面向中图法的学术文献自动分类研究

赵纪元　罗霄[1]

同方知网（北京）技术有限公司，北京，100084

E-mail: jy_zhao@hotmail.com

摘　要：本文结合自然语言处理中文本分类的理论，面向大量的学术期刊，研究了基于中图法的学术文献自动分类方法。该方法结合了 CHI 特征选择、后验概率训练以及 tf/idf 概率加权等方法，实现了对 500 余万篇学术期刊的自动分类。对中图法 37 个大类 5 万余子类的分类，在输出比例为 20%的情况下，准确率达到了 78%。同时研究了以二元词汇作为特征进一步修正上述结果，在保证正确率基本不变的情况下，使分类的输出比例大大提升。

关键词：文本分类，中图法，二元分类

Research on Academic Text Categorization Oriented to CLC

ZhaoJiyuan, LuoXiao

Tongfang Knowledge Network Technology (Beijing) Co.,Ltd., Beijing, 100084

E-mail: jy_zhao@hotmail.com

Abstract: This paper introduces methods in text categorization for large quantities of academic literature, which oriented to CLC and based on the theory of text categorization in natural language processing. The method combines the CHI feature selection, posterior probability training, as well as improved tf/idf method. It implements automatic classification for more than 5 millions academic journals. The precision rate has reached 78%, when the output ratio is 20%, according to 37 classes and more than 50,000 sub-classes in CLC. At the same time, binary terms are used in this method, for further amendments to the classification results. As a result, the output ratio has been greatly enhanced, with little change to the precision rate.

Keywords: text categorization, CLC(The Chinese library classification), text categorization using binary terms.

1 引言

《中国图书馆分类法》简称《中图法》，国内主要大型书目、检索刊物、机读数据库，以及《中国国家标准书号》等都著录《中图法》分类号[1]。目前，各图书馆及数字图书馆对于中文学术期刊的分类均按《中图法》实行。分类号的获取主要依靠编辑的人工审核，不但耗费了大量的人力物力，而且效率较低。

基于上述现状，本文针对学术文献的语料特点，设计了结合 CHI 特征选择、后验概率训练以及 tf/idf 概率加权等方法的自动分类系统。提出置信度定义，在分类结果中将高准确率结果提取出来，直接替代人工标注。希望以此提升学术文献数据加工的效率，减少人力浪费。

---

[1] 作者简介：赵纪元，女，1982年生，工程师，硕士，主要研究方向是文本分类和语义词典. 罗霄，男，1978年生，高级工程师，硕士，主要研究方向是自然语言处理.

2 特征选择算法

数量巨大的训练样本和过高的向量维数是文本分类的两大特点。为了兼顾运算时间和分类精度两个方面，我们不得不进行特征选择，力求在不损伤分类性能的同时达到降维的目的[3]。常用的特征选择方法有文档频率方法(DF)，信息增益方法(IG)，互信息方法(MI)，CHI 方法，期望交叉熵，文本证据权优势率[4]，基于词频覆盖度[5]的方法，主分量分析[6]方法等。

本研究中，通过对 500 万语料的统计分析，并结合中图分类号的特点，以及学术文献的存储特点，在特征选择时采取了 CHI 的方法。具体如下。

2.1 确定原始词典

本文采用短语作为分类特征项，从语义上来说，一般短语作为类别的特征项比词更好。本文的短语库是利用了同方知网（北京）技术有限公司以前的工作成果，该短语库是自动从中国知识基础设施工程语料库（CNKI）中自动抽取出来，并经过了清洗，规模约为 200 万左右的词条[7]。

2.2 利用 CHI 值进行特征项选择

CHI 方法：CHI 统计方法度量词条 t 和文档类别 c 之间的相关程度，并假设 t 和 c 之间符合具有一阶自由度的 χ^2 分布。具体公式如下：

$$\chi^2(t,c)=\frac{N(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \qquad \text{（公式1）}$$

N：训练语料中的文档总数；c：某一特定类别；t：特定的词条；A：属于 c 类且包含 t 的文档频数；B：不属于 c 类但是包含 t 的文档频数；C：属于 c 类但是不包含 t 的文档频数；D：既不属于 c 也不包含 t 的文档频数。

经过上述筛选，保留了约 180 万的专业词汇作为本研究中分类需要的特征词语。以下将通过对特征词语的训练得到词语与分类号的关系，并进一步应用到分类中。

3 分类算法

目前，常用的分类算法有：朴素贝叶斯分类；K 近邻（KNN）；决策树；支持向量机（SVM）等[8]。由于本文所用的语料均存储于数据库，研究中通过数据库检索，统计了训练语料中词语和分类号的概率分布，并结合词语在文中的位置进行了加权，形成了加权后的后验概率结果。

在分类阶段设计了利用 tf/idf 的概率加权的分类方法。该方法充分利用了学术文献的格式特点，考虑了特征词语在文献中不同位置的权重，能够充分利用特征词和类别的关系，从而达到了比较好的分类结果。具体介绍如下。

3.1 训练算法

对输入的特征词语 W，训练其对于不同分类号的权重。

在训练语料中，设词语 W 出现在：篇名、中文关键词、中文摘要或全文时，有 m 篇文章，

他们对应的分类号有 n 种，分别为：C1，C2…Cn。

把同一分类号在不同位置的出现的权重设为：篇名：posWeight=4；中文关键词：posWeight=2；中文摘要：posWeight=1.5；全文：posWeight=1。

每个分类号对所有文章、所有位置的权重取和，公式如下：

$$weight(Ci)=\sum_{k=1}^{m}\sum_{j=1}^{4}posWeight(Ci) \qquad \text{（公式 2）}$$

分类号 Ci 的最终权重计算如下：

$$WeightTrain(W,Ci)=\frac{weight(Ci)}{\sum_{i=1}^{n}weight(Ci)}*\ln 2 \qquad \text{（公式 3）}$$

其中，weight(Ci)是类别 Ci 的权重，分母是所有分类号的权重和，乘以 ln2 作平滑。此结果分类号训练的最终结果。它体现了对于词语 W，其可能出现的分类号的概率大小，WeightTrain(W,Ci)越大，说明词语 W 对应类别 Ci 的可能性越大。

训练完成后，形成了 180 万特征词语的分类词典，存储了词语及其可能对应的前几个分类号和分类号的权重，以便用于后续分类的查找和计算。

3.2 分类算法

1）词语权重的计算

首先，利用 tf/idf 计算词条 w 权重，研究中对标准 tf/idf 公式进行了改进，引入了词语长度和词语在文章中的位置信息，具体如下。

$$Weight(w)=\log(L+1)\times \mathrm{TF}_i\times\log(\mathrm{D}/DF(W_i)+0.01) \qquad \text{（公式 4）}$$

其中，L：词 W 的长度；TF_i：W 在待处理文档中出现的频度；D：训练总文档数目；$DF(W_i)$：词在其中出现至少一次的训练文档数目。

改进后的词语 w 权重计算公式：

$$Weight'(w)=\sum_{i=1}^{5}Weight(w)*\alpha \qquad \text{（公式 5）}$$

α 根据词语在文中的不同位置，不同种类的文献，取不同的值。例如，当词语出现在标题、关键词中，α=30；出现在摘要，α=20；出现在正文第一段中，α=2。由上述公式，即可结合词语 w 的词频、文档频度、长度以及在文中的位置信息，得到该词语的权重，而整篇文章对于某个分类号的权重就由文中每个词语的权重利用该词和分类号的对应概率加权得到。

2）文章权重的计算

一篇文章对于类别 C 的权重计算为：设文章在类别 C 下有 n 个特征词，分别为 w1,w2…wn，每个词语对应类别 C 的训练概率为 WeightTrain(wi)，则文章对于类别 C 的权重如下：

$$Weight(C)=\sum_{i=1}^{n}WeightTrain(wi,C)*Weight'(wi) \qquad \text{（公式 6）}$$

最后，利用总权重将各类别的权重归一化，取权重最大的类别作为文章分类结果，至此，便得到了待分类文章的类别。

3）置信度的计算

以往，对于学术期刊的中图分类号标注全部由编辑手工完成，在数据加工过程中耗费了大量的人力物力。为适应实际工作的流程，便于自动分类结果和人工标注结合，本研究设计了置信度的计算方法，把分类结果根据置信度大小分为了高准确率结果集和低准确率结果集。其中，高准确率结果的准确率要求达到 80%左右，这部分分类结果可直接输出，并存入期刊数据库，替代原来的人工操作，其余低准确率结果将由人工辅助完成。因此，高准确率结果的比例越大，在实际工作中能够替代的人力劳动就越多，节省的生产成本也就越多。

置信度定义如下：

$$\text{置信度} = \frac{\text{输出分类号权值}}{\text{所有结果分类号权值之和}} \qquad \text{（公式 7）}$$

每一篇文献，计算机都能给出多个中图分类号，每一个分类号都有相应的权重，对中图分类号按照权重由高到低排序，通过一系列的对比实验，我们设定如下规则：

a) 第一个分类号权重/所有分类号权重和>α

b) 第一个分类号权重/所有分类号权重和<=α 且 (第一个分类号权重+第二个分类号权重)/所有分类号权重和>α

当分类号权重满足上面两个规则之一时，我们认为该分类结果是较好的结果。放入高准确率结果集[7]。其中置信度阈值的选取是根据实验结果确定的，并根据个别类别加入了辅助规则。

4 二元分类的引入

为进一步提升机标分类号的标注比例，提高工作效率，本研究继续引入了“二元分类”的思想，希望在保证准确率基本不变的前提下，进一步增加高准确率结果，扩大标注范围[10][11]。

在对以往结果分析的基础上，发现有 20%的分类错误是由于对类别意义的区分不准确造成的，这些类别有很大的相似之处，仅使用一元的特征词很难区分，因此本研究考虑以二元词对作为特征，在上述一元分类的基础上，进一步提升准确率和输出比例。

由于由一元特征变为二元特征，将引起特征数量成平方级的增长，因此二元的特征选择变得尤为重要。目前，主要采取每篇文章的机标关键词进行两两组合，构成二元词对。

4.1 特征选择

特征选择考虑因素包括：机标关键词的权重；词语在类内的 df；词语在全部文献的 df。

$$weight(term) = \lambda 1 * weight(key) + \lambda 2 * \frac{\log(df(C) + 0.01)}{\log(df(all) + 0.01)} \qquad \text{（公式 8）}$$

其中，Weight(term)：特征词选择时的权重；Weight(key)：机标关键词的权重；df (C)：在本类所有文章中，该机标关键词一共出现的次数；df(all)：该机标关键词在训练语料中的文档频度。$\lambda 1$ 取 0.8，$\lambda 2$ 取 0.2。选择时将 df(all)=1 的词语过滤掉，如果 log (df(all)) / log(DF) >0.5，则是比较常用的词，也过滤掉。DF：指输入语料的全部文档数。

4.2 二元训练

基本思路：与原始算法类似，把训练字段设为只有机标关键词，查询条件由一个词语变为两个词语同现。利用公式 2，3 计算二元词对和分类号的关系。

在研究中，取 2007 年全年期刊约 135 万篇的机标关键词作为二元特征词的初始范围，根据二元特征选择算法的排序，每篇文章取前 5 个机标关键词，这样初始词汇约有 675 万。将这些词汇组合成二元词对，经过特征选择和二元训练的筛选，最终保留了约 900 万的二元特征词对，有效的控制了二元特征的维度，使分类计算便于应用。

4.3 二元分类

1）二元词对在文中的权重计算

使用两个机标关键词的权重相乘后开方，公式如下：

$$Weight(w1,w2)=\sqrt{keyWeight(w1)\times keyWeight(w2)} \qquad \text{（公式 9）}$$

其中，keyWeight(w1)表示机标关键词 w1 的权重，由机标关键词算法得到（对本研究是已知的），keyWeight(w2)同理。

2）分类公式

基本思路：与原始算法类似，将文中出现的二元词对权重，使用概率加权，再利用总的类别权重归一化。利用公式 6，7 计算，并把其中一个词语的权重变为由公式 10 计算的两个词语的权重，词语与分类号的概率关系也使用相应的二元训练结果。

3）一元二元加权

分别利用一元和二元分类算法，算出每个类别的权重后，将一元二元的结果加权，公式如下：

$$w=\alpha * w1+\beta * w2 \quad \text{，其中 } \alpha+\beta=1 \qquad \text{（公式 10）}$$

经实验，确定 α=0.7，β=0.3

5 实验及结论

5.1 评价标准定义

由于本研究目的在于，在保证自动分类正确率的基础上，尽可能的扩大标注比例，因此，引入了置信度的计算和高低准确率结果集的区分。本文把置信度大于阈值的结果叫做“高准确率结果”或“输出结果”。实验结果的评价主要集中在对输出结果的考察。采用了以下几个定义：

1）输出比例：置信度较高的结果集占总测试集的比例。

$$\text{输出比例}=\frac{\text{置信度大于阈值的结果数目}}{\text{所有测试文章数目}} \qquad \text{（公式 11）}$$

2）输出正确率：指在输出的高准确结果集中，分类的正确率

$$\text{输出正确率}=\frac{\text{输出结果中正确的数目}}{\text{所有输出的文章数目}} \qquad \text{（公式 12）}$$

可见输出比例越大，输出正确率越高，自动分类的效果就越好，其实用性就越强。

5.2 一元自动分类结果

实验使用2007年全年的期刊文献135万篇作为训练语料，使用2007年4万篇期刊作封闭测试，结果如表1所示。

表1 一元自动分类封闭测试结果

输出比例	输出正确率
20%	78%

由于中图分类号是有层次的，例如：A841.2，A为第一层，8第二层，4第三层，1第四层，2第五层。对上述结果按照中图分类号的不同层次统计，正确率如表2所示。

表2 一元自动分类按层统计结果

层级	输出正确率
第四层	78%
第三层	82%
第二层	85%
第一层	92%

可见，在仅使用一元分类的情况下，高准确率结果占20%，封闭测试的准确可达78%，其中第一层分类号的正确率最高，可达92%。

在以上统计结果中，选取准确率低于10%的分类号，按文章数降序排列，选取前30篇人工检查，总结错误原因如表3所示。

表3 一元分类的错误原因统计

错误原因	比例	举例	改进方法
被误分为数目较多的类	60%	F830，被误分为F224,F832；训练语料中F224,F832数目较多	平滑特征词的分布，改善特征选择
类别意义不好区分	20%	G444（学生心理学），被误分为B844（人类心理学）,B844.2（青少年心理学）	增加其他方向的区分词
年份不同，分类程度不同	20%	TB383被分为TB383.1，03-05年TB383比TB383.1多，06-08年TB383.1多	统一标准，按照标准执行

因此，一元分类出错的主要原因有：文章数和特征词语数较少的类别（低频类），被分为文章数较多的类别（高频类），因为高频类的特征词较多，造成了特征词的分布不均匀；另外，对于某些分类号，仅给出一个层面上的特征还不够使分类准确，需考虑从多角度选择特征。

以上这些问题，可通过二元特征的引入和特征的重新筛选进一步解决，具体实验如下。

5.3 二元自动分类结果

按照本文第四节介绍的算法进行二元实验，使用2007年全年的期刊文献135万篇作为训练语料，使用2007年前1万篇期刊作封闭测试，2005年前1万篇期刊作开放测试，结果表4所示。

表4 二元自动分类测试结果

	分类方法	总体正确率	输出比例	输出正确率
封闭测试	一元分类	15.46%	20.62%	76.81%
	二元分类	35.66%	42.76%	87.59%
开放测试	一元分类	12.82%	21.70%	60.60%
	二元分类	20.48%	37.74%	58.06%

由上述结果可见，引入了二元分类后，无论是封闭测试还是开放测试，输出比例有了较大提高，几乎翻倍；总体正确率也有较大提高；封闭测试输出正确率有所提升，开放测试的输出正确率稍有下降，但变化不大。因此，二元特征的引入能够保证在正确率基本不变的情况下，对输出比例的提升有较明显的效果，这对实际应用有着重要的意义。

此外，可以看到二元对封闭测试的正确率提升较大，对开放测试则不明显，这是由于二元特征的选取过分依赖训练语料造成的，在今后的研究中，可考虑扩大二元词对的选择范围，进一步精确特征选择算法，以期待达到更好的效果。

5.4 结论

本文描述了基于中图法的学术文献自动分类算法，实现了对中图法37个大类5万余子类的分类。其中一元分类的输出比例为20%，准确率达到了78%。同时，研究了以二元词汇作为特征进一步修正上述结果，在保证正确率基本不变的情况下，输出比例提升16%以上。今后的研究重点将主要集中在：二元特征的精确筛选，辅助规则的补充等方面，相信能够使自动分类的结果有进一步的提升。

参 考 文 献

[1] 刘延章等.简析《中图法》第四版之修订[J].郑州大学学报(社会科学版),2000, (3):113-116.

[2] Yang Yiming, Pederson J O. A Comparative Study on Feature Selection in Text Categorization [A]. Proceedings of the 14th International Conference on Machine learning[C]. Nashville :Morgan Kaufmann,1997:412-420.

[3] 周茜,赵明生.中文文本分类中的特征选择研究[J].中文信息学报,2004,18(3):17– 23.

[4] Mlademnic ,D., Grobelnik.M. Feature Selection for unbalanced class distribution and Nave Bayees[A] . Proceedings of the Sixteenth International Conference on Machine Learning[C]. Bled Morgan Kaufmann,1999 :258-267.

[5] 王梦云,曹素青.基于字频向量的中文文本自动分类系统[J].情报学报,2000,19(6):644-649.

[6] Y. Yang. Noise reduction in a statistical approach to text categorization[A].Proceedings of the 18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIRp95)[C]. Seattle:ACM Press,1995:256- 263.

[7] 孙雄勇,罗霄.中图分类法体系下的自动分类研究. 第四届全国信息检索与内容安全学术会议论文集(上)[C], 2008:604-609.

[8] 代六玲,黄河燕,陈肇雄.中文文本分类中特征抽取方法的比较研究[J].中文信息学报,2004 ,18(1) : 26 – 32.

[9] 寇莎莎,魏振军.自动文本分类中权值公式的改进[J].计算机工程与设计,2005,26(6):1616-1618.

[10] 王映,常毅,谭建龙,白硕.基于N元汉字串模型的文本表示和实时分类的研究与实现[J].计算机工程与应用,2005,5:88-91.

[11] 樊兴华,孙茂松.一种高性能的两类中文文本分类方法[J].计算机学报,2006,29(1):124-131.