领域术语自动抽取及其在文本分类中的应用

刘 桃,刘秉权,徐志明,王晓龙
(哈尔滨工业大学计算机科学与技术学院,黑龙江哈尔滨 150001)

摘 要: 本文提出了一种基于信息熵的领域术语抽取方法,在给定领域分类语料的前提下,该方法既考虑了领域术语在不同领域类别间分布的不均匀性,又考虑了其在特定领域类别内分布的均匀性,并针对语料的不平衡性进行了正规化. 人工评测显示该方法能更准确有效地抽取领域术语. 本文还将该算法应用于文本分类,用于代替传统特征选择算法,实验表明,该算法能够显著提高文本分类的精度.

关键词: 领域术语; 信息熵; 正规化; 文本分类; 特征选择
中图分类号: TP391.2 **文献标识码**: A **文章编号**: 0372-2112 (2007) 02-0328-05

Automatic Domain-Specific Term Extraction and Its Application in Text Classification

LIU Tao, LIU Bing-quan, XU Zhi-ming, WANG Xiao-long
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang* 150001, *China*)

Abstract: A statistical method based on information entropy is proposed for domain-specific term extraction from domain comparative corpora. It takes into account the distribution of a candidate word among domains and within a certain domain. Normalization step is added into the extraction process to cope with unbalanced corpora. The proposed method characterizes attributes of domain-specific term more precisely and more effectively than previous term extraction approaches. Domain-specific terms are applied in text classification as the feature space. Experimental results indicate that it achieves better performance than traditional feature selection methods.

Key words: domain-specific term; information entropy; normalization; text classification; feature selection

1 引言

领域术语自动抽取[1]是指从一定规模的语料中抽取出能反应某一领域文本特征或共性的词语,是自然语言处理中的一项重要任务. 术语抽取被广泛应用于本体构建[2,3]、自动摘要[4]、语言模型[5]等领域. 单纯靠语言学专家抽取领域术语费时费力,因而很难形成规模,开发一种自动化的方法来辅助术语抽取显得尤为必要,能为许多面向领域的应用提供重要语言学资源.

许多研究者在领域术语抽取方面做了不少工作. 通常被采用的方法可以分为基于规则的和基于统计的方法两大类. 基于规则[4]的方法是通过预先设定许多规则模版,然后把待处理语料中与规则模版相匹配的词语作为领域术语候选. 规则方法的不足在于很难制定一个完备的规则集来穷尽所有语言现象,并且当已有许多规则时,还需要考虑多个规则之间的冲突及解决. 基于统计的方法通常包括机器学习方法[5,6]和基于统计量度的方法[2,3]. Jianfeng Gao[5]和 Henri Avancini[6]分别采用词语聚类法和分类法获取领域术语. 陈文亮[7]采用 bootstrapping 方法逐步扩大领域词汇数量. 基于统计量度的方法是从领域分类语料中统计用词规律从而发现领域术语. 已有的统计量度包括 TFIDF[3]、KFIDF[8]、DR + DC[2]. KFIDF 量度与 TFIDF 相比,引入了词语的类别信息. DR 和 DC 分别表示词语的领域相关性和领域一致性,领域一致性是指词语在其相关领域的所有文档中分布的一致性,它对获取高质量领域术语起重要作用. 该方法被用于英文文本的领域术语抽取并取得了很好的效果,但该方法没有考虑不同领域语料的规模以及不同文档长度对领域术语抽取的影响.

本文在前人工作基础上,基于如下两个度量标准[9]提出了一种新的领域术语抽取方法,它能够更加准确、鲁棒地度量领域术语的属性. (1) 领域术语应该在不同领域类别间分布不均匀. (2) 领域术语在其相关领域的文档集中应尽可能分布均匀. 我们采用信息熵来衡量以上两个分布,并根据不同领域语料的规模及文档长度做正规化. 这样也就是从类间分布、类内分布及语料规模三方面来衡量每个词语. 本文还用该领域术语抽取方法代替传统特征选择方法应用于文本分类,使文本分类精度有了很大提高.

收稿日期:2005-10-21;修回日期:2006-11-23
基金项目:国家自然科学基金(No. 60673037)

2　领域术语的自动抽取

2.1　符号定义

方便起见,我们定义如下数学符号:

m:领域类别个数

$D_i(1 \le i \le m)$:第 i 个领域类别

$n_i(1 \le i \le m)$:类别 D_i 中的文档数目

$P(D_i|W)$:词语 W 出现在类别 D_i 中的概率

$d_{ij}(1 \le j \le n_i)$:类别 D_i 中的第 j 个文档

l_{ij}:文档 d_{ij} 的长度,即在该文档中出现的所有词语的词频之和

L_i:类别 D_i 包含的所有文档长度之和

WS_{D_i}:类别 D_i 的领域术语集合

WS_{rel}:领域相关词语集合

WS_{irre}:领域无关词语集合

WS:文本中所有词语集合

领域术语抽取的目标是给出集合 WS 的一个合理划分,满足 $WS_{\mathrm{rel}} \cup WS_{\mathrm{irre}} = WS$, $WS_{\mathrm{rel}} \cap WS_{\mathrm{irre}} = \varnothing$,同时求出 $WS_{D_i} \subseteq WS_{\mathrm{rel}}$,在领域术语不兼类时, $WS_{D_1}, WS_{D_2}, \cdots, WS_{D_m}$ 构成 WS_{rel} 的一个划分. 本文通过全面考虑词语在领域类内、类间分布情况来给出 WS 的合理划分.

2.2　正规化的类间分布

为了衡量词语在领域类别间的属性,本文利用信息熵[13]定义词语在领域类别间的分布为 corpus distribution(CD):

$$CD(W) = -\sum_{i=1}^{m} P(D_i|W)\log P(D_i|W) \qquad (1)$$

$CD(W)$ 值越小,说明词语 W 越有可能成为某个或某几个类别的术语. 与单纯考虑词语在类别出现比例($P(D_i|W)$)的方法[2]相比,该方法不仅考虑了词语 W 的类别出现比例 $P(D_i|W)$,同时考虑了 W 在不同类别间的分布,也就是出现类别数越少越好. 这是由于不同领域的领域术语可能会有交叉,一个词语可能同时为两个领域的领域术语,这就需要在考虑 $P(D_i|W)$ 的同时,考虑 W 在类别间的分布情况 $CD(W)$. 以 2003 年 863 文本分类评测的标准语料为例:词语"党性"和"知觉"在类别 A(马列主义)均以 0.5 的概率出现,但"党性"只出现在 A(马列主义)与 D(政治、法律)两个类别,而"知觉"共出现在 11 个类别的语料中,那么"党性"极有可能是 A 类的领域术语(当然可能兼 D 类),但以同样概率出现在 A 类的"知觉"则不是该类的领域术语. 可见传统方法中仅考虑词语的领域出现概率是不够的,还应考虑词语在多个领域间的分布.

不难发现,某个领域的语料越多,那么一个词语在该类语料中出现的可能性将越大,为了消除语料规模对词语的出现带来的影响,本文提出了基于领域规模的正规化方法. 词语 W 的正规化的类间分布为:

$$NCD(W) = -\sum_{i=1}^{m} P'(D_i|W)\log P'(D_i|W) \qquad (2)$$

其中,

$$P'(D_i|W) = \frac{P(D_j|W)/L_i}{\sum_{j=1}^{m}(P(D_j|W)/L_j)} \qquad (3)$$

2.3　正规化的类内分布

领域术语抽取的第一条标准是将 NCD 值尽可能小的词语 W 作为领域术语候选. 领域术语抽取的第二条标准是领域术语应该在其相关领域内分布尽可能均匀. 这条标准对于获取高质量的领域术语起着重要作用并首次被[2]提出,但原文没有考虑不同文档长度对分布的影响,本文定义了 $NDD(W, D_i)$ 来刻画词语 W 在类别 D_i 中的正规化类内分布:

$$NDD(W, D_i) = -\sum_{j=1}^{n_i} P'(d_{ij}|W)\log P'(d_{ij}|W) \qquad (4)$$

$$P'(d_{ij}|W) = \frac{P(d_{ij}|W)/l_{ij}}{\sum_{k=1}^{n_i}P(d_{ik}|W)/l_{ik}} \qquad (5)$$

$NDD(W, D_i)$ 值越大, W 越有可能成为类别 D_i 的领域术语. 如果 W 只在 D_i 的一篇文档中出现多次,在 D_i 的其他文档中没有出现,那么很有可能 W 在这篇文档中的出现是偶然的,不能代表该领域的普遍特征. 比如"蛔虫"在类别 G(文化、科学、教育、体育)的一篇介绍中小学生健康问题中蛔虫感染的文章中多次出现,但在该类别的其它文档中未出现,那么该词就不具有领域代表性,不能成为 G 类的领域术语. 另外领域分类文档很难百分百准确,难免会有一两篇文档被错分,这样更有可能出现上述某个非该领域的词汇多次出现在该领域的某篇文档中的情况. 因此通过衡量词语的类内分布会排除许多类似噪声.

2.4　算法

输入:领域分类的语料库 D,NCD 域值 α,NDD 域值 β

输出:每个类别的领域术语

(1) for $i = 1$ to m do
(2)　for $j = 1$ to n_i do
(3)　　依次读入 d_{ij} 中的每个词语 W 加入 WS 并记录频次,同时 l_{ij} 加 1
(4)　end for
(5)　计算 L_i
(6) end for
(7) for all($W \in WS$) do
(8)　计算 $NCD(W)$
(9)　if $NCD(W) < \alpha$ then
(10)　　求 $x = \arg\max_x(P'(D_x|W))$
(11)　　计算 $NDD(W, D_x)$
(12)　　if $NDD(W, D_x) < \beta$ then
(13)　　　将 W 加入 WS_{D_i} 中
(14) end for

3　在文本自动分类中的应用

下面我们给出领域术语在文本自动分类中的应用. 文本自动分类是指在给定的分类体系下,根据文本内容自动判别文本类别的过程. 其关键技术包括文本表示、特征选择和重

构、训练和分类算法几个部分. 本文的文本自动分类系统[10]是基于向量空间模型的 K 最近邻分类器,文本表示采用一个变形的 Okapi 权重计算公式,并对特征进行潜在语义索引从而达到特征重构与压缩的目的,下面重点介绍特征选择部分.

特征选择是从分类训练语料的所有词语中挑选出对分类贡献大的词语,以达到降低特征空间维数的目的,它在文本分类中具有重要作用. 我们知道,领域术语是某个领域最有代表性且最常涉及到的词汇,而人类区分不同类别的文档其实就是通过这些代表性的词语来区分的. 这就启发我们利用领域术语抽取算法代替传统特征选择方法[11]:词频与倒文档频度(TFIDF)、期望交叉熵(ECE)、χ^2 检验(CHI)、互信息(MI)、信息增益(IG)、文本证据权(WE)、文档频度(DF)等. TFIDF 和 DF 是最简单的评价函数,其优点是简单易计算,缺点是没有考虑词语与类别的关系. 其它公式虽然考虑了词语与类别之间的关系,但是并没有考虑词语在本领域内文本中分布的一致性. 本文的领域术语抽取算法不仅考虑了词语在不同类别间的分布,还考虑了词语在其相关类别内的分布一致性.

本文算法抽取出的词语数目远大于文本分类需要的特征数目,因此本文用公式 $RS(W, D_x)$ 表示词语 W 在其相关领域 D_x 中的排序权重,其中领域类别数 m 和 D_x 包含的文档数 n_x 用于将熵值 NCD 和 NDD 归一化. 本文经验最优化的 λ 取值为 0.5(图 5).

$$RS(W, D_x) = -\lambda NCD(W)/\log m + (1-\lambda)\, NDD(W, D_x)/\log n_x \tag{6}$$

4 实验

4.1 实验数据

文本分类系统的训练和测试语料分别采用 2003 年和 2004 年 863 文本分类评测的标准语料,采用中图法分类体系,不包括难以判定的"T 工业技术"和"Z 综合性图书"两类,共 36 类,每类 100 篇语料. 虽然训练语料中各类的文档数相同,但由于每篇文档长度相差比较大(图 1),导致各类的总文档长度相差比较大(图 2). 也就是说即使在不同语料文档数相同的情况下,语料的潜在不平衡性依旧比较大.

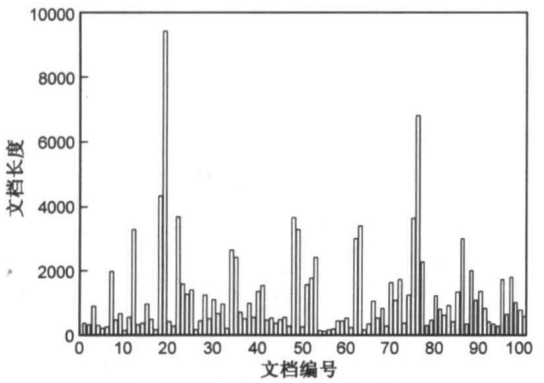

图 1 A 类文档规模分布图

4.2 领域术语抽取

在 2003 年的语料上运行领域术语抽取算法,将类间熵小于 $\alpha(2.5)$ 且类内熵大于 $\beta(0.5)$ 的词语作为领域术语. 表 1 给出了随机抽取的六个领域的部分结果. 由表可见,其中大部分词语均为本领域的领域术语,极少数词语(斜体部分)不具有领域相关性.

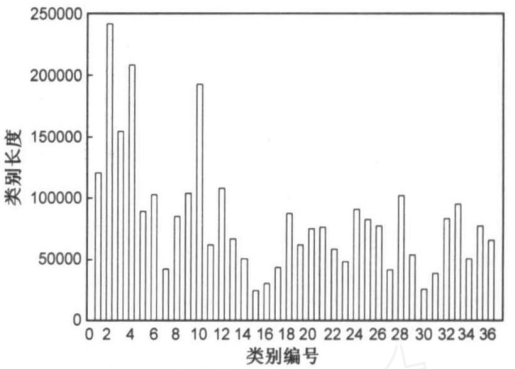

图 2 所有类别规模分布图

表 1 部分抽取结果

B 哲学、宗教	哲学 哲学史 存在论 德性 现象学 海德 形而上学 *独断* 对象化 哲学家 道德 佛陀 唯心 佛家 天国 佛学 相对主义 *重估* 哲学界 绝对性
E 军事	作战 军种 军事 军队 战争 兵力 事变 美军 新军 战法 我军 武器 军兵种 火力 军事科学 战场 打赢 航空兵 外军 陆军
H 语言、文学	汉语 语法 方言 动词 语言 语言学家 汉字 读音 书面语 英语 词汇 句法 普通 句子 词语 *交际* 音节 国际语 代词 副词
R 医药、卫生	患者 治疗 血管 临床 疗效 药物 病人 冠心病 并发症 动脉 冠状动脉 手术 症状 术后 口服 疗法 服用 心肌 疾病 血栓
TD 矿业工程	该矿 回采 煤层 工作面 赋存 矿体 掘进 矿井 尾矿 矿石 精矿 矿区 矿岩 巷道 矿山 采煤 黄铁矿 勘查 浮选 品位
TS 轻工业、手工业	包装 食品 调味 保质期 肉制品 玻璃瓶 品牌 肉食品 方便化 果汁 腥味 肉类 专卖店 草莓 货架 糖度 猪肉 番茄酱 杀菌 分割肉

表 2 给出该六个领域中被排除掉的部分词语,它们在其对应领域满足 $P(D|W)$ 大于 0.5,且 NCD 大于 2. 也就是在对应领域出现比例比较大,若用基于出现比例的判定方法,这些词语将被当作领域术语. 但由于 NCD 值过大,本文方法将这些词语滤掉. 这些词语中绝大多数不是领域术语,极少数词语(斜体部分)是领域术语,但由于这些词语较为通用,所以 NCD 值较大,导致用本文方法不能正确提取. 由此可见用类间熵来衡量词语的领域相关性比采用基于出现比例的方法更为合理.

表 2 标准一过滤掉的部分词语

B 哲学、宗教	反思 知觉 彼岸 有限性 万物 人权
E 军事	*打击* 变革 着眼 后勤 纵深 海湾
H 语言、文学	似的 埃及 *字面*太极 听话 口腔
R 医药、卫生	狭窄 每日 *医院*安慰 一例 陕西省
TD 矿业工程	考查 露天 倾角 混入 分期 花纹
TS 轻工业、手工业	粉末状 织成 居首 厂区 动物性 别致

表 3　标准二过滤掉的部分词语

B 哲学、宗教	笃实 知命 隐者 阴德 译稿 显摆 下篇 唤回 后王 法共
E 军事	像片 圣路易斯 肩章 公安部队 *大尉* 长机 巴解组织 流通性 塔山 北宁
H 语言、文学	硬腭 译员 娘娘 外务部 速记 舌尖 *舌根音* 平声 辣子 大姨
R 医药、卫生	珍奥 *药丸* 血肿 维生素 D 手术组 内窥镜 磨牙 门店 纯净水 清心
TD 矿业工程	紫金山 振磨机 砚台 砚石 探矿权 台板 三叶虫 南翼 辉铜矿 导水管
TS 轻工业、手工业	云锦 维生素 A 筒子 手袋 圣雪绒 三角区 乳粉 男式 华邦 金龙鱼

表 3 给出在其对应领域 NCD 值和 NDD 值均为零的部分词语. 它们频繁出现在某一个领域的一篇文档中,出现的最高频次多达 40 次. 如果不用第二条抽取标准,这些词语将被抽取出,但事实上它们中的大多数不是领域术语,只有如"大尉"、"舌根音"和"药丸"等少数词语是领域术语. 我们可发现,这少量词语在 R 医药和 TS 轻工业两个领域所占比例较大,原因是这两个领域的领域术语(比如药名、品牌名)比较多,100 篇文档构成的语料不足以枚举如此多的领域术语,因此会较多出现部分领域术语只出现在一篇文档中的情况,但这种情况会随着这些领域语料规模的上升而得到缓解. 由此可见,领域术语抽取第二条标准是合理并有效的.

4.3　领域术语人工评测

在不同的分类体系下或是面向不同应用,领域术语的定义会有很大差别. 通常有两种术语评测方法[2]:一种是人工评测,其缺点是受人的主观性影响. 另一种是在应用系统中评测,这种方法完全面向应用,不同的应用系统可能会产生不尽相同的结果. 本文首先人工评测了 NCD + NDD 与 DR + DC 方法[2]的抽取结果. DR + DC 方法中使用的域值(0.35,0.49)和原文中的一致,同时按此域值抽取的术语总数和本文 NCI + NDI 方法大体相当,这样保证了对比的公平性. 表 4 显示了两种方法在六个类别中抽取的词语数目. DR + DC 方法抽取词语个数会随着语料规模的变化产生较大变化,如从 B 类中提取出 1776 个词语,其原因是对不同规模的语料使用了相同的域值. 而 NCD + NDD 方法根据语料规模作了正规化,因而提取的词语数目不完全依赖于语料规模. 分别用式(7)和文献[12]中的 TW 公式对本文方法和 DR + DC 方法抽取的领域术语排序,图 3 和图 4 给出了两种方法的前 200 个词语的正确率和第200个以后的词语的正确率,对于语料规模很大的类

表 4　领域术语抽取数目

类别编号	词语总数	抽取词语个数	
		DR + DC	NCD + NDD
B	88830	1776	881
E	41030	621	677
H	38666	638	741
R	18182	444	571
TD	27925	318	162
TS	21792	257	358

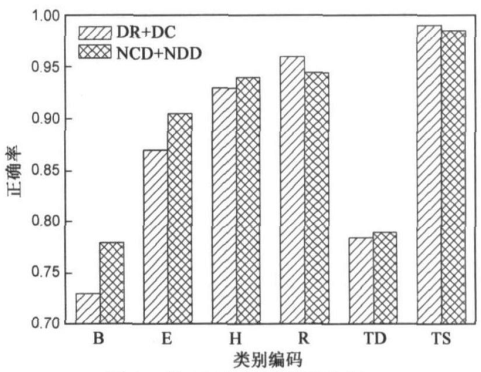

图 3　前 200 个词语的正确率

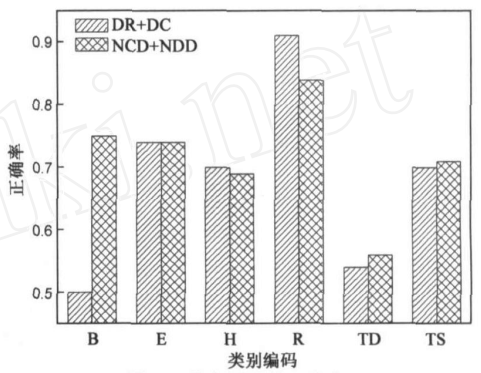

图 4　其余词语的正确率

别,本文方法正确率要明显高于 DR + DC 方法,在其它正确率相当的类别中,抽取的术语数目要明显高于 DR + DC 方法.

4.4　领域术语在文本分类中的应用评测

本文将术语抽取方法应用于文本分类,并在提取特征数量相同的前提下,和传统特征选择方法进行了对比. 图 5 给出了公式(7)中不同 λ 取值对应的文本分类 F1 值. 由表 5 可见,NCD + NDD 方法比以往特征选择中表现最好的 CHI 方法和以往术语抽取中表现最好的 DR + DC 方法在 F1 值上分别高了 4.9 和 3.8 个百分点.

表 5　文本分类对比实验

方法	准确率	召回率	F1 值
MI	0.419	0.409	0.414
DF	0.556	0.529	0.542
WE	0.564	0.541	0.552
IG	0.559	0.546	0.552
TFIDF	0.596	0.572	0.584
ECE	0.617	0.597	0.607
KFIDF	0.616	0.601	0.608
CHI	0.633	0.602	0.617
DR + DC	0.631	0.626	0.628
NCD + NDD	0.663	0.669	0.666

5　结论

本文提出了一种从领域分类语料中抽取领域术语的统计方法. 该方法通过词语的领域间分布和领域内分布来刻画词语的领域属性,并用信息熵来衡量分布的均匀性,同时考虑了

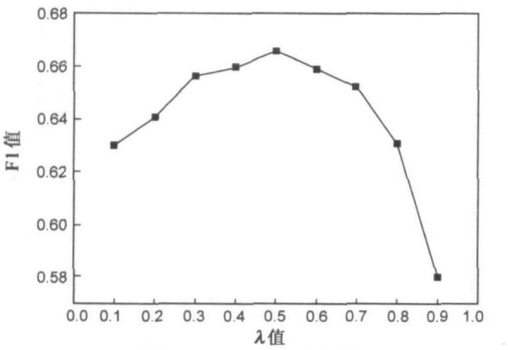

图 5 $RS(W,D_x)$中参数

语料的潜在不平衡性,使得领域术语抽取公式能准确反映分类语料所代表的信息.将本文 NCD + NDD 方法作为文本分类的特征选择算法,取得了比传统方法更高的分类精度和召回率.未来工作包括将本文方法与未登录词识别算法结合从而能抽取未被词典收录的领域术语.

参考文献:

[1] Boguraev B, Kennedy C. Applications of term identification technology: domain description and content characterisation [J]. Natural Language Engineering, 1999, 5(1): 17 - 44.

[2] Velardi P, Missikoff M, et al. Identification of relevant terms to support the construction of domain ontologies [A]. Proceedings of the Workshop on Human Language Technologies and Knowledge Management [C]. France: ACM Press, 2001. 1 - 8.

[3] Maedche A, Staab S. Ontology learning. Handbook on Ontologies in Information Systems [M]. Heidelberg: Springer-Verlag, 2004. 173 - 190.

[4] Oakes M P, Paice C. Term extraction for automatic abstracting. Recent Advances in Computational Terminology [M]. Amsterdam/ Philadelphia: John Benjamins Publishing Company, 2001. 353 - 370.

[5] Gao J, Goodman J, et al. The use of clustering techniques for language modeling-application to Asian language [J]. Computational Linguistics and Chinese Language Processing, 2001, 6(1): 27 - 60.

[6] Avancini H, Lavelli A, et al. Expanding domain-specific lexicons by term categorization [A]. Proceedings of 18th ACM Symposium on Applied Computing [C]. US: ACM Press, 2003. 793 - 797.

[7] 陈文亮,朱靖波等. 基于 Bootstrapping 的领域词汇自动获取[A]. 全国第七届计算语言学联合学术会议论文集[C]. 北京:清华大学出版社, 2003. 67 - 72.

[8] Xu F, Kurz D, et al. A domain adaptive approach to automatic acquisition of domain relevant terms and their relations with bootstrapping [A]. Proceedings of the 3rd International Conference on Language Resources and Evaluation [C]. Spain: LREC press, 2002. 224 - 230.

[9] Liu T, Wang X L, et al. Domain-specific term extraction and its application in text classification [A]. Proceedings of 8th Joint Conference on Information Sciences [C]. USA: World Scientific Press, 2005. 1481 - 1484.

[10] Wang Q, Wang X L, et al. A study of semi-discrete matrix decomposition for LSI in automated text categorization [A]. Proceeding of 1st International Joint Conference on Natural Language Processing [C]. China: Springer-Verlag, 2004. 606 - 615.

[11] Yang Y, Pedersen J O. A comparative study on feature selection in text categorization [A]. Proceeding of 14th International Conference on Machine Learning [C]. US: AAAI press, 1997. 412 - 420.

[12] Navigli R, Verlardi P. Learning domain ontologies from document warehouses and dedicated web sites [J]. Computational Linguistics, 2004, 30(2): 151 - 179.

[13] O Duda R, et al. Pattern classification (Second Edition) [M]. Beijing: China Machine Press, 2003. 321 - 322.

作者简介:

刘 桃 女,1981 年生于内蒙古.哈尔滨工业大学计算机科学与技术学院博士研究生.研究方向为专业领域知识库构建、文本分类、机器学习. E-mail: tliu@insun.hit.edu.cn

刘秉权 男,1970 年生于黑龙江,哈尔滨工业大学计算机科学与技术学院副教授.研究方向为移动计算、Web 知识挖掘、智能人机接口.

徐志明 男,1967 年生于黑龙江,哈尔滨工业大学计算机科学与技术学院副教授.研究方向为信息检索、文本挖掘.