



研究与开发

特征词选择与相似度融合的微博话题发现方法

陈红阳, 汪林林, 陈滢生, 鲁江坤, 左雪

(重庆人文科技学院计算机工程学院, 重庆 401524)

摘要: 微博短文本中存在一些相同或相近、但与主题关系不大的词项, 对准确度量文本之间的相似性具有较大的干扰作用, 影响微博话题被发现的质量。提出一种基于文本内容与结构化信息相结合的特征词选择算法, 能有效提取具有代表性的特征词, 并对文本、话题间相似度的计算策略进行改进, 然后将特征词选择算法与相似度计算方法融合, 应用于微博文本数据实现话题发现。实验结果表明, 本算法能有效降低话题发现的平均漏检率与误检率, 提高话题发现质量。

关键词: 微博; 特征词; 选择; 相似度; 话题发现

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2017258

A method of micro-blog topic discovery based on feature words selection and text similarity

CHEN Hongyang, WANG Linlin, CHEN Yingsheng, LU Jiangkun, ZUO Xue

College of Computer Engineering, Chongqing College of Humanities Science and Technology, Chongqing 401524, China

Abstract: Some words existing in micro-blog short text have a bad effect on the accuracy of text similarity calculation, further affecting the quality of topic discovery. And these words are the same in shape or semantic meaning, but remote from the topic. A novel method of feature words selection based on micro-blog short text content and structured information was proposed, which could effectively choose some important feature words from the text. Moreover, in computing the similarity between texts, an improvement on computing the similarity between the text and the topic was made. Finally, the methods were combined together and applied to discover micro-blog topics. Experimental results show that the new method of topic discovery can effectively reduce the average missing rate and false detection rate, and improve the quality of topic discovery.

Key words: micro-blog, feature word, selection, similarity, topic discovery

收稿日期: 2017-04-27; 修回日期: 2017-08-29

基金项目: 重庆市教委科技计划项目 (No.KJ1601601); 重庆市重点产业共性关键技术创新专项项目 (No.cstc2015zdcy-ztx40007); 国家自然科学基金资助项目 (No.61173184)

Foundation Items: The Commission of Science and Technology Plan Project of Chongqing (No.KJ1601601), The Key Industrial Generic Technology Innovation Special Project in Chongqing (No.cstc2015zdcy-ztx40007), The National Natural Science Foundation of China (No.61173184)

1 引言

微博话题发现技术可以有效组织和归类微博空间中散乱存在的海量文本数据以形成话题信息,从而帮助人们快速查看有价值的话题信息,并节约浏览时间^[1]。然而,微博文本具有文本长度短、用语不规范、口语化较严重、存在大量缩略语和错别字等特点,较大程度影响了文本特征词选择和相似度计算的准确性,也影响到文本聚类与话题发现的质量^[1,2]。目前的研究主要从文本特征词选择、相似性度量和文本聚类算法等方面出发,对微博话题发现展开研究,以提高微博话题发现的质量。

在特征词选择方面,参考文献[3]提出以TF-IDF(term frequency-inverse document frequency)方法来度量词项的权重,然后根据权重的高低进行特征词的选取。参考文献[4]在参考文献[3]的基础上,考虑了词项长度对词项重要性的影响,设计一种新的特征评估函数来实现文本的特征词选择。参考文献[5]也提出了一些用于对文本进行特征词选择的算法,如基于文档频率的特征提取法、信息增益法、互信息和开放检验等方法;参考文献[6]则综合文本内容多方面信息,基于统计和语义等多方面的信息来选取文本特征词,取得了一定的效果。参考文献[7]分析了微博短文本的结构,提出利用文本各部分信息的不同特征来进行词项权重计算。参考文献[8]提出挖掘主体标签(hashtag)的相关性,对于话题发现也具有一定帮助。然而,上述算法只考虑文本内容的特征来进行特征词选择,而对文本结构化信息利用不足,特征词挑选的准确性有待进一步提升。

在文本相似度计算方面,参考文献[9]提出的基于向量空间模型的余弦相似度方法存在一定缺陷,未从语义角度考虑词与词之间的关联性;参考文献[10]提出一种基于语义和最大匹配度的短文本相似度计算方法;参考文献[11]结合微博文本

特点,提出了基于时间、语义与社交关系的相似度,进一步提高了文本相似性度量度的准确性,应用于话题发现中,也取得了一定效果。但在上述相似度算法中,通常采用全部关键词参与运算。而实际上,权重低的部分关键词与文本话题的关系可能并不大,参与相似度运算反而会降低准确度。为解决上述问题,本文提出了基于特征词选择与文本相似度融合的微博话题发现方法。

2 微博话题发现流程及相似性度量方法

2.1 微博话题发现流程

微博话题发现的流程主要包含以下几个步骤:(1)微博数据获取,(2)微博文本预处理,(3)微博文本特征词选择,(4)微博文本相似性度量,(5)微博文本聚类、发现话题。微博文本预处理包含去噪、分词与停用词去除。其中,去噪(即微博信息过滤)主要是将文本中对话题内容表达没有贡献的数据信息清除掉,如表情符号、特殊字符和图片等,并过滤掉字数太少的微博文本^[4]。微博文本分词通常采用中国科学院所开发的汉语词法分析系统——ICTCLAS^[4],对净化后的微博文本进行分词。微博文本特征词选择方法通常采用词频、卡方统计、互信息与TF-IDF方法等以及相应的一些改进方法^[4]。微博文本相似性度量方面通常采用基于向量空间模型的余弦相似度方法,特征词集合与语义相似度方法等以及一些改进算法^[4]。微博文本聚类与话题发现常使用动态增量式的single-pass算法^[4],该算法实现比较简单,且具有较高的执行效率。

2.2 微博文本相似性度量方法

在计算微博文本相似度时,参考文献[11]提出了一种较好的计算方法,综合考虑了词与词之间的语义关联性、微博文本的结构化信息,对文本之间的相似性度量更为准确。该算法主要是:输入当前待比较的微博文本 D_1 、 D_2 ;判断两文本之间是否存在转发与评论关系,若是,则记两文



本之间的相似度值为 1, 即 $Sim(D_1, D_2)=1$, 否则, 引入微博文本发布时间来修正基于公共块序列的语义相似度形成新的相似度方法, 并据此来计算两微博文本之间的相似度值 $Sim(D_1, D_2)$, 具体计算式如下:

$$Sim(D_1, D_2) = Sim_{CB}(D_1, D_2) + \gamma_2 \times Sim_{time}(D_1, D_2) \quad (1)$$

$$Sim_{CB} = \alpha_2 \times Sim_{word}(D_1, D_2) + \beta_2 \times Sim_{order}(D_1, D_2), \alpha_2 > \beta_2 \quad (2)$$

$$Sim_{time}(D_1, D_2) = e^{-|time_1 - time_2|} \quad (3)$$

式 (1)、式 (2) 中, Sim_{CB} 则是基于公共块所含词项总数与公共块之间的组合顺序对微博文本之间的相似性的综合度量。 Sim_{word} 表示基于公共块所含词项总数来计算文本间的相似度, Sim_{order} 表示基于公共块之间的组合顺序计算文本间的相似度。在式 (3) 中, $time_1$ 、 $time_2$ 则是两文本所对应的发布时间, 而 $Sim_{time}(D_1, D_2)$ 则为两微博文本 D_1 、 D_2 之间的时间相似度。

3 基于文本内容与结构化信息相结合的特征词选择算法及微博话题发现

本文的主要思路为: 首先, 提出了一种有效的算法对微博文本进行特征词选择; 然后, 对参考文献[11]中的文本相似性度量策略进行改进, 进一步提高文本、话题间的相似度计算的准确度。并将这两个方法融合到话题发现过程中。

3.1 基于文本内容与结构化信息相结合的特征词选择算法

为更有效地从微博短文本中选取对微博内容表现力度大的特征词, 准确表示文本内容主题, 本文对文本结构化信息进行了充分的利用, 提出了一种有效的特征词选择算法。

该算法分为 3 个步骤对微博文本进行特征词的选择。

步骤 1 对微博文本内容进行两个方面的扩充: 利用 URL 指向的文本内容来丰富原始微博文

本; 根据微博文本的结构化特征, 搜集转发部分原始微博文本内容、对当前微博的评论信息和 URL 指向的文本内容, 以作为特征词权重计算的参考。

步骤 2 度量文本中词项权重。全面考虑文本内容与结构化信息等因素, 对微博各部分结构中的信息定义不同权重, 改进常用的权重度量方法 TF-IDF。其中, 词项的 TF (term frequency) 具体如下:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \quad (4)$$

$$n_{ij} = n_post_{ij} \times a + o_hashtag_{ij} \times b + o_url_{ij} \times c \quad (5)$$

式 (4) 中, tf_{ij} 指代词项 j 在文本 i 中出现的频率; n_{ij} 则指示词项 j 在与该文本 i 相关的各部分文字中出现的次数的加权和, 采用式 (5) 进行计算。式 (5) 中, n_post_{ij} 表示词项 j 在去除 URL 与主体标签之后的微博文本 i 以及相应转发信息和评论信息中出现的次数总和, o_url_{ij} 表示词项 j 在 URL 指向的文本中出现的次数, $o_hashtag_{ij}$ 表示词项 j 在主体标签中出现的次数, 并设置权重因子 a 、 b 、 c 分别对三部分出现次数进行加权计算。考虑到出现在主体标签中的词项通常比出现在微博文本其他位置的词项更能表现文本主题, 而 URL 所指向的文本内容通常只是作为当前微博内容的补充, 优先级设置相对较低。通常在根据实际情况选择 a 、 b 、 c 参数值时, 满足 $b > a > c$ 。

IDF 的计算方法为:

$$idf_{ij} = \frac{N}{n} + 0.01 \quad (6)$$

其中, N 表示文档集中所包含的微博总数, n 则表示文档集中包含词 j 的微博总数。

$$w_{ij} = tf_{ij} \times idf_{ij} \quad (7)$$

其中, w_{ij} 指代词项 j 在文本 i 中的权重值。

最后, 筛选微博文本特征词。设置词项权重阈值 α , 并将权重值超过该阈值的词项挑选出来作为微博文本的特征词构成特征词集合, 用以表示微博文本内容主题。

3.2 改进的文本与话题相似度计算方法

在微博话题发现过程中,需要进行的相似度计算通常包括文本与文本相似度计算以及文本与话题的相似度计算两个部分。本文将参考文献[11]中的相似度算法应用于文本与文本相似度计算,并对文本、话题间相似度计算策略进行了改进。

(1) 文本与文本之间相似性计算

在对微博文本进行特征词选取的基础上,以特征词集合来表示微博文本,即将一条微博文本 D 表示为 $D = \{term_1, term_2, \dots, term_n\}$, 其中, $term_i (1 \leq i \leq n, n < m)$ 表示微博文本中的第 i 个特征词, m 为原始微博中词项的总个数;然后使用参考文献[11]提出的基于语义、时间与社会化关系的微博文本相似性度量方法,进行文本之间的相似性计算。

(2) 文本与话题之间相似性度计算的改进

改进的计算思路为:首先,根据文本与文本

之间相似性计算当前文本与话题中已有微博文本的相似度值;然后,设置一定的阈值 β ,对超过该阈值的所有相似度值求平均,作为当前文本与话题的相似度值,从而降低相似度较低的文本(通常是无关文本)对微博文本聚类的干扰。

采用的微博文本 D 与话题 $Topic$ 的相似度值计算式如下:

$$Sim_{D, Topic} = \sum_{i=1}^k Sim_{D, T_i}, k < K \quad (8)$$

其中, T_i 为话题 $Topic$ 中与文本 D 相似度值大于或等于阈值 β 的第 i 个微博文本, K 则为话题 $Topic$ 中微博文本总数。

3.3 基于特征词选择与相似度融合的微博话题发现方法

将新的特征词选择算法与文本相似性计算改进算法融合,形成一种新的微博话题发现方法,该方法主要流程如图1所示。

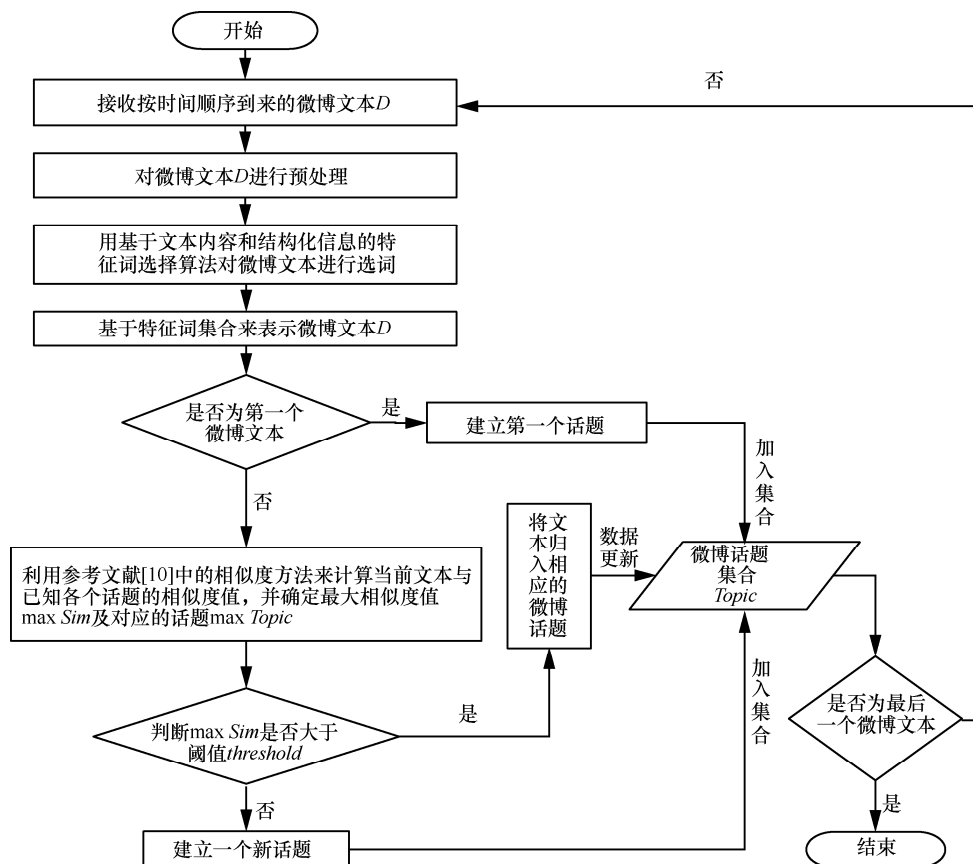


图1 基于特征词选择与相似度融合的微博话题发现流程



在图 1 中, 微博话题发现的实现具体如以下几个步骤所示。

步骤 1 文本预处理。接收依次按时间顺序到来的微博文本 D , 并对其进行预处理。

步骤 2 文本特征词选择与表示。基于本文所提出的特征词选择算法对微博文本 D 进行特征词选择, 并以特征词集合表示。

步骤 3 文本相似度计算与聚类 (Single-Pass 聚类算法)。首先, 判断当前微博文本是否为第一篇文本, 若是, 则将其作为第一个微博话题进行建立; 否则, 利用参考文献[11]中的相似度方法来计算微博文本 D 与已知话题的相似度值, 求得当前的最大相似度值 $\max Sim$ 及其所对应的微博话题 $\max Topic$; 然后, 再将 $\max Sim$ 与阈值 $threshold$ 作比较, 若 $\max Sim > threshold$, 则将该文本归入话题 $\max Topic$; 否则, 作为一个新话题进行建立。

步骤 4 判断微博文本 D 是否为最后一篇文本, 若是, 则结束, 否则, 返回步骤 1 重复上述流程继续处理下一个微博文本, 直至所有文本检测完毕。

4 实验及结果分析

4.1 实验数据

本实验所使用的数据主要来源于数据堂—新浪微博中所产生的数据^[12], 总共包含了 16 个微博话题, 共计 4 000 多条微博文本。其中, 每一条微博文本主要包含微博文本内容、文本发布时间、文本的回复数及转发数等信息。

4.2 实验的评价指标

本文采用话题检测与跟踪的常用性能评价指标——漏检率、误检率和归一化错误识别代价作为微博话题发现的性能评测指标^[9], 其计算方法具体如式 (9) 所示:

$$P_{\text{miss}} = \frac{\text{未检测到的与话题相关的微博文本数}}{\text{与话题相关的微博文本总数}}$$

$$P_{\text{fa}} = \frac{\text{所检测到的与话题不相关的微博文本数}}{\text{与话题不相关的微博文本总数}}$$

$$Norm(C_{\text{Det}}) = \frac{C_{\text{miss}} \times P_{\text{miss}} \times P_{\text{target}} + C_{\text{fa}} \times P_{\text{fa}} \times P_{\text{non-target}}}{\min(C_{\text{miss}} \times P_{\text{target}}, C_{\text{fa}} \times P_{\text{non-target}})} \quad (9)$$

其中, C_{miss} 与 C_{fa} 分别为漏检率和误检率的代价系数, P_{target} 与 $P_{\text{non-target}}$ 则代表先验目标出现的概率和先验目标不出现的概率, $P_{\text{non-target}} + P_{\text{target}} = 1$ 。通常, C_{miss} 与 C_{fa} 取值为 1.0 和 0.1, P_{target} 为 0.02, 则 $P_{\text{non-target}}$ 为 0.98。

4.3 实验结果及分析

4.3.1 实验 1

首先利用改进的特征词选择算法对各微博话题下的微博文本进行特征词选择, 并与原有基于统计与语义信息的特征词选择算法做对比。在实验中发现, 设置 $a=1$ 、 $b=2$; 如果原始微博文本 i 长度 < 30 , 则设置 $c=0.5$, 否则设置 $c=0.2$, 特征词选择效果较好。部分微博文本的原文内容以及经过特征词选择后的结果见表 1 (因篇幅有限, 原文内容只显示了部分)。

从表 1 中可以看出, 经过特征词选择后, 原始微博文本中一些无关、冗余的词项被去除掉, 而对文本内容表达贡献度大的词项则被保留下来; 改进的基于文本内容与结构化信息的特征词选择算法选取出的特征词更符合文本内容的主题表达, 扩充了新的词项, 丰富了文本内容主题表示。

4.3.2 实验 2

本文通过实验观察发现, 文本与话题之间的相似度值多在 0.1~0.7。为降低相似度较低的文本 (通常会是无关文本) 对微博文本聚类的干扰, 将阈值 β 的值设为 0.3。

然后, 将本文基于新特征词选择与相似度融合的微博话题发现方法与参考文献[11]中仅基于文本相似度算法的微博话题发现方法以及基于统计与语义信息相结合的特征词选择算法^[6]、文本相似性度量算法相结合的话题发现方法做对比,

表1 经两种特征词选择算法后各微博话题下的微博文本内容

微博话题	原始的微博文本内容	经原有特征词算法选取后的微博文本内容特征词	经改进特征词算法选取后的微博文本特征词
“昌都地震”中的微博文本1	【西藏昌都地震已造成21间房屋倒塌】中新网拉萨8月12日电,据西藏昌都地区的左贡县应急办消息:截至12日中午13时30分,发生于清晨5时23分的6.1级地震,已造成当地21间房屋倒塌。(分享自@凤凰网) http://t.cn/zQHRtJo	西藏,地震,造成,房屋,倒塌,拉萨,月,日,电,西藏,地区,应急,办,消息,午时,清晨,时分,级,地震,造成,当地,房屋,倒塌,凤凰,网	西藏,昌都,地震,造成,房屋,倒塌,拉萨,西藏,昌都,地区,左贡县,应急,办,消息,中午,发生,清晨,时分,级,地震,造成,当地,房屋,倒塌,通讯,交通,条件,落后,已经,派出,多股,救援,力量,前往,地震,震中
“昌都地震”中的微博文本2	【西藏昌都地震 暂无人员伤亡报告】西藏“左贡县与芒康县”交界处,8月12日发生地震后,西藏军区迅速“就近”派出部队官兵、医护人员及民兵350多人,赶赴灾区,展开救援	西藏,地震,报告,西藏,发生,地震,西藏,军区,迅速,就近,派出,部队,官兵,赶赴,灾区,展开,救援	西藏,昌都,地震,无,人员,伤亡,报告,西藏,发生,地震,西藏,军区,迅速,就近,派出,部队,官兵,医护,人员,民兵,赶赴,灾区,展开,救援

观测在不同的微博话题发现方法下,所有话题的平均漏检率、误检率以及归一化的错误识别代价等指标,具体见表2。

表2 各种微博话题发现方法下的性能对比

微博话题发现的方法	漏检率	误检率	归一化错误识别代价
参考文献[11]中的微博话题发现方法	0.154 8	0.020 4	0.254 8
基于原有特征词选择与相似度融合的微博话题发现方法	0.128 6	0.014 3	0.198 8
本文提出的微博话题发现方法	0.112 8	0.012 4	0.173 6

从表2可以看出,本文算法在话题的平均漏检率、误检率以及归一化错误识别代价这些指标均有所降低,具有更好的效果。主要是由于本文所提出的特征词选择算法综合考虑了文本内容和结构化信息,能更有效地选取出文本中具有代表性的特征,改进的相似度计算方法能在一定程度上减少低相似度文本对微博文本聚类的干扰。

5 结束语

本文提出了一种基于文本内容和结构化信息的特征词选择算法,对文本与话题相似度计算策略进行了改进,并将特征词选择和文本相似度相融合,实现微博话题发现。对算法实验结果进行了详细分析,以漏检率、误检率以及归一化错误识别代价作为话题发现效果评判指标进行算法比较。结果表明,相较于其他两种算法,本文算法具有更好的效果,能在一定程度上提高话题发现质量。

参考文献:

- [1] 丁兆云,贾焰,周斌. 微博数据挖掘研究综述[J]. 计算机研究与发展, 2016(7): 98-105.
DING Z Y, JIA Y, ZHOU B. Survey of data mining for micro-blogs[J]. Journal of Computer Research and Development, 2016(7): 98-105.
- [2] 据春华,鲍福光,戴俊彦. 一种融入公众情感投入分析的微博话题发现与细分方法[J]. 电信科学, 2014, 14(12):227-231.
JU C H, BAO F G, DAI J Y. Discovery and segmentation method in micro-blog topics based on public emotional engagement analysis[J]. Telecommunications Science, 2014, 14(12): 227-231.
- [3] 刘彦伟. 微博话题追踪系统的研究与实现[D]. 北京: 北京交通大学, 2013.
LIU Y W. Research and implementation of micro-blog topic tracking system[D]. Beijing: Beijing Jiaotong University, 2013.
- [4] 丁蕊. 微博热点发现技术的研究与实现[D]. 武汉: 华中科技大学, 2012.
DING J. Research and implementation of micro-blog hot topic detection[D]. Wuhan: Huazhong University of Science and Technology, 2012.
- [5] 姜芳,李国和,岳翔. 基于语义的文档特征提取研究方法[J]. 计算机科学, 2016, 43(2): 254-258.
JIANG F, LI G H, YUE X. Semantic-based feature extraction method for document[J]. Computer Science, 2016, 43(2): 254-258.
- [6] 黄贤英,陈红阳,刘英涛,等. 一种新的微博短文本特征词选择算法[J]. 计算机工程与科学, 2015, 37(9): 1762-1767.
HUANG X Y, CHEN H Y, LIU Y T, et al. A novel method of feature selection on micro-blog short text [J]. Computer Engineering and Science, 2015, 37(9): 1762-1767.
- [7] HUANG S, YANG Y, LI H, et al. Topic detection from microblog based on text clustering and topic model analysis[C]//2014 Asia-Pacific Services Computing Conference (APSCC 2014), Dec 4-6, 2014, Fuzhou, China. New Jersey:



IEEE Press, 2014:88-92.

- [8] 胡长龙, 唐晋韬, 王挺. 中文微博的 hashtag 话题相关性分析[J]. 计算机科学, 2013, 40(11A): 235-245.
HU C L, TANG J T, WANG T. Topic relevance analysis of hashtags in Chinese micro-blogging environment[J]. Computer Science, 2013, 40(11A): 235-245.
- [9] ZHENG J, LI Y. A hot topic detection method for chinese microblog based on topic words[C]//2014 2nd International Conference on International Conference on Information Technology and Electronic Commerce, Dec 20-21, 2014, Dalian, China. New Jersey: IEEE Press, 2014.
- [10] 孙建旺, 吕学强, 张雷瀚. 基于语义和最大匹配度的短文本分类研究[J]. 计算机工程与设计, 2013, 34(10): 3613-3618.
SUN J W, LV X Q, ZHANG L H. Short text classification based on semantics and maximum matching degree[J]. Computer Engineering and Design, 2013, 34(10): 3613-3618.
- [11] 黄贤英, 陈红阳, 刘英涛. 短文本相似度研究及其在微博话题检测中的应用[J]. 计算机工程与设计, 2015, 36(11): 3128-3133.
HUANG X Y, CHEN H Y, LIU Y T. Research on micro-blog short text similarity and its application in micro-blog topic detection[J]. Computer Engineering and Design, 2015, 36(11): 3128-3133.

[作者简介]



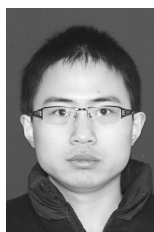
陈红阳 (1989-), 女, 重庆人文科技学院计算机工程学院助教, 主要研究方向为搜索引擎、文本挖掘及微博话题发现与跟踪等。



汪林林 (1945-), 男, 重庆人文科技学院计算机工程学院院长, 主要研究方向为数据库理论、应用及开发等。发表学术论文 90 多篇, 主持或参加科研项目 30 多项。



陈滢生 (1982-), 女, 重庆人文科技学院计算机工程学院讲师, 主要研究方向为文本挖掘、自然语言处理等。



鲁江坤 (1988-), 男, 重庆人文科技学院计算机工程学院助教, 主要研究方向为文本挖掘、自然语言处理等。



左雪 (1992-), 女, 重庆人文科技学院计算机工程学院助教, 主要研究方向为文本挖掘等。