基于文体学的中文 UGC 作者身份识别研究*

吕英杰[1]　范　静[2]　刘景方[3]

[1](北京化工大学经济管理学院　北京 100029)

[2](北京外国语大学国际商学院　北京 100089)

[3](上海交通大学安泰经济与管理学院　上海 200052)

【摘要】网络的开放性和虚拟性给发布信息的作者身份识别造成很大困难,因此探索性地提出通过对网上的用户生成内容(UGC)的写作特点进行分析来识别其作者身份的方法。在传统的文体学研究基础上,结合中文 UGC 的特点,提取出词汇特征、句法特征、结构特征和内容特征等 4 类能有效识别不同作者写作风格的特征,然后运用文本分类算法对作者身份进行有效识别。通过实验表明在 BBS 论坛文本和博客文本这两种典型的中文 UGC 环境中,本研究采用的方法均得到很好的识别效果。

【关键词】文体学　用户生成内容　作者识别

【分类号】TP391

Authorship Identification of Chinese UGC Based on Stylistics

Lv Yingjie[1]　Fan Jing[2]　Liu Jingfang[3]

[1](School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China)

[2](International Business School, Beijing Foreign Studies University, Beijing 100089, China)

[3](Antai College of Economics and Management, Shanghai Jiaotong University, Shanghai 200052, China)

【Abstract】The characteristics of information network such as openness and virtuality make it difficult for authorship identification. Therefore, this paper proposes the approach of authorship identification of Chinese UGC based on stylistics. The authors integrate four types of features including lexical, syntactic, structural and content - specific features to compose writing - style features, and then use text classification technologies for authorship identification. The experimental results demonstrate that the proposed approach can be used for authorship identification of Chinese UGC efficiently.

【Keywords】Stylistics　UGC　Authorship identification

1　引　言

近年来,随着以 Web2.0 技术为代表的社会化媒体的快速发展,传统的信息传播模式由网站发布信息、受众被动接受信息的单向传播模式逐渐转变为广大用户进行信息交流的互动模式,互联网已经成为供广大用户进行信息交流的平台,这些由用户交互产生的互联网内容被称为用户生成内容(User Generated Content, UGC)。社会化媒体作为一种用户参与创造价值的新模式,在其快速发展的同时,也产生了一系列问题亟待解决,其中一个突出的问题表现在对于很多网络上的 UGC 无法准确判断其作者的真实身份。在虚拟的网络环境下,网络用户可以自由发表意见而不受约束,由于个人所处的立场和看问题的角度不同,其发布的UGC可能带有更大的随意性甚至

收稿日期: 2013 -04 -18
收修改稿日期: 2013 -08 -08
*本文系国家自然科学基金项目"我国电子政务标准的产生机制及采纳扩散研究"(项目编号:71103021)和北京市哲学社会科学规划项目"北京市 G2G 电子政务业务协同的动力机制、推进方法与实证研究"(项目编号:13JGC085)的研究成果之一。

还有很多片面的个人观点和见解,盲目相信很可能会被误导,这也是对网络用户发布的 UGC 的科学性和可靠性缺乏足够信任的一个重要原因。而另一方面,一些专业人士通过 UGC 发表的权威的、科学的观点也很可能由于缺乏信任等原因得不到人们足够的重视,这些问题的产生都与网上 UGC 的匿名性有一定的关系。目前在大部分社会化媒体网站中人们要想发布 UGC 时,首先要通过网站注册表明其自身身份,也有部分社区网站会通过实名认证等方式来加强成员身份的真实性管理。虽然这种方式操作简单,但在很多情况下广大用户处于隐私保护的考虑,在网站注册个人信息时往往不会填写个人的有效资料,个人资料的缺乏给作者身份识别造成很大的困难。在这种情况下,本文将研究重点转移到用户发表在社会化媒体上的 UGC 中,通过对这些 UGC 体现出来的写作特点进行分析来识别其作者身份。近年来,基于文体学的研究方法在根据作者写作风格来识别作者身份方面取得了很好的效果,因此本文提出将文体学的分析方法引入到对网络中文 UGC 的分析中,为 UGC 作者身份识别提供一种新的思路和方法。

2　基于文体学的作者身份识别研究

所谓文体学,就是以统计方法分析作者的写作风格。文体学的研究方法基于一个基本的假设,即每一个作者都具有自己特定的写作习惯,这些写作习惯会在其所写的作品中表现出来,比如作者习惯使用的短语或助词、长短句的使用频率、文章整体的篇章结构特点等,而且作者的写作习惯是无意识表现出来的,即使作者出于某种目的故意掩饰其写作习惯也很难做到。基于这一假设,可以利用统计的方法提取出作品中能体现作者写作习惯的某些特征,然后根据这些写作特征判断作品的作者身份。

早期文体学的研究方法主要集中在对文学作品的作者身份识别上[1]。比较经典的例子是有研究者对英国作家莎士比亚的写作风格进行研究[2],他们统计了莎士比亚作品中出现的词汇总数及其关键词汇的频率,以此作为判断一篇作品是否出自莎士比亚的凭据,通过这种方式他们成功地推断出于 1985 年发现的一首未署名的诗是由莎士比亚所作。国内也有不少研究者借鉴文体学的方法来研究文学作品的作者身份归属问题。比如针对《红楼梦》这部名著,其后 40 回是高鹗续写还是曹雪芹原著,一直是众多学者争论不休的话题,有研究者采用文体学的研究方法将《红楼梦》前后两部分的内容提取一些写作特征进行比较,得出前后两部分为不同作者所著的结论[3,4]。也有研究者选取中国现当代文学代表人物的作品进行作者身份识别研究,通过选取以词汇为基础的多种统计量作为识别特征,利用 SVM 分类模型在跨文体的作品的作者身份识别中取得了非常优异的识别性能[5]。除了采用词汇特征、句法特征等传统的文体风格特征外,有研究者提出了新的基于语义分析的文体特征,通过利用 HowNet 知识库,有效地改进了作者身份识别的性能[6]。

随着互联网技术的快速发展,网络上的文本信息大量增加,比如 BBS、博客、电子邮件等,有研究者开始将文体学的分析方法运用到网络文本信息作者身份的识别上。最早也是最广泛被应用的研究对象是电子邮件,De Vel 等[7]从电子邮件中抽取了语言特征和结构特征作为作者的写作特征,采用支持向量机等机器学习方法,对电子邮件作者身份进行分类识别,并在邮件主题对分类识别结果的影响方面进行了研究。在其他类型的网络文本的作者身份识别方面近几年也有不少研究,代表性的有 Zheng 等[8]设计了一个对网络文本的作者进行识别的框架,通过选取词法特征、句法特征、结构特征和内容特征等构成特征集,采用 SVM 等三种不同的分类算法对实验语料进行作者身份识别的实验,实验结果表明这几类特征对不同网络信息作者均起到明显的区分效果;美国亚利桑那大学(University of Arizona)的 Abbasi 等[9]为有效监控互联网上恐怖分子发布的危害公众安全的恐怖信息,提出运用文体学的方法对网络论坛发布信息的作者身份进行识别,抽取这些信息中的词汇、句法、结构、内容等相关特征,采用支持向量机和决策树作为分类算法,将论坛中发布恐怖信息的不法分子和普通用户进行区分,并通过对一些英语和阿拉伯语的网络论坛语料进行实验测试,结果证明了该方法在作者身份识别中的有效性。

虽然基于文体学的研究方法在传统的文学作品作者身份识别中取得了重要的成果,但将其运用于网络上 UGC 的作者身份识别仍然有很多问题需要解决。这主要是因为 UGC 大都具有很明显的口语化特点,不注重语法,大量使用省略句式,以便于理解和方便沟通

为原则,讲究沟通的时效性,因此如何从中有效地提取出作者的写作特征,是需要重点研究的问题。国外对于英文 UGC 的作者身份识别取得了一定的成果,但中文与英文在语法和写作习惯等各方面都存在显著差异,所以针对英文 UGC 提取的作者写作特征并不一定适用于中文环境,因此本文针对中文 UGC 的特点,在借鉴以往研究的基础上,建立适用于中文 UGC 作者身份识别方法。

3 中文 UGC 的写作特征提取方法

结合以往研究成果和中文 UGC 的特点,本文将中文 UGC 的写作特征分为以下 4 个部分:词汇特征、句法特征、结构特征和内容特征。词汇特征主要是基于字和词语的相关统计特征;句法特征包括词性、功能词、标点等的统计特征;结构特征则主要面向网络文本,比如网络文本的布局,所用的字体、字号和颜色等方面;内容特征主要是指能有效反映文本主题内容的相关特征。

3.1 词汇特征

词汇特征在英文中常被用来作为识别文本作者身份的有效特征,例如 De Vel 等[7]在对电子邮件的作者识别研究中,将大小写字母、数字、空格、制表符、各种常见标点符号等词汇特征作为区分不同作者的有效特征,识别准确率高达 88.2%。而由于中文与英文在写作习惯上有很大不同,最大区别在于英文是以字母组成的单词为基本单位,单词之间以空格隔开,而中文的书写是以字为单位,字与字之间是连续书写的,另一方面很多中文特殊字符与英文字符也存在不少区别,因此本研究在借鉴英文词汇特征提取的基础上,根据中文 UGC 的特点提取了包括数字、英文字母、空格和其他一些常见的特殊字符在内的中文词汇特征。本文统计的特殊字符共 17 种,如表 1 所示:

表 1 特殊字符

特殊字符	描述	特殊字符	描述
~	波浪字符	_	下底线
@	at 字符	=	等号
#	数字符号	+	加号
$	美元符号	>	大于号
%	百分号	<	小于号
^	抑扬音符号	|	竖线
&	与字符	*	星号
/	斜线	-	连接号
\	反斜线		

而对于词的特征提取首先要进行中文分词,然后再提取相关的词汇特征,本文提取的所有词汇相关的特征如表 2 所示:

表 2 词汇特征

特征类型	特征	描述
基于字符的特征	总字数(C)	
	平均的字母个数	总字母个数/C
	平均的数字个数	总数字个数/C
	平均的空格数	总空格数/C
	23 个特殊字符的频率	每个特殊字符数/C
基于词的特征	总词数(M)	
	平均词长	C/M
	平均每个句子中的字数	C/句子数
	平均每个句子的词数	M/句子数

3.2 句法特征

网上 UGC 文本主要出于交流的需要,因此与其他传统文本有很大区别,为节省时间,突出重点,提高交流效率,UGC 在句法结构上大都倾向于使用短句子、不完全句子、省略句子和不规则句子。每一个句子字数较少,有的甚至只有几个字就构成一个句子。并且在 UGC 中频繁使用问号、感叹号、省略号等标点符号,如赞同或欣赏对方观点时常常增加一大串感叹号,在不理解对方观点时,会加一大串问号或者是用省略号等。根据以上这些句法特点,本文选取三方面的句法特征:基于标点符号的特征、基于功能词的特征和基于词性的特征。

(1)基于标点符号的特征。基于标点符号的特征主要是统计在社区帖子中常见标点符号的使用频率,本文选取的常见标点符号有 12 种,如表 3 所示:

表 3 标点符号

标点符号	描述	标点符号	描述
,	逗号	、	顿号
。	句号	‘,’	单引号
?	问号	“,”	双引号
!	叹号	()[]	括号
:	冒号	—	破折号
;	分号	……	省略号

(2)基于功能词的特征。中英文文体学的相关研究均已证明[10,11],功能词的使用是有效区分不同作者写作风格的重要特征,例如张凯等[4]在对《红楼梦》的文体学研究中,就以“之”、“了”、“的”等 8 个功能词作为识别特征进行分析,得出《红楼梦》前后章节为不同作者所著的结论。由于研究的领域不同,功能词的选择并没有一个普遍被认可的标准。而对于中文来说,

功能词主要是以连词、介词和助词为代表的虚词。由于虚词众多,本文主要选取常见的能突出作者主观感情色彩的语助词来代表不同作者的用词习惯,选取的 15 个常见功能词如下:啊、吧、呗、地、的、啦、了、么、吗、嘛、哪、呢、哇、呀、耶。

(3)基于词性的特征。不同作者对于各类词性的运用也有所不同,本文通过对文本中的词进行词性标注,统计平均每个句子中各类词性的词频,作为区分不同作者的重要特征。基于词性的特征也在众多文体学相关研究中被广泛采用[12],根据中文的特点,本文共统计 12 种词性,其中实词分别有名词、动词、形容词、数词、量词、代词;虚词分别有副词、介词、连词、助词、拟声词和叹词。

3.3　结构特征

结构特征反映了作者是如何组织整个文章的篇章结构,不同作者对整体文本的呈现效果有不同的偏好,主要包括整篇文章有多少段落、多少行数;每段文字大约有多少内容;段落与段落之间是否有空行等。这些篇章级的结构特征也是识别不同作者身份的有效特征,有研究结果表明加入结构特征能使作者识别的准确率至少提高 5% 以上[7]。本文用到的所有结构特征如表 4 所示:

表 4　结构特征

特征	描述
文本的总行数(L)	
文本总的句子数(S)	
文本总的段落数(P)	
平均每段的句子数	S/P
平均每段的字数	C/P
平均每段的词数	M/P
段落之间是否有空行	
段首是否有缩进	
是否有引用内容	例如在论坛中作为回复关系的帖子是否在发帖之前引用原主题帖
引用的位置	例如论坛中引用的主题帖的位置是在发言贴内容之前还是之后

3.4　内容特征

基于内容相关的特征对区分作者身份也有很好的效果,有研究就以网络新闻组[8]和网络论坛[9]等几种网络文本为语料对内容特征的区分效果进行测试,通过 t 检验结果证明内容特征在提高作者身份识别的准确率上有显著效果。其原因在于尽管 UGC 中涵盖的主题可能涉及到很多方面,但对个人而言所感兴趣的或者擅长的主题总是相对较少的,文本内容相关的特征可以在一定程度上反映出不同作者感兴趣的主题类别,因此也成为有效区分不同作者身份的重要特征。

内容相关的特征分析主要是从 UGC 中抽取能有效表达主题的关键词,关键词的提取方法有很多种,最常见的是采用一些统计方法例如 TF - IDF[13]、词频[13]、互信息[14]、信息增益[15]等。本文选择目前应用最广泛且最有效的一种关键词提取算法——基于信息增益的方法进行关键词抽取。算法描述如下[15]:

假设$\{c_i\}_{i=1}^{m}$代表已有的分类集合,而 t 代表其中的一个特征项,那么特征项 t 的信息增益定义为:

$$G(t) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + P(t)\sum_{i=1}^{m} P(c_i|t)\log P(c_i|t) + P(\bar{t})\sum_{i=1}^{m} P(c_i|\bar{t})\log P(c_i|\bar{t}) \quad (1)$$

其中,$G(t)$表示特征项 t 的信息增益,$P(c_i)$表示类别 c_i 的概率,$P(c_i|t)$表示给定 t 的情况下 c_i 的条件概率,$P(t)$表示特征项 t 出现的概率,而 $P(\bar{t})$表示 t 没有出现的概率。

4　实验与结果分析

4.1　实验数据

目前,中文 UGC 中最具代表性的类型是 BBS 论坛、博客以及微博等,由于微博受到字数的限制往往很难充分体现其作者的写作特征,因此对于微博等超短网络文本的身份识别可以通过实名认证等其他方式解决,本文采用的基于文体学的方法更适用于具有一定文本容量的论坛及博客等 UGC 类型,因此在实验语料的选择上,笔者分别从国内某著名 BBS 论坛和博客中搜集中文 UGC 数据,并从中分别选取了发文量最多的 5 位作者发表的 UGC 作为实验语料并对所有语料进行人工标注其作者。然后对所有的实验语料进行预处理:去掉转载的 UGC,由于这些 UGC 并非由转载者本人所写,因此也就无法从中体现出转载者的写作特征;过滤掉内容较少的文本,这些文本包含的信息不足,无法从中提取到足够的作者写作特征,因此将字数少于 100 的文本过滤掉。经过筛选后,本实验共选取 586 篇论坛帖子和 402 篇博文,分布如表 5 所示。

4.2　实验设计

从实验语料的每个文本中分别提取词汇特征、句法特征、结构特征和内容特征共 4 部分的写作特征,各个特征值的计算方法需要根据特征类型的不同分别以该特征出现的次数或者频率等方式来表示。

表5　实验选用的 UGC 的统计分布

实验语料	作者	发文量
论坛	P1	205
	P2	163
	P3	98
	P4	66
	P5	54
	合计	586
博客	B1	121
	B2	87
	B3	75
	B4	63
	B5	56
	合计	402

在分类算法的选择上,以往研究表明,朴素贝叶斯算法(Naïve Bayes)[16]、决策树算法(C4.5)[17]和支持向量机(SVM)[18]等广泛应用于文本分类中并且表现出良好的分类性能,因此也采用这三种分类算法,并且由于实验语料是5位作者发表的UGC,因此在分类算法中分类数目设置为5。

在结果评估上,将分类正确率作为分类效果的评价指标。由于所有的实验语料均已事先标注其所属作者,因此如果某文本通过分类算法自动识别的作者身份与人工标注相一致,则说明该文本的作者识别正确。在测试方法的选择上,采用十折交叉验证法,将实验语料分成10份,轮流将其中9份作为训练数据,1份作为测试数据,进行实验得到分类正确率,然后将实验重复10次得到的分类准确率取平均值,作为最终的分类正确率。

4.3　实验及结果分析

实验环境为 Eclipse 集成开发环境,利用 Java 编程语言实现对实验语料各项特征的自动提取,然后利用 Weka 数据挖掘平台实现本实验中用到的三种文本分类算法。为了验证本文提出算法的有效性,需要对实验结果进行分析,以下分别从特征选择和分类算法选择两方面进行结果分析。

(1)特征选择方法评价

本文整合了多种特征类型来表示作者的写作特征,为进一步验证这些特征是否对识别中文 UGC 作者身份都具有效果,需要对每一种特征类型进行有效性评估。采用的评估方法是首先选取一种类型的特征来构成特征集进行分类,对得到的分类结果进行精确度评价,然后在特征集构成中依次加入其他类型的特征后,重新计算精确度是否比之前有所改进,以此可以有效地验证新加入的特征能否有效地提高识别效果。在实验中,将待评估的特征类型分为4类,分别为词汇特征(F1)、句法特征(F2)、结构特征(F3)和内容特征(F4),将以上4部分特征依次加入特征集后,采用 SVM 分类器进行分类,结果如表6所示:

表6　分类精确度

特征集	BBS	博客
F1	52.27%	55.67%
F1 + F2	54.77%	58.23%
F1 + F2 + F3	62.29%	64.41%
F1 + F2 + F3 + F4	80.34%	85.31%

可以看到在 BBS 论坛文本和博客文本两种中文 UGC 的作者识别结果中,随着新特征的不断加入,精确度值均有所增长,当所有4类特征都被加入到特征集后,分类的精确度达到最高值,由此可以证明本文选取的4种类型的特征对于识别中文 UGC 作者身份都起到明显的区分作用。

为进一步验证各个特征的有效性,本文采用 t 检验进行有效性分析。首先假设随着特征类型的不断增加,分类精确度也会不断提高,然后对于每一次比较,通过随机选取样本然后执行十折交叉验证计算两种特征选择方式下的分类精确度,重复10次后可以得到 t 检验值结果,如表7所示:

表7　对各特征的有效性进行 t 检验

特征集	P 值	
	BBS	博客
F1 < F1 + F2	0.0022	0.0034
F1 + F2 < F1 + F2 + F3	<0.0001	<0.0001
F1 + F2 + F3 < F1 + F2 + F3 + F4	<0.0001	<0.0001

结果表明,当使用特征集 F1 + F2 与特征集 F1 比较时,P 值小于 0.01,通过 t 检验,表明加入特征 F2 后,分类精确度的提高是显著的。同理,当比较特征集 F1 + F2 + F3 与特征集 F1 + F2 时,P 值小于0.01,表明加入特征 F3 后,分类精确度的提高也是显著的,以此类推可知,特征类型 F4 加入后分类精确度的提高也是显著的,最终本模型选取的4类特征的有效性都得以验证。

(2)分类算法评价

为检验不同的分类算法在中文 UGC 作者识别中的性能,本文分别采取 Naïve Bayes、C4.5 和 SVM 三种分类算法计算分类精确度,结果如图1所示。

从图1可以看出,无论是 BBS 论坛文本还是博客文本,SVM 分类算法在文本作者识别中的性能均优于 Naïve Bayes 算法和 C4.5 算法,且采用 SVM 分类算法

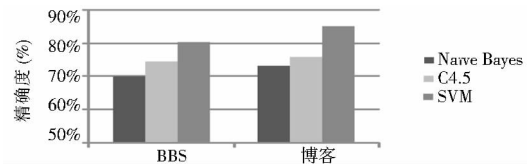

图 1　不同分类算法的分类精确度

的结果精确度均达到 80% 以上，表明 SVM 分类算法在中文 UGC 作者识别研究中具有良好的效果。

5　结　语

网络的开放性和虚拟性在提供便利的同时，由于匿名性导致的缺乏信任等问题也带来了很多困扰，在目前缺乏有效的身份识别机制下，本文探索性地提出利用用户发表的中文 UGC 的写作特点来识别其作者身份的方法，借助于已有的文体学研究方法，并结合网络文本的特点，提取出能有效识别不同作者写作风格的特征，然后运用文本分类算法对作者身份进行有效识别。通过实验表明在 BBS 论坛文本和博客文本这两种典型的中文 UGC 环境下，本文的方法均得到了很好的实验效果。由于中文 UGC 还存在很多不同类型的应用模式，针对不同的应用环境还有更多的能体现作者写作风格的特征有待深入挖掘。

参考文献：

[1] 孙晓明，马少平. 基于写作风格的作者识别[C]. 见：*中国中文信息学会二十周年学术会议论文集*. 北京：清华大学出版社，2001：198－204.（Sun Xiaoming, Ma Shaoping. Author Identification Based on Stylometric Approach[C]. In: *Proceedings of the 20th Anniversary Chinese Information Processing Society of China*. Beijing: Tsinghua University Press, 2001: 198－204.）

[2] Efron B, Thisted R. Estimating the Number of Unseen Species: How Many Words did Shakespeare Know? [J]. *Biometrika*, 1976, 63(3): 435－447.

[3] 张运良，朱礼军，乔晓东，等. 基于句类特征的作者写作风格分类研究[J]. *计算机工程与应用*，2009，45(22)：129－131，223.（Zhang Yunliang, Zhu Lijun, Qiao Xiaodong, et al. Research on Text Authorship Categorization Based on Sentence Category Features[J]. *Computer Engineering and Applications*, 2009, 45(22): 129－131, 223.）

[4] 张凯，张明允. 基于 SVM 的《红楼梦》写作风格研究[J]. *贵阳学院学报：自然科学版*，2011，6(1)：55－57.（Zhang Kai, Zhang Mingyun. Research on the Writing Style of "Dream of the Red Chamber" Based on SVM[J]. *Journal of Guiyang College: Natural Sciences*, 2011, 6(1): 55－57.）

[5] 年洪东，陈小荷，王东波. 现当代文学作品的作者身份识别研究[J]. *计算机工程与应用*，2010，46(4)：226－229.（Nian Hongdong, Chen Xiaohe, Wang Dongbo. Research on Authorship Attribution of Contemporary Literature[J]. *Computer Engineering and Applications*, 2010, 46(4): 226－229.）

[6] 武晓春，黄萱菁，吴立德. 基于语义分析的作者身份识别方法研究[J]. *中文信息学报*，2006，20(6)：61－68.（Wu Xiaochun, Huang Xuanjing, Wu Lide. Authorship Identification Based on Semantic Analysis[J]. *Journal of Chinese Information Processing*, 2006, 20(6): 61－68.）

[7] De Vel O, Anderson A, Corney M, et al. Mining E－mail Content for Author Identification Forensics[J]. *ACM SIGMOD Record*, 2001, 30(4): 55－64.

[8] Zheng R, Li J, Huang Z, et al. A Framework for Authorship Identification of Online Messages: Writing－style Features and Classification Techniques[J]. *Journal of the American Society for Information Science and Technology*, 2006, 57(3): 378－393.

[9] Abbasi A, Chen H. Identification and Comparison of Extremist－group Web Forum Messages Using Authorship Analysis[J]. *IEEE Intelligent Systems*, 2005, 20(5): 67－75.

[10] Holmes D I, Forsyth R S. The Federalist Revisited: New Directions in Authorship Attribution[J]. *Literary and Linguistic Computing*, 1995, 10(2): 111－127.

[11] Juola P, Baayen H. A Controlled Corpus Experiment in Authorship Identification by Cross－entropy[J]. *Literary and Linguistic Computing*, 2005, 20(S): 59－67.

[12] Abbasi A, Chen H. Writeprints: A Stylometric Approach to Identity－level Identification and Similarity Detection in Cyberspace[J]. *ACM Transactions on Information Systems*, 2008, 26(2): 1－29.

[13] Salton G, Buckley C. Term－weighting Approaches in Automatic Text Retrieval [J]. *Information Processing and Management*, 1988, 24 (5): 513－523.

[14] Battiti R. Using Mutual Information for Selecting Features in Supervised Neural Net Learning [J]. *IEEE Transactions on Neural Networks*, 1994, 5(4): 537－550.

[15] Yang Y, Pederson J O. A Comparative Study on Feature Selection in Text Categorization [C]. In: *Proceedings of the 14th International Conference on Machine Learning*. 1997: 412－420.

[16] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers [J]. *Machine Learning*, 1997, 29 (2－3): 131－163.

[17] Quinlan J R. C4. 5: Programs for Machine Learning [M]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[18] Cortes C, Vapnik V. Support－Vector Network[J]. *Machine Learning*, 1995, 20 (3): 273－297.

（作者 E－mail：luyingjie982@163.com）