DOI:10.3969/j.issn.1673-4785.201205045
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20120917.1632.001.html

基于词共现图的中文微博新闻话题识别

赵文清,侯小可
(华北电力大学 控制与计算机工程学院,河北 保定 071003)

摘 要:针对传统的话题检测算法主要适用于新闻网页和博客等长文本信息,而不能有效处理具有稀疏性的微博数据,给出一种基于词共现图的方法来识别微博中的新闻话题.该方法首先在微博数据预处理之后,综合相对词频和词频增加率2个因素抽取微博数据中的主题词.然后根据主题词间的共现度构建词共现图,把词共现图中每个不连通的簇集看成一个新闻话题,并使用每个簇集中包含信息量较大的几个主题词来表示微博新闻话题.最后在微博数据集上进行实验,实现了对微博中新闻话题的识别,验证了该方法的有效性.

关键词:微博;新闻话题;新闻话题识别;主题词;词共现图
中图分类号:TP391.1 **文献标志码**:A **文章编号**:1673-4785(2012)05-0444-06

News topic recognition of Chinese microblog based on word co-occurrence graph

ZHAO Wenqing, HOU Xiaoke
(School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China)

Abstract: The traditional topic detection algorithm is applied to longer texts such as: news website pages or blogs, causing it to be hard to deal with sparse microblog data effectively. In this paper, a method based on the word co-occurrence graph was provided to detect news topics of microblogs. Firstly, the relative word frequency and the word frequency increase rate were considered to extract new keywords from microblog text after pretreatment. Secondly, a word co-occurrence graph was built by co-occurrence degrees of keywords; each unconnected cluster in a word co-occurrence graph was taken as a news topic by calculating several keywords. These keywords contain much more information in each cluster, was used to represent a news topic of microblog. Finally, data analysis provided evidence on how the approach is most effective and also revealed the microblog data set recognized news topic recognition.

Keywords: microblog; news topics; topic recognition; keywords; word co-occurrence graph

随着微博的飞速发展,微博作为一种通过关注机制分享简短实时信息的广播式的社交网络平台,吸引了越来越多的网民参与.微博改变了人们获取信息的方式,是一种能够观察和了解中国正在发生什么的实时民意调查系统.中国微博由一种单纯的社交工具,变成舆论监督的利器,参与并且影响着整个世界.因此从海量微博数据中检测出当前热点新闻话题,并对新闻话题进行情感分析,及时把握人们普遍关心的问题以及人们对热点新闻话题的看法,对事件监测、民意调查、行业调研等都有重要作用.

传统的针对普通网络信息(如新闻网页和博客等长文本信息)新闻话题识别的研究较早且相对成熟[1-2].一般将长文本中的词视为特征,首先利用特征向量来表示文本,并采用TF-IDF方法度量向量每一维(即每个特征)的权重;然后采用一定的聚类方法,将叙述相同或相似新闻话题的长文本聚类到同一类中[3].但对于微博来说,其文本长度短、信息量少,特征关键词不足以表示文本.而现有的文本聚类算法都是基于向量空间模型,利用词向量表示文本特征,文本相似度量依赖于2个文本中词语重叠的数量.当2个文本较长时,其重叠的词语可能足够描

收稿日期:2012-05-26. **网络出版日期**:2012-09-17.
基金项目:国家自然科学基金资助项目(70671039);中央高校基本科研业务费专项资金资助项目(12MS121).
通信作者:侯小可. E-mail: houxiaoke2008@163.com.

述文本的内容;但是当文本比较短时,文本间匹配的词数减少即相关词集规模较小,不足以准确描述文本内容,使得相似度发生漂移,大大地影响短文本聚类效果.

针对微博数据的稀疏性问题研究者们做了很多方法的尝试. 路荣等利用 LDA 模型对微博数据集进行隐主题建模,进而通过隐主题模型计算文本之间的相似度,处理微博数据稀疏的特点[3]. LDA 模型的缺点是它的计算量很大,这是因为需要模拟 Dirichlet process 对主题反复抽样,导致速度较慢. Liu 等提出基于 part of-speech 和 HowNet 来扩展单词的语义特征,进而改进分类和聚类效果[4]. 金春霞等针对短文本相似度漂移问题,提出了一种基于 HowNet 扩充相关词集来构建动态文本向量的方法,利用动态向量计算中文短文本的内容相似度,进而发现短文本之间的内在关联,从而缓解特征词词频过低、存在变形词以及新词对聚类的影响,实验表明该算法的聚类效果较好[5]. 郑斐然为了提取出新闻主题词综合考虑短文本中的词频和增长速度而构造复合权值,用以量化词语是新闻词汇的程度,在话题构造中使用了上下文的相关度模型来支撑增量式聚类算法,相比于语义相似度模型,其更能适应该问题的特点[6]. 杨震等将每个短文本文档看成一个由文字、数字和标点构成的字符串,并基于字符串自身的特性直接计算其相似性,在此基础上进行短文本层次化聚类,进而发现网络舆情热点[7]. 由于这种方法不使用特征提取和文本表示过程,在一定程度上避免了传统方法在短文本表示时特征向量稀疏的不足,较好地解决来了短文本的聚类问题.

针对微博数据稀疏性、实时性、不规范性的特点,本文给出一套完整的微博数据处理和新闻话题识别方法. 在向量空间模型的基础上,从微博主题词的时域分布中筛选出信息量最大的新闻主题词;根据微博的主题词共现度构建词共现图,以词共现图为基础,把不连通的簇集看成一个新闻话题,进而完成微博新闻话题识别.

1　微博新闻话题识别

1.1　数据准备

虽然主流微博都提供了 API 接口供第三方访问,但所有微博服务商都不会无条件将完整 API 开放给普通用户,通常 API 服务商对用户的 API 接口调用频率与查询范围也会根据用户权限的不同有所限制,因此使用 API 的方式并不能完全解决微博数据获取问题. 为了本文的研究工作,采用自然语言处理与信息检索共享平台公开共享的 NLPIR 微博内容语料库[8](23 万条数据)作为本文的实验数据,该语料库是由张华平博士从新浪和腾讯两大主流微博中公开采集并抽取而获得.

1.2　文本的预处理及词频统计

在进行主题词抽取之前,需要对微博数据进行预处理,预处理主要包括文本分词、词性过滤、停用词过滤等,本文把停用词过滤放在词频统计之后,过滤掉词频很高但作用很小的词语. 预处理完之后便可对得到的文本数据抽取主题词. 预处理的过程如图 1 所示.

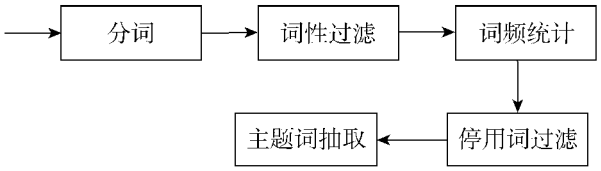


图 1　预处理过程

Fig. 1　Preprocessing process

本文的实验数据已经剔除了大量的冗余数据,可以直接进行分词. 本文采用中科院张华平等开发的 ICTCLAS[9](Institute of Computing Technology, Chinese lexical analysis system)分词工具,它的主要功能包括中文分词、词性标注、命名实体识别、新词识别,是目前文本处理中经常使用且分词效果最好的中文分词软件之一.

每条微博经过分词之后,并不是所有词都可以作为主题特征词,有许多的词包含的信息量很少,将它们包含进来不仅不能提高反而会降低正确率,所以只考虑对新闻话题识别作用较大的词. 词性的不同对主题的表达是有区别的,名词描述性较强,故能更好地表达主题,动词也可以作为衡量标准;因此,根据分词的词性标注,在词性过滤时对名词和动词保留,其他词性的词全部过滤,这样可以降低计算的复杂度,为下文的主题词汇抽取做准备.

统计词频时,先将微博消息按正文发布时间划入不同的时间窗口. 如按照天进行划分,然后对同一窗口中的词频进行统计,得到一个该时间窗口内的总的词语列表. 根据文献[10-11],在一段时间窗口的微博词语具有长尾现象,即绝大多数的词只出现了很少的次数,只有少数词语有较高的出现频率. 将列表按词频排序,按比例保留频率最高的词语用于主题词抽取,而把长尾部分去掉. 当然,并非所有的高频词都是有用的词,例如“图片”一词,很多微博中都包含图片;所以在微博中该词出现的频率很高,但对新闻话题识别贡献很小,并不适合作为主题词. 因此,在经过分词后的文本中适当减少停用词,可显

著提高主题词的密度,让微博中的主题词更加突出.

1.3 主题词抽取

通常新闻话题的形成有一定的特点,它的时域性较强,且影响力较大,也就是说新闻话题讨论的内容在某个时间窗口之前出现的概率很小,而在一个时间窗口中突然大量出现,从而能够引起社会的高度关注. 根据新闻话题的这个特点,判断一个词语是否为新闻话题中的主题词. 本文把相对词频和词频增加率作为抽取主题词的2个影响因素.

1)相对词频.

当一个词汇在某时间段内频繁出现,且出现的频率比该时间段内其他的词汇明显更大,一定程度上意味着它和当前一些关注度较大的热点话题相关联. 因此采用相对词频的方法,对主题词的词频贡献度进行量化:

$$T_{ij} = \frac{f_{ij}}{f_{\max}}.$$

式中:T_{ij}是词汇 i 在 j 时间窗口的相对词频,f_{ij}是词汇 i 在 j 时间窗口的频率,$f_{\max}$表示当前时间窗口的最高词频.

2)词频增加率.

当一个词汇在某时间段内频繁出现,且出现的频率要比上一个时间段内明显增加,则在一定程度上意味着它和当前一些比较新的新闻话题关联.

$$G_{ij} = \frac{f_{ij} - f_{i(j-1)}}{1 + f_{i(j-1)}}.$$

式中:G_{ij}表示词汇 i 在 j 时间窗口的增加率,$f_{i(j-1)}$是词汇 i 在 $j-1$ 时间窗口(即上一个时间窗口)的频率.

对微博数据进行分词、词性过滤、词频过滤等预处理之后,有选择性地留下那些有意义的动词和名词,在此基础上考察相对词频和词频增加率2个方面的复合权值来评价一个特征词的主题表现力 W_{ij}:

$$W_{ij} = \alpha \ln T_{ij} + \beta \ln G_{ij}. \tag{1}$$

式中:W_{ij}值越大说明该特征词是主题词的概率越大;α 和 β 参数用来调节相对词频和词频增加率的比重关系,α 一定时,β 越大则词频增加率起主要作用,相反 β 一定时,α 越大则相对词频优先考虑.

对每个时间窗口内的词计算其 W_{ij}值,按照阈值 T 选出其中权值较大的特征词得到一个主题词表. 这个主题词表的特点是其中的词语在当前时间窗口出现次数较多,并且在之前的时间窗口出现次数较少. 选出主题词之后,就可以对这些主题词进行词共现分析来构建词共现图,通过对图的划分来实现新闻话题识别.

1.4 基于主题词共现图的微博新闻话题识别

词的共现分析是自然语言处理技术在信息检索中的成功应用之一,它的核心思想是词与词之间的共现频率在某种程度上反映了词之间的语义关联. 最早有学者利用词共现来计算文档的相似性[12],也有利用词共现模型来计算词之间的相关度[13]. 耿焕同等提出了一种基于词共现图的文档自动摘要算法,他们先运用词共现图的主题提取技术得到各个主题,然后根据各个主题的重要性来提取主题词、主题句、生成摘要[14]. 常鹏等提出一种基于词共现的文档聚类算法,利用文档集上的频繁共现词建立文档主题向量表示模型,从而准确地反映文档之间的主题相关关系[15].

所有词共现的研究都基于这样一个假设:如果在一个大规模文本语料中,2个词频繁出现在同一窗口单元中(例如一句话、一个自然段、一篇文档等),就可以认为这个词汇组合是比较稳定的,在意义上相互关联,并且共现的频率越高,其相互关联越紧密. 它们表示了一定的语义概念,表达了某个潜在的主题信息.

为了从理论上进一步阐述基于词共现图的微博新闻话题识别的原理,参考文献[16]给出了下面的定义.

定义1 词汇 w_x 相对于词汇 w_y 的相对共现度 $R(w_x|w_y)$ 定义为

$$R(w_x \mid w_y) = \frac{f(w_x w_y)}{f(w_y)}. \tag{2}$$

式中:$f(w_x w_y)$为单位时间段窗口中词 w_x 与词 w_y 在同一条微博中共同出现的次数,$f(w_y)$为词 w_y 在单位时间窗口中出现的次数. 可知,$R(w_x|w_y)$通常不等于 $R(w_y|w_x)$.

定义2 词汇 w_x 与词汇 w_y 之间共现度则定义如式(3):

$$C(w_x, w_y) = \frac{R(w_x \mid w_y) + R(w_y \mid w_x)}{2}. \tag{3}$$

故有 $C(w_x, w_y) = C(w_y, w_x)$.

按照词共现原理,当2个主题词经常出现在同一条微博中,则认为这2个主题词在意义上相互关联,表达了某个潜在的主题信息,与当前微博中的新闻话题有一定关联. 本文根据主题词之间的共现度构建词共现图,在词共现图的基础上,将每个连通的子图看成一个簇集,簇集内部是连通的,而不同的簇集之间是不连通的,此时每个不连通的簇集对应微博中一个新闻话题,通过对词共现图中簇集的划分来完成微博新闻话题的识别. 根据上述思想,下面给

出识别微博新闻话题的基本步骤.

1)主题词共现图中点集 N_S 的生成. 根据上文的分词、停用词过滤、复合权值计算后最终得到主题词表,将主题词表中的主题词作为词共现图 G 的点集,如图 2 ~ 3 中那些黑色的实心圆点.

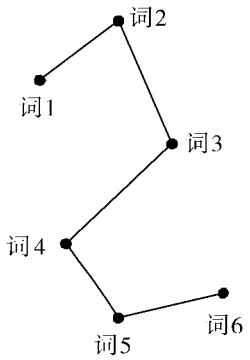


图 2　单连通图 G

Fig. 2　Single-connected graph

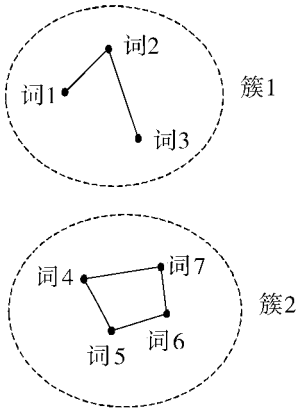


图 3　多连通图 G

Fig. 3　Multiple connectee graph

2)对词共现图中的点集连边. 根据点集 N_S 中 2 个词之间的共现度值的大小决定是否进行连边,如果与结点相对应的 2 个词之间的共现度达到一定阈值,那么就对它们进行连边.

3)基于词共现图的各个微博新闻话题的确定. 若词共现图 G 是一个单连通图,表示该时间窗口的微博消息中只包含一个热点话题(例如图 2). 如果词共现图 G 是非单连通图,那就相当于把图 G 分割为多个连通区域,即构成簇(例如图 3 中的 2 个簇),每个簇与一个热点话题对应.

4)基于词共现图的各个微博新闻话题表示. 如果一个词汇与越多的词汇形成共现词组合,则这个词汇具有较为积极的主题意义,它很可能是某个主题的领域词汇. 同样,在词共现图中,一个主题词连的边越多,那么它包含的信息量越大,能更好地表示潜在的主题信息. 利用式(4)来计算每个簇中主题词的信息量大小,其表示对簇集的贡献程度大小.

$$G(w_i) = \sum_{(w_i,w_j)\in E(G)} C(w_i,w_j). \qquad (4)$$

式中:$E(G)$ 是图 G 中的边集;通过对主题词 w_i 的信息量 $G(w_i)$ 进行排序,选出 K 个对话题簇贡献度较大的主题词,作为该新闻热点话题的表示.

2　实验结果及其分析

实验采用自然语言处理与信息检索共享平台公开的 NLPIR 微博内容语料库[8](23 万条数据)作为本文的实验数据,实验中将时间窗口的长度设定为 1 d,并对 2012-02-01—02-09 的微博数据进行人工标注,该时间段内微博热议的主要新闻话题有“吴英案”、“香港双非问题”等事件.

2.1　主题词抽取的参数确定

为了评估式(1)中的参数对主题词抽取结果的影响,把 2012-02-01—02-09 共 9 天的微博数据分成 8 组(其中有 4 天的微博数据较少不予考虑),对每个时间窗口中的数据进行主题词抽取,然后找出每组阈值 T 较大的前 100 个词中与当前标注的主要新闻话题相关的主题词数,最后求平均值. 其中设定 α 为1.0,比较 β 取不同值时对相关主题词数的影响,如图 4 所示.

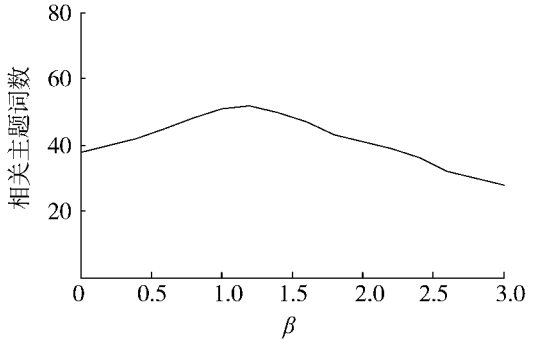


图 4　β 对相关主题词数的影响

Fig. 4　Effect of β on the quantity of related keywords

当 β 取 0 时,此时只考虑词频对主题词的作用,阈值 T 排在前 100 的主题词中平均有 38 个是相关主题词;当 β 取到 1.2 左右时,阈值 T 较大的前 100 个主题词中平均有 52 个相关主题词. 从图 4 中可以看出,如果词频增加率的比重足够大时(即 β 较大)相关主题词数反而下降,甚至少于只考虑词频时的情况.

2.2　实验过程与结果

对 NLPIR 微博内容语料库中 2012-02-01 的 1 432条微博数据进行话题识别,首先经过预处理及词频统计后,对微博数据进行主题词抽取,其中抽取

主题词的参数 α 取 1.0，β 取 1.2，从而得到满足合适阈值的主题词表；然后采用 1.4 节的基于词共现图识别新闻话题的步骤来完成新闻话题识别，这其中需要用到式(2)～(3)来计算主题词之间的共现度. 本文列出了部分主题词间的共现度，如表 1 所示.

表 1 部分词共现度

Table 1 Some keywords co-occurrence degrees

共现词	共现度	共现词	共现度
香港-内地	0.3432 2	民间-融资	0.5671 3
内地-孕妇	0.2151 3	金融-垄断	0.4940 9
吴英-集资	0.4551 0	经济-犯罪	0.2447 6
吴英-死刑	0.4229 8	方舟子-韩寒	0.6830 2
集资-非法	0.5614 6	行为-违法	0.4778 8
死刑-判决	0.3497 2	城管-执法	0.3363 9

对共现度足够大的主题词结点之间连边，并把孤立点(也就是没有连边的点)去除之后得到词共现图，如图 5 所示.

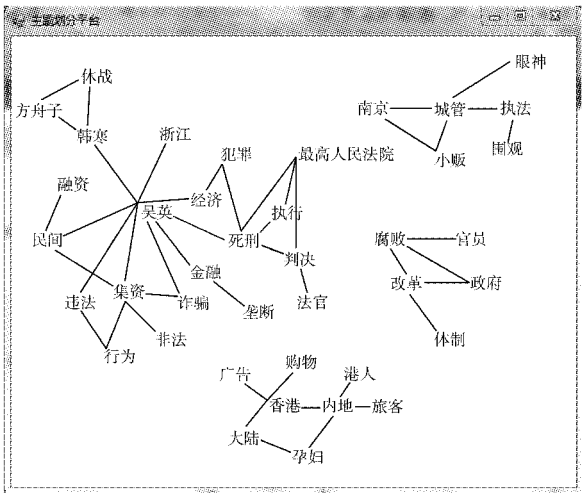


图 5 主题词共现图

Fig. 5 Keywords co-occurrence graph

通过图 5 可以发现词共现图中包括多个簇，也就意味着当前时间窗口中包括多个热点话题，当然最大的热点新闻话题也就是包含节点最多的簇. 通过式(4)可以得到每个簇中信息量最大的 K 个主题词用来表示新闻话题，本文 K 取 5. 表 2 显示了实验中 2012-02-01 当天的热点话题. 通过实验可以发现，其中“吴英案”是当天最大的热点话题，实验结果表明本文提出的基于词共现图的划分识别微博新闻话题的方法是有效的.

表 2 2012 年 2 月 1 日当天热点话题表示

Table 2 Keywords represent news topics on February 1, 2012

编号	热点话题表示
1	吴英、死刑、集资、判决、民间
2	香港、内地、孕妇、旅客、广告
3	城管、小贩、执法、南京、围观
4	改革、腐败、政府、官员、体制

3 结束语

微博热点新闻话题的识别研究有着重要的应用背景，本文在分析了一些短文本热点话题发现研究的基础上，结合微博数据本身的特点提出了一种基于词共现图的微博新闻话题识别的方法. 该方法通过预处理、抽取主题词、构建词共现图等步骤来识别微博中的新闻话题. 实验结果证明了提出的方法是有效的，而且该方法简单，易于实现. 同时，在接下来的工作中将进一步对词共现图的微博话题识别的方法进行优化和提高，在此基础上开始对微博中的热点新闻话题进行情感分析的研究.

参考文献：

[1] MORI M, MIURA T, SHIOYA I. Topic detection and tracking for news web pages[C]//Proceedings of the 2006 ACM International Conference on Web Intelligence. Washington, DC, USA, 2006: 338-342.

[2] ALLAN J, CARBONELL J, DODDINGTON G, et al. Topic detection and tracking pilot study: final report[C]//Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. San Francisco, USA: Morgan Kaufmann Publisher Inc, 1998: 194-218.

[3] 路荣，项亮，刘明荣，等. 基于隐主题分析和文本聚类的微博客新闻话题发现[J]. 模式识别与人工智能，2012，25(3)：382-387.
LU Rong, XIANG Liang, LIU Mingrong, et al. Discovering news topics from microblogs based on hidden topics analysis and text clustering[J]. Pattern Recognition and Artificial Intelligence, 2012, 25(3): 382-387.

[4] LIU Zitao, YU Wenchao, CHEN Wei, et al. Short text feature selection for microblog mining[C]//The 4th International Conference on Computational Intelligence and Software Engineering. Wuhan, China, 2010: 1-4.

[5] 金春霞，周海岩. 动态向量的中文短文本聚类[J]. 计算机工程与应用，2011，47(33)：156-158.
JIN Chunxia, ZHOU Haiyan. Chinese short text clustering based on dynamic[J]. Computer Engineering and Applications, 2011, 47(33): 156-158.

[6]郑斐然,苗夺谦,张志飞,等. 一种中文微博新闻话题检测的方法[J]. 计算机科学, 2012, 39(1): 138-141.
ZHENG Feiran, MIAO Duoqian, ZHANG Zhifei, et al. News topic detection approach on Chinese microblog[J]. Computer Science, 2012, 39(1): 138-141.

[7]杨震,段立娟,赖英旭. 基于字符串相似性聚类的网络短文本舆情热点发现技术[J]. 北京工业大学学报, 2010, 36(5): 669-673.
YANG Zhen, DUAN Lijuan, LAI Yingxu. Online public opinion hotspot detection and analysis based on short text clustering using string distance[J]. Journal of Beijing University of Technology, 2010, 36(5): 669-673.

[8]张华平. NLPIR 微博内容语料库—23 万条[EB/OL]. (2012-02-14)[2012-05-20]. http://www.nlpir.org/?action-viewnews-itemid-231.2012,02,14/2012,02,18.

[9]张华平. ICTCLAS2012 版本 SDK 发布(u0106 版本修正了 UTF8 下的 bug)[EB/OL]. (2011-12-31)[2012-05-20]. http://www.nlpir.org/?action-viewnews-itemid-229.2011,12,31/2012,02,18.

[10]彭泽映,俞晓明,许洪波,等. 大规模短文本的不完全聚类[J]. 中文信息学报, 2011, 25(1): 54-59.
PENG Zeying, YU Xiaoming, XU Hongbo, et al. Incomplete clustering for large scale short texts[J]. Journal of Chinese Information Processing, 2011, 25(1): 54-59.

[11]常鹏,马辉. 高效的短文本主题词抽取方法[J]. 计算机工程与应用, 2011, 47(20): 126-128, 154.
CHANG Peng, MA Hui. Efficient short texts keyword extraction method analysis[J]. Computer Engineering and Applications, 2011, 47(20): 126-128, 154.

[12]TRIVISON D. Term co-occurrence in cited/citing journal articles as a measure of document similarity[J]. Information Processing & Management, 1987, 23(3): 183-194.

[13]乔业男,齐勇,侯迪. 一种高稳定性词汇共现模型[J]. 西安交通大学学报, 2009, 43(6): 24-27.
QIAO Yenan, QI Yong, HOU Di. A highly stable term co-occurrence model[J]. Journal of Xi'an Jiaotong University, 2009, 43(6): 24-27.

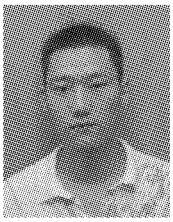
[14]耿焕同,蔡庆生,赵鹏,等. 一种基于词共现图的文档自动摘要研究[J]. 情报学报, 2005, 24(6): 651-656.
GENG Huantong, CAI Qingsheng, ZHAO Peng, et al. Research on document automatic summarization based on word co-occurrence[J]. Journal of The China Society for Scientific and Technical Information, 2005, 24(6): 651-656.

[15]常鹏,冯楠,马辉. 一种基于词共现的文档聚类算法[J]. 计算机工程, 2012, 38(2): 213-214, 220.
CHANG Peng, FENG Nan, MA Hui. Document clustering algorithm based on word co-occurrence[J]. Computer Engineering, 2012, 38(2): 213-214, 220.

[16]耿焕同,蔡庆生,于琨,等. 一种基于词共现图的文档主题词自动抽取算法[J]. 南京大学学报:自然科学, 2006, 42(2): 156-162.
GENG Huantong, CAI Qingsheng, YU Kun, et al. A kind of automatic text keyphrase extraction method based on word co-occurrence[J]. Journal of Nanjing University: Natural Sciences, 2006, 42(2): 156-162.

作者简介:

赵文清,女,1973 年生,副教授,中国人工智能学会粗糙集与软计算专业委员会委员. 主要研究方向为机器学习、数据挖掘、贝叶斯网络学习等. 获河北省科技进步三等奖 1 项,国家发明专利 1 项. 发表学术论文 30 余篇,出版教材 3 部.



侯小可,男,1985 年生,硕士研究生,主要研究方向为人工智能、数据挖掘.