
02.12 - 检察	0	0	5.56	1.39	3.33	12.5	22.78
02.13 - 审判	91.33	10.15	7.83	1.59	4.89	54.8	170.6
02.14 - 社会治安综合治理	61.83	23.19	3.79	1.71	3.67	8.43	102.62
02.15 - 社会公共安全	14.78	26.6	6.33	1.4	6.05	19	74.16
02.17 - 国家安全	0	0	3.4	1.21	5.67	0	10.28
02.21 - 犯罪、案件	27.17	81.5	13.58	1.39	7.76	54.33	185.74
02.23 - 刑罚、刑罚执行	59.83	16.32	2.14	1.14	2.76	6.41	88.6
02.98 - 法学研究	28.43	12.98	5.28	1.31	4.09	22.11	74.21
02.99 - 法律、司法其他	34.14	15.13	10.83	1.42	7.5	21.48	90.5

(2)当新闻标题太短,或者标题中的词语不具有明显类目特征,也会造成分类错误。

比如,在新闻语料库中有一篇标题为『7月份买的房10月份就要拆』的新闻,原始分类属于"02.21 - 犯罪、案件",用改进的类目加权法进行处理,被分到"02.13 - 审判"类目下,如表10所示。通过察看新闻的具体内容得知,该篇新闻主要讲述因拆迁纠纷产生的诉讼案件,而从新闻标题很难得到这样的信息,因此造成分类错误。

表10 『7月份买的房10月份就要拆』关联度表

待分类标题	7月份买的房10月份就要拆					
原始分类	02.21 - 犯罪、案件					
	7月份	买	房	10月份	拆	合计
02.01 - 法制建设	0	12.45	18.87	0	83.57	114.89
02.03 - 法律服务	0	8.56	12.45	0	68.5	89.51
02.05 - 知识产权保护	0	12.45	49.8	0	0	62.25
02.07 - 法律、法规、法令	0	5.66	9.21	0	44.48	59.35
02.08 - 国际法	0	21.82	13.74	0	92.75	128.31
02.11 - 司法	0	5.21	8.41	0	44.35	57.97
02.12 - 检察	0	25	25	0	0	50
02.13 - 审判	0	23.25	46.5	0	93	162.75
02.14 - 社会治安综合治理	0	5.26	9.59	0	18.11	32.96
02.15 - 社会公共安全	0	7.39	12.09	0	22.17	41.65
02.17 - 国家安全	0	8.5	8.5	0	0	17
02.21 - 犯罪、案件	0	51.5	34.33	0	51.5	137.33
02.23 - 刑罚、刑罚执行	0	9.97	21.12	0	71.8	102.89
02.98 - 法学研究	0	5.69	10.66	0	25.96	42.31
02.99 - 法律、司法其他	0	4.13	7.93	0	24.89	36.95

(3)新闻语料库中原始手工分类错误,但自动分类结果正确。

例如,在新闻语料库中有一篇标题为『雇工摔伤后工头携款逃跑』的新闻,原始分类属于"02.01 - 法制建设",用改进的类目加权法进行处理,被分到"02.21 - 犯罪、案件"类目下,如表11所示。

通过察看该篇新闻的内容后发现:原始手工分类结果明显错误,自动分类结果正确。

(4)新闻语料库中原始手工分类错误,自动分类结果也错误。例如,在新闻语料库中有一篇标题为『盗窃窨井盖处罚难让人忧心』的新闻,原始分类属于"02.01 - 法制建设",用改进的类目加权法进行处理,被分到"02.21 - 犯罪、案件"类目下,如表12所示。

通过察看新闻的具体内容后发现:该篇新闻主要讨论盗窃窨井盖应该判何罪,并对国内的相关判例作出分析,因此,应该分到"02.98 - 法学研究"类目中,即原始手工分类和自动分类结果均错误。

表 11　『雇工摔伤后工头携款逃跑』关联度表

待分类标题	雇工摔伤后工头携款逃跑							
原始分类	02.01 – 法制建设							
	雇工	摔	伤	工头	携	款	逃跑	合计
02.01 – 法制建设	585	73.13	16.71	292.5	117	7.31	83.57	1 175.22
02.03 – 法律服务	274	45.67	9.79	274	91.33	3.56	91.33	789.68
02.05 – 知识产权保护	249	0	49.8	0	249	7.11	0	554.91
02.07 – 法律、法规、法令	645	58.64	12.06	430	161.25	5.4	80.625	1392.97
02.08 – 国际法	46.5	0	0	0	23.25	13.29	13.29	96.32
02.11 – 司法	0	14.71	17.17	0	103	5.42	0	140.3
02.12 – 检察	0	0	0	0	0	4.55	0	4.55
02.13 – 审判	0	185.5	14.84	371	0	9.76	41.22	622.32
02.14 – 社会治安综合治理	0	81.5	23.29	163	163	7.41	54.33	492.53
02.15 – 社会公共安全	133	22.17	9.5	133	44.33	8.87	22.17	373.03
02.17 – 国家安全	0	0	8.5	0	0	0	0	8.5
02.21 – 犯罪、案件	1331.5	38.04	9.68	190.21	61.93	6.36	88.77	1 726.49
02.23 – 刑罚、刑罚执行	0	0	15.61	359	44.88	6.77	27.62	453.87
02.98 – 法学研究	597	66.33	10.12	597	66.33	3.93	45.92	1 386.64
02.99 – 法律、司法其他	925.75	48.41	12.04	370.3	60.7	4.99	44.08	1 466.27

表 12　『盗窃窨井盖处罚难让人忧心』关联度表

待分类标题	盗窃窨井盖处罚难让人忧心							
原始分类	02.01 – 法制建设							
	盗窃	窨	井盖	处罚	难	让人	忧心	合计
02.01 – 法制建设	39	780	1 170	9.44	5.09	1.64	2 632.5	4 640.53
02.03 – 法律服务	54.8	0	0	12.45	5.59	1.63	0	76.8
02.05 – 知识产权保护	249	0	0	9.58	4.88	1.55	0	269.09
02.07 – 法律、法规、法令	21.15	245.71	224.35	6.94	5.16	1.4	3 870	4 377.08
02.08 – 国际法	23.25	0	372	13.29	11.63	1.16	0	424.06
02.11 – 司法	20.6	0	0	9.36	4.48	1.37	0	38.46
02.12 – 检察	12.5	0	0	5	6.25	1.61	0	28.49
02.13 – 审判	24.73	0	0	8.83	6.87	1.77	0	45.77
02.14 – 社会治安综合治理	54.33	0	0	7.76	4.94	1.43	0	70.61
02.15 – 社会公共安全	14.78	266	266	10.23	9.5	1.29	0	570.41
02.17 – 国家安全	0	0	0	0	5.67	1.06	0	8.15
02.21 – 犯罪、案件	15.86	1 742.59	1 410.67	7.59	8.21	1.32	9 522	12 710.79
02.23 – 刑罚、刑罚执行	10.88	0	0	4.27	25.64	1.55	0	47.01
02.98 – 法学研究	19.26	1 194	2 388	7.11	3.26	1.21	5 373	8 988.05
02.99 – 法律、司法其它	49.31	3 550.67	1 521.71	13.18	3.84	1.379	3 994.5	9 136.61

5　结　语

本文借鉴 tf – idf 的思想构建基于标题的中文新闻网页自动分类方法，并设计多个实验对各种基于标题的自动分类方法进行评测。实验结果表明：基于标题对中文新闻网页进行分类可以大大缩短判断处理的时间，且准确率较高，特别是改进的类目加权法效果最好。

当然，该方法也有不足及需要进一步改进与优化的地方，主要包括以下两个方面：

（1）创建新闻分类词典：参考《新闻英语分类词典》，按照《中文新闻信息分类与代码》中的新闻类目创建常用词语统计表，利用这个统计表作为新闻分类的一项依据，如此应该可以提高自动分类的准确率。

（2）进一步优化中文新闻网页分类标注语料库：本文的分类标注语料库包含的新闻网页数量有限，而且分布不均衡，手工分类的质量也有待提高。例如，可以请相关行业的专家对分类标注语料库进行二次加工，并对网页数量较少的类目进行扩充，从而将其对自动分类准确率的影响降到最低。

（致谢：感谢为本文加工新闻分类标注语料库的各位老师和同学，他们是：大众点评网的黄卫堂、南京财经大学的安艳杰、金陵科技学院的许剑颖、南京工业大学的冯桂珍、南京大学的梁勇、高霄云、顾婷婷、宋伟萍。）

参考文献：

[1] Fuchun P, Schuurmans D, Shaojun W. Augmenting Naive Bayes Classifiers with Statistical Language Models [J]. *Information Retrieval*, 2004(7):317 - 345.

[2] 秦兵，郑实福，刘挺，等. 可分性判据在中文网页分类中的应用[J]. *微处理机*，2002(1):26 - 28.

[3] Joachims T. Text Categoriztion with Support Vector Machine: Learning with Many Relevant Features [C]. In: *Proceedings of the European Conference on Machine Learning* (*ECML* - 98), Chemnitz. Germany, 1998: 137 - 142.

[4] Joachims T. Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms [M]. Boston: Kluwer Academic Publishers, 2002:1 - 176.

[5] Rung - Ching C, Chung - Hsun H. Web Page Classification Based on a Support Vector Machine Using a Weighted Vote Schema[J]. *Expert Systems with Applications*, 2006, 31(2): 427 - 435.

[6] Yiming Y, Liu X. A Re - Examination of Text Categorization Methods[C]. In: *Proceedings of the* 22*th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999: 42 - 49.

[7] Jyh - Jong T, Wang Jing - Doo. Improving Automatic Chinese Text Categorization by Error Correction[C]. In: *Proceedings of the* 5*th International Workshop Information Retrieval with Asian Languages*, 2000: 1 - 8.

[8] 邓茜，林红. 中文新闻信息自动分类标引的构想与实现[J]. *中国传媒科技*，2005(9):19 - 21.

[9] 侯汉清，薛鹏军. 基于知识库的网页自动标引和自动分类系统的设计[J]. *大学图书馆学报*，2004，1(9):50 - 55,64.

[10] 何琳，侯汉清，白振田，等. 基于标引经验和机器学习相结合的多层自动分类[J]. *情报学报*，2006，25(6):725 - 729.

[11] 姜远，周志华. 基于词频分类器集成的文本分类方法[J]. *计算机研究与发展*，2006，43(10):1681 - 1687.

[12] 搜狗实验室. 文本分类语料库[EB/OL]. [2008 - 07 - 20]. http://www.sogou.com/labs/dl/c.html.

[13] 北京大学网络实验室. 中文网页分类训练集[EB/OL]. [2008 - 07 - 20]. http://www.cwirf.org/2006WebTrack/YQ - CCT - 2006 - 03.tgz.

[14] 中华人民共和国国家质量监督检验检疫总局，中国国家标准化管理委员会. GB/T 20093 - 2006 中文新闻信息分类与代码[S]. 北京：中国标准出版社，2006.

[15] 高惠璇. 应用多元统计分析[M]. 北京：北京大学出版社，2005:183 - 191.

[16] 吕震宇. SharpICTCLAS 分词系统[EB/OL]. [2008 - 04 - 10]. http://www.cnblogs.com/zhenyulu/category/85598.html.

[17] 中国科学院计算技术研究所. 汉语词法分析系统 ICTCLAS [EB/OL]. [2008 - 04 - 10]. http://www.i3s.ac.cn/index.htm.

[18] 詹卫东. 中文信息处理基础[EB/OL]. [2008 - 04 - 10]. http://ccl.pku.edu.cn/doubtfire/Course/Chinese%20Information%20Processing/2002_2003_1.htm.

[19] Apache. Lucene [EB/OL]. [2008 - 04 - 10]. http://lucene.apache.org/.

[20] Apache incubator. Lucene .Net [EB/OL]. [2008 - 04 - 10]. http://incubator.apache.org/lucene.net/.

[21] Dell Z, Yisheng D. Semantic, Hierarchical, Online Clustering of Web Search Results[C]. In: *Proceedings of the* 6*th Asia Pacific Web Conference* (*APWEB*), Hangzhou. 2004: 69 - 78.

（作者 E - mail:happyfate2001@yahoo.com.cn）