DOI:10.3969/j.issn.1673-4785.201301012
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20130515.0839.002.html

一种基于情感的中文微博话题检测方法

方然[1,2],苗夺谦[1,2],张志飞[1,2]
(1.同济大学 计算机科学与技术系,上海 201804;2.同济大学 嵌入式系统与服务计算教育部重点实验室,上海 200092)

摘　要:针对微博这种特殊的文本形式的话题检测,传统的算法并不能取得很好的效果.为了提高其查全率,根据微博这种带有结构化特点的信息,提出了一种带有情感内容加权的话题检测方法.该方法基于含有负面情感的词语往往携带了更多的信息量这一论点,在现有短文本话题检测的算法中,通过加大含有负面情感的短文本在话题检测中的权重,之后再根据一种基于自查询的聚类方法进行话题聚类,将情感倾向融合到短文本话题检测中.在真实数据集上的实验表明,此方法能有效地进行话题聚类并检测话题,并提高了查全率.

关键词:中文微博;话题检测;聚类;情感
中图分类号: TP391　　**文献标志码**:A　　**文章编号**:1673-4785(2013)03-0208-06

中文引用格式:方然,苗夺谦,张志飞. 一种基于情感的中文微博话题检测方法[J].智能系统学报, 2013, 8(3): 208-213.
英文引用格式:**FANG Ran, MIAO Duoqian, ZHANG Zhifei. An emotion-based method of topic detection from Chinese microblogs [J]. CAAI Transactions on Intelligent Systems**, 2013, 8(3): 208-213.

An emotion-based method of topic detection from Chinese microblogs

FANG Ran[1,2], MIAO Duoqian[1,2], ZHANG Zhifei[1,2]
(1. Department of Computer Science and Technology, Tongji University, Shanghai 201804, China; 2 The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, China)

Abstract: Previous research studies have laid the foundation in the area of traditional topic detection and shown there are some effective ways to detect topics. However, the traditional algorithms do not work well in special situations for Chinese microblogs. In order to raise the recall ratio, the focus of this paper proposes to examine methods for detecting topics. The key to topic detection method, examines how to handle the structure of microblog with emotional content weighting, which is based on the argument that the negative words tend to carry more information. The existing topic detection methods for short messages merge emotional incination into the topic detection by first raising the weight of short messages containing negative emotion in the topic detection, then clustering the topics by a clustering method based on self-inquiry. The experiment on a real microblog dataset show that the approach provided in this paper can cluster topics and detect topics effectively, and also increase the recall ratio.

Keywords: Chinese microblogs; topic detection; clustering; emotion

在信息爆炸时代,从海量数据中挖掘出有用的信息显得格外重要.随着 Web2.0 的兴起,微博客即微博,一种基于用户关系的信息分享、传播以及获取的平台也随之兴起.微博用户可以通过网络、手机、其他客户端进行实时的短文本信息分享与传播.美国著名的微博网站 Twitter 用户数达到 5.17 亿[1],最高峰时达到 6 939 条每秒.用户通过这些微博平台发布生活中的所见所闻,以及对于一些事件的态度和评论等.

目前对于微博的研究大多是用户关系结构的分

收稿日期:2013-01-09.　**网络出版日期**:2013-05-15.
基金项目:国家自然科学基金资助项目(60970061,61075056,61103067);中央高校基本科研业务费专项资金资助项目(基于云计算的高效数据挖掘算法研究).
通信作者:方然. E-mail: ufo2243@ gmail.com.

析,但对于微博内容的分析并不多[2],而对其进行话题检测在舆情控制、自然灾害预警等方面又具有重要的实际意义.在话题检测与跟踪(topic detection and tracking, TDT)领域,传统的算法主要面向于文本和语音形式的新闻报导[3],针对的目标并不是这种短文的微博形式,很多现有的方法如凝聚层次聚类算法[4]、UMass 和 Dragon 等方法[5]并不能直接在微博上使用.近年来,很多学者也在基于微博的话题检测方面做了一些研究,如在地震监控方面,Takashi 等[6]提出一种基于关键字为证据的贝叶斯决策方法,可以实时地通过 Twitter 监控地震发生的情况.郑斐然等[7]提出的一种中文微博新闻话题检测方法,也通过实验证明了其方法可以从大量消息中检测出新闻话题.而在情感词语所表达的信息量方面,Garcia 等的研究[8]表明包含积极内容的词语的使用次数相对于包含消极内容的词语要少,通过自信息量(self-information)的比较,这些消极的词语包含了更多的信息量.

本文通过分析微博自身的文本特点[7,9],提出了一种基于情感内容加权的话题检测方法,该方法在向量空间模型的基础上,在微博话题检测的主题词选取时,通过对具有负面或消极含义的词语进行加权的方法筛选出最适合的主题词,再进行聚类.

1　微博话题检测方法

本文提出的话题检测方法以中文微博为处理对象,分为预处理、分词、主题词检测、话题聚类几大部分.中文微博在格式上有着其自己独特的特点,每条微博是由不超过 140 个中文字长度的文本与图片组成,这里不考虑其图像的含义只考虑文本的内容.在文本中还包含一些微博的特殊格式,例如用“#主题#”来表示这条微博是属于某一些特定主题的,这里的主题是人为设定的,大多数情况下是一些活动或商品的推广,会给话题检测造成一定的影响.用“@用户”来表示这条微博与某些制定的微博用户有关,一般情况下是转发微博的时候系统会默认“@用户”指被转发的用户,还有一些情况是向特定用户发的对话性质的微博内容.这些特殊的格式都必须在预处理中进行相应的处理,以防止其对话题检测造成不良的影响.

在话题检测过程中,为话题建立相应的模型也是其中的一个基础性问题,常见的模型有空间向量模型、词汇链模型、图模型等.本文使用空间向量模型,其中计算文本相似度的方法包括 Okapi 公式、Clarity、WeightSumt、余弦相似度[10]等,这里采取了一种自查询的方法来计算文本的相似度.

1.1　数据预处理

在预处理这一步,目标是将原始的微博数据根据其自身的特殊格式进行相应的处理,排除一些可能对话题检测的影响.这里的处理规则大体上可以分为 2 类:一类是针对微博本身的文本内容的预处理规则;另一类是针对微博文本内容以外包括发微博者的一些其他数据的预处理规则.

1)针对微博的文本内容.

①对于带有“#主题#”格式的微博,由于这个主题的词大多数是人为设定的,大部分带有商业目的而且转发数量大,这对话题检测会有不利的影响.于是删除所有带有这种格式的文字内容,但仅删除“#主题#”格式的字段,保留其他的文本内容,因为这部分内容是用户关于这些人为设定的主题的讨论,可以作为提取话题的文本.

②对于带有“@用户”格式的微博,大多数情况下是在转发微博时被使用,根据其格式删除“@用户”的字段.这是因为微博的用户名不会给话题检测带来帮助,相反在统计词频的时候还会带来很多干扰,所以删除所有能确定是用户名字的字段.

2)针对发微博用户.

由于微博存在一些称为“僵尸账号”的微博账号,这些账号大多数是有名无实的微博账号,它们通常是由系统自动产生的恶意注册用户,这些账号会发布大量重复的内容用于一些商业目的,会对话题检测造成不利的影响,因此在预处理这一步要对其进行判断.判断帐号是否为僵尸账号是一件较为复杂的工作,由于需要进行大量的判断,因此将其尽量简化,主要根据帐户的收听人数来判断,收听人数少于阈值 F 的用户,将其判断为僵尸账号,这样能避免大量的僵尸账户,但会把一些不活跃的正常账户也剔除掉.

1.2　分词

汉语中词是最小、能独立活动、有意义的语言成分,但不像英语或者其他语言中词语之间有明显的标记来加以区分.因此分词也是中文信息处理的关键,分词的方法有很多,如基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法等.由于本文的重点在于微博的话题检测,这里直接采用中国科学院计算技术研究所的 ICTCLAS 分词系统[11],对经过预处理之后的文本语料进行分词处理.ICTCLAS 分词系统在分词的同时会进行词性的

标注,在分词和词性标注之后微博文本会变成如图1所示的形式.

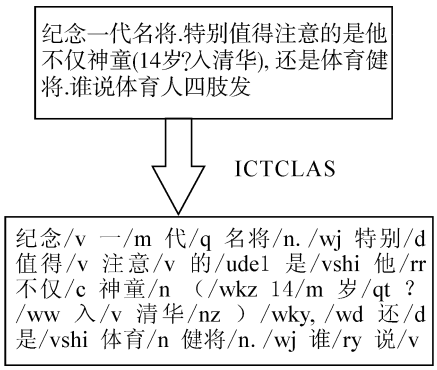

图1 ICTCLAS中文微博分词示例

Fig.1 Word segmentation example using ICTCLAS

由于在对微博进行话题检测时面对的是海量的微博数据,因此需要进行一定的删减,再进行主题的检测.在各个词性中,名词和动词对表达主题的贡献最大,故最后会保留下来的是每个微博中的动词和名词.对文本的情感倾向判断分析主要分为3个级别:词汇级别、句子级别和文档级别[12].本文采用HowNet免费对外的褒贬义词表来简单地判断短文本的情感倾向,并对其进行情感倾向加权.记 e 为一条微博的情感倾向值,假设该条微博分词后共有 m 个词,则有

$$e = \frac{\sum_{i=1}^{m} \text{emotion}(i)}{m}.$$

式中:根据HowNet的褒贬义词表,贬义词语的emotion(i)记为1,褒义词语的记为-1,不在中文情感词库中的词语记为0.这里的 e 将用于下一步的主题词检测.

1.3 主题词检测

由于微博数据的特殊性,不同于传统话题检测面向的对象,重要的一点是它还具有很强的时序性,传统的TF-IDF等方法无法利用微博文本的时序性特点,因此无法使用用于静态长文本的传统方法来计算主题词.本文采用兼顾被检测词在短时间内的增长速率和当前词频,当然更重要的是对待选词使用情感倾向加权的评价方法来挑选出适当的主题词.

将微博的文本按照固定的时间窗口划分成若干块,每块都固定一个时间长度 T,这和具体实验时微博的采样频率有关.时间窗口确定之后,可以得到在最近的一个时间窗口,即当前时间窗口中某词的频率 F:

$$F = \frac{F_i}{F_{\max}}. \tag{1}$$

式中:F_i 是该词在当前窗口中的出现次数,$F_{\max}$ 为当前时间窗口中的最高词频.

假设包含该词的微博在当前时间窗口有 n 个,则该词的情感倾向的加权 E 为

$$E = \frac{\sum_{i=1}^{n} e_i}{n}.$$

式中:e_i 为当前时间窗口中所有包含该词语的微博情感倾向.由于待处理的数据量很大,因此采用相对简单的算法来判断微博情感倾向.

本文再引入一个增长系数 G 来表示一个词在当前窗口时出现频率的增长速度,同时设定一个回顾时间窗口 B,来限定该增长系数考察的范围.由于是在一定时间窗口内,并不要求该词的词频在之前的回顾时间窗口的范围内持续增长,因此增长系数 G[7] 并不考察这一点,而需要考察的是相对于当前时间窗口的增长速度:

$$G = \frac{F_i \times B}{\sum_{j=1}^{B} F_j}.$$

式中:F_i 为该词在当前窗口中的出现次数.G 的值越大越说明该词在当前时间窗口中出现了突增的情况,就越有可能是主题词.

考虑使用上述3种权值来获取主题词列表,因此构造了一个综合的权值 V 来评价一个词是否为主题词的程度:

$$V = \log G + \alpha \log F + \beta \log |E|.$$

式中:V 值与主题词程度正相关,α 与 β 用于调剂三者之间的比例关系,从实际结果来看 α 取1.0~1.5最适当[7],β 取±0.5左右较为合适,且 β 与 E 不同正负,这样负面情感的词语就会增加其主题词权值.

1.4 话题聚类

聚类的目的是为了将主题词列表中的候选主题词聚类成若干个词为一组的话题.完成聚类后的主题词将会得到若干类,每类都由一个或多个主题词组成,这样的一类就形成了一个新闻话题.本文聚类算法的核心思想是 K 均值聚类算法的改良,是一个增量的聚类算法,由于该方法并不预先假设话题数量,因此初始状态为只有第1个词为初始类.大体步骤如下:

1)以第1个词为初始类;

2)读入下一个词,判断它与已有每一个类的距离(类的位置取其所包含词的平均位置);

3)设定一个阈值 D,如果这个词与每一个现有类的距离都大于 D,那么认为该词为一个新的类;

4)重复 2)~3),直到所有词处理完毕.

计算一个词与其他词之间的距离的方法大体上有 2 种:一种是预先确定词与词之间的距离;另一种是增量式的,随着文本不断的读入不断调整词与词之间的距离关系.传统的话题检测算法中大多数采取的是第 1 种方法,因为传统话题检测所面对的检测对象大多数是长篇的文章,对于词与词之间的相似度有大量的预先经验,所以更适用于此种方法.本文采用第 2 种方法,如果 2 个词出现在同一条微博中,就认为这 2 个词语更为相似.具体的一个词到一类的距离公式为

$$D(a,C)=\frac{\sum_{i=1}^{n} d(a,C_i)}{n}.$$

式中:$D(a,C)$为词 a 到类 C 的距离,C_i 为 C 类中的一个词,$d(a,C_i)$为事先维护的词与词之间相似度的表,即两者出现在同一微博中的次数.此处距离理论上的意义是指,如果一个词与某一类中的词出现在同一微博中的次数较多,那么该词就与这一类的距离较近.

2　实　验

本次实验数据通过新浪微博 API 进行抓取,抓取了 2012-7-31—2012-8-2 之间 200 万条微博数据,具体的数据格式如表 1 所示.

表 1　API 抓取微博样例

Table 1　Weibo example using API

序列号	文本内容	发布时间
3473823024605352	都江堰夜景还是保持住,没有给地震搞了	2012-07-31 12:03:31
3473823028249919	波士顿红袜对底特律老虎,开场走起,Go Red SOX Go!	2012-07-31 12:03:33
3473823028800377	朴泰桓狂言:孙杨来奥运是个错误,只会衬托我的速度	2012-07-31 12:03:33

表 1 中只列举了最主要的几项内容,实际抓取的数据还包括转发次数、评论次数以及发微博者的一些相关数据,如他的粉丝数、微博数等,用于判断是否被过滤掉.另外还对这段数据人为标注了 7 个事件进行评价,如“伦敦奥运会”、“体操男团冠军”、“爱情公寓 3 开播”等.

2.1　预处理

预处理主要包括两部分:1)根据发布微博的用户 id 及其数据过滤掉部分僵尸微博;2)对微博内容进行分词和词性标注.然后对其分词进行统计,会发现词列表近似服从帕累托分布,少数常用词语大量反复出现而绝大多数词语所占的总比重很低.在总共统计的8 000多个词中,前1 000个词占总数量的50%以上,而后1 000个词占了不到 1.5%.分词的同时进行词性标注,只保留动词和名词作为主题词的候选,统计的结果如图 2 所示.从图 2 中可以看出,前面少量词语占了很大的比重,而后面会有大量相似的词语,这也是对微博进行分词统计词频后发现的一个特性.

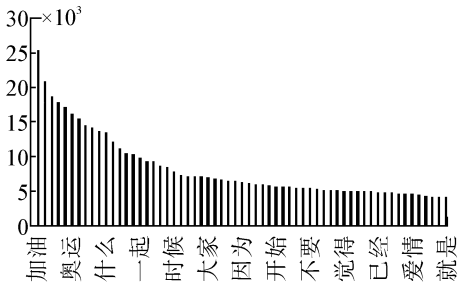

图 2　词频示例

Fig.2　Word frequency example

2.2　时间窗口 T 以及聚类阈值 D 的选取

实验发现时间窗口 T 和聚类阈值 D 这 2 个经验值的选取很大程度上依赖于原始数据的采样频率.实验数据中新浪微博 API 的流量限制为每 20 min取 5 万条,而当采样频率发生变化的时候,时间窗口与聚类阈值的选取也要随之变化.

聚类阈值 D 从实验中可以看出,当时间窗口固定为 3 h 时,随着 D 值的增大,查全率下降但查准率上升.在上述固定采样频率的条件下 D 取 20 时,能取得相对理想的查准率和查全率.

时间窗口 T 的选取也格外重要,而且更大程度上依赖于采样频率,这是因为在某些特定的时间段微博上会爆发大量相同的词语.这些词语大部分是和特定时间有关,而非重要的主题词,如“吃饭”、“睡觉”等一般与时间不相关,这就使得时间窗口不宜选取得过短,经过实验发现选取在2~3 h较好.

话题检测结果如表 2 所示,可以看出该方法可以成功地进行聚类,得到相应的微博话题,并且通过

情感倾向加权能够取得一定的效果.不过实验过程也发现了一些问题,如分词不准导致的不利于后续话题检测,部分微博围绕一些人名,但有些人名的分词效果并不好,会对话题检测产生部分噪音,而当围绕这些人名的微博大量产生的时候,就会对话题检测产生较大影响.

表2 部分话题聚类结果

Table 2 Part of the topic clustering results

时 间	话题内容
2012-07-31 18:00:00	男篮、伦敦、小组赛、俄罗斯、不敌
2012-08-01 04:00:00	失利、无缘、体操、团体、女子
2012-08-01 06:30:00	游泳、记录、混合、奥运会、叶

同时由于此次实验所取的数据是在奥运期间,部分词语如"奥运"大量产生,而其理论上应该分属很多个子话题,如"奥运篮球"、"奥运体操"等,但当前算法并未考虑这种情况,这也是后续需要改进的方向之一.在查全率方面,试验做了一次自身的对比,即在同样的数据下进行有无情感加权的查全率对比,实验显示,在有情感加权的情况下查全率从71.4%提升至85.7%,说明该算法能够在一定程度上提高查全率.

3 结束语

在总结前人在微博话题检测工作的基础上,提出了包含情感倾向加权的一种微博话题检测方法,并通过在新浪微博上的实验说明了其可用性.同时,需要指出的是该方法在很多方面还需要改进,例如实验中所取的时间窗口 T 以及话题聚类中的阈值 D 很大程度上依赖于实验数据,评价标准也相对缺乏,同时缺乏中文的微博语料库和标注话题,当然如何更好地提高查准率和查全率也是后续研究的重点.

参考文献:

[1] LUNDEN I. Analyst: Twitter passed 500M users in June 2012, 140M of them in US[EB/OL]. [2013-03-26]. http://techcrunch. com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/.

[2] RAMAGE D, DUMAIS S, LIEBLING D. Characterizing microblogs with topic models[C]//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, DC, USA: The AAAI Press, 2010: 130-137.

[3] 洪宇,张宇,刘挺,等.话题检测与跟踪的评测与研究综述[J].中文信息学报, 2007, 21(6): 71-85.
HONG Yu, ZHANG Yu, LIU Ting, et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing, 2007, 21(6): 71-85.

[4] YANG Y M, PIERCE T, CARBONELL J. A study of retrospective and on-line event detection[C]//Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 1998: 28-36.

[5] ALLAN J, CARBONELL J, DOODINGTON G, et al. Topic detection and tracking pilot study final report[C]//Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop. Lansdowne, USA, 1988: 194-218.

[6] SAKAKI T, OKAZAKI M, MATSUO Y. Earthquake shakes Twitter user: real-time event detection by social sensors [C]//Proceedings of the 19th International Conference on World Wide Web. New York, USA: ACM, 2010: 851-861.

[7] 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): 138-140.
ZHENG Feiran, MIAO Duoqian, ZHANG Zhifei, et al. News topic detection approach on Chinese microblog[J]. Computer Science, 2012, 39(1): 138-140.

[8] GARCIA D, GARAS A, SCHWEITZER F. Positive words carry less information than negative words[J]. EPJ Data Science, 2012, 1(1): 1-16.

[9] 印桂生,张亚楠,董宇欣.基于提升系数的微博异常排名检测方法[J].哈尔滨工程大学学报, 2013, 34(4): 488-493.
YIN Guisheng, ZHANG Ya´nan, DONG Yuxin. A boost factor based detection method for abnormal rank of microblogging[J]. Journal of Harbin Engineering University, 2013, 34(4): 488-493.

[10] 张晓艳,王挺.话题发现与追踪技术研究[J].计算机科学与探索, 2009, 3(4): 347-357.
ZHANG Xiaoyan, WANG Ting. Research of technologies on topic detection and tracking[J]. Journal of Frontiers of Computer Science & Technology, 2009, 3(4): 347-357.

[11] ZHANG Huaping, YU Hongkui, XIONG Deyi, et al. HHMM-based Chinese lexical analyzer ICTCLAS[C]// Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing. Stroudsburg, USA, 2003, 17: 184-187.

[12] 陈岳峰,苗夺谦,李文,等.基于概念的词汇情感倾向识别方法[J].智能系统学报, 2011, 6(6): 489-493.
CHEN Yuefeng, MIAO Duoqian, LI Wen, et al. Semantic orientation computing based on concepts[J]. CAAI Transactions on Intelligent Systems, 2011, 6(6): 489-493.

作者简介:

方然,男,1988年生,硕士研究生,主要研究方向为自然语言处理、智能信息处理、数据挖掘.

苗夺谦,男,1964 年生,教授,博士生导师,中国计算机学会高级会员、中国人工智能学会理事、上海市计算机学会理事.主要研究方向为智能信息处理、粗糙集、粒计算、网络智能、数据挖掘等.已主持完成国家级、省部级自然科学基金与科技攻关项目多项,并参与完成国家“973”计划项目 1 项、“863”计划项目 2 项等.曾获国家教委科技进步三等奖、教育部科技进步一等奖、上海市科技发明一等奖、重庆市自然科学一等奖等.发表学术论文 160 余篇,其中被 SCI、EI 检索 80 余篇,出版教材及学术著作 9 部,授权专利 9 项.

张志飞,男,1986 年生,博士研究生,主要研究方向为文本挖掘、自然语言处理.

第 8 届中国生物识别学术会议(CCBR2013)

The 8th Chinese Conference on Biometric Recognition (CCBR2013)

生物识别是模式识别、图像处理、人工智能等学科领域的前沿方向,同时也是保障国家和公共安全的战略高新技术、电子信息产业的新增长点.中国生物识别学术会议从 2000 年开始在北京、杭州、西安、北京、广州先后成功主办过 7 届,有力推动了我国生物识别的学科发展和应用推广,同时为国内生物识别学术界和产业界同行提供了一个交流与合作的平台.第 8 届中国生物识别学术会议(CCBR2013)由山东大学、中国科学院自动化研究所和中国人工智能学会联合主办,将于 2013 年 11 月 16—17 日在济南举行.本届会议向广大科技工作者公开征集优秀学术论文(英文),大会录用的稿件将由 Springer 出版社的 Lecture Notes in Computer Sciences(LNCS)图书系列出版,并被 EI 和 ISTP 检索.

征文范围

生物特征获取装置	指纹识别	静脉识别
生物识别过程的人机交互	虹膜识别	其他生物特征的识别与处理
生物特征质量评价	说话人识别	多模态生物识别与信息融合
生物特征信号质量增强	笔迹(含签名)识别	生物特征数据库建设与合成
基于生物特征的情感计算	步态识别	生物特征识别应用与系统
人脸检测、识别与跟踪	掌纹识别	其他相关内容

重要日期

投稿截止日期:2013 年 7 月 5 日

录用通知日期:2013 年 8 月 20 日

会议召开日期:2013 年 11 月 16—17 日

联系我们

联 系 人:袭肖明

通信地址:山东济南市舜华路中段 山东大学 计算机学院

电　　话:15069056021

邮　　箱:ccbr2013@ sdu.edu.cn

网　　址 http://ccbr2013.sdu.edu.cn