文章编号: 1003 - 0077(2006) 06 - 0061 - 08

基于语义分析的作者身份识别方法研究①

武晓春, 黄萱菁, 吴立德

(复旦大学 计算机科学与工程系, 上海 200433)

摘要: 作者身份识别是一项应用广泛的研究, 身份识别的关键问题是从作品中提取出代表语体风格的识别特征, 并根据这些风格特征, 评估作品与作品之间的风格相似度。传统的身份识别方法, 主要考察作者遣词造句、段落组织等各种代表文体风格的特征, 其中基于标点符号和最常见功能词频数的分析方法受到较为普遍的认同。本文依据文体学理论, 利用 HowNet知识库, 提出一种新的基于词汇语义分析的相似度评估方法, 有效利用了功能词以外的其他词汇, 达到了较好的身份识别性能。

关键词: 计算机应用; 中文信息处理; 身份识别; 语义分析; 文档相似度

中图分类号: TP391 文献标识码: A

Authorship Identification Based on Semantic Analysis

WU Xiao-chun, HUANG Xuan-jing, WU Li-de

(Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

Abstract: Authorship identification techniques are popular in various research areas. The key problems of authorship identification include extracting style-marks and evaluating the document similarity in terms of writing style. Traditional methods examine features revealing the author's writing habits, such as the author's style of using words, constructing sentences and organizing paragraphs, among which analyzing the frequency of punctuations or function words is prevalent. Consulting theoretical stylistics, this paper proposed a new similarity evaluation method based on semantic analysis using HowNet. Experimental results show that content words can also be used as style-marks to discriminate among various authors.

Key words: computer application; Chinese information processing; authorship identification; semantic analysis; document similarity

1 引言

根据文体风格识别文章作者身份是一个应用广泛的研究领域, 不仅在中外名著作者的考证领域引起关注, 而且受到法律界和信息安全部门的重视可应用于恶意邮件识别、网络用户身份识别等信息安全领域。

身份识别的关键问题是从作品中提取出代表语体风格的识别特征, 根据这些识别特征, 评估作品与作品之间的风格相似程度。各国学者通过对英文、中文、日文、荷兰语、希腊语等各种文本的实验, 提出了多种区分文体风格的识别特征, 例如平均句长、词汇丰富性度量、罕见词的比例、特定类型单词(例如元音字母开头)的出现频率等词法特征; 标点符号、词性标注、被动

① 收稿日期: 2005-12-11 定稿日期: 2006-05-17
作者简介: 武晓春(1981—), 女, 硕士研究生, 主要研究方向为自然语言处理.

态使用等句法特征;以及段落长度、行首空格、折标等段落特征。

在目前已经提出的风格识别方法之中,利用标点符号频数、功能词频数作为写作风格的识别特征,不仅受到较为普遍的认同[1,2,7~10],而且较少地依赖于语言学专家或专门 NLP工具的辅助统计。除了拥有实验结果的支持,用标点符号和功能词频数作为风格特征的主要理论依据是:标点符号和功能词的使用,反映了作品句法结构的特点,体现了作者组织句子的习惯,与文章描述的内容,讨论的主题无关[1,7,8]。

传统观点认为区分作者写作风格的特征,应该"独立于内容"[11],依赖于特定内容的特征将无法在不同的上下文中区分作者[1],因此在识别文章作者身份的时候,通常不考虑描述内容的词汇。虽然有学者认为文章风格和内容之间的结合程度比想象中要高[9],也有人把"内容相关"的词汇用来对文章进行作者分类,但有些是针对特定话题的[7],而有些没有取得特别好的效果[8]。

一部作品中"内容相关"的词在数量上通常超过功能词,统计上应更为可靠,但"内容相关"的词汇是否可以用来有效区分不同作者的写作风格,如果利用"内容相关"的词汇来区分写作风格,是否可以说明区分结果确实是作者写作风格的差异,而不是描述内容的差异呢?

基于上述思考,本文从语义角度分析词汇,在将词汇按照语义归类之后,把归类后的每一组词作为风格特征,并结合信息论中熵的概念,对各项特征加权,提出一种新的作品风格相似度评估方法。下文第 2节提出基于语义归类的风格识别特征;第 3节说明利用 HowNet知识库实现词汇语义归类的具体过程;第 4节描述语义归类基础上的特征加权策略和新的作品相似度评估方法;第 5节分析各项实验结果;第 6节总结新方法的优势和缺陷。

2 基于语义归类的风格识别特征

理论文体学的作家决定论认为,作品风格产生于作者对自己思想的合理安排。即知识、事实和发现属于语言使用者的"外部因素",而作品的形式由语言使用者的"内部因素"决定,风格应从"内部因素"去找[6]。根据作品的语言特征识别作者身份,需要假设"每个人都有其自己独特的使用语言的特征"。这个假设符合作家决定论。在接受这个假设的前提之下,关键问题在如何寻找每个人"独特的使用语言的特征"。

作品中独特的词汇可以体现作者使用语言的特征,但是这种形式上的"独特性",需要排除外部因素的影响——即排除作品需要借这个词汇所表达的实质内容的影响。同一个作者表达不同内容时也会采用不同的词汇,如不考虑内容而直接将独特的词汇作为特征,那么这项特征区分的可能仅仅是不同的内容。因此,要比较"表达形式"上的差异,从中寻找每个人的"个性",需要将比较建立在相同或相近内容的基础上。换言之,对于相同相近的内容采取不同的表达形式,才能体现出语言使用者之间"风格"的差异。

但是在比较两部作品的风格差异时,如何寻找"内容相同或相近"的部分呢?一个可行的解决方法是通过分析作品中出现的词汇的语义,用词汇语义之间的"相同相近",来近似表示内容的"相同相近"。即比较两部作品在语义相同之处用词的差异。

文献[1]中,在考证《红楼梦》前 80回与后 40回作者时,除了使用 47个虚字以外,也考察了多组同义词在前 80回和后 40回中的不同分布,这些同义词是文章作者根据阅读经验总结得到,区分了《红楼梦》前 80回和后 40回不同的写作风格,这恰是比较两部作品在语义相同之处用词差异的一个成功例子。但是,如果换了别的研究对象,在没有专家根据作品本身的特点指出合适的同义词作为"风格特征"的情况下,需要有一种自动的方法,从候选的语义相近

的词的集合中,灵活根据作品本身的特性,挑选合适的“风格特征”,自动评价每一个“风格特征”对于比较两部作品相似性的重要程度。

3 利用 HowNet知识库实现词汇语义归类

为了找到一种普适于一般中文作品,并且自动化的风格相似度评估方法,首先需要确定候选的语义相近的词的集合,本文利用了 HowNet2005[①] 知识库。

HowNet2005知识库中共收录了 158 850条记录,每条记录表示一个词语的概念及其描述。由记录号、中文词语、中文词性、中文词语用法示例、英文词语、英文词性、英文词语用法示例和概念等 8项组成,共涉及 82636个中文词语, 24658项概念,中文词语和概念之间可能是多对多的关系:若一个中文词语对应多个概念,表明这个中文词语具有多种意义;若一个概念和多个不同的中文词语有联系,表明这个概念所代表的意义可用多个中文词语来表达。

因此,按照每个中文词语对应的概念的数量,可将所有的中文词语分为两类:对应多个概念的词语是多义词,仅对应一个概念的中文词语是无歧义的词。为避免一词多义对最终识别效果的影响,目前仅考虑所有无歧义的中文词语。

所有无歧义的中文词语涉及的记录中,每个中文词语仅对应一个概念,将每个概念对应的所有中文词语聚合在一起,得到初步归类结果。但是,同一概念的不同中文词语所对应的英文词语之间可能没有交集,这恰好可以进一步反映中文词语之间的语义差别。于是按照以下方法在词语按概念归类结果的基础上,对语义初步归类结果进一步细分。

把任意一个概念对应的所有不同的记录看作一个点集 V,以点集中的点为顶点构造简单图:

如果记录 v_i 与记录 v_j 的中文词语或英文词语相同,那么记录 v_i 与记录 v_j 相邻;否则记录 v_i 与记录 v_j 不相邻。

根据图的基本概念,有

邻接矩阵 $\boldsymbol{M}=(m_{ij})$　　路径矩阵 $\boldsymbol{P}=(p_{ij})$

$$m_{ij}=\begin{cases}1, \text{若 } v_i \text{ 与 } v_j \text{ 相邻}\\ 0, \text{若 } v_i \text{ 与 } v_j \text{ 不相邻}\end{cases}\qquad p_{ij}=\begin{cases}1, \text{若 } v_i \text{ 与 } v_j \text{ 之间至少存在一条路径}\\ 0, \text{若 } v_i \text{ 与 } v_j \text{ 之间不存在路径}\end{cases}$$

并且路径矩阵和邻接矩阵之间满足

$$\boldsymbol{P}=\boldsymbol{M}^{(1)}\vee\boldsymbol{M}^{(2)}\vee\cdots\vee\boldsymbol{M}^{(n)}$$

其中 n 为顶点集中元素总数,即 $n=|V|$,$\boldsymbol{M}^{(k)}$,($1\leqslant k\leqslant n$)是 $\boldsymbol{M}$ 在布尔意义下的 k 次幂[13]。

对于每个概念生成的简单图,根据路径矩阵,求出所有的连通分支,于是每个连通分支构成一个语义类,包含了语义更为接近的中文词语[②]。

由于本文致力于考察不同作者“表达同一语义的时候,选择不同词语”的现象,因此暂不考虑那些仅包含一个中文词语的类,即在简单图中排除所有的悬挂点——如果一种语义仅有一个词语可以表达,没有其他选择,就无法体现风格的差异。

---

① http://www.keenage.com

② 例如,HowNet中的概念 DEF={Age|年龄:host={human|人:modifier={MiddleAge|中年}}}总共涉及到 9条记录,其中包含了 6个不同的中文词语:四十多岁、五十多岁、六十多岁、不惑之年、知命之年、中年。按照上述方法,“四十多岁”和“不惑之年”归为一类;“五十多岁”和“知命之年”归为一类;“中年”独立成一类;“六十多岁”从语义来看,可以和“耳顺之年”归为一类(两者对应的英文词语也是一致的),但“耳顺之年”的概念为 DEF={Age|年龄:host={human|人:modifier={aged|老年}}},两者不属于同一个概念,因此最终没能将它们归为一类。

HowNet知识库中所有概念项包含“FuncWord|功能词:”的记录所涉及的中文词语,作为中文功能词。HowNet总共定义了99个与功能词相关的概念,涉及938个不同的中文功能词。分别对功能词和其他词语按照上述原则归类并排除不予考虑的项,最后得到106个的功能词类,涉及507个无歧义的功能词;10365个其他词类,涉及35123个无歧义的其他词语。以词语的语义归类结果为新的风格特征,可以在每一项特征中观察到不同作者表达同一语义时对于不同形式的选择。

除了HowNet以外,其它分类词典,例如《同义词词林》也提供类似的同义词集合。

例如,《同义词词林》[①]中表达“年轻人”这个语义,提供了“青年人”、“青年”、“小伙子”、“青少年”、“后生”、“弟子”、“子弟”、“初生之犊”、“年青人”、“小伙”、“小青年”和“年轻人”等同义词;而在HowNet2005提取出来的同义词表中,这个语义进一步被细分为4类:第一类包含“年青人”、“年轻人”、“青年人”、“青壮年”和“小青年”5个词语;第二类包含“儿郎”和“后生”两个词语;第三类包含“弟子”、“门生”、“门徒”、“门下”、“徒弟”、“徒工”、“学徒”、“学徒工”、“艺徒”、“学员”、“学子”和“门生”11个词语;第四类包含“初生牛犊”和“初生之犊”两个词语。

又如,《同义词词林》中表达“普通人”这个语义,提供了“匹夫”和“个人”两个词语,而在HowNet2005提取出来的同义词表中,提供了“百姓”、“苍生”、“草野小民”、“大众”、“老百姓”、“黎民”、“黎庶”、“氓”、“民众”、“匹夫匹妇”、“平民”、“平头百姓”、“黔首”、“世人”、“庶民”、“庶人”和“子民”17个不同词语。但与此同时“匹夫”和“个人”两个词语没有包括在HowNet的这个语义类内,说明HowNet和《同义词词林》对同一个词语的语义规定是有所不同的。另外也存在《同义词词林》中提供的同义词集合,在HowNet2005按照上述方法提取的同义词表中,没有对应的项。

两种同义词集各有长处,本文仅在HowNet2005提取的同义词集上进行了实验。

4 特征加权与新的作品相似性评估方法

评估任意两部作品之间的风格相似度时,首先统计每部作品各自所涉及的语义类的数量,以及各个语义类中每个词的词频;其次观察同一个语义类中,两部作品词频的差异;最后综合所有语义类的词频比较结果,评估两部作品之间的相似性。

设由 m 个语义类构成的风格特征集合为 $C=\{C_1, C_2, \cdots, C_m\}$,对于第 i 个语义类 C_i,设其中总共涉及的不同词语个数为 n_i,又设这 n_i 个不同的词,在作品 A 中相应的词频分别为 $f_{A_{i1}}$, $f_{A_{i2}}$, …, $f_{A_{in_i}}$。

求每个词频在语义类 C_i 中的 *zscore* 标准得分:

$$z_{A_{ij}}=\frac{f_{A_{ij}}-mean(f_{A_{ij}})}{stdev(f_{A_{ij}})}, \text{其中 } 1\leqslant j\leqslant n_i$$

于是语义类 C_i 中所有的词语可表示为 $z_{A_{i1}}, z_{A_{i2}}, \cdots, z_{A_{in_i}}$

以每个语义类为单位,对其中的所有的词求 *zscore* 标准得分,最终将作品 A 表示成如下向量:

$$OpusA=(C_1, C_2, \cdots, C_m)=(z_{A_{11}}, z_{A_{12}}, z_{A_{1n_1}} z_{A_{21}}, z_{A_{22}}, \cdots, z_{A_{2n_2}}, \cdots, z_{A_{m1}}, z_{A_{m2}}, \cdots, z_{A_{mn_m}})$$

在衡量两个作品风格之间的相似程度时,为了便于进一步对不同的语义类赋予不同的权重,用欧氏距离来衡量两个作品之间的相似度。

---

① 本文中《同义词词林》的例子,来源于《同义词词林(扩展版)》的样例
http://ir.hit.edu.cn/phpwebsite/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=162

$$Dis(OpusA, OpusB) = \sqrt{\sum_{1 \le i \le m} SumSquare(C_i)} = \sqrt{\sum_{1 \le i \le m} \sum_{1 \le j \le n_i} (z_{A_{ij}} - z_{B_{ij}})^2}$$

距离越大，差异越大，相似度越小；反之则相似度越大。

事实上，由于不同作品实质内容的差异，它们所涉及的语义类互不相同，虽然完全没有交集的可能性不大，但即便有不少交集，不同的语义类上比较结果的可靠性互不相同，因此对不同的语义类赋予不同的权重。

对于一个包含 n 个词的语义类 C，假设这 n 个词在作品 A 中的词频分别为 f_{A1}，f_{A2}，…，f_{An}，在作品 B 中的词频分别为 f_{B1}，f_{B2}，…，f_{Bn}，则作品 A 在语义类 C 上的总词频为 $F_A = \sum_{i=1}^{n} f_{Ai}$，作品 B 在语义类 C 上的总词频为 $F_B = \sum_{i=1}^{n} f_{Bi}$，则语义类 C 相对于两个作品的熵为

$$Entropy(C) = -\frac{F_A}{F_A + F_B} \log \frac{F_A}{F_A + F_B} - \frac{F_B}{F_A + F_B} \log \frac{F_B}{F_A + F_B}$$

熵衡量了语义类 C 中的词在两部作品中分布的不确定性，可以作为语义类 C 的权重。如果作品 A 在语义类 C 上的总词频 F_A 和作品 B 在语义类 C 上的总词频 F_B 之间的差异很大，用语义类 C 来衡量两个作品之间风格差异的可靠性较差，相应的熵的值较小，一个极端的情况是，作品 A 在语义类 C 上的总词频为 0，而作品 B 在语义类 C 上的总词频非零，此时，按照熵定义中 $\log 0 = 0$ 的规定，语义类 C 的熵为 0，表示语义类 C 不可用于衡量两个作品之间风格的差异——这在实际情况下是合理的，因为此时作品 A 和作品 B 在语义类 C 上的词频差异，反映的是两部作品之间语义的差异，而不是相同语义下不同表达形式的差异。如果作品 A 和作品 B 在语义类 C 上的总词频之间差异较小，那么语义类 C 所反映的作品之间的差异，较少涉及到语义本身的差异，更接近表达形式的差异，此时语义类 C 的熵恰好比较大。

因此，在评估作品 A 和作品 B 之间差异度时，先计算每一个语义类的熵 $Entropy(C_i)$，设语义类 C_i 在评估作品 A 和作品 B 之间差异度中所占的权重为 w_i，

定义 $w_i = \dfrac{Entropy(C_i)}{\sum_{1 \le j \le m} Entropy(C_j)}$

于是权重向量 $W = (w_1, w_2, \cdots, w_m)$ 的值，对应了每个语义类在比较作品 A 和作品 B 时的相对重要程度。

按照上述公式，熵较大的语义类所占权重较大，在语义类加权的基础上，重新定义作品 A 和作品 B 之间的差异：

$$wDis(OpusA, OpusB) = \sqrt{\sum_{1 \le i \le m} w_i \times \frac{SquareSum(C_i)}{n_i}} = \sqrt{\sum_{1 \le i \le m} w_i \times \frac{\sum_{1 \le j \le n_i} (z_{A_{ij}} - z_{B_{ij}})^2}{n_i}}$$

上述公式可以看作加权欧氏距离的一个变体，一般加权欧氏距离在比较任意两个向量的时候，各项权重 w_i 是固定不变的，这就保证了任意两个向量的差异度计算结果之间具有可比性，即 $Dis(A, B)$ 和 $Dis(C, D)$ 的值具有可比性。但根据上文提到的特征加权策略，比较不同的作品对时，各语义类所占的权重是变化的，即比较作品 A、B 与比较作品 C、D 时权重向量 W 是不同的。在这种情况下，仍然需要保证不同作品对计算得到的差异度之间具有可比性。首先要保证被加权的项与项之间具有可比性，所以每个语义类上两个作品在各个词语标准得分上的差的平方和 $\sum_{1 \le j \le n_i} (z_{A_{ij}} - z_{B_{ij}})^2$ 被除以语义类各自包含的总词数 n_i；其次要保证所有权重的总

和固定，这里所有语义类的权重的总和永远为 1，即满足 $\sum_{1 \leq i \leq m} w_i = 1$。

5 实验分析

实验一在表 1所示的语料上分别以功能词和非功能词集合作为风格特征，对比了基于语义分析的不同词汇集合及语义类加权方法对分类性能的影响；实验二在限制特定话题的情况下，考察新方法对不同作者作品的分类性能。

实验一：基于语义分析的不同词汇集合以及语义类加权方法对相似度的影响

表 1 分类语料情况

作 者	文章数	文章发表时间段	总词数
陈一鸣	36	2005. 09. 21 ~ 2006. 04. 11	24807
黄培昭	54	2005. 11. 06 ~ 2006. 04. 14	32501
廖先旺	34	2005. 05. 26 ~ 2006. 03. 24	24552
任 彦	42	2005. 06. 16 ~ 2006. 03. 24	30590
施晓慧	36	2005. 07. 12 ~ 2006. 04. 11	24875
总 计	202		137325

为了验证功能词和其他词汇对于区分不同写作风格的贡献大小，分别以按照语义归类的功能词集合和非功能词集合表示文本，并分别以不加权和加权两种方法，计算作品两两之间的相似度。

实验语料由 5位《人民日报》记者近期发表的作品组成。每位记者的文章数量及其概况如表 1所示。

在进行作者分类的过程中，对每位作者随机抽取一定数量的作品作为训练语料，剩余作品作为测试语料。判断一个作品作者归属的时候，比较这个作品与每一位作者所有训练作品的平均相似度，以最近作者为系统答案。每组实验重复 1000次，对每位作者求取 1000次实验的平均查全率和平均查准率，并在此基础上求出调和平均值（F值）。最后求取 5位作者的宏品均 F值，作为最终的评估指标。

由于相似度分析对作品所包含的词语数量有一定要求，如果两部作品之间没有任何可以比较的语义类，则最终无法按上述方法得出相似度分析结果，而《人民日报》上的文章本身较短，即便不考虑千字左右的新闻，剩余每篇文章包含的词语数平均也只在 600 – 700左右，通常可用于对照的语义类很少，甚至无法得到相似度分析结果，因此将每位作者的文章按照发表时间先后排序，将连续发表的若干个作品合并为 1个考察单位，保证每个考察单位包含的词语数量不少于一定阈值。

当每个考察单位的词语数量不少于 1500词的时候，5位作者总共得到的 72个考察单位，不断扩大训练集，进行分类测试，得到如下结果：

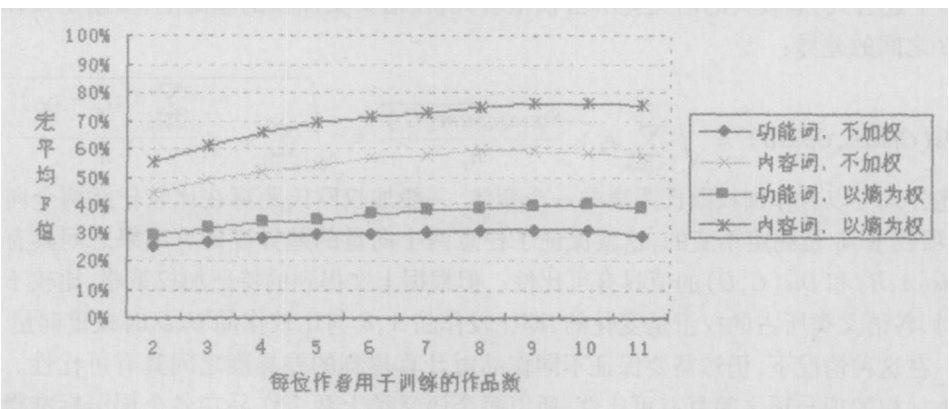

图 1 不同特征集合以及加权方法的分类性能比较（5作者 72个考察单位）

从上图来看，以非功能词为特征时，效果优于功能词特征；利用熵对各语义类进行权重评估时，效果优于不加权；训练集增大时，分类性能有所提高。

在非功能词集合上，以熵为权的方法性能最好，但宏平均 F值不到 80%。这可能是每个考察单位中真正用于相似性比较的词语数量较少引起的。对 72个考察单位的（72 -1）×72 / 2=2556次相似度比较进行分析，发现以非功能词为特征时，每一对作品两两比较时，每个考察单位中，平均只有 146.7词实际用于作者身份的判定。

将 5位作者的作品重新合并，分别使每个考察单位包含的词语数量保持在 3000词左右，得到 42个考察单位，重复上述分类实验，得到如表 2所示结果。

表 2 非功能词集合，以熵为权的分类性能（5作者 42个考察单位）

每位作者用于训练的作品数量		2	3	4	5	6
宏平均	查全率	71.35%	78.04%	82.61%	86.26%	88.99%
	查准率	71.33%	77.96%	82.17%	85.22%	86.52%
	F值	70.19%	76.72%	81.19%	84.41%	86.23%

对 42个考察单位的（42 -1）×42 /2=861次相似度比较进行分析，每一对作品两两比较时，平均每个考察单位有 324个词实际用于作者身份的判定。有效词频的提高改善了分类性能。

实验一表明，非功能词用于作者分类可以取得比功能词更好的效果，对语义类加权可以提高分类性能，以熵为权重的相似度比较方法对作品长度有一定要求。

实验二：特定话题上不同作者的作品相似度分析

以往研究中通常使用功能词作为风格特征，而很少使用非功能词，主要由于非功能词和文本内容相关。新的相似度比较方法能否区分讨论“同一话题”的不同作者的风格呢？为此，本文收集了《人民日报》上两位不同作者关于“巴以问题”的 50篇文章，用于分类测试。特定话题语料的收集情况如表 3所示。

表 3 特定话题语料情况

作 者	文章总数	文章发表时间段	总词数
黄培昭	26	2005.08.22 ~ 2006.04.14	16657
岳麓士	24	2000.07.17 ~ 2006.04.12	15884
总 计	50		32541

将每位作者连续发表的若干个作品合并为 1个考察单位，每个考察单位不少于 3000个词语，最后得到 10个考察单位，对此进行相似度分析，对每种方法每种语料划分进行随机实验的分类结果如图 2所示。

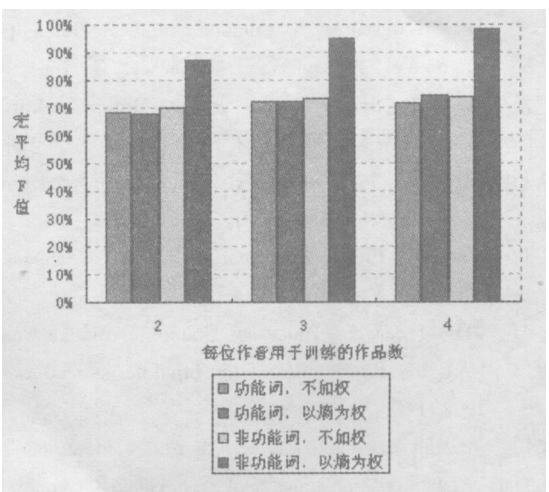

图 2 特定话题上不同特征集合以及加权方法的分类性能比较（2作者 10个考察单位）

从分类效果来看，在非功能词集合上，以熵为权的方法取得比其他方法更好的性能。说明新的相似度比较方法，虽然使用了“非功能词”，但在“话题相同”的情况下仍能区分不同作者的写作风格。

对于特定话题，通常有特定的关键词（一般是非功能词）与之相关，在不同作者的文本中都以高频率出现。例如，同是讨论“巴以问题”的两位作者的作品中，都出现“以色列”、“巴勒斯坦”、“加沙”等词语，这些词语都是“非功能词”，但因为它们不属于任何一个语义类，所以计算相似度的时候不予考虑，也就是说，从 HowNet2005提取语义类的时候，同时已经排除了一

些非功能词,这些非功能词所表达的语义,没有其它词可以替代。

如果和话题内容相关的某个“非功能词”也出现在某个语义类中,那么必然存在其它可以代替它的词,通过观察作者对用词的选择,正可以区分不同的写作风格。例如,同是讨论“巴以问题”的两位作者的作品中,都出现了“执政”一词,与之同义的词语中,一位作者还使用了“当政”、“掌权”等词,另一位作者没有用到这些词语,体现了用词的差异。新的相似度比较方法通过考察这种“差异”,区分不同作者的作品。

实验二表明,通过有选择地使用“非功能词”,比较不同作者用词的差异,可以区分内容相近的不同作者的作品。

6 结论

本文提出一种基于词的作品风格相似性分析方法,发现除了普遍认同的功能词以外,其他词汇经过合理筛选也可以有效地区分不同作者写作风格的特征。其次本文采用的基于熵的特征加权方法,在理论上解释了不同词汇在作品比较过程中的重要程度,在实验中提高了分类性能。最后,本文提出的方法,对作品长度有一定的要求,同时也发现作品中特有的一些词汇在HowNet知识库中没有定义,所以扩充知识库或改进知识库的利用方法,更精确更完整地定义“语义相同,形式不同”的风格特征集合,将进一步提高结论的可靠性。

参 考 文 献:

[1] 陈大康.从数理语言学看后四十回的作者——与陈炳藻先生商榷[J].红楼梦学刊,1987,(1):293-318.

[2] 李贤平.《红楼梦》成书新说[J].复旦学报(社会科学版),1987,(5):3-16

[3] Ward Elliott, Robert Valenza. Was the Earl of Oxford the true Shakespeare? [J] A Computer Aided Analysis. Notes and Queries, 38, 501-506, April 1991.

[4] Efstathios Stamatatos, Nikos Fakotakis, George Kokkinakis. Computer-Based Authorship Attribution Without Lexical Measures[J]. Computers and the Humanities, Volume 35, Issue 2, 193-214, May 2001.

[5] O. de Vel, A. Anderson, M. Corney, G. Mohay. Mining E-mail Content for Author Identification Forensics. SIGMOD Special Section on Data Mining for Intrusion Detection and Threat Analysis[C], 55-64, 2001.

[6] 胡壮麟(编著).理论文体学[M].外语教学与研究出版社,2000,50-63.

[7] Jiexun Li, Rong Zheng, Hsinchun Chen. From Fingerprint to Writeprint[J]. Communications of the ACM, 47(3): 70-76, March 2004.

[8] Carole E. Chaski. Empirical evaluations of language-based author identification techniques[J]. Forensic Linguistics 8(1): 1-65, 2001.

[9] Harald Baayen, Hans van Halteren, Anneke Neijt, Fiona Tweedie. An experiment in authorship attribution [A]. In: Proceedings of the 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002)[C]: 29-37.

[10] Niamh McCombe. Methods Of Author Identification[D]. B.A. (Mod.) CSLL Final Year Project, 2002.

[11] Efstathios Stamatatos, Nikos Fakotakis, George Kokkinakis. 2001. Automatic Text Categorization in Terms of Genre and Author[J]. Computational Linguistics, Vol.26, Issue 4-December 2000, 471-495.

[12] Yuta Tsuboi. Authorship Identification for Heterogeneous Documents[D]. Master's Thesis, Nara Institute of Science and Technology, 2002.

[13] 朱洪,胡美琛,张霭珠,赵一鸣(编著).离散数学教程[M].上海科学技术文献出版社,1999,89-94.