·人工智能及识别技术· 文章编号:1000-3428(2017)06-0177-05 文献标志码:A 中图分类号:TP18

融合统计信息与语义相似度的特征扩展算法

李晓红,曹 林,宿 云,马慧芳

(西北师范大学 计算机科学与工程学院,兰州 730070)

摘 要: 通过分析短文本的高维性和稀疏性,提出一种融合特征词间统计信息与语义相似度的短文本特征扩展算法。根据词的贡献度对候选特征集进行筛选,得到扩展集合初始值。计算特征词之间的统计相关度,构建二元相关词对集合。利用外部知识库知网中的语义关系获取相关词对的义项集合并计算语义相似度,将满足条件的义项扩展为短文本的特征词,得到扩展后的特征集。实验结果表明,使用该算法对短文本进行特征扩展后,可显著提升分类器的分类效果。

关键词: 短文本;统计相关度;语义相似度;知网;特征扩展

中文引用格式: 李晓红,曹 林,宿 云,等. 融合统计信息与语义相似度的特征扩展算法[J]. 计算机工程,2017,43(6):177-181.

英文引用格式: Li Xiaohong, Cao Lin, Su Yun, et al. Feature Extension Algorithm Fusing Statistical Information and Semantic Similarity[J]. Computer Engineering, 2017, 43(6):177-181.

Feature Extension Algorithm Fusing Statistical Information and Semantic Similarity

LI Xiaohong, CAO Lin, SU Yun, MA Huifang

(College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China)

【Abstract】 By analyzing high dimension characteristic and sparsity of short text, this paper proposes a feature extension algorithm fusing statistical information feature words between concepts and semantic similarity for short text. Firstly, it selects reasonable feature set through the contribution degree of word and constructs initial feature extension set. Then it calculates statistical correlation between feature words and constructs a binary word correlation pair set. Finally, by using the semantic relations of external knowledge base, HowNet, it obtains synsets of relevant words, calculates the semantic similarity, extends the synsets which meet the conditions to the feature words of the short text and obtains the extend feature set. Experimental results show that, after using the proposed algorithm to extended features, the classification results of classifiers can be greatly improved.

【Key words】 short text; statistical correlation; semantic similarity; HowNet; feature extension

DOI: 10.3969/j.issn.1000-3428.2017.06.028

0 概述

随着网络新媒体的兴起,互联网已经成为人们进行信息交互和处理的有效平台,以短文本形式呈现的数据更是以极高的速度增长,如微博、短信、博客评论、新闻标题、图片标题等。这些信息具有包含的词语数量稀少、词频低、数据量大等特征,如何从海量的短文本中发现有用的信息,已经成为信息处理领域亟待解决的关键难点之一。

目前,已有很多研究者围绕短文本展开了研究。例如:文献[1]抽取很少的具有主题指示性的查询词来表示文本并进行分类;文献[2]提出引入背景知识的方法来提高分类效率;文献[3]通过语义词典对词汇的语义相关性分析并结合统计方法实现特征降维,从而进行短文本聚类。但由于关键词特征稀疏,以上几种方法的效果均不够理想。面向短文本分类的特征扩展问题随之成为一个重要的研究方向。例如,文献[4]提出基于词对共现关系进行特征扩展;文献[5]利用主题-关键词图上的链接分析进行短文本特征扩展。除此之外,还有其他研究者也进行了

基金项目: 国家自然科学基金(61163039);甘肃省青年科技基金(1606RJYA269,145RJYA259);甘肃省高等学校科研项目(2015A-008);西北师范大学青年教师科研能力提升计划骨干项目(NWNU-LKQN-14-5,NWNU-LKQN-16-20)。

作者简介: 李晓红(1978—),女,讲师,主研方向为数据挖掘、智能信息处理;曹 林,硕士研究生;宿 云,讲师、博士研究生;马慧芳,副教授、博士。

收稿日期: 2016-04-25 **修回日期:** 2016-06-27 **E-mail:** xiaohongli@nwnu.edu.cn

短文本的特征扩展研究[6-7]，其中，文献[7]提出基于频繁二元词集并利用背景知识对特征进行扩展的方法，该算法搜索空间大，致使时间复杂度过高。特别地，当背景知识库的规模增大时，特征词集维度也会急剧增大，导致算法效率大幅下降。

针对上述算法的不足，本文融合特征词之间的统计相关度和基于知网概念的语义相似度，提出一种新的短文本特征扩展算法 FEASS。首先利用贡献度对候选词集合进行筛选，在所得到的特征集合上根据特征词之间的统计相关度构建相关词对集合；然后以迭代的方式依次获取相关词对在知网中的义项集合，并利用知网提供的语义相似度计算功能计算词之间的相似度，将满足特定条件的义项扩展为短文本的特征词；最后得到扩展的特征词集。

1 相关知识

给出本文所涉及的部分符号的含义，如表 1 所示。

表 1 符号含义

符号	含义
$D=\{C_1,C_2,\cdots,C_k\}$	数据样本集合
$C_k=\{d_{k1},d_{k2},\cdots,d_{kN_k}\}$	第 k 个类别
N_k	类 C_k 所包含的短文本数
d_{kj}	第 k 类中第 j 个短文本
CS	w_i 和 w_j 对应的义项集合的交集
$set(w_i)=\{s_{i1},s_{i2},\cdots,s_{in}\}$	概念 w_i 的义项集合

1.1 词的贡献度

设 $f(w_i,d_{kj})$ 表示词项 w_i 在文本 d_{kj} 中的出现次数，$f_{\max}(d_{kj})$ 表示文本 d_{kj} 中词的最大出现次数，则词 w_i 对文本 d_{kj} 的贡献度[8]可表示为：

$$contr(w_i,d_{kj})=\frac{f(w_i,d_{kj})}{f_{\max}(d_{kj})} \tag{1}$$

词 w_i 对类别 C_k 的贡献度可定义为该词对所属类中所有文本的贡献度之和，计算公式为：

$$CONTR(w_i,C_k)=\frac{\sum_{j=1}^{N_k}contr(w_i,d_{kj})}{N_k} \tag{2}$$

根据式(2)可以计算得到候选特征集 F 中所有词的贡献度，k 取不同值时，$CONTR(w_i,C_k)$ 表示同一特征词对不同类别的贡献度，值越大表示该词倾向于某类别的程度越高，越能表示某一类别；i 取不同值时，$CONTR(w_i,C_k)$ 表示不同的特征词对某特定类别的贡献度，差异越大，词的类别区分度越强。

1.2 词的相关度

定义 1 基于信息检索反馈机制，定义词 w_i 相对于词 w_j 的相关性[9]为：

$$\begin{aligned}R(w_i,w_j)&=\mathrm{lb}\frac{f(w_i,w_j)/[f(w_i)-f(w_i,w_j)]}{[f(w_j)-f(w_i,w_j)]/[N-f(w_i)-f(w_j)+f(w_i,w_j)]}\\&\times\left|\frac{f(w_i,w_j)}{f(w_j)}-\frac{f(w_i)-f(w_i,w_j)}{N-f(w_j)}\right|\end{aligned} \tag{3}$$

其中，$f(w_i,w_j)$ 表示词 w_i 和词 w_j 共同出现的文本数；$f(w_j)$ 表示只包含词 w_j 的文本数；N 为所有短文本数，根据上文，$N=N_1+N_2+\cdots+N_K$。显然，$R(w_i,w_j)\neq R(w_j,w_i)$。

定义 2 词 w_i 与词 w_j 的相关度定义为[10]：

$$Rel(w_i,w_j)=\frac{R(w_i,w_j)+R(w_j,w_i)}{2} \tag{4}$$

词语相关性描述的是 2 个词语互相关联的程度，可用这 2 个词语在同一个语境中的共现概率来衡量。2 个词在同一个文本中出现的频率越高，相关程度就越高，关联得就越紧密；而 2 个词单独出现的频率越高，相关程度就越低。因此，$Rel(w_i,w_j)$ 越大，词 w_i 和词 w_j 之间的相关性越高。

2 特征词扩展算法及权重计算

知网是一个揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库。由于汉语中“词”的含义非常复杂，一个词在不同的语境中往往会表达不同的语义，因此知网中每条记录都是由词语的一条义项及其描述所组成。

2.1 语义相似度

假设知网中任意 2 个词 w_i 和 w_j，w_i 有 n 个义项：$s_{i1},s_{i2},\cdots,s_{in}$，$w_j$ 有 m 个义项：$s_{j1},s_{j2},\cdots,s_{jm}$。词 w_i 的义项集合记为 $set(w_i)=\{s_{i1},s_{i2},\cdots,s_{in}\}$。

如“打”这个词在知网中有 3 个义项，即词 $w=$“打”，$s_{i1}=$“买”，$s_{i2}=$“编制”，$s_{i3}=$“通讯”，则 $set(w)=\{$买，编制，通讯$\}$。

定义 3 若 w_i 和 w_j 之间的相关度大于 0，则 $CS=set(w_i)\cap set(w_j)$，且 $CS\neq\varnothing$。

该定义所表示的关联信息如图 1 所示，其中，黑色三角形表示 w_i 的义项；白色圆圈表示 w_j 的义项；白色三角形表示 w_i 和 w_j 共有的义项，即定义 3 中的 CS。

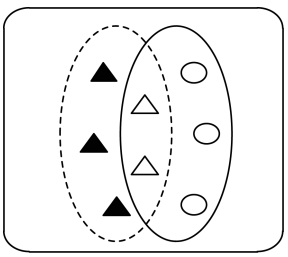

图 1 相关词对间的义项关系

本文算法使用文献[11]中的词汇语义相似度计算方法实现词语之间语义相似度的计算。另外,只关注每一条义项中的基本义原。

定义 4　w_i和 w_j的语义相似度是两词中相似度最大的 2 个义项的相似度,即:

$$sim(w_i,w_j)=\max_{\substack{k=1,2,\cdots,n\\h=1,2,\cdots,m}} sim(s_{ik},s_{jh}) \tag{5}$$

$sim(s_{ik},s_{jh})$的值依赖于义原在树状层次体系中的语义距离[12]。

2.2　特征扩展算法描述

对短文本的特征词集合进行扩展的目的是使其尽可能精确无误地描述短文本要表达的主题和内容,即提高特征词集合的准确性和完备性。

本文基于以下思想进行短文本特征词集的扩展:对于相关词对(w_i,w_j),首先计算 $sim(w_i,w_j)$,如果 $sim(w_i,w_j)>\delta$,视为 w_i和 w_j相似,说明这2 个词既相关又相似,则将使 $sim(w_i,w_j)$取得最大值的 2 个义项扩展进来;否则,认为 w_i与 w_j只相关不相似,由定义 3 推知它们的义项集合的交集必然非空,将公共义项扩展进来。本文算法流程如图 2 所示。

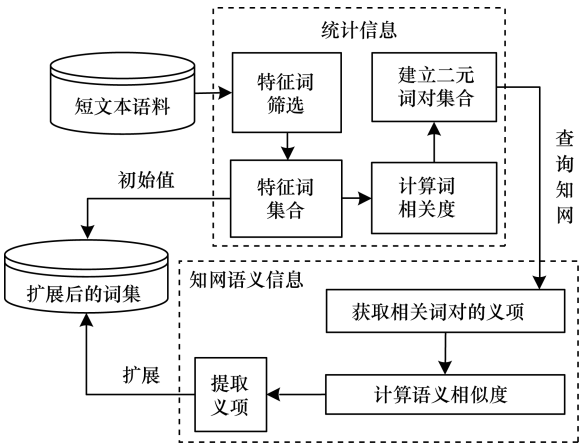

图 2　FEASS 算法流程

FEASS 的算法时间复杂度为 $O(n^2)$。算法伪代码如下:

算法　FEASS

输入　候选特征集 $F=\{w_1,w_2,\cdots\}$

输出　扩展特征集合 EF

1. 初始化:EF = Ø,U = Ø,k = 0;
2. F≠Ø 时利用式(2)计算词的贡献度 CONTR(w_i,C_k);
3. 特征词筛选:
   if(CONTR(w_i,C_k) > α):EF = EF ∪ {w_i};
4. 利用下列操作构建二元词对集合 U:
   for(i = 1;i < | EF |;i + +)
   　for(j = i + 1;j < | EF |;j + +)
   　　if (Rel (w_i,w_j) > β) then U = U ∪ {(w_i,w_j)};
5. 扩展特征:k < |U|时重复执行以下操作:
   5.1. 取第 k 个词对(w_i,w_j)$_k$,由式(5)计算 sim(w_i,w_j);
   5.2. if sim(w_i,w_j) > δ　then EF = EF ∪ (s_{ik} ∪ s_{jh});
   　else {
   　　CS = set(w_i) ∩ set(w_j);
   　　EF = EF ∪ CS;
   　　}
   5.3. k ++,转步骤 5.1;
6. return EF

2.3　特征权值的计算

特征词的权重体现了特征词在文本中的重要程度。由本文算法描述可知,短文本特征词集 EF 中有 2 类词:一类是短文本中原有的词,另一类是从外部知识库知网扩展进来的词。显然,这两类词在表达文本主题信息时的重要程度有所不同,因此,其权重计算方法也不同。

对于短文本原有的特征词,权值定义为词 w_i对第 k 类中的文本 d_{kj}的贡献度:

$$weight(w_i)=contr(w_i,d_{kj}) \tag{6}$$

在特征扩展中,短文本原有特征词都是最重要的,扩展词的重要性不会高于原有特征词。而扩展词又是通过词对(w_i,w_j)之间的关联关系及其在知网中的语义相似度关系引入。综合考虑以上 2 个方面,对于扩展词 w_k,给出一种更全面的权值计算方法[13]:

$$weight(w_k)=\lambda Rel(w_i,w_j)+(1-\lambda)sim(w_i,w_j) \tag{7}$$

其中,λ 是调节因子,$\lambda\in[0,1]$。与传统的权值计算方法相比,新的权值函数综合了统计特性和语义特性,能在一定程度上揭示短文本原有特征和引入特征的重要性区分。

3　实验结果与分析

为验证本文所提出的短文本特征词扩展方法的有效性及相应分类结果的准确性,首先设计实验对本文的方法进行验证,然后采用相应的评价指标对实验结果进行评价。

3.1　数据描述

本文实验分别采集了 2013 年 - 2015 年期间公开发表的、收录在 CNKI(中国期刊全文数据库)上的期刊论文的标题,共计 33 285 条。人工将其分为 10 个类别,然后将每类文本集合随机地平均分为 3 份,用其中一份作为测试样本,共 11 094 条,另外 2 份作为训练样本,共 22 187 条。一个标题即为一个短文本。实验数据统计信息如表 2 所示。

表 2 实验数据统计信息

类别	训练样本	测试样本	类别	训练样本	测试样本
经管	2 262	1 131	地质	1 626	961
海洋	2 454	1 227	数学	2 608	1 304
天文学	2 242	1 121	农林	1 864	932
生物	2 100	1 050	物理	2 160	1 080
医药学	2 208	1 104	计算机	2 367	1 183

3.2 实验设置

预处理:采用 ICTCLAS[14] 方法对其进行分词,并去除停用词,获得候选特征集 F。

分类器:朴素贝叶斯(Naive Bayes)分类器和支持向量机(Support Vector Machine,SVM)分类器,其中 SVM 用 LibSVM 包、径向基核函数。

评价指标:分类结果的度量采用常见的指标,即准确率(P)、召回率(R)和 F1 值[15]。

3.3 结果分析

为验证本文算法的有效性,共设计以下 3 个实验:1)对参数 α,β,δ 取不同的值,来测试它们对扩展算法性能的影响;2)在 2 个分类器上分别比较不同特征扩展算法之间的性能差别;3)比较扩展前后的特征集合在 SVM 分类器上的性能。

3.3.1 α,β 和 δ 对扩展算法的影响

本实验对 α,β 和 δ 3 个参数取不同值分别在 Naive Bayes 和 SVM 上进行,并选取具有代表性的结果,如表 3 所示。可以看出,当 $\alpha=0.15$,$\beta=0.10$,$\delta=0.25$ 时,在 2 个分类器上取得的结果都是最好的(即表 3 中的加粗数据)。当 α,β 的值较小时,分类效果较低,其主要原因有:1)冗余特征没有被筛掉,特征数目过多;2)相关词对相对泛滥,导致引入了很多无用的新特征。反之,当 α,β 的值过大时,分类效果也呈现出下降趋势,主要是由于部分重要特征被删除,而且从外部引入的特征数量也急剧下降的原因导致的。从实验结果来看,δ 的取值控制在 0.25 左右时引入的特征是高质量的。

表 3 参数值对算法性能的影响

参数值			Naive Bayes 分类器			SVM 分类器		
α	β	δ	P	R	F1	P	R	F1
0.10	0.05	0.20	70.24	61.25	65.43	68.35	65.38	66.83
0.15	0.05	0.20	72.86	69.58	71.18	69.83	70.24	70.03
0.15	**0.10**	**0.25**	**78.05**	**73.09**	**75.49**	**75.48**	**73.75**	**74.25**
0.20	0.20	0.25	73.51	68.24	70.28	70.35	61.26	65.49
0.20	0.20	0.30	70.23	65.26	67.65	68.27	59.38	63.59
0.25	0.25	0.30	62.42	55.69	58.36	63.23	57.41	60.23

3.3.2 不同特征扩展算法的性能对比

本文分别选择了文献[5]提出的 SCTCEFE(Short Chinese Text Considering Effective Feature Expansion)算法、文献[8]提出的 FEMFTS(Feature Extension Method using Frequent Term Sets)算法与本文算法(FEASS)在 2 个分类器上进行分类测试,结果如图 3 ~ 图 6 所示。

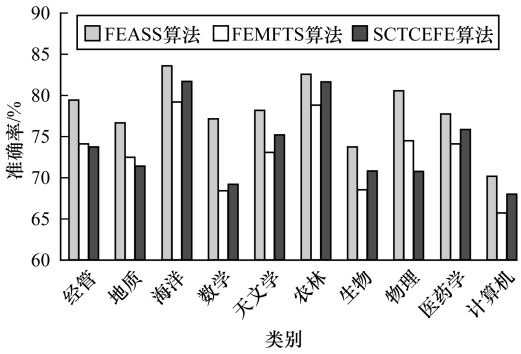

图 3 在 Naive Bayes 分类器上的准确率比较

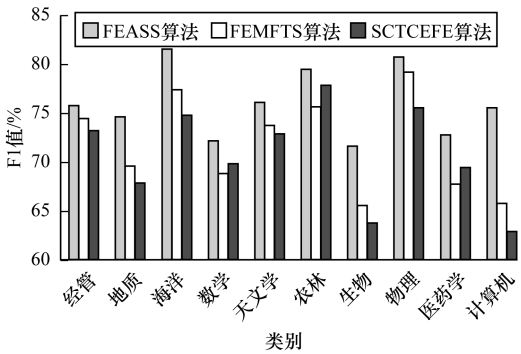

图 4 在 Naive Bayes 分类器上的 F1 值比较

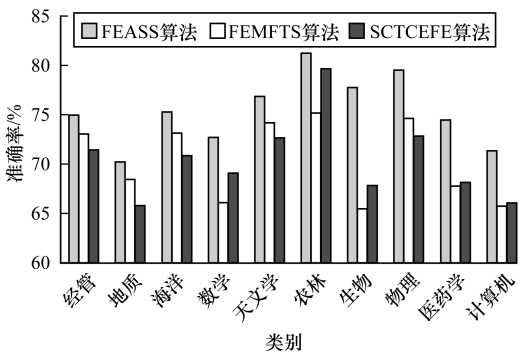

图 5 在 SVM 分类器上的准确率比较

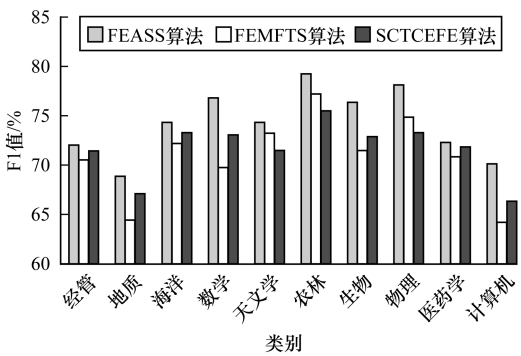

图 6 在 SVM 分类器上的 F1 值比较

由图 3 ~ 图 6 可以看出,10 个类别的短文本数据使用本文算法得到的特征集合在 2 个分类器上均

获得了最好的分类效果,这足以说明 FEASS 算法的优越性。而 FEMFTS 算法效果不好,并且在不同的类别上分类成绩波动较大,主要是因为该算法处理大批量文档的能力较差导致的;SCTCEFE 算法相对稳定,但是分类效果并不好,这是由于算法基于语义和统计约束,使得扩展进来的新特征数目较少,对主题内容的补充不充分造成的。本文算法刚好弥补了这两个算法的缺点,综合了统计信息和外网的语义信息,取得了较好的分类效果。

3.3.3　特征扩展前后对分类性能的影响

图 7 和图 8 显示了在 SVM 分类器上进行特征扩展前后的实验结果。可以看出,短文本特征被扩展后,不论是准确率还是 F1 值均得到了较大的提升,特别是数学类和生物类达到了 7.78% 和8.66% 的改善率,表明本文方法可以有效地解决特征稀疏问题,由此提高短文本分类性能。

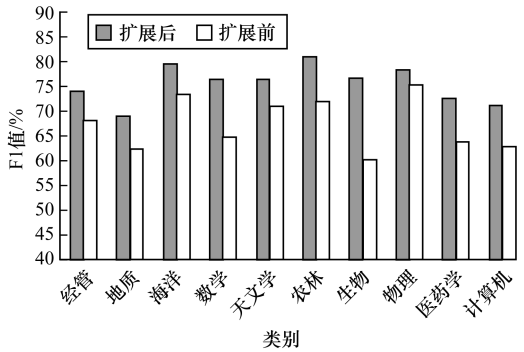

图 7　特征扩展前后在 SVM 分类器上的 F1 值比较

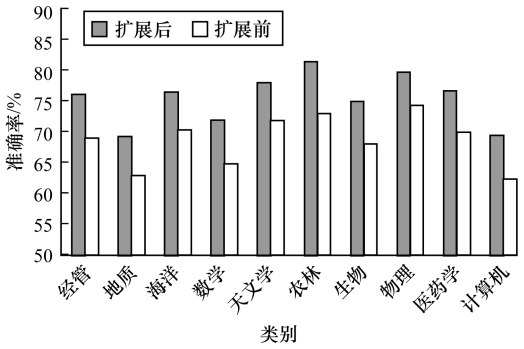

图 8　特征扩展前后在 SVM 分类器上的准确率比较

4　结束语

本文针对短文本特征稀疏、高维的特点,将特征词之间的统计相关度和知网中义项的语义相似度相结合,提出一种扩展短文本特征词集合的算法,旨在引入一些对文本主题贡献大的词。实验结果表明,对于短文本,借助外部知识库进行特征信息的补充方法是可行的。下一步工作将研究如何发现扩充语料中的关键信息,并减少补充信息包含的噪音,达到有效扩展的目的。同时,也将对外部知识网补充特征信息与原数据结合的方法进行改进。

参考文献

[1] Sun Aixin. Short Text Classification Using Very Few Words[C]//Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2012:1145-1146.

[2] Zelikovitz S, Marquez F. Transductive Learning for Short-text Classification Problems Using Latent Semantic Indexing[J]. International Journal of Pattern Recognition and Artificial Intelligence,2005,19(2):143-163.

[3] 杨婉霞,孙　理,黄永峰. 结合语义与统计的特征降维短文本聚类[J]. 计算机工程,2012,38(22):171-175.

[4] Yan Tao,Wang Xiwei. Feature Extension for Short Text[C]//Proceedings of the 3rd International Symposium on Computer Science and Computational Technology. Jiaozuo,China:[s.n.], 2010:338-341.

[5] Liu Mingxuan, Fan Xinghua. A Method for Chinese Short Text Classification Considering Effective Feature Expansion [J]. International Journal of Advanced Research in Artificial Intelligence,2012,1(1).

[6] Wang Peng, Zhang Heng, Xu Bo. Short Text Feature Enrichment Using Link Analysis on Topic-keyword Graph[C]//Proceedings of NLPCC'14. Berlin, Germany: Springer,2014:79-90.

[7] Man Yuan. Feature Extension for Short Text Categorization Using Frequent Term Sets[J]. Procedia Computer Science,2014,31:663-670.

[8] 陈羽中,方明月,郭文忠. 面向微博热点话题发现的多标签传播聚类方法研究[J]. 模式识别与人工智能,2015,28(1):1-10.

[9] Cataldi M, di Caro L, Schifanella C. Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation [C]//Proceedings of the 10th International Workshop on Multimedia Data Mining. Washington D. C. ,USA:[s. n.],2010:1-10.

[10] Chen Mengen, Jin Xiaoming, Shen Dou. Short Text Classification Improved by Learning Multi-granularity Topics[C]//Proceedings of the 22nd International Joint Conference on Artificial Intelligence. Barcelona, Spain: [s. n.],2011:1776-1781.

[11] 刘　群,李素建. 基于《知网》的词汇语义相似度的计算[C]//第三届汉语词汇语义学研讨会. 台北,中国:[出版者不详],2002:59-76.

[12] Pan Liqiang, Zhang Pu, Xiong Anping. Semantic Similarity Calculation of Chinese Word [J]. International Journal of Advanced Computer Science and Applications,2014,5(8): 205-214.

[13] Liu Wenyin, Quan Xiaojun, Feng Min, et al. A Short Text Modeling Method Combining Semantic and Statistical Information[J]. Information Sciences, 2010, 180(20):4031-4041.

[14] Zhang Huaping, Yu Hongkui, Yi De. HHMM-based Chinese Lexical Analyzer ICT-CLAS[C]//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Sapporo, Japan:[s. n.],2003:184-187.

[15] Peat H J, Willet P. The Limitations of Term Co-occurrence Data for Query Expansion in Document Retrieval Systems[J]. Journal of American Society for Information Science,1991,42(5):378-383.

编辑　金胡考