面向微博客的热点事件情感分析方法

宋双永 李秋丹 路冬媛

(中国科学院自动化研究所 北京 100190)

摘 要 微博客是一种新兴的网络信息交互平台,近年来受到越来越多的用户的关注。信息的简洁性以及传播渠道的多样性使得微博客成为广大网民浏览热点事件相关信息和发表个人观点的重要途径。分析和监测微博客内容中所包含的情感信息,能够了解民众对特定热点事件的关注程度和情感变化,从而辅助评估和掌握事件的发展状况。因此,提出一种面向微博客的热点事件情感分析方法,该方法首先自动挖掘用户对某热点事件的多个关注点,并针对不同关注点进行情感分析以及情感趋势监测,最终实现一个可视化的热点事件情感趋势分析原型系统。通过实例验证了微博客信息在网络热点事件的情感分析和监测中的有效性。

关键词 微博客,热点事件,情感挖掘,情感趋势监测

中图法分类号 TP391 **文献标识码** A

Hot Event Sentiment Analysis Method in Micro-blogging

SONG Shuang-yong LI Qiu-dan LU Dong-yuan

(Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract Micro-blogging, as a new form of social communication, is attracting more and more attention during the last few years. Simplicity of information and diversity of communication modes in micro-blogging make it an important channel of seeking information about hot events and expressing personal views. By analyzing and monitoring sentiment information in micro-blogging posts, we have opportunities to gain insights into users' emotion trend on hot events, which can help us evaluate and grasp the current situation of hot events. Therefore, we proposed a hot event sentiment analysis method in micro-blogging. This method first automatically detects different aspects of an event from the user perspective, and then performs sentiment analysis and emotional monitoring on each aspect. Additionally, we built a novel hot event sentiment analysis prototype system. The experimental results show that micro-blogging is effective in monitoring hot events on World Wide Web.

Keywords Micro-blogging, Hot events, Sentiment mining, Emotional trend detection

1 引言

随着互联网的发展和网民数量的迅猛增长,网络给社会带来前所未有的变革,全面改变了人类社会的生态模式[1,4]。近年来,基于Web2.0的互联网技术更是提高了网络终端用户的个性化程度。微博客是新近兴起的互联网热门服务,方便快速的信息传播方式使其成为广大网民浏览热点事件相关信息和发表个人观点的重要渠道。在热点事件发生之后,人们往往能够通过微博客第一时间获取事件信息,并进行反馈和传播[2]。微博客中蕴含着用户对于事件的情感,其信息传播模式具有开放性、实时性和自由选择性的特点[8,10],对微博客中用户所产生的舆论信息进行情感分析和情感趋势分析,能够很好地挖掘网民群体的行为规律,从而为用户决策提供服务[3,5]。

本文提出一种面向微博客的热点事件情感分析方法,该方法能够自动分析微博客数据中用户帖子所包含的情感倾向,从而监测用户群整体的情感变化趋势。首先抽取事件中所包含的不同方面的关注点,然后检测不同关注点相关的帖子中所包含的用户情感信息,接着统计用户群对各个关注点的情感变化趋势。最后,基于所提方法,本文实现了一个热点事件的情感趋势分析原型系统,并通过实验证实了该系统在发现微博客用户情感及其变化规律中的有效性。

本文第2节概要介绍与本文内容相关的研究工作;第3节详细介绍本文提出的微博客中针对热点事件的情感分析方法流程;第4节给出该方法的实验结果分析,并通过对热点事件实例的情感趋势分析,表明了本文设计系统的有效性;最后进行总结并对未来工作进行展望。

2 相关工作

微博客信息的时间特性和空间特性[11]使其成为一种高效的热点事件传感器,可以帮助用户及时、准确地了解事件发展状态[12]。文献[12]设计了一个面向微博客信息的事件监

本文受国家重点基础研究发展计划(973)(2007CB311007),北京市自然科学基金(4112062)资助。

宋双永(1986—),男,博士生,主要研究方向为信息检索等,E-mail:shuangyong.song@ia.ac.cn;李秋丹(1976—),女,博士,副研究员,主要研究方向为信息检索、数据挖掘、移动电子商务等;路冬媛(1984—),女,博士生,主要研究方向为信息检索等。

测系统，用于检测例如地震等自然事件发生的地理中心以及事件在空间上的扩散轨迹。文献[13]将微博客信息所反映出的用户对事件的兴趣表示成类似图像中的'像素'形式，并通过帖子包含的时间信息，将事件相关帖子表示成类似视频文件的动态结构，用来监测事件发展的形势变化情况。文献[14]通过对俄克拉何马草原大火和雷德河洪灾发生后微博客信息的分析，阐述了微博客对于自然灾害事件发展情况所具有的良好监测作用。

当前，微博客平台已经成为了一种重要的社会传播媒介，蕴含着丰富的用户观点和情感信息。微博客内容简短，表意直接，使阅读者能方便了解其中包含的用户情感。文献[6]以微博客作为用户对商品品牌的口碑平台，从用户对各个品牌的评论中发现用户情感及其随时间的变化情况，以此来了解微博客用户的品牌爱好，为商家提供了良好的用户兴趣收集渠道。文献[7]将微博客中的用户发表信息作为民意测验数据，通过比较用户情感变化与美国总统选举的发展状态，发现二者的变化趋势具有很大程度的相似性，证明了微博客在监测民意上的积极作用。文献[9]通过对热点事件发生之后微博客信息中包含的用户情感进行监测，分析整个事件与网络用户的情感变化规律之间的关联关系。事实上，热点事件发生之后，用户往往会对事件有不同的关注点，针对事件不同关注点的情感分析和情感趋势监测能够更细致地发现与热点事件相关的网络用户的情感活动。因此，本文设计了一个可视化的热点事件情感趋势分析原型系统，以从热点事件的不同关注点中发现用户的情感变化。

3 面向微博客的热点事件情感分析

3.1 方法流程

图1给出了本文提出的热点事件情感分析流程框图。其中，情感分析模块包括事件关注点抽取、情感极性判断和情感趋势分析3个步骤。首先对事件相关信息中所包含的主题词语进行聚类分析，从中发现事件发生和发展过程中所包含的一些关注点；其次，针对每个关注点的内容进行句子级的情感极性判断：抽取主题词和情感词的搭配组合，并根据语言习惯删除掉不合理的组合搭配，之后对保留的组合进行基于情感词典的情感极性判断，从而实现对包含该组合的原句子的极性判断；最后，利用对情感极性频率的时间序列统计，得到关于热点事件的情感趋势分析结果。基于本文所提方法，利用Lucene和JSP技术设计了一个热点事件情感趋势分析原型系统。

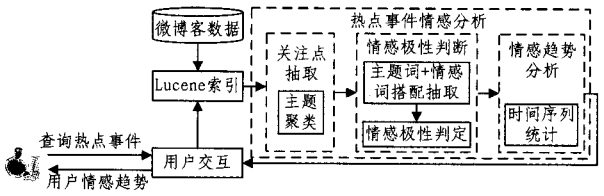

图1 热点主题情感分析方法流程图

3.2 关注点抽取

热点事件发生后，人们针对该事件往往会产生多个关注点，不同的关注点会引发人们不同的态度，因此对事件的多个关注点分别进行情感挖掘，能够更好地理解用户对事件的关注方式。本文借鉴文献[16]中的主题聚类方式，从热点事件中抽取事件包含的关注点。

主题聚类一般包括主题抽取、聚类以及聚类结果描述3个部分[16]。设定热点事件中包含的主题均为名词，并利用开源中文分词软件ICTCLAS将微博客帖子内容进行分词，从中抽取高频名词作为候选主题词。然后，根据主题词在每个帖子中出现的次数，将主题词映射到帖子文档集合上，表示成N维向量：$topic=\{d_1,d_2,\cdots,d_N\}$，其中，$N$表示热点事件相关帖子数量，$d_i(1\leqslant i\leqslant n)$表示主题在第$i$个帖子中出现的频率。

在主题聚类部分，我们使用WEKA数据挖掘工具实现对主题词的K均值聚类。K均值算法以K为输入参数，将对象的集合分为K个簇，使得结果簇内的相似度高，而簇间的相似度低。K均值聚类算法在准则函数收敛时结束聚类，其中准则函数如式(1)所示。

$$E=\sum_{i=1}^{K}\sum_{p\in C_i}|p-m_i|^2 \tag{1}$$

式中，E是数据集中所有对象的平方误差和；p是主题向量；m_i是簇C_i的中心点(均值)，通过式(2)计算得到：

$$m_i=\sum_{p\in C_i}(w_p.\ p) \tag{2}$$

式中，w_p为聚类样本p在簇C_i中的权重。假设簇C_i中聚类样本个数为n_i，则文中设定：

$$w_p=\frac{1}{n_i} \tag{3}$$

对主题词进行聚类的过程中，分别在K取不同值($K\in[2,5]$)的情况下对聚类结果进行分析，选取主题分类明确、各个类别主题所占帖子数量比例较为均衡的情况进行保留。

3.3 情感分析

情感分析过程主要包含两个步骤：1)主题词与情感词的搭配关系抽取；2)基于极性词典的搭配极性判断。下面给出这两个步骤的详细介绍。

3.3.1 词语搭配抽取

本文中的词语搭配是指主题词与情感词的搭配。此过程是为了将主题词与能够合理修饰它的情感词关联到一起，提高情感分析的准确性。在研究基于不同关注点的情感分析的过程中，我们发现关注点中包含的主题词主要为名词，所以在文献[15]提出的几种词语搭配模式的基础上，本文考虑$a+n$(形容词+名词)，$v+n$(动词+名词)，$n+a$(名词+形容词)3种模式的搭配。

首先将微博客帖子数据分割成单个的句子，此处的句子是指由点号[1)]分割而成的语言单位。首先检测句子中是否包含3.2节中抽取的主题词语，若包含，则进一步匹配上面提出的3种词语搭配，并将匹配成功的3种搭配保存在临时搭配库中。对临时搭配库中的每种搭配基于出现频率进行排序，对高频搭配进行人工标注后，形成最终的搭配库。

3.3.2 极性判断

以极性词典为基础，对主题词-情感词搭配的情感极性进行判断。本文实验使用的极性词典包括文献[17]相似度计算软件包中提供的xsimilarity中文情感词典和文献[18]中建立的极性词汇表。其中，前者包含4566个正极性词汇和4370

1) http://baike.baidu.com/view/845777.htm

个负极性词汇，后者包含 2764 个正极性词汇和 7778 个负极性词汇。在进行基于极性词典的词语搭配极性判断过程中，将两个极性词典结合使用。需要指出的是，将两个极性词典结合的过程中，如遇到某极性词语在两个极性词典中的极性标注结果不同时，会对该词语的情感极性进行人工判定。将两个极性词典结合的方式，能够使得极性词典中词语覆盖范围进一步扩大，更多地识别出搭配的极性，并且通过在结合词典的过程中对带有极性歧义的词语的进一步处理，更好地保证了极性判别结果的准确性。

3.4 情感趋势分析

为了检测用户群对某一特定热点事件的情感变化趋势，以天为单位，对得到的情感极性判别结果进行统计分析。以此为基础，设计了一个面向微博客数据的热点事件情感趋势分析原型系统。该系统利用 Lucene[2] 为微博客数据建立索引，用户可以输入需要查询的热点事件主题和相应的时间信息，系统根据用户输入信息查找该段时间内包含用户查询主题的微博客数据，并将其作为系统当前的输入数据。

对查找出的微博客帖子数据中包含的句子进行极性判断和统计分析之后，利用 ChartDirector[3] 实现在系统中的情感趋势变化曲线的显示。显示结果包括热点事件各个关注点相关的用户所表达的正极性情感和负极性情感的统计结果随时间变化的趋势图。该系统的实现，能够方便用户浏览查询时间段内微博客用户群对该热点事件的情感倾向变化。

4 实验结果及分析

4.1 数据描述

从国内某主流微博客网站上收集用户的帖子信息，对本文方法进行试验分析。我们从中收集了 2011 年 3 月－4 月中‘日本地震’事件专题下的数据，并抽取每条帖子中的帖子内容和发表时间作为实验数据。我们对数据做了以下预处理：1)删除帖子中包含的网址链接信息；2)删除帖子中的非中文语句或词语；3)删除在进行前两步预处理之后内容为空的帖子；4)对帖子内容进行中文分词处理。最终得到的数据集总共包含 5021 条帖子。

4.2 实验结果分析

4.2.1 关注点抽取结果

从日本地震事件对应的数据集中，抽取其包含的关注点，结果如表 1 所列。

表 1 日本地震主题中包含的关注点抽取结果

Cluster1	地震，海啸，东京，世界，国家，电，时间
Cluster2	核电站，核，放射性，机组，事故
Cluster3	中国，核辐射，新闻，灾难，政府

从表 1 结果可以看出，基于主题聚类的关注点抽取方法能够将微博客上该段时间内日本地震的相关内容分为 3 个方面。第一个方面描述了地震、海啸这些自然灾害本身的情况，第二个方面描述了自然灾害导致的核泄漏问题的相关内容，而第三个方面则更多地阐述了这一事件对中国方面的影响。通过该过程，能够更清楚地了解在对应事件中所包含的用户关注的不同角度，使得我们能够针对不同关注角度更细致地挖掘用户所表达出的情感。

4.2.2 情感极性判断结果

在对主题词＋情感词搭配进行基于极性词典的极性判断之后，我们以此为基础对帖子数据集中包含的句子进行极性判断，表 2 中列举了几个句子情感判断结果。

表 2 句子极性分类结果举例

句子	主题词	情感词	极性
日本地震如此严重	地震	严重	－1
日本媒体评论称“自卫队害怕核辐射放弃洒水”	核辐射	害怕	－1
支持中国救援	中国	支持	1

4.2.3 情感趋势分析结果

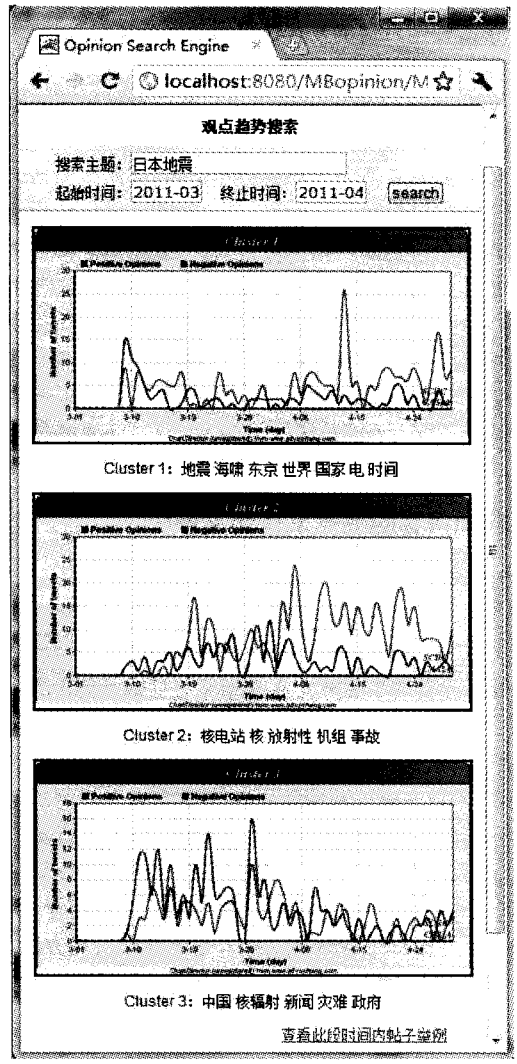

图 2 日本地震事件的情感趋势分析结果，查询时间为 2011-03-01 至 2011-04-30

图 2 给出了日本地震事件中的用户情感趋势的系统显示结果。其中，蓝色曲线代表用户正面情感的变化趋势，红色曲线代表用户负面情感的变化趋势。从图 2 给出的结果可以看出，从事件发生开始，用户对海啸以及地震这一系列突发的自然灾害都给予了很强的关注，不管是正面情感还是负面情感，

（下转第 260 页）

2) http://lucene.apache.org/. Apache 软件基金会 4 jakarta 项目组的一个子项目，是一个开放源代码的全文检索引擎工具包

3) http://www.advsofteng.com/. 统计绘图工具

[3] 袁平波，俞能海，疏晓葵，等. 论坛帖子热度变化模型的研究[D]. 中国科技论文在线，2009
[4] 邱立坤，程葳，龙志祎，等. 面向 BBS 的话题挖掘初探[C]//全国第八届计算语言学联合学术会议(JSCL-2005). 北京：清华大学出版社，2005：401-407
[5] 高俊波，王晓峰，等. 一种新的主题影响力模型研究[J]. 计算机工程与应用，2007，43(25)：182-185
[6] Matsumura N，Ohsawa Y，Ishizuka M. Influence diffusion model in text-based communication[C]//Poster of the Eleventh International World Wide Web Conference. 2002
[7] 任晓东，张永奎，薛晓飞. 基于 K-Modes 聚类的自适应话题追踪技术[J]. 计算机工程，2009，35(9)：222-224
[8] 张虹，钟华，赵兵. 基于数据挖掘的网络论坛话题热度趋势预报[J]. 计算机工程与应用，2007，43(31)：159-161
[9] 张虹，赵兵，钟华. 基于小波多尺度的网络论坛话题热度趋势预测[J]. 计算机技术与发展，2009，19(4)：76-79
[10] 高辉，王沙沙，傅彦. Web 舆情的长期趋势预测方法[J]. 电子科技大学学报，2011，43(3)：440-445
[11] Mitchell T M. Machine Learning[M]. 曾华军，张银奎，等译. 北京：机械工业出版社，2009
[12] 刘里，何中市. 基于关键词语的文本特征选择及权重计算方案[J]. 计算机技术与发展，2009，19(4)：76-79
[13] Liu Bing . WEB DATA MINING[M]. 俞勇，薛贵荣，韩定一，等译. 北京：清华大学出版社，2009
[14] 程辉. 基于时间序列的网络舆情预测模型[J]. 网际网路技术学刊，2008，9(5)：16-17
[15] 大旗网. 口碑观察：上海大众刘坚车祸[DB/OL]. http://shehui. daqi. com/article/2951609. html，2011-08-20
[16] ICTCLAS 官网. ICTCLAS 汉语分词系统[DB/OL]. http://ictcla s. org/ictclas_about. htm，2011-08-20

(上接第 228 页)

都出现了高峰。但是随着时间的推进，两种情感都有所缓和，这主要是由于用户对该关注点整体关注度的降低。然而，随着'此次地震的起因是人祸而非天灾'的说法越来越受到网民的关注，网络上广泛掀起了对此次事故的批判声音。针对该事件的第二个关注点——核辐射相关内容，网民则一直表示出了很强的负面情感，不管是核辐射导致的海水和鱼虾蔬菜的污染，还是其导致的中国购盐热，都产生了很多的负面影响。而对于日本地震中所包含的第三个主题，即关于该次事故对中国本身的影响，以及国内新闻的报道、中国政府的应对措施等等，从图中曲线可以看出，国内民众的相关情绪逐渐得到稳定。

结束语 随着人们越来越多地通过网络平台发布自己对热点事件的观点和看法，针对热点事件主题的用户情感趋势挖掘能够快捷地反映网络用户在事件发生后的情感变化。而用户情感变化往往与事件本身发展的情况相关联，因此监测民众情感趋势能够对事件的发展状态进行实时评估，从而帮助决策部门及时制定相关措施，控制事件发展的不利影响。本文提出了一种面向微博客的热点事件情感分析方法，并实现了微博客背景下的情感趋势分析原型系统。突发事件的预测和异常事件的检测则是下一步工作的研究方向。

参 考 文 献

[1] 贺德方. 我国科技情报行业发展战略与发展路径的思考[J]. 情报学报，2007，26(4)：483-487
[2] 王飞跃. 开源情报与网络时代的国家安全[J]. 科学新闻杂志，2007，2(1)：3
[3] 王飞跃. 基于社会计算和平行系统的动态网民群体研究[J]. 上海理工大学学报，2011，33(1)：8-17
[4] 曾大军，王飞跃，曹志冬. 开源信息在突发事件应急管理中的应用[J]. 科技导报，2008，26(16)：27-33
[5] Savage N. Twitter as medium and message [J]. Commun ACM，2011，54(3)
[6] Jansen B J，Zhang M，Sobel K，et al. Twitter power：Tweets as electronic word of mouth [J]. Journal of the American Society for Information Science and Technology，2009，60(11)：2169-2188
[7] O'Connor B，Balasubramanyan R，Routledge B R，et al. From tweets to polls：Linking text sentiment to public opinion time series [C]//Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Washington，DC，2010：122-129
[8] Bollen J，Pepe A，Mao H. Modeling public mood and emotion：Twitter sentiment and socio-economic phenomena [J]. CoRR，2009，abs：0911-1583
[9] Thelwall M，Buckley K，Paltoglou G. Sentiment in Twitter events [J]. Journal of the American Society for Information Science and Technology，2011，62(2)：406-418
[10] Shen Y，Li S，Zheng L，Ren X，et al. Emotion mining research on micro-blog[C]// Proceedings of 1st IEEE Symposium on Web Society. Lanzhou，China，2009：71-75
[11] Song S，Li Q，Zheng N. A spatio-temporal framework for related topic search in micro-blogging [C]//Proceedings of the 2010 International Conference on Active Media Technology. Toronto，Canada，2010：63-73
[12] Sakaki T，Okazaki M，Matsuo Y. Earthquake Shakes Twitter Users：Real-time Event Detection by Social Sensors [C]//Proceedings of the 19th International World Wide Web Conference. Raleigh，NC(USA)，2010：851-860
[13] Singh V K，Gao M，Jain R. Situation Detection and Control using Spatio-temporal Analysis of Microblogs [C]// Proceedings of the 19th International World Wide Web Conference. Raleigh，NC(USA)，2010：1181-1182
[14] Vieweg S，Hughes A L，Starbird K，et al. Microblogging During Two Natural Hazards Events：What Twitter May Contribute to Situational Awareness [C]// Proceedings of the 28th International Conference on Human Factors in Computing Systems. Atlanta，Georgia，USA，2010：1079-1088
[15] 王素格，杨安娜. 基于混合语言信息的词语搭配倾向判别方法[J]. 中文信息学报，2010，24(3)：69-74
[16] 章成志，梁勇. 基于主题聚类的学科研究热点及其趋势监测方法[J]. 情报学报，2010，29(2)：342-349
[17] 夏天. 中文信息相似度计算理论与方法[M]. 郑州：河南科学技术出版社，2009
[18] Ku L，Liang Y，Chen H. Opinion extraction，summarization and tracking in news and blog Corpora[C]// Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs. Stanford University，California，USA，2006