基于文本分类的林业 Web 黄页分类系统①

王 欢[1]，武 刚[1]，杨 抒[1,2]

[1](北京林业大学 信息学院，北京 100083)
[2](新疆农业大学 计算机与信息工程学院，乌鲁木齐 830001)

摘 要：将文本分类技术应用于林业 Web 黄页的分类，实现了林业 Web 黄页信息的高效应用和管理。讨论了林业 Web 黄页多层次分类体系，并给出了分类系统的设计方案和关键技术，详细介绍了类别区分词特征选择算法。实验结果具有较好的准确率和查全率。

关键词：文本分类；林业 Web 黄页；多层次分类；类别区分词

Forestry Web Yellow Page Category System Based on Text Classification

WANG Huan[1], WU Gang[1], YANG Shu[1,2]

[1](Information Science and Technology College, Beijing Forestry University, Beijing 100083, China)
[2](Computer and Information Science Department, Xinjiang Agriculture University, Urumqi, 830001, China)

Abstract: The paper employs text classification technology to forestry Web yellow page field, and realizes a high application and management in forestry Web yellow page information. The paper also Discusses a multi-level text classification system, which provides the design and the key technologies of category system as well as the feature selection of category discriminating. The system shows good precision and recall ratio.

Key words: text categorization; forestry Web yellow page; multi-level text classification system; category discriminating word

1 引言

随着我国林业信息化的发展，林业 Web 信息呈现出信息资源海量化、数据结构多元化、存储异构化等特征。作为发布林业信息的源头，林业 Web 黄页也在急速的膨胀，这些海量信息往往未经组织分类或只经简单分类就呈现给 Web 用户，从而影响林业 Web 黄页信息高效的利用和检索。如何才能在海量信息中获取到适时、准确、有价值的林业 Web 黄页信息是领域内研究人员面临的一个重要研究问题。

Web 黄页是纸上黄页在互联网上的延伸和发展，其内容更广泛，包括机构地址、邮编、电话、传真、邮件、网址、机构介绍、产品介绍、产品图片、人才招聘等[1]。在当前信息快速增长的现实条件下,文本分类是解决海量数据高效管理与利用问题的重要环节，是实现高效信息检索系统的基础。文本分类技术已经成功的应用在很多领域，但是目前在林业信息分类领域还缺乏研究。本研究选取林业 Web 黄页中能够体现林业机构性质类别的机构名称、机构简介、产品介绍等文本内容作为研究对象，采用文本分类相关技术，将林业 Web 黄页进行多层次的组织分类。从而有效地把信息管理起来，使用户能够快速检索到适时、准确、有价值的林业 Web 黄页信息。

2 文本分类技术的研究与现状

文本分类(Text Categorization)技术是信息检索和文本挖掘的重要基础,是指指在给定分类体系下，根据文本的内容，自动为文档集合中的每一个文档确定其类别的过程[2]。早期的文本分类主要是知识工程方法。

① 收稿时间:2011-04-26;收到修改稿时间:2011-05-30

到 20 世纪 90 年代，基于统计和机器学习的自动文本分类方法日益受到重视，已经成功的应用于信息检索、搜索引擎、数字图书馆等领域。

目前在信息处理方面，文本的表示主要采用向量空间模型（VSM）。其基本思想是文本的内容由一些特征项（字、词、词组等）来表示，用向量来表示文本，则文本可表示为 D =（t1 ,t2 ,…, tn）,其中 ti 表示各个项。在这个文本向量中，每个特征都被赋予一个权重 W，以表示这个特征项在该文本中的重要程度。

在文本特征选择方面，目前比较成熟的文本特征选择方法包括文档频率(DF)、互信息(MI)、信息增益(IG)、CHI 等[3]。这些方法的基本思想都是对每一个特征（即中文词），计算某种统计度量值，然后设定一个适当的阈值 T，过滤掉那些度量值小于 T 的特征，剩下的即为有效特征。

在分类算法方面，目前机器学习的文本分类方法逐渐替代了知识工程的分类方法。一般是基于特征独立性算法：忽略了文本内词语之间的语义关系，文本被表现为分量间关系独立的向量，主要利用数学统计方法将文本分类问题转换为数学分析。基于机器学习的自动分类方法主要有贝叶斯分类、最近邻分类、决策树和支持向量机等[4]。典型的分类器有朴素贝叶斯分类器，用支持向量机建立的分类器等。

我国在文本分类技术方面的研究存在着不足，如中文分词算法不足、专业分类词库缺乏、缺少统一的中文语料库以及测试标准不统一等。这些问题将成为现阶段我国文本分类相关研究和应用的重点和主要突破的方向。

3　林业Web黄页多层次分类体系

林业组织机构多层次分类体系的确定为林业 Web 黄页分类及其检索提供了先决条件。按全国组织机构代码管理中心的分类，组织机构可以分为企业单位、事业单位、社会团体、机关、工会、民办非企业六种。由于林业组织机构中社会团体、工会等机构类型数量较少，可以统一划分到同一类别中，因此把林业组织机构划分为林业政府机构、林业事业机构、林业企业机构和其他林业机构四大类。同时借助林业领域内的专业报告、研究成果、书籍，对各大类进行逐层类别划分。以企业单位为例，其层次可进一步划分为资源、种苗、木质林产品、林业机械、林业服务、非木质林产品 6 个子类别。这些子类别又可继续分类。图 1 给出了林业组织机构多层次分类体系结构图，图 1 中只以部分类别为例给出其层次关系。

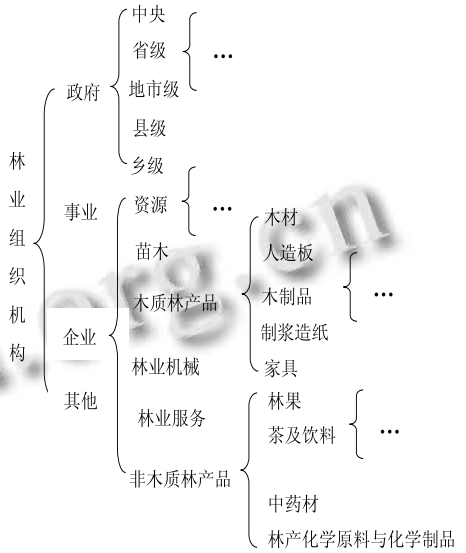

图 1　林业组织机构多层次分类体系

4　林业Web黄页分类系统的设计与实现

4.1 系统结构框架

本系统将基于机器学习的文本分类分为训练和分类两个阶段，如图 2 所示。

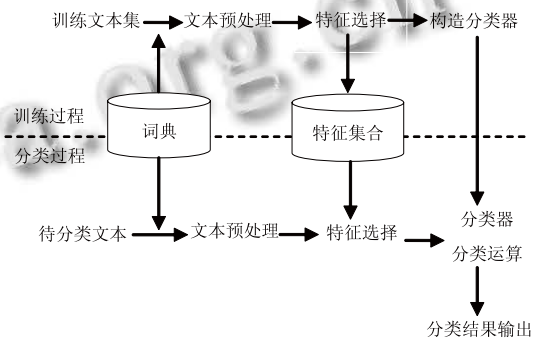

图 2　系统结构框架图

在训练阶段，将一组预先标注过类别的文档作为训练集，对这些训练文本进行文本预处理、中文分词、文本特征选择以及分类器的构建等部分。林业 Web 黄页信息文本收集和预处理是文本分类的基础。由于预先没有现成的可分类文档，因此需要设计和实现林业 Web 黄页文本收集器，对收集来的林业 Web 黄页进行文本预处理。中文分词需要把文本都分成中文词汇，

并记录其相关的权重。文本特征选择需要把中文分词的结果中能代表分类特征的词汇提取出来，形成向量，并计算其权重，为训练做准备。分类器的构建是对文本特征向量的训练过程。

在分类阶段，将待分类的林业 Web 黄页进行文本预处理、中文分词、特征选择及分类和输出结果等。同样，为了能够对新的林业 Web 黄页文本进行分类，首先也必须对其进行预处理，得到其特征向量的权重，然后使用已经获取的训练模型来对林业 Web 黄页文本进行分类。

4.2 系统实现

本系统采用 Visual C#在 Windows 环境下实现

4.2.1 文本分词预处理模块

文本分词预处理模块主要功能是从训练文本和测试文本提取词条，并过滤停用词和常用词。常用的方法包括 3 大类：基于词典的分词方法，基于统计的分词方法和基于 AI 的分词方法[5]。常用的基于词典的分词算法包括正向最大匹配法(FMM)、逆向最大匹配法(BMM)、双向匹配法(BM)等。本系统采用基于词典的逆向最大匹配分词算法(BMM)，其算法描述如下：

① 设读入句子长度为 length。设置要截取的词的最大长度 max；

② 从句子中取 length-max 到 length 的字符串 subword，去词典中查找是否有这个字符串。如果有就执行③,没有就执行④；

③ 记录 subword，将 length-max 付值给 length，继续执行③，直到 length=0；

④ max-1，执行③。

选择合适的分词词典是分词的基础与关键，向基础词库中添加林业 Web 黄页领域相关词汇，使分词模块从文本中辨别出更多林业专业领域特征词汇，为接下来的特征选择做好准备。

4.2.2 特征选择模块

目前比较成熟的文本特征选择方法有一个共同的特点：它们并不按类别计算统计值，选出的是那些全局意义上的“强类别意义”的词，这些词可能有着多类的指示意义。对于不兼类的文本分类问题来说，选用这些词作为分类特征，将使得某些文本向量位于两类的分界线附近，自动分类极易发生错误。在实验过程中发现这样一种现象，有些词具有很明显的单类类别意义，比如“胶合板”，“打印纸”，“木门”，等等，它们几乎就只出现在某一类文档之中。如果要把木质林产品的文档分为木材、人造板、木制品、制浆造纸、家具这五大类，那么“打印纸”在文章中的出现就使我们有理由猜测该文章属于制浆造纸类。这些类别区分词有着极强的类别指示意义，类别区分性相当好。如果根据词出现的统计信息，选出对应每类的“类别区分词”作为分类的特征表示，那么有可能在大大缩减特征空间的同时，选出那些最具类别指示意义因而也最利于分类的特征。因此，设计“类别区分词”的选取方法如下：

首先，定义词 t1 的类间概率分布如下：

$$Distribute(t)=(P(C_1|t_1),P(C_2|t_1),...P(C_n|t_1),)$$

其中 $P(C_i|t_1)=\frac{P(t_1|C_i)P(C_i)}{P(t_1)}$ 为贝叶斯后验概率，

$P(C_i|t_1)=\frac{P(t_1|C_i)P(C_i)}{P(t_1)}$ 。 $P(t_1|C_i)=\frac{1+\sum_{k=1}^{d_i}tf(t_{1k})}{|V|+\sum_{j=1}^{|V|}\sum_{k=1}^{d_i}tf(t_{jk})}$，

$tf(t_{jk})$表示词 tj 在 Ci 类的第 k 篇文档中出现的次数。|V|为总次数，d_i 表示 C_i 类的总文档数。

其次，定义区分词挑选标准 CDW(t)=Max1-Max2，其中 Max1 为 中 $P(C_i|t_1),i=1,2,...m$ 的最大值，Max2 为次大值。用后验概率最大值和次大值之间的差距来衡量该词的类间区分性。

最后，设置一个 0 到 1 之间的阈值，CDW(t)值大于 T 的那些词被挑选出来作为类别区分词。这种做法用词与类别之间的后验概率来衡量该词的类别指示意义。最大类别后验概率越大，与其余类别后验概率之间的差越大，那么该词关于某一类的指示意义就越强。阈值的取值与训练样本集有关，阈值越大，选出词的类别区分性越强，但这样的词也越少，会导致特征太少。因此，需要根据实际情况在类别区分性和结果词的数量上做一个折衷。

4.2.3 分类模块

分类模块主要功能是实现待分类文档的正确分类，本系统采用朴素分类算法，构造朴素贝叶斯多层分类器。算法基本思想是利用特征和分类的联合概率来估计给定文档的分类。首先将经过预处理和特征选取后的待分类文档通过第一层分类器进行分类运算，根据计算结果将文档划分到最接近的类别，然后再通过该类别的子类分类器进行分类运算，这样逐级计算

最终实现文档的多层次分类。多层次分类器的算法描述如下：

Input：一个待分类的文本 D；

分类层次数 L；L 个层次主题类别为；

$C_1\{C_{11},C_{12},\cdots C_{1n}\},\cdots\cdots,C_L\{C_{L1},C_{L2},\cdots C_{Lm}\}$

Output：文档 D 的多层次主题类别

For i=1 to L do

1) 选择第 i 层特征集 $F_i\{f_{i1},f_{i2}\cdots\cdots\}$ 和特征权重集 $W_i\{w_{i1},w_{i2},\cdots\cdots\}$ ，该层有 K 个主题类别

2) For i=1 to K do

① 根据 F_i 和 W_i，利用领域特征集聚计算公式 计算文档 D 在 i 层的主题值。

② 记录最大主题值所对应的主题类别，作为文本 D 在该层的主题。

3) 输出文本 D 在各个层次的主题类别。

5 实验结果与结论

文本分类从根本上说是一个映射过程，因此评价分类系统的标志是映射的准确程度和映射的速度。评价分类效果的标准很多，本系统采用准确率和查全率作为评价标准。实验过程中，把采集整理来的林业 Web 黄页文本通过构造的多级分类器，针对每一层次特点选取相应特征构造分类器并进行逐层分类。

以林业企业二级分类的木质林产品企业为例，可再分为：木材企业、人造板企业、木制品企业、家具企业、制浆造纸企业这 5 个三级类别。在实验中我们选择木质林产品黄页文本信息共 600 篇，平均每类 120 篇。每一类选择训练集和测试集的方法如下：将这些分类好的数据平均分成 10 份，选择其中 1 份作为开放测试集，剩余的 9 份作为训练集和封闭测试集。这样每一份都依次轮流作为开放测试集，运行分类算法，共执行 10 次分类操作，计算其平均值。封闭性测试结果的平均准确率和平均查全率分别达到了 86.27%和 85.33% ,说明了特征选择算法较好的抽取出了训练集的模式与特征，具有较好的分类效果。表 1 给出有实际意义的开放测试结果。

表 1　分类结果

类别	人工归入	机器正确归入	机器实际归入	准确率(%)	查全率(%)
木材企业	120	96	103	93.20%	80.00%
人造板企业	120	111	138	80.43%	92.50%
木制品企业	120	98	116	84.48%	81.67%
家具企业	120	105	138	76.09%	87.50%
制浆造纸企业	120	102	105	97.14%	85.00%

实验结果表明林业 Web 黄页分类系统具有较好的分类效果。系统的建立为开展林业 Web 数据挖掘、实现垂直搜索引擎、向专业用户提供高水平林业主题分类与检索等提供技术支持。进一步的工作主要是提高分类准确度，尝试对比几种不同分类算法，完善分类系统的评估体系。

参考文献

1 杨扬.网络黄页下一个金矿.中国电子商务,2005,7:22–23.
2 高洁,吉根林.文本分类技术研究.计算机应用研究,2004,7:28–30.
3 周茜,赵明生.中文文本分类中的特征选择研究.中文信息学报,2004,3:28–30.
4 肖可,奉国和.1999–2008 年国内文本分类研究文献计量分析.情报学报,2010,29:679–687.
5 黄昌宁,赵海.中文分词十年回顾.中文信息学报,2007,21(3):8–19.
6 刘冬雪.文本分类技术在信息检索中的应用.信息技术,2010,18;11–12.
7 王继成,潘金贵,张福炎.Web 文本挖掘技术研究.计算机研究与发展,2000,37(5):513–520.
8 于娟.领域特征词的提取方法研究.情报学报,2009,28(3):368–373.
9 申红,吕宝粮,内山将夫,井佐原均.文本分类的特征提取方法比较与改进.计算机仿真,2006,3:222–227.
10 苏新宁,章成志,卫平.论信息资源整合.Web 资源与建设,2005,9:54–61.
11 崔瑞琴,孟连生.数字信息资源整合问题研究.图书情报工作,2007,(7):35–37.