一种基于文本特征提取的版权保护方法

雷军程[1,2] 黄同成[2] 柳小文[2]
(长沙理工大学计算机与通信工程学院 长沙 410076)[1] (邵阳学院信息工程系 邵阳 422000)[2]

摘 要 互联网上,文本非法复制和盗版现象日益严重,因此迫切需要有效的文本版权保护方案。借助于特征提取方法和文本分类技术,针对具有版权争议的文字作品,提出了一种基于文本特征提取的作者识别方法。实验结果表明,提出的算法能够区别出不同作者的文字作品,能有效地把一个具有争议的文字作品进行分类,并识别出作者。因此该方法可以辅助解决争议作品(特别是著名作者的争议作品)的版权纠纷,打击盗版,维护诚信。

关键词 特征提取,数据挖掘,版权保护
中图法分类号 TP309.2 **文献标识码** A

New Copyright Protection Method Based on Text Feature Extraction

LEI Jun-cheng[1,2] HUANG Tong-cheng[2] LIU Xiao-wen[2]
(Institute of Computer and Communication Engineering, Changsha University of Sciences and Technology, Changsha 410076, China)[1]
(Department of Information Engineering, Shaoyang University, Shaoyang 422000, China)[2]

Abstract On the Internet, the illegal copying and piracy of text are becoming increasingly serious, so there is an urgent need for effective text copyright protection method. This paper proposed a new text copyright protection method based on text feature extraction and text classification techniques. The experiments show that the proposed algorithm can distinguish the different authors and can effectively classify a controversial literary works, and identify the real author. Therefore the method can assist to resolve the copyright disputes of dispute works (especially the famous works) to maintain integrity.

Keywords Feature extraction, Data mining, Copyright protection

1 引言

计算机存储技术和网络技术的迅速发展,使海量的信息涌现在人们面前。这些信息通常以图像、视频、音频、动画[1]和文本等为主要表现形式,其中尤以文本的传播范围最广且使用频率最大。信息的大量传播给人们的工作和生活带来便捷,但也存在缺点,比如会产生很多版权纠纷和非法复制等问题,这就迫切需要能够解决版权纠纷的作者识别方法。

通过研究发现,不同的作者或著者所撰写的文本具有较大的风格差异,同一作者或著者所撰写的不同文本具有相同的写作手法和惯用的语句结构、词汇等。作者识别方法就是,首先对不同作者所撰写的大量文本进行特征提取和统计,并训练分类器;然后针对有争议的文本,利用有效的特征提取方法获取统计向量,并输入到已训练的分类器中;最后输出具体分类类别或者具体作者。作者识别方法能够辅助解决争议作品(特别是著名作者的争议作品)的版权纠纷,打击盗版,维护诚信。

文本作者识别方法的关键环节是训练和建立分类器。分类是一种典型的有教师的机器学习方法,也是数据挖掘领域一个重要的研究课题。分类函数或分类器是通过不断学习训练数据而得到的,在需要分类的时候,测试数据可以利用已经得到的函数或分类器输出一个给定的类别。在应用中如何选择一个合适的分类模型[2]是一个重要的问题。文本分类技术可以广泛应用在自然语言处理与理解、信息管理、数据评估、信息过滤等领域。比较常见的文本分类方法有支持向量机、K近邻法、贝叶斯分类法以及神经网络及决策树分类法等。支持向量机(Support Vector Machines, SVM)主要应用于模式识别等领域,它是一种基于统计学习理论的模式识别方法,它的特点是能够同时最大化几何边缘区与最小化经验误差。根据已知样本情况,最近邻算法可以判断新样本与已知样本是否在同一个类别里。最近邻算法有很多发展和改进,但其大致思想都是先存储全部或部分训练样本,然后通过相似函数计算测试样本与所训练样本的距离,最终判断测试样本的类别。最近邻算法能够快速实现分类,特别是在基于统计的模式识别领域非常高效。神经网络的原理是通过模拟人类大脑的结构,将样本看成一个相连的输入/输出单元,训练样本通过调整单元值来进行学习。

本文针对争议作品提出了一种文本作者识别方法,即通过提取文本特征向量,然后利用分类器进行训练,形成具有一定区别度的作者风格分类器,最后对有争议的文本进行分类。

2 作者识别方法

作者识别方法主要包括两个模块:训练模块和分类模块。

到稿日期:2011-12-04 返修日期:2012-03-15 本文受湖南省教育厅基金项目(09C890)资助。
雷军程(1977—),男,硕士,讲师,主要研究方向为网络安全、算法,E-mail:ljc_zy@126.com;黄同成(1964—),男,博士,教授,主要研究方向为数字图像处理、计算机视觉、数据挖掘;柳小文(1978—),女,硕士,讲师,主要研究方向为信息安全。

训练模块的功能主要有原始语料预处理、文本关键特征提取和训练得到分类器等过程；争议文本分类模块的功能具体包括对争议文本的预处理、从争议文本中提取统计特征向量，然后输入到已训练分类器，最后从分类器中输出作者类别。这两个模块的前两个阶段所采用的方法完全相同，训练模块的主要功能是建立训练分类器，如果是有争议作品，那么从中提取关键统计特征输入到训练过的分类器中，最后根据相似度的值来判断作者类别。训练模块和分类模块的流程图分别如图 1、图 2 所示。

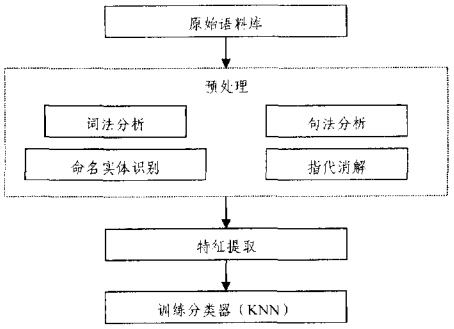

图 1　训练模块图

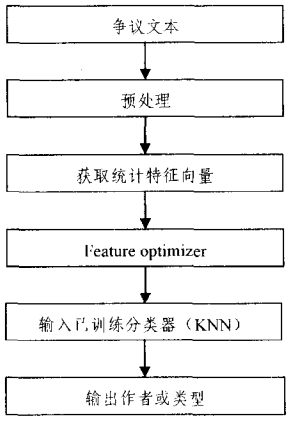

图 2　分类模块图

2.1　预处理

中文语料必须首先进行文本预处理，因为中文词语没有分隔符，不像英文等单词后面都有一个空格等分隔符。中文文本预处理过程具体包括文本语料标准化表示、中文句法分析处理、中文句子分词处理、中文命名实体识别、中文指代消解技术等。针对需要训练的原始语料库，首先需要把它们处理成计算机可以理解和计算的形式，即规范化标准化处理。规范化后的文本才能够进行自然语言处理。

(1)中文词法分析技术

中文语料首先要经过文本规范化处理，将其表示为计算机可以处理的形式以后，再针对这些规范化的文本分词进行处理。分词处理就是把中文句子按照一定的分词标准分割为具有实际意义的词语组合，这一步的目的是为文本关键特征提取工作做准备。英语等语言文字因为有空格等可以直接用来分词，所以不需要专门进行分词处理。中文汉字词语之间都是相连的，没有空格等标记，因此需要研究汉语词法分析。国内针对中文分词技术的研究有很多，评价分词技术好坏的指标一是处理速度，二是分词的准确率。本文引用了在这两个指标中都表现很好的中科院计算技术研究所的分词成果，即 ICTCLAS 词法分析系统，它采用了多层隐马尔科夫模型，系统结构如图 3 所示。

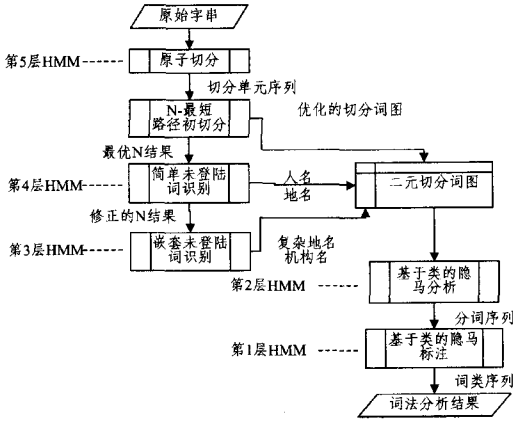

图 3　ICTCLAS 系统结构

(2)句法分析

句法分析技术的目的是通过读取单词的序列来分析句子的语法结构，它是在分词技术的基础上完成的。句法分析在自然语言处理中起一个桥梁的作用，它的基础是词法分析，它的上层可以识别篇章语义，所以良好的句法分析能使计算机真正地处理好中文句子。目前句法分析技术的研究还不是很深入，英语句法分析技术研究起步较早，取得的成果也较多，中文句法分析技术研究起步较晚，准确率也落后于英文句法分析。国内句法分析技术的代表性成果有两个，一个是中科院研究的概率句法分析器，另一是哈工大研究的依存句法分析技术。中科院研究的概率句法分析器首先需要建立一个概率模型，将中文语句的线性结构转换为树形结构，利用依存弧来反映词语之间的关系。如“武汉取消了 49 个收费项目”，该句子分析树如图 4 所示。

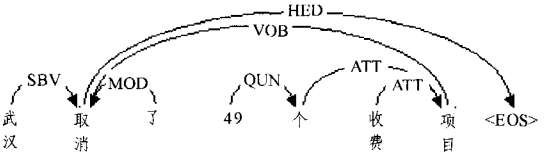

图 4　依存分析树

(3)中文文本命名实体识别

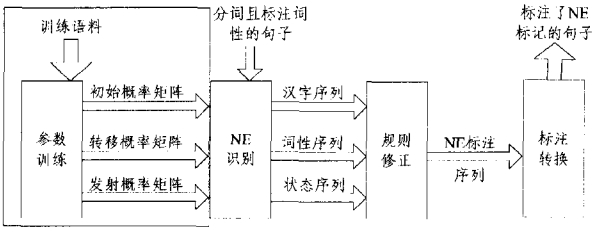

图 5　命名实体模块的结构图

命名实体是指中文文本句子中所表达的具有实际意义的实体内容，比如单位名称、人物、地理名称、机构名等。自然语言处理技术中的基础性工作之一就是进行命名实体识别，它在分词、句法分析、借助于机器的自动翻译等技术中都扮演着重要角色。目前中科院和哈工大所研究的词法分析技术成果中就具有中文文本句子命名实体识别的模块。该模块原理如图 5 所示。该成果的中文文本命名实体识别 F 值为86.93%，在计算机上运行效率为 27.2k/s，测试语料为人民日报语料库。

(4)中文句子指代消解技术

指代消解的研究难度较大，针对英文的指代消解技术早有研究，但针对中文的指代消解技术在公开见刊的参考文献中很少，目前能检索到的只有一个成果，即哈工大研究的Language Technology Platform(LTP)技术平台。该技术平台中有一个模块可以实现中文句子的指代消解。本文利用LTP指代消解技术，对测试文档中的第三人称代词进行消解，测试结果显示其能准确识别出指代实体。

2.2 特征提取

本文建议方法所指的特征提取是指针对原始语料。首先进行分词处理，然后利用特征提取方法来计算每个关键词的得分，根据得分结果选取最能代表原始语料的特征词，并借助于向量空间模型来形式化表示这些特征词条。特征提取方法实际上是一个函数映射，即对每个分割的词语进行综合评估，并按得分高低进行排序，选取得分高的前 n 个词语作为原始语料的特征词条。

本文引入了一种新的基于TFIDF的特征提取方法，我们对之进行改进，即增加了一个新的加权因子来反映词条所具有的类与类的区别能力。这种新的文本特征加权方法是在考虑原有的TFIDF方法针对训练集语料中类与类的区分能力不强的基础上，加强了训练文本集中类与类的区分能力，即加了一个新的权值来加强类与类的区分能力。增加该加权因子后，所提取的词条更能代表本文档的特征，且能区分于其它文档。根据我们改进的TF·IDF所提取的词条比较准确地代表了文本内容的特征。

2.3 分类器

分类器是在训练语料的基础上建立起来的，它实际上是一个自动分类模型。针对有争议的文本，输入分类器后能够自动输出分类结果。从训练语料中提取关键特征时，由于包含了大量的非关键词，对分类结果没有帮助，因此需要首先去除大量的这些无用词汇，利用提出的新的权重计算方法(新TF·IDF)对文本中词语权重进行计算，将词权重的所有值按比例缩放，使得它们落入较小的指定空间，并进行数据规范化，最后建立类模板。

3 实验结果

实验样本包括测试样本和训练样本，我们从互联网上收集了8个著名作者的共318篇文章，这些文章包括诗歌、小说、散文等类型，然后随机地把它们分为测试样本和训练样本，具体数量如表1所列。

表1 样本数量

作者姓名	训练样本数	测试样本数	样本总数
1 巴金	21	11	33
2 朱自清	20	19	39
3 鲁迅	13	13	26
4 张爱玲	25	15	40
5 许地山	21	15	36
6 孙犁	36	16	52
7 余秋雨	18	28	46
8 韩寒	27	19	46
合计	181	136	318

实验中我们根据特征词性的不同把提取的特征词分成了4个小组，每个小组的实验结果如下。

实验一 利用介词、助词作为作者的写作特征，提取这些词性并进行训练，然后利用测试样本进行分类，所得实验结果如表2所列。从表2可以看出，联合使用介词、助词作为特征，所获得的结果要好于单独使用一个特征所获得的结果。

表2 利用介词和助词作为特征所获得的分类精度

作者	介词	助词	介词+助词
1	88.374	91.36	93
2	90.457	92.65	93
3	88.425	91.892	94
4	89.689	92.806	93.8
5	88.032	90.335	93.9
6	86.452	89.782	92.7
7	87.325	89.961	93.1
8	87.6	89.742	92.6

实验二 利用标点符号、依存关系作为作者的写作特征，提取这些特征并进行训练，然后利用测试样本进行分类，所得实验结果如表3所列。从表3可以看出，联合使用标点符号、依存关系作为特征，所获得的结果要好于单独使用一个特征所获得的结果。

表3 利用标点符号和依存关系作为特征所获得的分类精度

作者	标点符号	依存关系	标点符号+依存关系
1	83.374	80.0	86.7
2	84.0	79.4	85.3
3	85.7	78.5	87.7
4	83.9	80.6	84.5
5	84.2	79.5	86.3
6	85.2	79.7	86.7
7	84.5	79.9	86
8	85.7	78.7	87.1

实验三 利用数词、代词作为作者的写作特征，提取这些词性并进行训练，然后利用测试样本进行分类，所得实验结果如表4所列。从表4可以看出，联合使用数词、代词作为特征，所获得的结果要好于单独使用一个特征所获得的结果。

表4 利用数词和代词作为特征所获得的分类精度

作者	数词	代词	数词+代词
1	81	87.1	88.9
2	79.56	83	85.5
3	79.14	84.6	88.7
4	80.3	82.3	85.4
5	81.4	82.9	87.4
6	82.4	84.1	89.2
7	80.1	84.6	87.6
8	79.8	83.8	88.3

实验四 利用量词、副词作为作者的写作特征，提取这些词性并进行训练，然后利用测试样本进行分类，所得实验结果如表5所列。从表5可以看出，联合使用量词、副词作为特征，所获得的结果要好于单独使用一个特征所获得的结果。

表5 利用量词和副词作为特征所获得的分类精度

作者	量词	副词	量词+副词
1	71.5	79.6	79.9
2	74.6	76.4	77.8
3	75.2	77.8	78.6
4	75.9	78.9	79.1
5	71.9	75.2	77.4
6	72.5	75.4	76.8
7	75.9	75.9	77.5
8	70.4	72.6	73.4

以上实验说明不同的作者具有自己独特的风格，即有自己惯用的词汇表达方式。借助于我们提出的算法可以有效地把一个具有争议的文字作品进行分类，并识别出作者，从而给版权保护等法律判定提供依据。

结束语 本文针对具有版权争议的文字作品，提出了一

种有效的原始作者识别方法，即借助于特征提取方法和文本分类技术，先对原始语料进行训练学习，然后把从争议作品中提取的特征输入到已训练分类器中，最后输出作者或作品类别。该方法能够辅助解决争议作品（特别是著名作者的争议作品）的版权纠纷，打击盗版，维护诚信。该方法也可以扩展应用于其它文本分类中。同一作者在不同时候的行文风格可能不同，这就给基于学习的分类方法带来了困难，未来的工作重点是研究如何提取出更能代表原始作品的特征来提高文本分类的精度。

参 考 文 献

[1] Yang L, Wang S. Fast text categorization approach based on distribution character of class[J]. Computer Engineering and Design, 2009, 30(5): 1267-1269

[2] Mehdi H A, Nasser G A. Text feature selection using ant colony optimization[J]. Expert Systems with Applications, 2009, 36(3): 6843-6853

[3] Tsoumakas G, Katakis I. Multi-label classification: An overview [J]. International Journal of Data Warehousing and Mining, 2007, 3(3): 1-13

[4] Zhang H P, Liu Q, Cheng X Q, et al. Chinese Lexical Analysis Using Hierarchical Hidden Markov Model[C]// Proc of 2nd SIGHAN workshop affiliated with 41th ACL. Sapporo, 2003: 63-70

[5] Zhang H P, Liu T, Ma J S, et al. Chinese Word Segmentation with Multiple Postprocessors in HIT-IRLab[C]// Proc of SIGHAN'05. 2005: 172-175

[6] 中科院计算所数字化研究室. 概率句法分析器[OL]. http://mtgroup.ict.ac.cn/ictparser/parser_1.php, 2011-07-10

[7] 刘挺，马金山，李生. 基于词汇支配度的汉语依存分析模型[J]. 软件学报, 2006, 17(9): 1876-1883

[8] 孙星明，殷建平，陈火旺，等. 汉字的数学表达式研究[J]. 计算机研究与发展, 2002, 9(6): 707-711

[9] 哈尔滨工业大学信息检索研究室. LTP[OL]. http://ir.hit.edu.cn/demo/ltp/, 2012-01-12

[10] Shi Tong-nian, Lu Zhong-liang. Research on the Chinese text categorization of multi-classification and multi-label[J]. Journal of Chinese Information Process, 2003, 22(3): 306-309

[11] Trohidis K, Tsoumakas G. Multilabel classification of music into emotion[C]// Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR). 2008: 76-85

[12] Han Jia-wei, Pei Jian, Yin Yi-wen. Mining Frequent Paterns without Candidate Generation[C]// Proc. 2000 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'00). 2000: 1-12

[13] Huang Xuan-jing, Wu Li-de. Language independent text categorization[J]. Journal of Chinese Information Process, 2000, 14(6): 1-7

[14] Wang Zhen, Musajan W. Text Feature Selection Method Based on Improved TFIDF[J]. Modern Computer, 2009, 7(10): 34-36

[15] Lin Yong-min, Lv Zheng-yu, Zhao Shuang, et al. Analysis and improvement of feature weighting method TF·IDF in text categorization[J]. Computer Engineering and Design, 2008, 29(11): 2923-2925

[16] 王茜，张鲲鹏. 隐私保护数据挖掘算法 MASK 的改进[J]. 重庆理工大学学报：自然科学版, 2012, 26(6): 63-66

（上接第 89 页）

[2] The working group for WLAN standards. Wireless LAN Medium Access Control (MAC) and physical layer (PHY) specifications[R]. IEEE 802.11 standards, Part 11. Technical report, IEEE, 1999

[3] Polastre J, Hill J, Culler D. Vesatile low power media access for wireless sensor networks[C]// Proc. of the 2nd ACM Conf. on Embedded Networked Sensor Systems (SenSys). Baltimore, 2004: 95-107

[4] Rajendran V, Obraczka K, Garcia-LunaAceves J J. Energy-Efficient, Collision-Free Medium Access Control for wireless sensor network[C]// Proc. of the ACM SenSys 2003. Los Angeles, 2003: 181-192

[5] Rhee N, Warrier A, Aia M, et al. ZMAC: A hybrid MAC for wireless sensor networks[C]// Pro. of the 3rd ACM Conf. on Embedded Networked Sensor Systems (SenSys 2005). San Diego: ACM Press, 2005: 90-101

[6] Gupta N, Kumar P R. A performance analysis of the 802.11 wireless LAN Medium access control[J]. Communications in information and systems, 2004, 3(4): 279-304

[7] Ye Wei, Heidemann J, Estrin D. An Energy-Efficient MAC Protocol for Wireless Sensor Networks[C]// Twenty First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings (INFOCOM 2002). IEEE, 2002: 1567-1576

[8] Wei Ye, Heinemann J, Estrin D. Medium Access Control with Coordinated Adaptive Sleeping for Wireless Sensor Networks [J]. IEEE, Acll Transactions on Networking, 2004, 12(3): 493-506

[9] Singh S, Raghavendra C S. PAMAS: Power aware multi-access protocol with signaling for ad hoc networks[J]. ACM Comput. Commun. Rev., 1998, 28(3): 5-26

[10] Lin P, Qiao C, Wang X. Medium access control with a dynamic duty cycle for sensor networks[C]// Proc. of IEEE Wireless Communications and Networking Conference (WCNC). 2004: 1534-1539

[11] Dam T V, Langendoen K. An adaptive energy-efficient MAC protocol for wireless sensor networks[C]// Proceedings of the 1st International Conference on Embedded Networked Sensor System (SenSys'03). 2003: 171-180

[12] 蹇强，龚正虎，朱培栋，等. 无线传感器网络 MAC 协议研究进展[J]. 软件学报, 2008, 19(2): 389-403

[13] 任丰原，黄海宁，林闯. 无线传感器网络[J]. 软件学报, 2003, 14(7): 1282

[14] 王辛国，张信明，陈国良. 时延受限且能量高效的无线传感网络跨层路由[J]. 软件学报, 2011, 22(7): 1626-1640

[15] 谭长庚，肖渊，王建新. 无线传感器网络中节点睡眠调度机制研究[J]. 计算机科学, 2008, 35(5): 51-54

[16] 苏金钊，刘丽艳，吴威. 适用于周期休眠 MAC 协议的分簇时间同步算法[J]. 计算机研究与发展, 2010, 47(11): 1893-1902

[17] 罗俊，蒋铃鸽，何晨. 一种多跳无线传感器网络中基于 SMAC 协议的性能分析模型[J]. 中国科学：信息科学, 2010, 40(11): 1464-1472

[18] 孙利民，等. 无限传感器网络[M]. 北京：清华大学, 2005