基于《中图法》的多层自动分类影响因素分析

何　琳　刘　竟　侯汉清

摘　要　系统总结基于《中图法》知识库的多层自动分类项目的研究经验,分析训练数据、特征词选择、分类算法、类目体系和评估方法等因素对多层自动分类的影响。围绕《中图法》,对自动分类的适应性、稀有类别的处理、知识库更新、明显正确或错误数据的标注、标准数据集的制定等进行探讨。图4。表2。参考文献9。

关键词　中图法　多层分类　文本分类　影响因素

分类号　G254.24

ABSTRACT　In former automatic text classification research, most of the prevalent classification technologies divided texts into several classes at one layer. However, with the increased quantities of information retrieval, this flat organizational classification is more and more unsuitable to the information retrieval task. This paper tries to analyze the impact factors in the multi-layer classification, including training data, classification algorithms, classification hierarchy systems and evaluation method. It also discusses the main difficulties and potential solutions in the multi-layer text classification. 4 figs. 2 tabs. 9 refs.

KEY WORDS　CLC.　Multi-layer hierarchy.　Automatic classification.　Impact factors.

CLASS NUMBER　G254.24

文本自动分类是信息检索与数据挖掘领域的研究热点与核心技术,近年来得到了广泛的关注和快速的发展。文本分类技术应用广泛,在分类粒度较大、类目数量不多的行业分类中已逐渐趋于实用;但随着相关应用的发展及需求的不断提升,对于分类粒度小、类目数量庞大的多层分类仍存在很多值得研究的问题,而其恰恰又是图书馆和信息机构处理文本更需要的。本文结合基于《中国图书馆分类法》(以下简称《中图法》)的多层自动分类项目组多年来的研究成果对多层自动分类的影响因素作一些实践探讨。

1　多层自动分类的意义

以往的文本分类体系大多都是面向行业的浅层粗略分类体系,随着网络信息资源数量的激增,文本的种类越来越多,人们对多层次信息组织方式的需求也随之提高。与此同时,图书情报部门每天需要对大量的图书、报纸、期刊分类和标引,这些工作如果能够自动完成,将会节省大量的人力和物力,提高信息服务的效率。

在文本分类领域,大多数研究都集中在浅层的粗略分类体系,所定义的类别数量有限,类别之间基本是孤立的,没有任何结构关系。当类别数量激增时,该分类方法则无法满足要求。尤其是像《中图法》这样类目详尽的分类体系,类目之间的语义相关性很大,基本类目多达五六万个,复分、仿分后所派生出来的类目更是数量庞大,浅层的粗糙分类体系无法满足应用的需求。

采用机器学习方法对以《中图法》为分类体系的信息资源分类,需要对《中图法》的每个类目进行统计训练,必然会使得训练的特征维数过大,导致运算量庞大;此外,这些文献资源在各个类目间的数量分布极不均衡,使得有些文献分布较少的类目难于分类。由于《中图法》分类体系的类目数目庞大和类下文献分布极不均衡,导致机器学习算法难于适应以《中图法》为分类体系的信息资源的自动分类。

2 基于《中图法》的多层自动分类原理[1-2]

《中图法》是我国使用最为广泛的大型综合性分类法,图书情报部门每天标引大量的书目记录,需要巨大的人力物力投入。基于为图书情报部门解决实际工作困难的考虑,本项目开发的以《中图法》为分类体系的多层自动分类系统已投入到上海图书馆《全国报刊索引》的自动标引和自动分类工作中。

基于《中图法》的多层自动分类所采用的分类方法是基于情报检索语言中分类号、主题词和关键词三者之间兼容互换的原理,将《中图法》所有类目的众多标引实例进行数据挖掘,采用基于人工标引经验和机器学习相结合的分类算法,利用图书情报部门的标引实例构建分类知识库(规则库),利用词串定类的方法,通过计算待分类文本的关键词串与分类知识库的相关度度量来为文本进行分类。分类过程是将待分类文本经过主题标引行程的标引词串与分类知识库中的规则进行最佳匹配的过程。系统的大致流程如图1所示。

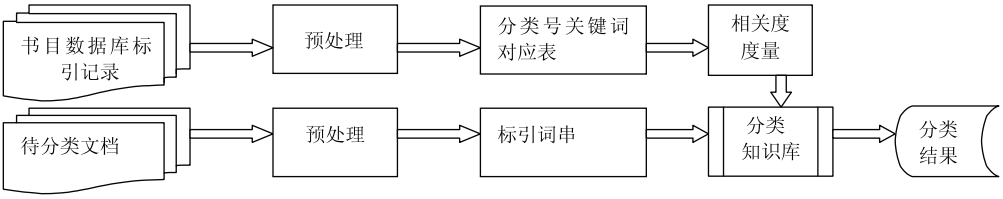

图1 基于标引经验的自动分类系统训练和分类流程

3 自动分类的影响因素

3.1 数据集

3.1.1 数据集的质量

本项目采用《中图法》为分类体系,其训练数据来自各图书情报部门的实际标引数据。目前这部分标引数据有两种类型:一种是基于《中国分类主题词表》的受控标引;另一种是不依靠词表的关键词自由标引,这种自由标引数据不同于简单的关键词抽取,它是由作者或专业标引人员经过智力分析所得到的概念分析结构,只是没有对照词表赋予主题词而已。受控标引和自由标引的标引质量各自有不同的特点(表1)。

表1 受控标引和非受控标引质量对比

	受控标引数据	自由标引数据
标引一致性	高	低
通用词数量	少	多
标引深度	略低	高
标引专指度	低	高
词长	略短	略长
词语更新速度	慢	快
词量	少	多

若构建知识库的语料多采用受控数据,标引一致性高,标引词同分类号之间的语义对应关系明确,但同时会带来匹配的难题,因为从文本中抽取的关键词同叙词相比在词形匹配上存在较大差异,语义匹配虽然可通过建立概念空间的方式解决,但在匹配过程中无法做到等值匹配,使得语义丢失影响分类的正确性[3]。

若构建知识库采用非受控数据语料,则在分类相似度计算时,语义匹配困难较小,然而因自由标引数据的表达差异的问题,造成分类特征词过多。目前汉语新词丰富,增加速度较快,采用自由标引数据可以使得知识库收词量大,特征丰富,且标引质量相比关键词抽取的质量要高,因此分类文本的特征抽取正确率较高。从近年的多次测试来看,使用自由标引数据为主的训练数据得到的知识库比采用受控标引数据为主得到的知识库的正确率高5%左右。

3.1.2 数据集的分布

在本项目的研究过程中,发现数据集关于类别的分布往往是偏斜(skewed)或不均衡的,即类别间样本的数量可能存在数量级的差距,这是导致分类效果很不理想的一个重要因素。在数据

偏斜的情况下,样本无法准确反映整个空间的数据分布,分类器容易被大类淹没而忽略小类,从而导致部分小类的分类正确率极差。

在数据偏斜的情况下,本项目采用重取样(re-sampling)的方法,适当屏蔽一些大类的信息量,同时,对小类的分类特征的提取予以优化,或通过提高权重来获得分类器对小类别特征的重视,从而能够在有限范围内提高小类的分类正确率。

3.1.3 知识库规模

知识库是对原始数据进行预处理、兴趣度过滤以及相关数据挖掘后得到的训练结果,是进行自动分类的主要依据。如果对原始数据进行训练过程中,各项过滤以及度量指标设置过高,则得到的知识库的规模较小,扩展的类目数量小,但数据严格,用这样的知识库进行自动分类,得到的结果是正确率较高,但会出现一部分类目的数据漏分。如果知识库规模过大,虽然得到的扩展类目较多,但同时也会因为使用弱规则致使知识库质量下降,而庞大的知识库将会使得分类的时间大大增加。如何平衡知识库(训练数据)的规模与正确率和分得率是进行自动分类研究必须要考虑的问题。表 2 是 4 种规模知识库的分类性能比较。

表 2 4 种规模知识库的分类性能比较

知识库规模	最小支持度	最小置信度	Dice 测度	类目数量	正确率(%)	分得率(%)	F1(%)
501789	1	0.0	否	9901	73.7	99	86.4
81765	1	0.0	是	5208	79.8	95	87.4
41052	2	0.5	是	2805	79.4	93	87.2
14753	4	0.5	是	1771	75.8	90	82.9

注:此类目为 F 类,在《中分表》的类目数量为 1358,词串总数为 5342。

3.2 特征词选择

特征词选择是指从初始特征集合中抽取出比较重要的、能够表达文献主题内容的标引词,它是影响自动分类正确与否的重要基础因素。

目前特征词选择的方法有很多,如文献频率法(TF-IDF)、信息增益法(IG)、互信息法(MI)、开方拟合实验法(CHI)等。研究表明,特征词选取方法与实验数据集合分类算法等密切相关,需要在实际应用中比较各种方法的优劣[4]。在自动分类系统的研究和开发过程中,课题组对各种特征词选择的方法进行了研究,发现各种特征词选择的方法都有其自身的不足和优势,因此提出“投票法”,即基于上述方法的集成特征选取,通过实验表明,集成方法对分类的正确率有 5% 左右的提高。

文本不同部位所表达文献主题内容的能力也有所不同,因此在特征抽取时,文本不同部位的权值设定,对自动分类的正确率也有一定的影响。文本的题名、文摘、作者和关键词是反映主题概念的主要标引源,其中题目的标引能力最强,关键词次之,文摘再次之,从这三个部位提取特征词时,设置的不同权值比例对分类影响较大,图 2 和图 3 分别是特征词选择方法和不同标引源的权值对分类正确率的影响比较。

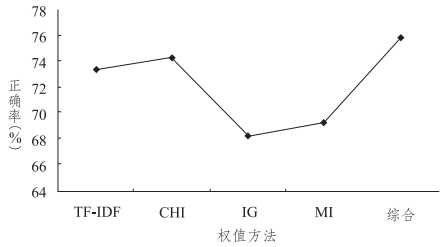

图 2 特征词选择方法对分类正确率的影响

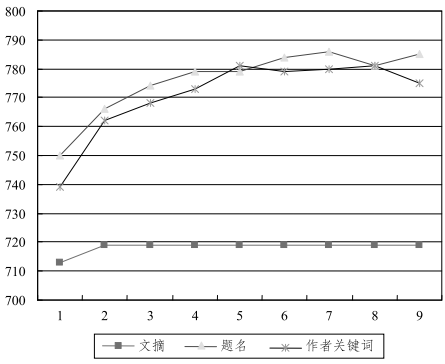

图 3 不同标引源的权值对分类正确率的影响

3.3 分类算法

3.3.1 层次分类(二次分类)方法[5]

《中图法》是一部详尽专深的综合性分类法,仅社会科学11个大类就有上万个类目,其中几个大类中,如经济、法律和教育等几个类目存在着一定的语义概念交叉、特征词相近的情况,在分类上较难把握和定量衡量特征词的区分能力。若待分类文献集由某一社科期刊(包括多类论文)混杂构成时,调用某类知识库分类时,就会把许多文献误分到相近类目中去,极大影响分类的正确率,这种情况在社会科学的政治(D)、法律(D9)、文化(G)、艺术(J)等大类中尤为明显。

针对上述情况,本项目提出采取灵活的二次分类方法,即先粗略分类再进一步细分类。具体做法是:首先对类目进行一次粗略分类,确定文献所在的一、二级大类,然后再根据大类筛选结果逐个调用相应知识库进行详细分类,这种做法可减少相近类目的误判,提高分类的准确率。通过实验表明,采用二次分类的方法在分类正确率上会提高5%左右。

3.3.2 集成分类方法

目前存在着多种分类方法,如KNN、朴素贝叶斯、SVM等等,基本可以划分为基于知识规则和基于统计处理两种派别,各自均存在优势和不足。近年来多分类器集成学习的方法也得到了学者的认可。本项目尝试将这两种分别具有"经验主义"和"理性主义"的分类方法进行集成,以期能够充分发挥各自特色,提高正确率。采用的策略是保留分类系统的前三个分类结果,同时通过基于粗糙集的属性约简和规则提取方法,从训练语料中凝炼分类规则,我们称之为"强知识规则"。利用提取的强知识规则与分类文本向量匹配,得到的分类结果与系统给出的前三个分类结果取交集,以确定最终的分类结果,通过实验证明,这种集成分类方法能够将原有的分类正确率提高近3%。

3.4 分类体系

分类表,也称分类体系、分类架构,有单层、复层和多层分类两种。单层和复层基本是在类别较少的情况下,目前很多的自动分类系统的研制都是建立在这种分类体系之上。多层分类则类别较多且类别间关系复杂,区分度小,这样的分类表类别越往下位类分,类别间的主题就越接近,越难作出区分,分类难度与分类架构的设计有很大关系[7]。

本项目采用的分类表是《中图法》第四版,该分类法类目非常详细,每个大类下平均有3000多个小类。笔者通过对分类结果的分析发现分类体系的设计结构会极大程度地影响分类效果。以《中图法》社科大类为例,很多分类都是先按照地区分,再按照主题分,这样就导致了文本内容相似却分在不同大类的情形。例如,中国的金融制度与美国的金融制度是两个相同的主题,但是在分类法中却要先按照地区分在不同类(F832和F837.12),然后再在各自类别下细分。很多分类错误都是由于地区选择错误而导致的。因此合理的适合于自动分类的分类体系应该是先按照主题分,再按照地区分。对比而言,《中图法》自然科学在类别设置上先按照主题来分,就相对比较合理。

此外,由于社科中的许多大类,如政治、经济、文化等类别,由于其包含的特征项在别的类别中也会出现,而且目前的文本分类方法都是基于词的方法,无法正确描述特征词之间的语义信息,使得社科中许多类别相似的类目的分类正确率大大降低。

3.5 评价方法

目前对自动分类的测评大多是沿用英国Cranfield项目的测评方法,采用类似信息检索领域的检全率和检准率来评估的。即用计算检全率的方法计算自动分类的召回率(或称其为分得率),用计算检准率的方法计算自动分类的分准率,采用分准率、分得率和F1三个参数来评价,这样的评价方式在实际应用中产生了许多问题[8-9]。

分准率 = 分类正确数/含有该类记录总数

分得率 = 该类记录未分数/含有该类记录总数

F1 = (分准率 + 分得率)/2

(1)正确率很大程度受不相关数据(非本类数据、干扰数据)的影响,因此若一个测试集包含本类数据较多,而包含其他杂类数据较少,作出的测试结果会比较乐观;反之,作出的测试结果有可能会相对下降。

(2)分准率、分得率、F1 也存在一定的问题,在实际中会有很多高分准率、低分得率,和高分得率、低分准率的情况,这样的数据用 F1 来综合衡量时,得出的数据同样会比较乐观。此外,采用 F1 值综合考量分类效果也不是非常恰当,因为对于特定的系统和用户来说,分准率和分得率不一定都是用户需要的。例如像应用在文本分类、文本过滤这样的系统,需要较高的分准率,尽量少出错;而应用在浏览、检索的时候,有可能需要较高的分得率,尽量少遗漏。《中图法》不同类别的平均正确率分布见图 4。

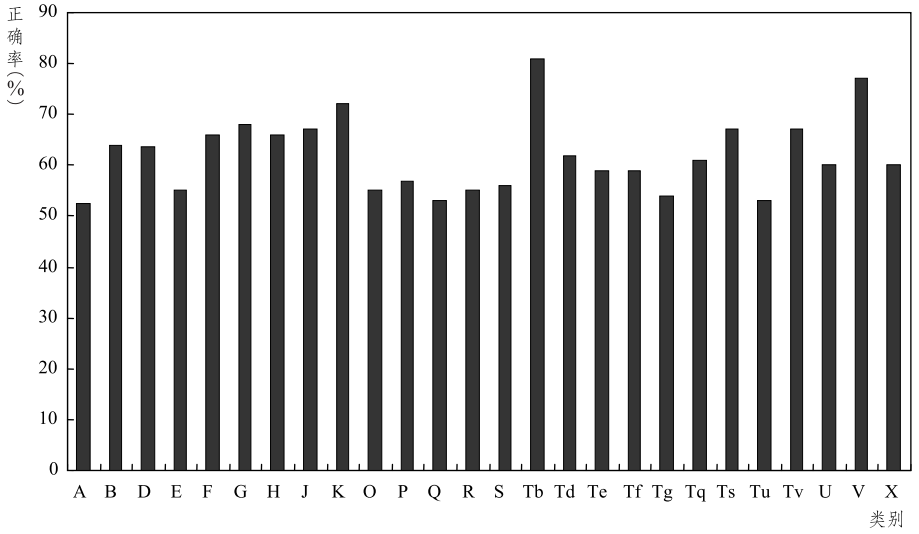

图 4 《中图法》不同类别的平均正确率分布

3.6 其他因素

3.6.1 抽词词典

抽词词典是针对某一领域有检索意义的词所构成的集合,是文本分类的基础,抽词词典的规模是影响分类质量的重要因素之一。对分类文本采用同一个抽词词典还是为每个类别分别做一个核心类别的抽词词典? 通过实验,笔者发现为每个类别做抽词词典的分类效果优于使用同一个大规模的抽词词典。

3.6.2 词长

信息检索领域与自然语言处理领域的不同之处在于,前者只选取有检索意义的实词。因此构建抽词词典时选取的词汇尽量要专指,词长要稍长,这样在抽取的时候可以避免一些通用词的干扰,提高匹配效果。而且词长不仅在抽词过程,在分类匹配的过程中也是需要考虑的重要因素。通过实验发现,在分类阶段加入词长因素分类效果可以有小幅度提高。

3.6.3 同义词

目前的分类算法基本都是把特征词看作独立的个体,不考虑词间的语义关系,但是自然语言中却存在着大量的同义词、准同义词。这些词在语义上是等同的,但在词形上却存在很大的表述差异,计算机难以识别。在本项目中,同义词的控制方面采用基于模式匹配和基于语义相似度识别的同义词抽取方法,可以解决相当一部分的同义词识别问题,在很大程度上提高了自动分类的正确率。

语义网的提出给同义词的研究提供了新的契机,知识本体是关于概念和概念关系的详细描述模型。利用领域本体对概念的描述,可以对分类文本的描述从基于关键词的评价提升到概念评价的层次,从而可以较好地解决同义词问题以及词形匹配问题。

4 讨论

4.1 《中图法》对自动分类的适应性

由于《中图法》自身分类体系的诸多因素影响,使得该分类体系用于自动分类等文本自动处理中产生了一些问题,削弱了《中图法》在网络信息资源组织中的作用。尤其是对社会科学部类中先按照地区分类再按照主题分类的做法,不仅会造成许多主题含义相同的资源的分散,同时会影响计算机自动分类处理这些数据的正确性。笔者建议在《中图法》修订过程中,应减轻体系分类法的一些弊端,适当引入分面分类法的某些特性来提高《中图法》对计算机自动处理中的适应性。

4.2 稀有类别的处理

有些热点类目的标引数据多达数万条,有些类目仅有寥寥一条或几条数据,这就使得训练数据在各类目上分布严重不均衡,这是影响分类质量的主要因素之一。

目前绝大多数较成熟的分类方法在进行训练的时候,都是将各个类目训练数据平等对待,这就使得某些"稀有类别"的数据被"湮没"。看似平等的原则实际对待各个类别是不平等的。而且,这些"稀有类别"的实际标引数据很少,就使得训练过程的准确度下降。

本文所提出的方法是通过训练数据类别的分布统计,标注出不常用的"稀有类别"。结合《中图法》上下位类的含义,将这些类别合并到上位类或单独组成一个类别以形成一个分布比较均衡的训练体系。若将来这些类别的文献大量增长,可以对合并后的类别再做一个二级分类器。

4.3 知识库更新

汉语词汇更新速度非常快,未登录词的识别一直是汉语信息处理中的研究热点和难点。在基于《中图法》的多层自动分类的多年研究中,发现训练知识库的更新问题是影响分类正确率的重要因素之一。因为每隔一段时间,有些类别就会产生一定数量的新词,这些新词如果不能及时反映到知识库和抽词词典中就会影响分类的正确率,尤其是在社会科学大类中。本项目的做法是累计当年的图情部门的标引数据,对这些数据进行统计合并形成一个小的知识库后,同原有的知识库进行去重合并以达到更新知识库的目的,这种方法虽然有效,但效率较慢。研究高效的知识库智能自学习机制将是本项目今后的发展方向之一。

4.4 明显正确或错误数据的"标注"

虽然基于《中图法》的多层自动分类系统在同类研究中,分类正确率较高,但对分类结果标引人员仍需进行全部浏览检查。对于自动分类的结果,如果能够将明显错误或明显正确的数据游离出来将会大大减轻人工干预,使得自动分类系统更加实用化。对自动分类明显正确或错误的数据进行大量的统计,分析和寻求分类号与知识库中最匹配的分类号的相似度值的规律将是解决这一问题的途径之一。

4.5 标准数据集

本项目以基于《中图法》的多层自动分类系统为例,从训练数据、训练过程、分类体系以及评价因素几个方面分析影响自动分类效果的因素。通过对比分析,发现标准的数据集或语料库(训练集和测试集)是影响分类的最大因素之一,而目前还没有像 Reuters 这样大型的、有影响力的中文标准数据集。北大天网建立的中文 WEB 测试集 CWT100g [6] 只是基于信息检索目的构建。标准分类数据集的建立可以解决训练集的数据质量、类别分布以及测试集的质量和评价等问题,用一个统一的标准来正确评价中文自动分类的效果和进展情况。

参考文献:

[1] 侯汉清, 薛鹏军. 中文信息自动分类用知识库的设计与构建[J]. 情报学报, 2003(6): 681-686.

[2] 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):22:50-55.

[4] 白振田, 侯汉清. 基于词典约简及多分类算法的文本分类系统的设计与开发[J]. 情报学报, 2008(3):337-343.

[5] 何琳,侯汉清. 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, 2007(12): 725-729.

[3] 刘竟,朱玉梅,侯汉清. 网络环境信息标引的测评与比较研究[J]. 中国图书馆学报, 2008(1):73.

[6] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展. 软件学报[J]. 2006(9):17.

[7] 曾元显. 文本主题自动分类成效因素探讨[J]. 中国图书馆学会会报, 2002(6):68.

[8] [2008-01-09]. http://www.univs.cn/newweb/news/campus/whss/2004-06-25/6251.

[9] Sbeastinai, F. A tutorial on automated text categorization[C]. In Proceedings ofTHAI-99, European Symposium on Telemetric, Hypermedia and Artificial Intelligence. Varese: IT,1999. 1-25.

何 琳 南京农业大学信息管理系讲师。通讯地址:南京市童卫路6号南京农业大学信息科技学院。邮编210095。

刘 竟 江苏大学图书馆馆员。通讯地址:苏州。邮编212013。

侯汉清 南京农业大学信息管理系教授,博士生导师。通讯地址:南京。邮编210095。

(收稿日期:2009-05-04)

(上接第9页)

[8] 李晓新,李婷,朱艳华. 公共图书馆社会和谐使命的再认识——以社会资本理论作为研究视角[J]. 图书与情报,2008(5):28-33.

[9] 黄纯元. 关于《电子图书馆的神话》[J]//黄纯元. 黄纯元图书馆学情报学论文集. 上海:上海科学技术文献出版社,2001:164-170.

[10] 于良芝. 公共图书馆存在的理由:来自图书馆使命的注解[J]. 图书与情报,2007(1):1-9.

[11] 于良芝,李晓新,朱艳华,等. 公共图书馆的使命与服务:基于内容分析法的国内外比较研究[J]. 图书馆论坛,2007(6):21-28.

[12] [美]罗伯特·帕特南. 使民主运转起来:现代意大利的公民传统[M]. 王列,等,译. 南昌:江西人民出版社,2001:195.

[13] 罗曼. 美国图书馆政策体系及其带来的思考[J]. 中国图书馆学报,2005(1):78-81.

[14] IFLA/FAIFE. IFLA 图书馆与知识自由声明[R]//程焕文,潘燕桃. 信息资源共享[M]. 北京:高等教育出版社,2004:381-382.

[15] 彭定光. 政治伦理的现代建构[M]. 济南:山东人民出版社,2007:116.

[16] 李国新. 日本的"图书馆自由"述论[J]. 图书馆,2000(4):12-16,20.

[17] 张文贤. 企业社会责任的指标体系设计[J]. 新资本,2006(5):20-23.

[18] [日] 川崎良孝. 美国公立图书馆的存在目的、历史、现状与问题[J]. 图书馆杂志,2008(3):57-61.

蒋永福 黑龙江大学信息资源管理研究中心主任,信息管理学院副院长,教授,硕士生导师。通讯地址:哈尔滨市南岗区学府路74号。邮编150080。

(收稿日期:2009-01-29;修回日期:2009-03-31)