基于微博文本的舆情分析和研究

Analysis and Study on Public Opinion based on Microtext

曾星宇* 李淑琴** 陈斌*
ZENG Xing-yu LI SHu-qin CHEN Bin

微博是一个基于用户关系的信息分享、传播以及获取的平台，它的影响力引起了多方政府部门的高度重视。做好对微博舆情监测，是做好网络舆情监测工作的一项重点。本文就微博获取，文本处理，文本聚类做了一定的研究，实现了一种基于统计规则的舆情发现的方法，实验表明该方法是有效的。

微博舆情 关键字 文本聚类

Abstract Microblog is platform to assist user in sharing and gaining information, whose influence has been attached great attention by government department gradually.So to do well in monitoring public opinion in microblogging is the key to perfecting management of internet public opinion. This article did some research on microblogging access, text processing and text clustering which finds a solution to exposing public opinion based on Statistical rules. The experiment result shows that this solution is workable.

Keywords Microblog public opinion Keyword Text clustering

doi: 10. 3969/j. issn. 1672 - 9528. 2014. 01 .23

1 现状分析

微博（Microblog）作为新型的全网互动平台，深受广大网民的宠爱。微博已经成为新的舆论产生和民意反应的区域，它的影响力引起了多方政府部门的高度重视。然而，作为如今最流行的网络社交工具之一，微博带来便捷的同时，也正在成为虚假信息滋生和泛滥的温床。在商业微博利益的驱使下，网络水军的涌入、虚假信息的传播和不合法的网络传销，特别是一些失真的违背民意方面的舆论一定程度的引发社会恐慌和网民愤怒，严重威胁着社会的稳定和安全，网络媒体正日益成为危机舆情的首发媒体[1]。因此，做好对微博舆情监测是做好网络舆情监测工作的一项重点。

如何对网络微博进行监测，是摆在国家政府及网络监管部门面前的一个大课题，由于微博传递的爆炸式特性，仅仅靠人工进行搜集和浏览，浪费过多的人力、物力不说，要想及时、准确、全面把握微博舆论的敏感内容和传播趋势也是难以实现的。因此，通过计算机技术找到有效的方法，及时、准确、全面把握微博舆论的敏感内容和传播趋势是十分值得研究。

2 本文设计思路与方法

本文主要基于微博文本对舆情进行分析和研究。在思路上借鉴了文献[2]和文献[3]的思想，还借鉴了社会学、新闻传播学、统计学、管理学等方面的研究视角和研究方法，并以互联网和计算机技术为支撑，把理论和实际相结合，结合应用，环环相扣，以求得到最佳效果，能反映特定时段的网络舆情。本文设计实现的系统主要包括微博信息的采集，微博文本的预处理，微博舆情关键词的分析与提取，微博舆情内容聚类，微博舆情的判别显示五个模块，相应的解决方案与步骤如表1所示。

表1 微博舆情分析流程与相应解决方案

步骤	微博信息分析流程	相关解决方案与步骤
①	微博信息的采集	基于API进行新浪微博信息数据的采集

* 北京信息科技大学计算机学院 北京 100085
** 网络文化与数字传播北京市重点实验室 100085

②	微博信息的预处理	基于统计和规则方法进行数据除噪等
③	微博舆情关键词的分析与提取	基于分词的统计和加权计算，来提取微博中与舆情相关的关键词
④	微博舆情类容聚类	运用统计方法，基于内容进行关键词聚类，根据类中多个关键词大致了解所发生的舆情事件.
⑤	微博平台软件的设计	利用图形界面，显示实验结果

3 相关技术的实现与研究

3.1 微博信息的采集

常用的获取微博方式有三种，基于 API 的微博抓取，基于页面分析的微博抓取，向公司索要数据。这三种方法各有好处也有不足。

基于 API 的微博抓取，这种抓取的数据的方式操作简单，只需要向指定的 API 接口发送请求，就可以得到相应的 JSON 数据，再通过处理数据进行下一步工作。但这种方法有一定的次数限制，获取数据较少。

基于页面分析的微博抓取，这种抓取数据的方式比较灵活，通过处理网页信息得到相应微博数据，但无法大量的爬取微博数据。

向公司要数据，这种获得数据的方式省时省力，但相对应需要支付一定的费用，并且数据也不够灵活。

本文采用基于 API 的微博抓取，本文同时用多台机器，不同账号向新浪微博提供的 API 接口 urlhttps://api.weibo.com/2/statuses/public_timeline.json?access_token 发送 web 请求。

3.2 微博信息的预处理

对采集的微博信息进行预处理。由于微博是一个公共平台，任何人都可以在这个平台上发布信息，因此会有许多与舆情无关的数据如广告等，这些数据本文应该去除掉，以减少系统对舆情判断的影响。本文采用基于统计的信息预处理方法，即忽略文本的语言学上的特征，将文本作为特征项集合来看，利用加权特征项构成向量进行文本表示，利用词频信息对文本特征进行加权。根据字符串匹配，给出了一个微博文本信息的过滤模型，其核心语料库包括词库和规则库，通过词库和规则库算出每个微博对应的权值，当这个权值大于先前定义的一个阀值，本文就可以认为这条微博是一个无用的微博，可以舍去。

3.3 微博舆情关键词的分析与提取

3.3.1 关键词的提取方法

通常舆情事件检测方法是将一段连续时间发布的微博划分到一些独立的时间段内，并对某个时间段内的微博进行舆情事件检测。本文将每个独立的时间段简称为“时间窗”（t）。在某个时间窗内，若一个实词被大量地使用，且在之前的时间窗内很少被使用，则该实词被视为一个关键词 (key word)。判断一个词是否为关键词需设立一个衡量标准，为此本文为每个词计算一个值，称为关键度（keydegree）。关键度越高代表该词成为关键词的可能性越大。

词的关键度主要与三个因素有关。一是，词本身在某个时间窗内所有微博中所出现的频度，一般频度高的词成为关键词的概率较高。二是，该词所在微博的影响力，由于微博是全网性互动平台，常常伴随着表态和回复，表态和回复数越多表明微博的影响力越大，该词成为关键词的概率较高。三是，该词所在微博的作者对该词的影响力。在微博中，通常不同用户所发的微博的影响力（U(inf)）不同，通常名人的关注度高，发出微博内容的影响力通常应比普通人高。这样，一个用户的影响力越大，则其发布的微博中所描述的事件更有可能成为微博的舆情事件，其中出现的词语更适合用以描述舆情事件。

本文提取关键字的基本步骤为：

（1）对某个时间段中的所有实词计算所出现的频度，在此称为基础权重（$w(\text{base})$。计算公式如下公式 1 所示。

$$w(\text{base})_{j,i}=\frac{w_{j,i}}{Max(w_i)} \quad (1)$$

其中，$w(\text{base})_{j,i}$ 是微博 i 中实词 j 的基础权重，$w_{j,i}$ 是时间窗 t 中是微博 i 中实词 j 的出现频次，$Max(w_i)$是时间窗 t 中微博 i 中实词的最高词频。

（2）计算加权权重（$W(weight)$）。在计算出词的基础权重（$w(\text{base})$基础上，考虑该词所在微博的影响力，在此主要考虑对此微博的表态和回复数。计算公式如下公式 2 所示。

$$W(\text{weight})_j=\sum_{p_i\in P_j}(w(\text{base})_{j,i}\times F(attitude_{p_i},reply_{p_i},reposting_{p_i})) \quad (2)$$

其中，$w(\text{base})_{j,i}$是词语 j 在微博 i 中的基础权重，P_j^t 是时间窗 t 内包含词语 j 的所有微博，$attitude_{p_i}$是微博 i 的表态数（用户点击“赞”的次数），$reply_{p_i}$是微博 i 的回复数，$reposting_{p_i}$是该微博的转发数，$F\left(attitude_{p_i}, reply_{p_i}, reposting_{p_i}\right)$表示$attitude_{p_i}$，$reply_{p_i}$，$reposting_{p_i}$三者的函数关系，共同决定了加权值。

（3）计算用户对该词的影响力，用U(inf)_u表示。用户的影响力主要考虑用户的粉丝数（$Fans_u$）、微博数（Twu）、是否VIP用户（Vip_u）及用户的活跃度（Act_u），用 γ 表示 VIP 用户的影响因子，$0<\gamma<1$ 。计算公式如下公式 3 所示。

$$\text{U(inf)}_u = Fans_u \times Total_u \times (1+\gamma \times Vip_u) \times Act_u \quad (3)$$

（4）提取关键词。在计算出每个词的加权权重后，这时候需要算出衡量这个词是不是关键词的关键度（KD），由于将时间分成多个时间片，通常需要比较多个时间片的词的变化才能得到较合理的结果，在多个时间片综合处理即可得到相应的关键度。在统计的过程中，要加上用户的性质即用户影响力对关键度的影响。计算公式如下公式 4 所示。

$$KD_j^t = \frac{\sum_{i=t-N}^{t-1} \left(W(\text{weight})_j^t \times \sum_{p_n \in P_j^t} F\left(\text{U(inf)}_{p_n}\right) - W(\text{weight})_j^t \times \sum_{p_b \in P_j^t} F\left(\text{U(inf)}_{p_b}\right)\right)}{N} \quad (4)$$

其中 KD_j^t 是时间窗 t 内的词语 j 的关键度，N 是存的时间窗的个数，$W(\text{weight})_j^t$ 是时间窗 t 内词语 j 的加权权重，P_j^t 是时间窗 t 内包含词语 j 的所有微博，P_i^t 是时间窗 i 内包含词语 j 的所有微博。$F\left(\text{U(inf)}_{p_n}\right)$是 P_n 这条微博用户的用户影响力。根据 F(X) 函数多的，F(X) 可以为对数函数等函数，不同的函数对结果也不一样。

这样本文可以得到该时间片关键词的集，基于统计规则本文可以选取符合标准的词作为关键词。

3.4 微博舆情内容聚类

目前用于主题聚类方法主要有集成聚类、半监督聚类、样本加权聚类等。

集成聚类是综合考虑各种聚类器的聚类结果，综合投票得到最终聚类结果。通过集成聚类，可以提高聚类的精度。

半监督聚类是在已知聚类对象有一定的约束关系时，进行约束条件下的聚类。常规的聚类算法忽略了这些约束关系，因此聚类的结果有不合理的情况出现。通过半监督聚类，可以提高聚类结果的合理性[2]。

样本加权聚类是一种高质量的聚类的方法，由于不同的样本或对象对最后聚类结果的影响是不同的，所以往往实际情况中需要加上样本或对象的权重。由于这种思想符合在微博文本聚类的特征，例如，提取出的关键词的关键度是不同的，往往在聚类的时候需要考虑到关键度对结果的影响，本文采用了样本加权的层次聚类，同时参考了文献 [4] 的思想。

在关键词聚类时，本文先将每个关键词看做一个类，接下来算出两个类的距离（Dist），在算距离的过程中，需要算出两个类的相似性（Similarity），相似性表现了两个类在同一条微博同时出现的概率，相似性的计算公式如公式 5 所示。

$$\text{Simi}(\text{K}w_k, \text{K}w_l) = \frac{\text{n I } (\text{V}(\text{K}w_k), \text{V}(\text{K}w_l))}{All} \quad (5)$$

公式 5 公式表示时间窗 t 内的关键词的相似度的计算，Kw_k,$\text{K}w_l$是两个关键词，ALL 是时间窗 t 内的所有微博集，$\text{n I } (\text{V}(\text{K}w_k), \text{V}(\text{K}w_l))$ 是时间窗 t 内同时包含关键词 Kw_k,$\text{K}w_l$ 的微博。$\text{Sim}(\text{K}w_k, \text{K}w_l)$表示两个的词性的相似度，当 $\text{Sim}(\text{K}w_k, \text{K}w_l) > 0$ 时，说明两个类之间有联系，可以算出他们的类间距离，当类间距离小于某一个阀值后，就可以把这两个类合并成一个类，当$\text{Sim}(\text{K}w_k, \text{K}w_l)=0$ 时，说明两个类没有出现在相同的微博中，两个类之间的距离无限远。

在得相似性之后，可以继续计算出任意两个类（Ca，Cb）的类间距离，类间距离和这两个类中的关键词的关键度有关，通常和每个类之间的突发词的乘积和成正比关系，具体计算公式 6 所示，其中$\text{Dist}(c_a, c_b)$表示类 Ca，Cb 之间的类间距离。

$$\text{Dist}(c_a, c_b) = \begin{cases} \dfrac{|c_a \times c_b|}{\sum_{w_k \in c_a, w_l \in c_b} \text{Sim}(\text{K}w_k, \text{K}w_l)} \\ \infty \\ \sum_{w_k \in c_a, w_l \in c_b} \text{Sim}(\text{K}w_k, \text{K}w_l) > 0 \\ else \end{cases} \quad (6)$$

当计算出任意两个类的距离后，本文就可以将距离最近的两个类合并，最终得到聚完类的类，也就是本文的最终结果，根据多个关键词本文就可以大致了

解到所发生的舆情事件。

4 微博舆情监控系统平台设计与实现

本平台是基于C# 语言开发，C# 是一种安全的、稳定的、简单的、优雅的，由C 和C++ 衍生出来的面向对象的编程语言。在平台里本文实现了基于本地数据和服务器数据的处理，由于实验环境的限制，只做出对本地数据的分析。本系统可以导入导出处理数据，可以以时间或数量作为独立的分析单元，同时还实现了分词功能，以及配套的词典管理工具和统计功能。平台的界面如图1所示。界面中，左侧为不同时间段聚集到的热点关键词集合，右侧显示每一个热点的相关微博。整个平台功能基本完善，运行情况良好。

以下是对三个不同时间段的微博数据进行分析得出的结果。实验获得了“天安门，吉普车，金水，桥”、“十一，长假，国庆”和“十八，三中全会，改革，学习”这三个热点，如图1、图2、图3所示。“天安门，吉普车，金水，桥”对应最近在北京天安金水桥发生的吉普车撞击事件；“十一，长假，国庆”对应着今年国庆节七天假期；“十八，三中全会，改革，学习”对应着中国共产党的十八届三中全会。

图1 “天安门，吉普车，金水，桥”舆情事件

图2 “十一，长假，国庆”舆情事件

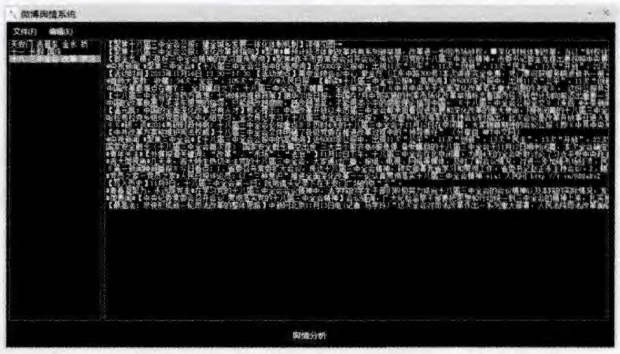

图3 “十八，三中全会，改革，学习”舆情事件

5 小结

本文通过对微博中的词计算权重来对不同时间段的各种微博信息进行了发掘和处理，对微博舆情实现了监测与分析，并从中提取出了相关的舆情事件，但是对关键词的提取还需要进一步研究。因为，微博空间内存在一类词语，他们的词频增量较高，但却不具有明显的实际意义。例如，微博空间内存在大量抒发用户个人情感的信息，这些信息用词相对集中，且可能在一定时间段内大量被发布，从而导致该类信息中出现的词语在该时间段内的词频增量相对较高，然而，这类词语与舆情事件并无显著关系，将其作为关键词用于事件检测，往往会对检测结果造成一定的干扰，如何露出这些干扰将作为本文进一步的研究内容。

参考文献：

[1] 彭作文．微博舆情监测已成紧要需求．中国经济时报，2012.06

[2] 许鑫，章成志．互联网舆情分析及应用研究〔J〕．情报科学，2008(08).1194-1198

[3] 钱爱兵．基于主题的网络舆情分析模型及其实现 [J]. 现代图书情报技术，2008（4）：50-52

[4] 王继成，潘金贵，张福炎．Web 文本挖掘技术研究 [J]. 计算机研究与发展，2000,37(5):513-518

（收稿日期：2013-11-19）