一种中文微博新闻话题检测的方法

郑斐然 苗夺谦 张志飞 高 灿
(同济大学计算机科学与技术系 上海 201804)
(同济大学嵌入式系统与服务计算教育部重点实验室 上海 201804)

摘 要 微博的迅猛发展带来了另一种社会化的新闻媒体形式。提出一种从微博中挖掘新闻话题的方法,即在线检测微博消息中大量突现的关键字,并将它们进行聚类,从而找到新闻话题。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。在真实的微博数据上运行的实验表明,本方法可以有效地从大量消息中检测出新闻话题。

关键词 微博,新闻,话题检测,聚类
中图法分类号 TP391 **文献标识码** A

News Topic Detection Approach on Chinese Microblog

ZHENG Fei-ran MIAO Duo-qian ZHANG Zhi-fei GAO Can
(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)
(The Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China)

Abstract The popularity of microblogging brings another form of social news media. The paper proposed an approach of news topics mining from microblog. News topics were formed by finding the emerging keywords in large numbers and clustering them. To extract news keywords, a compound weight was introduced combining the word frequency and the growth, to measure the likelihood of a word to be a news keyword, and to construct the topic, contextual relevance model was used to support incremental clustering, which is more suitable to the problem compared with semantic similarity. The experiments on real world microblog data show the effectiveness of the approach to detect news topic out of massive messages.

Keywords Microblog, News, Topic detection, Clustering

1 引言

微博是近年来发展非常快且影响非常大的网络全民媒体形式。自从 2006 年 Twitter.com 在美国上线以来,其注册用户已超过 1.6 亿[1]。而国内与之相似的本地化微博服务也在最近几年大为流行,各大互联网服务商相继推出各自的中文微博平台,如新浪微博(weibo.com)、腾讯微博(t.qq.com)、饭否(fanfou.com)等网站,受到很多网民的欢迎。这些微博平台都有消息集成化和多平台的特点,用户可以通过网页、移动客户端、IM 软件和开放 API 等多种途径,随时随地记录生活见闻、表达个人观点、关注亲友状态,或者了解最新时事等。其中,跟踪和分享新鲜事,是用户使用微博的一个很重要的目的[2]。由于微博的即时性很强,突发新闻在微博上的传播速度很快;而且对于影响力较大的新闻事件,参与报道、转发、评论的用户数量也很大,往往能够先于传统新闻媒体做出反应,这进一步证明了全民媒体在消息传递方面的功能不容忽视[3]。事实上,Twitter 被研究用于辅助各种突发事件的应对,例如重大火警、交通路况、自然灾害等,并且取得了一些进展。

针对微博的实时性,对微博内容进行分析和整合具有重要的实际意义,不仅可以帮助过滤无用信息、提高内容质量、改善用户体验,更能起到事件监测、观点挖掘、舆情控制等重要作用。另一方面,微博是一个信息流量相当大的平台,而内容和格式又非常散乱,数据噪声较大[4]。人工审视或者基本统计方法很难有效地从海量数据中提炼出精确和有用的信息,因此引入文本挖掘的方法对信息进行去重[1]、筛选、聚类、分类[5]非常必要。突发事件检测作为微博文本挖掘的一大方向,在国内外也逐渐受到关注[6]。

本文分析了中文微博系统中的用户习惯和数据特征,提出了一套完整的微博数据处理方法和新闻话题的检测算法。在向量空间模型的基础上,从文档主题词的时域分布中,筛选出信息量最大的新闻主题词,并进行聚类。

到稿日期:2011-02-21 返修日期:2011-06-15 本文受国家自然科学基金项目(60970061,61075056,61103067)资助。
郑斐然(1988-),男,硕士生,主要研究方向为文本挖掘、智能信息处理等,E-mail:famcool@gmail.com;苗夺谦(1964-),男,教授,博士生导师,CCF 高级会员,主要研究方向为粗糙集理论、粒计算、Web 智能、模式识别等;张志飞(1986-),男,博士生,CCF 学生会员,主要研究方向为文本挖掘、机器学习和自然语言处理等;高 灿(1983-),男,博士生,CCF 学生会员,主要研究方向为粗糙集理论、粒计算、数据挖掘和机器学习等。

2 相关工作

话题检测与跟踪(Topic Detection and Tracking,TDT)是文本挖掘的一个方向,旨在帮助人们应对信息过载问题[7,8]。传统的话题检测技术以新闻专线、广播、电视等媒体信息流作为处理对象,将语言形式的信息流,通过文本聚类算法分割为不同的新闻报道。话题检测有一些广为人知的方法,例如CMU、UMass 和 Dragon 等[8]。这些算法各有特点,但是主要的处理对象是新闻媒体的长篇报道,不能有效地适用于微博消息。

针对微博的新闻检测需要引入一些新的观察角度和处理方法。Takeshi 等提出的基于 Twitter 的实时地震监控系统[9],采用了以关键字为证据的贝叶斯决策方法,其在实际应用过程中成功检测到了 80%以上的地震发生,并且在时效性上能够做到比当地的地震告警机构更快。该系统不仅能检测到地震的发生,还可以根据微博数据中包含的地理信息,估计出地震发生的大概位置。Sasa 和 Miles 等提出了一种改进的新话题检测算法[10],能够在不失精度的前提下,快速地处理大于 1.6 亿条 Twitter 消息。Zitao Liu 等人提出了一种更适用于短文本的新的特征选择方法[4],它基于 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。

3 话题检测系统

中文微博数据,实质上是一系列独立的短文本,它们随时不断产生。每条文本的字数不超过 140 个,而且文本中可能含有一些特殊格式,以表示公共主题和用户间的互动关系。例如用"@用户"的格式来表示"提到"某个用户,用"RT"来表示转发,用"#主题#"来表示参与某个特定主题的讨论。同时文本还具有一些附加属性,如发送时间、来源、地理信息、发送者的用户信息等。微博数据增长的速度非常快,每个小时可达数万条之多,涉及的话题也非常分散。突发事件检测的目标是在这些话题当中筛选出与现实中发生的新闻事件有关的话题。

话题检测系统主要分为数据获取、预处理、分词和词频统计、主题词检测,以及话题聚类这 5 个步骤。本节主要描述预处理策略、主题词检测和主题词聚类这 3 个步骤。

3.1 预处理

在获取到的微博文本被分词之前对其进行预处理,可以提高后续检测环节的计算速度和准确性。预处理的主要效果是尽可能消除噪声数据,屏蔽无关数据,过滤掉文本中的无用信息;预处理的方法主要是机械化的规则。具体的处理规则如下。

(1) 忽略收听人数小于阈值 F 的用户的消息。收听人数(或称"粉丝"数量)接近于 0 的用户 ID 很可能不是正常的微博用户,而是一些广告账户或者僵尸账户,其所发的微博条数少,且有用信息少而噪声大,会对聚类算法产生干扰。

(2) 忽略带有"@用户"格式的消息。包含"@用户"形式的消息多数是具有指向性的话题,即类似于对话式的互动。而我们的检测目标是新闻事件,应属于一般性的话题,用户的报道或观点很少会指向另一个特定用户,所以带有@格式的消息直接描述新闻的可能性很小。因此,这里对指向性消息进行过滤,以增加主题词检测的精度。

(3) 删除消息中以"#话题名#"为格式的部分。该格式是一个微博话题标记,在中文微博中往往是由微博平台给定的一些主题词,人为因素很大,且由于多次出现,会强烈影响基于词频的统计算法。

3.2 分词和词频统计

预处理以后,对文本进行分词。中文分词有多种不同的算法和工具。本文采用了 ICTCLAS[11] 分词系统,它是中文文本处理中经常使用的一个工具。其分词效果较好且支持人名识别、地名识别、组织机构名识别等特殊词类。最重要的是它在分词的同时支持词性标注,所标出的词性用于辅助词频统计。

对一条微博消息文本进行分词后,得到一个词向量,其中每个词都带有词性标记,如名词、动词、形容词、方位词等类型。不同词性的词对主题表达的贡献程度不同,其中对主题表达和辨识作用最大的是动词和名词,所以在词频统计中我们只考虑这两种词性,其它词性的词忽略。统计时,先将消息按产生时间划入不同的单位窗格,如按照 1 小时进行划分;然后对同一窗格中的词频进行统计,得到一个该时间段内的总的词语列表。在第 4 节实验中可以看到,该列表具有长尾特征,即绝大多数的词只出现了很少的次数,只有少数词语的出现频率较高。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检测,而把长尾部分去掉。

3.3 主题词检测

新闻主题词的检测有别于静态的文档主题词选择,具有很强的时域特征,所以无法直接应用 TF-IDF 来计算主题词权值。而且,希望检测算法是一种增量式算法,能够以顺序输入词频数据并且不断产生最近的主题词,这样就可能实现在线检测话题的目的。这里引入增长系数 G_{ij} 来表示词 i 在某一窗格 j 的词频增长速度,定义为当前窗格中该词的频率除以之前 K 个窗格中的频率平均值:

$$G_{ij}=\frac{F_{ij}}{\overline{F}_i}=\frac{F_{ij}\cdot K}{\sum_{u}^{K}F_{iu}} \tag{1}$$

式中,F_{iu} 表示词汇 i 在 u 时间窗的出现频率,K 是回顾窗的大小,即回顾窗由 v 时间窗之前 K 个时间窗组成,并在该范围内计算回顾的频率平均值。G_{ij} 的值越大,说明该词越有可能是突然出现的热议词汇,即有可能是新闻的主题词。

为了合理地选出表示新闻主题的主题词,同时考虑词频和增长速度,构造一个复合的权值来评价一个词是主题词的程度:

$$w_{ij}=\log G_{ij}+\alpha\log\frac{F_{ij}}{F_{\max}} \tag{2}$$

式中,$F_{\max}$ 表示在时间窗内的最高词频。若 w 的值越大,则该词越有可能是新闻主题词。α 参数可以调节词频和增长速度的比重关系:当 α 较大时词频起主要作用,α 较小时则优先考虑增长快的词。因此,α 的实验值在 1.0 到 1.5 之间最好,太大或太小都难以得到准确的结果。

对每个时间段内的词计算其 w 值,按照阈值 T 选出其中较大的词汇,即得到一个主题词表。这个主题词表的特点是,其中的词语在本时段出现次数较多且在之前时段出现次数较少。这些选出的词将会被聚类产生出各个新闻话题。

3.4 主题词聚类

按照 w 的大小对主题词表进行降序排列,然后对排序后的词增量聚类。

增量聚类的算法。

输入：带有权值的主题词列表。

输出：簇列表。

步骤 1　以第一个词作为初始簇；

步骤 2　输入下一个词，判断它与每个已有簇的距离；

步骤 3　如果离它最近的簇的距离大于阈值 D，那么把这个词作为一个新簇；否则，把这个词放入该簇；

步骤 4　继续输入下一个词，重复步骤 2 到步骤 4，直到所有的词都处理完毕；

步骤 5　输出结果。

为了判断步骤 2 中的词是否属于某个簇，要有一种适当的相似度定义。词语相似度的确定方法有两种：一种是查相似度表的方式，也就是说任意两个词的相似度都是确定不变的，不随着运算而更新。这也是大多数文本聚类和分类算法的思路，而且词语的相似度取值往往符合词义近似的规律[12]，即意义相近的词相似度也大，意义无关的词则认为不相似。事实上这种判断依据不适应话题聚类的需要。例如出现“今日水果全面涨价，其中苹果涨幅最大”这样的话题时，按照预定相似度的概念，“水果”和“苹果”相似，而“水果”和“涨价”则不相似，得到的聚类结果很可能是{“水果”，“苹果”}和{“涨价”}，而不是{“水果”，“涨价”，“苹果”}这样的理想话题。所以另一种相似度定义方法——上下文式的相似度更符合这里的需要。根据两个词语同时出现在一条微博消息里作为词语相似的依据，然后利用统计的方法来量化得出相似度。这里引入两个词(a,b)的条件概率：

$$P(a|b)=\frac{F_{a,b}}{F_b} \tag{3}$$

即 a，b 同时出现的消息条数除以 b 出现的消息条数。基于条件概率的相似度解决了含义相似与同属于话题的相似的矛盾。为了定义词与簇（一个词的集合）的相似度，把条件概率的最大值 $\max\{P(c_i|w)|c_i\in C\}$ 当作词 w 到簇 C 的距离的倒数，因此词 w 到簇 C 的距离定义为：

$$d_{w,C}=\begin{cases}\dfrac{1}{\max\{P(c_i|w)|c_i\in C\}},(\max\{P(c_i|w)|c_i\in C\}>0)\\ \infty,(\max\{P(c_i|w)|c_i\in C\}=0)\end{cases} \tag{4}$$

也就是说，当簇 C 里存在词 c_j，该词在含有 w 的消息里出现的概率很高，那么 w 到 C 的距离较近，应把 w 归入 C 簇；反之则 w 距 C 较远，即 w 不归入 C 簇。

完成主题词的增量聚类以后，就得到若干个簇。每一个簇都是含有一个或多个主题词的新闻话题，例如{“水果”，“涨价”，“苹果”}这样的簇，就表示一个描述“水果或者苹果涨价”的话题。

4　实验

从新浪微博（http://open.weibo.com）所提供的开放平台采集到 2011 年 4 月 24 日到 5 月 5 日之间的约 300 万条原始微博数据，人工标注了该时间段内热议的主要新闻话题，有“清华校庆”，“本拉登被击毙”和“欧冠决赛”等 8 个事件。

4.1　检测结果说明

为了评估参数对检测结果的影响，把数据平均分为 10 组进行话题检测，然后求出平均值。其中设定 α 为 1.0，分别比较时间窗长度和阈值 T 对结果的影响。例如，图 1 所示的是在时间窗长度为 1 小时、阈值 T 为 25 的结果中，一个时间窗内的词频分布，可以明显地看到其长尾的特点；其输出的话题片段如表 1 中前两列所列，并且在表中第 3 列注明了当时主要的新闻事件。分隔符“|”表示聚类得到的簇所对应的主题词的划分。从表中可以明显看出话题主题词与实际发生的新闻事件的对应关系。

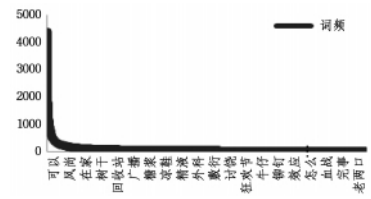

图 1　词频的长尾特性

表 1　检测结果片段

消息时间	微博话题	对应新闻事件和发生时间
2011/4/21　3:30	上半场|佩佩头球|西班牙语|黄牌|比利亚|射门|门柱	2011/4/21　凌晨 3 点，西班牙国王杯决赛：巴塞罗那对阵皇家马德里
2011/4/21　23:29	晚安|月亮|死刑　执行　民意　量刑	2011/4/21 药家鑫案将于次日宣判，网友呼声高涨
2011/4/22　12:40	判处　剥夺　杀人罪	2011/4/22　上午宣判，药家鑫被判处死刑
2011/4/22　13:29	打雷|春雷|雷声|雷鸣	2011/4/22　北京地区午后突然打雷
2011/5/2　11:29	本拉登|拉登|巴马　击毙　白宫　军方|本·拉登|巴基斯坦	2011/5/2　当日下午，美国宣布本拉登被击毙

4.2　时间窗对结果的影响

使用不同的时间窗长度，比较检测结果，如图 2 所示。时间窗较小时，容易受到噪声数据干扰，查准率和查全率都较低；时间窗较大时，选出的主题词较精确，但是由于粒度较大，有的话题被漏掉。例如在以 1 小时为窗的时候，像“晚安”、“午饭”之类在一天中的特定时间频繁出现，而其它时间出现较少的词语，就被误判为新闻话题。在窗格过宽的时候，如果新闻被关注的时间较短，那么其主题词与该窗格内其他词语的数量就不会相差太多，w 值就不够大。

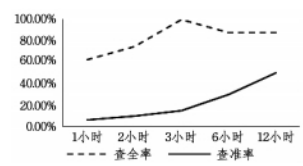

图 2　时间窗长度对查全率、查准率的影响

4.3　阈值 T 对结果的影响

以 3 小时为时间窗，用不同的阈值 T 进行检测，得到的结果如图 3 所示。在 T 值为 5 的时候查全率很高，但也会得到大量的无关结果，导致噪声很大。随着 T 值的增大，查准率变高，查全率则逐渐降低。可见，阈值更大时对主题词的判断较为谨慎，虽然减少了噪声，但增大了遗漏新闻话题的风险。

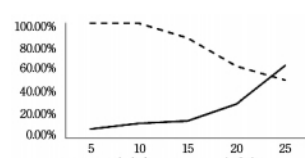

图 3　阈值 T 对查全率、查准率的影响

结束语　本文描述了一种中文微博的新闻话题的检测方

法，包括预处理、中文分词、主题词检测、话题聚类等步骤。在主题词检测中，构造了权值 w 来综合考虑词频和增长速度，以决定其是否为新闻主题词；根据上下文相关性定义相似度，以增量聚类来建立新闻话题。最后，整个过程用程序实现并在真实的新浪微博数据集上进行了实验，验证了方法的有效性。

同时，实验所得结果的查准率和查全率有相当大的改进空间，而且检测效果容易随计算参数值的选取而浮动。因此进一步提高算法速度和精度，增强稳定性或者引入自适应特性，是后续研究工作的重点方向。

参 考 文 献

[1] 曹鹏，李静远，满彤，等. Twitter 中近似重复消息的判定方法研究[J]. 中文信息学报，2010，25(1)：20

[2] Kwak H，Lee C，Park H，et al. What is Twitter，a Social Network or a News Media？[A]// WWW'10 Proceedings of the 19th International Conference on World Wide Web，2010[C]. Raleigh，North Carolina，USA：ACM，2010：591-600

[3] 蔡晓婷. 突发性事件中的微博客传播[J]. 新闻爱好者(上半月)，2010，(6)：78-79

[4] Liu Zi-tao，Yu Wen-chao，Chen Wei，et al. Short Text Feature Selection for Micro-blog Mining[A]// Computational Intelligence and Software Engineering，2010[C]. Wuhan，China：Wuhan University，2010：1-4

[5] 崔争艳. 基于语义的微博短信息分类[J]. 现代计算机(专业版)，2010，(8)：18

[6] Pak A，Paroubek Pa. Twitter as a Corpus for Sentiment Analysis and Opinion Mining[A]// Proceedings of LREC，2010[C]. Valletta，Malta：European Language Resources Association (ELRA). 2010：1320-1326

[7] 洪宇，张宇，刘挺. 话题检测与跟踪的评测及研究综述[J]. 中文信息学报，2007，21(6)：71-85

[8] Allan J，Carbonell JG，et al. Topic Detection and Tracking Pilot Study Final Report[A]// Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop，1998[C]. 1998：194-218

[9] Sakaki Ti，Okazaki M，Matsuo Y. Earthquake Shakes Twitter User：Real-time Event Detection by Social Sensors[A]// Proceedings of the 19th International Conference on World Wide Web，2010[C]. Raleigh，North Carolina：ACM Press，2010：851-861

[10] Petrovi S，Osborne M，Lavrenko V. Streaming First Story Detection with application to Twitter[A]// Proceedings of HLT-NAACL，2010[C]. Stroudsburg，PA，USA：Association for Computational Linguistics. 2010：181-189

[11] Zhang H P，Yu H K，Xiong D Y，et al. HHMM-based Chinese lexical analyzer ICTCLAS[A]// Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17，2003[C]. Sapporo，Japan：Association for Computational Linguistics，2003：184-187

[12] 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研究[A]// 第六届全国信息检索学术会议，2010[C]. 2010：291-298

(上接第 133 页)

检查每个非功能性特征的作用域是否是功能性特征集合的一个子集。

通过多次对以上 8 个活动的反复迭代，最终得到一个一致的、完整的需求模型。

结束语　本文将特征分为功能性特征和非功能性特征，并分别对其进行详细的讨论和分析，以支持通过特征组合的方式建模软件需求；并在此基础上，提出基于特征组合的软件需求建模过程。本文提出的方法可以通过特征运算的方式来建模功能性特征的组合，从而建模软件的功能性需求，并通过定义非功能性特征的作用域的方式把功能性特征和非功能性特征统一成整体。下一步将研究特征的形式语义，并在此基础上证明本文提出的 23 条公理的可靠性，建立关于特征运算的公理系统。

参 考 文 献

[1] Pressman R S. Software Engineering：a Practitioner's Approach (Fifth Edition)[M]. McGraw Hill，New York，2000

[2] 金芝，何克清，王青. 软件需求工程：部分研究工作进展[J]. 中国计算机学会通讯，2007，3(11)：25-34

[3] 何克清，彭蓉，刘玮，等. 网络式软件[M]. 北京：科学出版社，2008

[4] 陈小红，尹斌，金芝. 基于问题框架的需求建模：一种本体制导的方法[J]. 软件学报，2011，22(2)：177-194

[5] Kang KC，Cohen S G，Hess J A，et al. Feature-Oriented domain analysis (FODA) feasibility study[R]. CMU/SEI-90-TR-21. Pittsburgh：Software Engineering Institute，Carnegie Mellon University，1990

[6] Zhang Wei，Mei Hong，Zhao Hai-yan，et al. Transformation from CIM to PIM：A Feature-Oriented Component-Based Approach[C]// Proceedings of the 8th International Conference on Model Driven Engineering Languages and Systems. Heidelberg：Springer Berlin，2005：248-263

[7] Zhang Wei，Mei Hong，Zhao Hai-yan. A Feature-Oriented Approach to Modeling Requirements Dependencies[C]// 13th IEEE International Conference on Requirements Engineering. 2005：273-284

[8] Schobbens P Y，Heymans P，Trigaux J C，et al. Feature Diagrams：A Survey and a Formal Semantics[C]// Proceedings of the 14th IEEE International Conference on Requirements Engineering(RE' 06). 2006：136-145

[9] 王忠杰，徐晓飞，战德臣. 基于特征的构件模型及其规范化设计过程[J]. 软件学报，2006，17(1)：39-47

[10] 张俊，刘淑芬，姚志林. 一种基于角色的特征模型构件化方法[J]. 电子学报，2011，39(2)：304-308

[11] 吕建，马晓星，陶先平，等. 网构软件的研究与进展[J]. 中国科学 E 卷，2006，36(10)：1037-1080

[12] 李长云，李莹，吴健，等. 一个面向服务的支持动态演化的软件模型[J]. 计算机学报，2006，29(7)：1020-1028

[13] 王璞巍，金芝，刘红岩. 网构软件实体的功能描述及其发现[J]. 中国科学 F 辑：信息科学，2009，39(12)：1271-1287

[14] 吴映波，王旭. 一种面向服务的领域特征模型[J]. 计算机科学，2011，38(6)：180-182，194

[15] 张伟，梅宏. 面向特征的领域建模技术研究[J]. 中国计算机学会通讯，2008，4(3)：34-42