doi: 10.3969/j.issn.1001-893x.2015.06.005
引用格式: 彭浩,周杰,周豪,等. 微博网络中基于主题发现的舆情分析[J]. 电讯技术,2015,55(6):611-617. [PENG Hao, ZHOU Jie, ZHOU Hao, et al. Public Opinion Analysis Based on Topic Detection in Micro-blog Network[J]. Telecommunication Engineering, 2015, 55(6): 611-617.]

微博网络中基于主题发现的舆情分析*

彭 浩,周 杰,周 豪,赵丹丹**

(浙江师范大学 计算机科学与工程系,浙江 金华 321004)

摘 要: 针对现有微博网络舆情分析的研究中没有从全局层面考虑舆情文本特征的情况,结合微博网络舆情的主题及趋向性分析,提出了基于主题发现的微博网络舆情分析模型,从文本预处理、微博文本特征提取、微博舆情的主题发现及趋向性分析三方面进行了具体描述。仿真结果表明,基于该模型实现的微博网络舆情分析方法在微博网络舆情的分析处理中检测效果良好,说明该模型有效。相关内容可为该领域的进一步研究提供有价值的参考。

关键词: 微博网络; 舆情分析; 主题发现; 文本特征
中图分类号: TN915.9 **文献标志码:** A **文章编号:** 1001-893X(2015)06-0611-07

Public Opinion Analysis Based on Topic Detection in Micro-blog Network

PENG Hao, ZHOU Jie, ZHOU Hao, ZHAO Dandan

(Department of Computer Science and Engineering, Zhejiang Normal University, Jinhua 321004, China)

Abstract: For the problem that current research on public opinion analysis in micro-blog network does not consider public opinion text features from the global level, a public opinion analysis model based on topic detection is presented according to micro-blog network public opinion topic and trend analysis. The model is described from three aspects, text preprocessing, micro-blog text feature extraction, topic detection and trend analysis. The simulation results show that the detection performance of the model is satisfying and the model is effective in the analysis of micro-blog network public opinion. This exploration in this paper can provide some valuable reference for further research in this field.

Key words: micro-blog network; public opinion analysis; topic detection; text feature

1 引 言

微博网络作为社交网络的一种重要方式,以其简短、便捷的特点呈现爆发式增长势态,截止到2014 年 7 月微博用户已突破 2.75 亿。由于微博网络的信息能够即时分享,使信息传播时间趋向于零,已成为热点舆情产生、传播的重要源地,微博的影响力也呈现几何式倍增态势,并以惊人的速度渗透到社会和行业的各个方面,在极大地满足人们发布和获取信息便利的同时,也给用户带来很好的时空便利。同时我们看到,社交网络上存在各种各样对各

* 收稿日期: 2014-12-17; 修回日期: 2015-04-16 Received date: 2014-12-17; Revised date: 2015-04-16
基金项目: 国家自然科学基金资助项目(61170108); 浙江省自然科学基金资助项目(LQ13F020007); 上海市信息安全综合管理技术研究重点实验室开放课题(AGK2013003)
Foundation Item: The National Natural Science Foundation of China (No. 61170108); The Natural Science Foundation of Zhejiang Province (LQ13F020007); The Open Research Project of Shanghai Information Security Key Laboratory of Integrated Management of Technology (AGK2013003)
** 通讯作者: ddzhao@zjnu.edu.cn Corresponding author: ddzhao@zjnu.edu.cn

种社会事件和各行各业评价的舆论信息，这些舆论评价信息既包含正面评价信息，也包含负面评价信息，这些舆论信息的传播有可能对社会和一些行业产生重要的影响。面对微博网络数据的不断增多，如何设计相应的舆情分析模型，使其能快速有效地收集和分析这些数据，并产生有用的舆情分析报告，是许多学者关注的焦点。

目前，国内外学者在舆情分析方面做了许多有意义和相关的工作。李岩等[1]基于短文本聚类及用户评论情感分析，解决了微博文本呈现的不完整性、稀疏性及碎片化等问题，在一定程度上解决了因关键词稀疏带来的相似度漂移问题。唐晓波等[2]将共词网络分析和复杂网络的思想与方法拓展到微博舆情分析中，设计了基于网络可视化的微博舆情分析模型，为基于微博的网络舆情分析提供了有效的可视化途径。Yu 等[3]基于舆情信息扩散过程中的用户交互过程，考虑到用户的交互历史、相互作用的类型和频率，提出了一种有向树模型，该模型可以描述信息的扩散，更精确地表达舆情传播的影响，在数据集中识别垃圾邮件更有效。曾振东等[4]基于现代统计学理论，专门针对小样本、不确定性预测问题，提出了一种基于灰色支持向量机的网络舆情预测模型，相对于传统预测模型，该模型提高了网络舆情的预测精度。然而，上述研究工作多集中在单一技术应用层面，缺乏系统性、全局性的微博网络舆情分析方法的研究和分析；同时，上述舆情分析模型中，结合社交网络舆情特点的分析模型较少，不具有一般适用性。

基于上述分析，本文提出了微博网络中一种基于主题发现的舆情分析模型，对微博网络中的热点话题，采用热点分析和趋向性分析两种思路，能够进一步了解用户对社会或行业的热点话题的观点与态度，识别出其情感倾向及演化规律，更好地理解用户的行为，分析热点的舆情主题，从而为政府、企业或其他机构的决策提供重要的参考依据。

2 微博网络的分析模型

2.1 模型分析

微博网络的本质就是微博用户及用户之间所建立起来的稳定关系所构成的社会网络，微博中信息和资源的传播都在这一社会网络上完成。在传统的舆情分析研究中[5-7]，研究对象主要基于不同网站之间网页的联系，这种联系的建立难以表现出以用户为导向的信息自由流动特征。为了克服基于链接的方法在微博网络舆情分析中的不足，我们需要对微博网络的舆情分析模型进行相应分析，确定该模型需要实现的基本功能，具体包含以下四个方面：

(1) 文本抓取模块：微博网络具有用户基数大、数据量巨大的特点，所以要实现自动抓取功能，能对指定主题的微博自动抓取；

(2) 文本预处理模块：微博网络的内容可能包括声音、文字、图片以及视频，多而复杂，所以要有文本预处理的功能，对文本进行简单的分类；

(3) 微博信息跟踪记录模块：微博转发速度非常快，需要实现对指定微博的转发及评论用户信息进行抓取；

(4) 舆情分析模块：要进行舆情分析，就要对抓取的内容进行分析，发现其中存在的微博主题，并进行趋向性分析。

2.2 模型设计

微博网络以传播广度为主，聚合度非常高。同时，微博网络上聚集了大量的用户群体，加速和扩展了信息的传播。结合微博网络的这些特点，给出其舆情分析模型的框架，如图 1 所示。该模型包括微博网页的信息抓取、文本预处理、微博特征表示及提取、舆情分析等主要模块，其中，舆情分析模型是本文研究的核心，包括主题发现、热点分析、趋向性分析等三个方面。

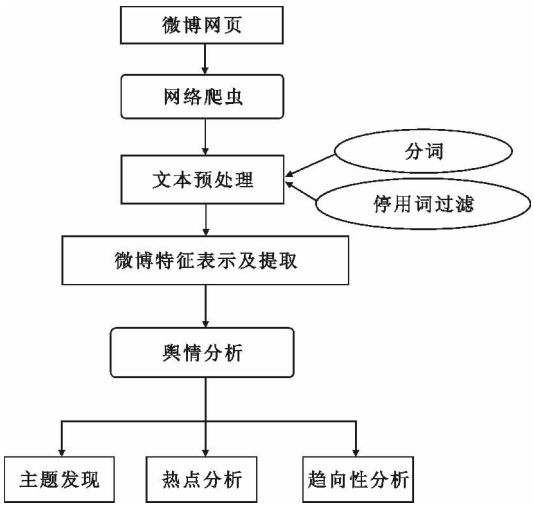

图 1 微博网络的舆情分析模型设计图

Fig. 1 The design model of public opinion analysis in the micro – blog network

3　基于主题发现的舆情分析

3.1　文本预处理

文本预处理包含三个过程：信息自动抓取、分词和过滤停用词。

第一，通过网络爬虫实现对微博的抓取。网络爬虫是一个自动提取网页的程序，它为搜索引擎从万维网上下载网页，是搜索引擎的重要组成。网络爬虫的基本搜索过程如图 2 所示。

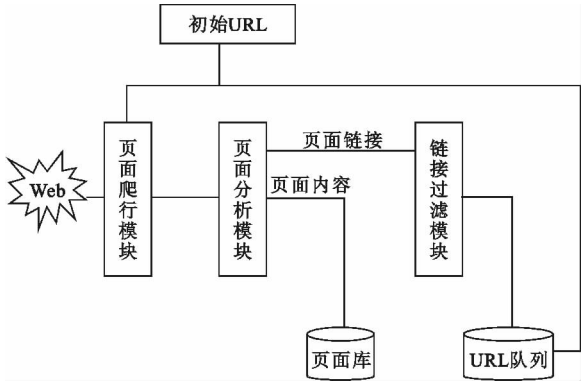

图 2　网络爬虫搜索过程

Fig. 2 The search process of web crawler algorithm

传统爬虫从一个或若干初始网页的统一资源定位器(Uniform Resource Location，URL)开始，获得初始网页上的 URL，在抓取网页的过程中，不断从当前页面上抽取新的 URL 放入队列，直到满足系统的一定停止条件。由于聚类算法在处理微博网络碎片信息方面[1]具有很好的性能，本文采用改进的增量聚类方法，下面具体描述。

对于每一个抓取到的微博文本，我们都可以从主标题和正文中提取 m 个关键字，并根据关键字的属性和权重构成向量来表示微博文本的主题 X_i，即主题向量 $\boldsymbol{X}_i=\{W_1,W_2,\cdots,W_m\}$，$i=1,2,\cdots,n$，$W_j(j=1,2,\cdots,m)$ 表示每一个关键字的属性和权重。这样 n 个抓取到的微博文本就可以构成由 n 个特征向量组成的数据集合 $C=\{X_1,X_2,\cdots,X_n\}$。进行微博文本的聚类，需要知道两个微博文本的主题向量 $\boldsymbol{X}_i$、$\boldsymbol{X}_j$ 的相似度 $Y(X_i,X_j)$。假设 $\boldsymbol{X}_i$ 和 $\boldsymbol{X}_j$ 有 k 个共同的关键字属性，而 $\boldsymbol{X}_i$ 有 k_1 个关键字属性，$\boldsymbol{X}_j$ 有 k_2 个关键字属性，那么两个主题向量 $\boldsymbol{X}_i$、$\boldsymbol{X}_j$ 的相似度表示如下：

$$Y(\boldsymbol{X}_i,\boldsymbol{X}_j)=\frac{k}{\sqrt{k_1k_2}}。\tag{1}$$

根据两个主题的相似度就可以完成微博文本的聚类，得到话题簇。

第二，需要对抓取的舆情信息进行分词。分词可以分为中文分词[8]和英文分词[9]。英文分词相对简单，一般通过空格分开，然而中文分词就相对麻烦，词与词之间没有明显的分隔符。

如图 3 所示，中文分词的算法主要包含三大类，分别为基于字典匹配、基于统计和基于规则的分词。基于字典匹配的方式由于实现简单、执行效率高，目前使用较普遍。本文采取了其中的逆向最大匹配的方法。在分词过程中，只需从后向前不断进行匹配，看词库中是否有该词，而对于没有在词库中出现的单词就无需进行分离。

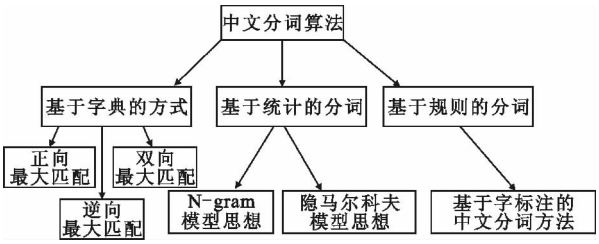

图 3　中文分词方法

Fig. 3 Chinese segmentation method

第三，停用词[10]的过滤，可以提高系统的运行效率。本文停用词的过滤用了算法 1 和算法 2 两个算法。

(1) 算法 1

输入：所取词语 a；

输出：词语 a 在查询语料库中和标准库中的频率之和；

计算：$f_b(a)$ 为词 a 在标准库中的频率，$f_q(a)$ 为词 a 在查询语料库中的频率：

for each a you input{

　　if($f_b(a)>p$ &&$f_q(a)>p$)

　　　　$\text{sum}_1=f_b(a)+f_q(a)$;

}

Output: sum_1

其中，p 设定的一个频率的阀值，防止词语 a 在某类词料中频率极高使得误差偏大，sum_1 就是词语 a 在查询语料库中和标准库中的频率之和。

(2) 算法 2

输入：所取词语 a；

输出：词语 a 在不同语料的左右熵之和；

计算：$S_{W_i}(a)$ 为词语 a 的左右熵，$W=\{\text{left},\text{right}\}$，$i=\{b,q\}$，即 S_{left_b}，S_{left_q}，S_{right_b}，S_{right_q}：

for each a input{

$sum_2 = 0$;
$S_{W_i}(a) = 0$; $W = \{left, right\}$, $i = \{b, q\}$
while(d in D) {
$S_{W_i}(a) -= P(d|a) * \lg P(d|a)$;
}
$sum_2 += S_{W_i}(a)$;
}
Output: sum_2

其中，D 为词语 a 的邻接词语集，而 d 为词集 D 其中的任意一个词语，sum_2 就是词语 a 在不同语料的左右熵之和。

根据算法 1 和算法 2，sum_1 和 sum_2 的取值越大就越可信。算法 1 减少了因为词频相差不大而排序等级相差较大造成的误差，而算法 2 降低了词语在不同语料左右邻接熵值带来的误差。

3.2 微博文本特征提取

微博信息进行文本预处理后，需要对抓取的微博文本信息进行特征提取。这里微博文本的表示，我们主要基于向量空间模型(VSM)[11]进行构建。首先，以向量($Weight_1$, $Weight_2$, $Weight_3$, …, $Weight_n$)来表示微博文本，其中 $Weight_j$ 为第 j 个特征项的权重($j = 1, 2, \cdots, n$)。对于每一个提取的微博文本进行特征提取，将分词后的词的集合向量化，使得每一条微博文本都转化为一个高维空间向量。同时，通过微博文本相似度的分析，得到两个微博文本之间的关联性。

在微博文本向量中用 1 表示该微博文本中有该词，用 0 表示微博文本中没有该词。为了使微博文本特征提取的准确度更高，后面渐渐用词频替代了原来的 0 和 1，目前一般通过 TF - IDF(Term Frequency - Inverse Document Frequency)[12]方法计算得到。其核心思想为：如果某个特征项在大多数的微博文本中出现的频率都很高，那这个特征项对微博文本的分类贡献不大，不能表示该微博文本的特征。基于此，我们在系统中采用了一种改进的 TF - IDF 公式：

$$Weight(w,a) = \frac{p(w,a) \times \lg(N_w/n_w + x)}{\sqrt{\sum_{w \in a} [p(w,a) \times \lg(N_w/n_w + x)]^2}} \quad (2)$$

式中，$Weight(w,a)$ 为词 w 在文本 a 中的权重；$p(w,a)$ 为词 w 在文本 a 中出现的频率；N_w 为所有文本的总数；n_w 包含词 w 的文本总数；x 为一个系数，在这里可以改变 x 的大小来调整准确度，比如取 0.01。将微博文本向量化以后，我们要确定该微博信息的重要性。一个微博文本可能包含大量的词汇，导致对应的向量可能包括很多维，因此我们需要减少维数从而提高效率和精度。为了使分类精度更高，应去除那些和主题相差较远的一些词，筛选出与该文本主题最相近的一定个数些词作为该类的特征项集合。鉴于此，这里利用信息量判断的标准进行特征向量抽取，其算法过程如下：

(1) 计算出特征集合中每一个词在所有微博文本中出现的频率之和：

$$sum_1 = \sum_{i=1}^{n_w} \sum_{j=1}^{n_a} p(w_i, a_j)。$$

式中，n_w 为词的总数，n_a 表示微博文本的总数，$p(w_i, a_j)$ 表示词 w_i 在微博文本 a_j 中的频率；

(2) 对于每一个特征集合中的词，计算该词在每一个微博文本中的频率之和：

$$sum_2 = \sum_{i=1}^{n_a} p(w, a_i);$$

(3) 计算该词在微博文本中的比重：

$$P(w|D_t) = \frac{1 + sum_2}{n_w + sum_1};$$

(4) 计算得到该特征词的互信息量：

$$I = \lg\left(\frac{P(w|D_t)}{P(w)}\right)。$$

对于特征集合中的每一个词，重复步骤 2 ~ 4，算出所有词对应的互信息量；

(5) 对同一类的词根据互信息进行排序，最后取前面特定数量的词组成特征向量，代表该微博文本的特征文本向量。

3.3 微博舆情的主题发现及趋向性分析

在微博网络中，当某一个话题的参与者超过某一值时，该话题就成为了热点。要发现热点，首先要将微博信息分类，将主题相近的一些微博归类到一起，然后还要统计参与各个微博的用户数，将所有同类的用户数相加，就可以表示该主题在网络的动态情况。参与该类微博信息的用户数越多，就说明该微博当前时间段内越热。一条舆情的热度等于关于此舆情的微博关注的热度之和加上关于此舆情的评论热度之和，其具体的算法实现如下：

输入：该类微博中各条微博的听众数 n_l，微博被转发的次数 n_s，微博被评论的次数 n_p，微博发布时间 T，微博第一条评论时间 T_f，微博最后一条评论时间 T_l；

输出：该类微博舆情的热度 High：

High = 0;
for(i in D) do
$\text{High}_g = \lg(1 + n_l) + \sqrt{n_s} + n_p$;
$\text{High}_p = \dfrac{1 + T - T_f}{(1 + T_l - T_f) \cdot \sqrt[3]{n_p}} \cdot \lg(1 + n_l)$
$\text{High} = \text{High} + \text{High}_1 + \text{High}_2$
End
Output: High

其中，i 表示所有该类微博 D 中的一条微博。对于每一类微博舆情，都可以通过上述算法算出对应的热度 High。当发现某个主题以后，我们不仅需要知道它是否为热点话题，还要了解它未来的发展趋势，从而能对未知的微博舆情信息进行及时感知和响应。

分析一个主题的趋向性需要统计各个时间段内该主题参与的用户数的动态变化，如果该主题的参与者在该时间段内参与的用户数呈爆炸式的增长，说明该主题将更快地在网络上传播; 如果用户传播的数处于减少状态，说明该主题已经接近尾声。同时我们看到，微博用户观点不仅会随时间而变化，也会随微博网络环境而变化，观点演化结果直接影响微博舆情危机的预警。预测舆情的趋向性本文采用了马尔科夫[13]预测模型，具体描述如下:

第 t 个时期的状态概率向量可表示为 $S(t) = (S_1^{(t)}, S_2^{(t)}, \cdots, S_n^{(t)})$，其中 S_i^t 表示第 t 个时期预测系统处于状态 i 的概率，所以有 $S_i^{(t)} \geqslant 0 (j = 1, 2, \cdots, n)$ 和 $\sum_{i=1}^{i=n} S_i^t = 1$。特别地，初始化状态概率为 $S_1^0, S_2^0, \cdots, S_n^0$，相应的初始状态概率向量为 $S(0)$，最后状态转移方程为

$$S(t) = S(t-1) * P = S(0) * P^t。 \quad (3)$$

根据上面的转移方程就可以建立主题趋向性分析的马尔科夫预测模型，其中 P 为初始的概率转移矩阵。可以根据初始的概率向量和初始状态转移矩阵推测出未来第 t 个时期的概率 $S(t)$。可以看出，通过热点分析和趋向性分析两种维度，微博网络的当前主题和未知主题都能进行分析和预测。这将会为有关部门提供有意义的舆情分析结果，从而能对微博网络系统的舆情现状进行感知和应对。

4　仿真与分析

本仿真实验，硬件平台是 Intel i5 - 4200U 2.3 GHz的双核处理器、8GB RAM 和 64 位的 Windows7 操作系统的 PC。软件平台中 Internet 信息服务是 IIS6.0，数据库为 SQL Server 2008，在此基础上基于 C#语言实现微博舆情分析系统。该系统会周期地对新浪微博、腾讯微博、网易微博、搜狐微博等进行网页抓取，并对其内容实行解析后得到热点话题。通过大量的实验结果和真实的情景进行比较，不断地调整参数。本实验随机对抓取的部分热点话题进行分析以及趋向性分析。

本实验的实验数据是采用开源的网络爬虫软件 Locoy Spoder[14] 从新浪微博上抓取的，通过3 h的数据采集共抓取了83 571条有效微博数据。然后，采用 3.1 节的方法进行会话抽取之后对采集到的信息进行数据预处理、数据转化、分词处理等加工，针对微博文本相对较长的内容，采用了中文词法分析器 ICTCLAS[15] 进行分词操作，根据聚类的统一主题，收集日期、回帖量、点击量等信息，对各个热点话题进行舆情分析，做出趋向性分析。根据前面的一些预处理，对网络的一些热点如马航事件、乌克兰事件等热点事件进行抓取分析，得到一个舆情主题关注度的表格，若以周为统计单位，则每个舆情主题关注度都有 8 个统计数据，用 M_i 表示一个统计周期，满足 $i \in [1, 8]$，如表 1 所示。

表 1　舆情主题关注度
TabLe 1 The attention degree of public opinion topic

舆情主题	关注度/次							
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
马航客机失联事件	0	0	0	0	20 4321	337 958	253 541	153 726
乌克兰事件	41 237	54 032	107 725	156 235	179 043	258 464	153 546	133 452
昆明火车站暴恐事件	0	0	0	0	23 0321	78 321	24 304	25 607
文章姚笛出轨事件	0	0	0	0	0	0	0	308 743
春晚	352 684	223 451	89 834	11 356	2355	268	53	26

根据以上数据，对应的二维关注度趋势分析图如图 4 所示。

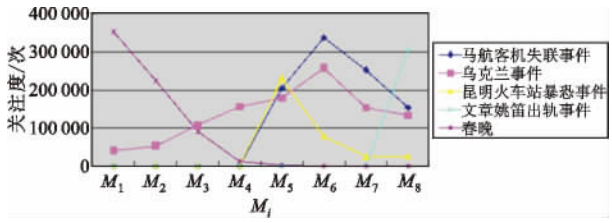

图 4 关注度趋势分析图

Fig. 4 The trend analysis graph of concern

从图 4 可以看出，春晚在 2014 年 2 月初关注度非常高，处于峰值，随后关注度逐渐减少，呈下降的趋势，而乌克兰事件从 2014 年 2 月初到 3 月中期都处于不断上升的状态，在后面则慢慢呈减小的趋势。从图中还可看出，马航自 2014 年 3 月初发生以后关注度快速飙升，直到 3 月中旬，关注度冲到峰值，后面慢慢下降，但关注度仍然较高。前面那些事件在 3 月末以后都呈下降趋势，预测下面的关注度也会不断减小，但像乌克兰事件可能仍然会持续一段时间。而对于文章、姚笛出轨事件，在 3 月末关注度一下子暴增，很可能成为后面的一个热点话题。

为了完成对采集的微博舆情进行主题发现，可以按照以下步骤:

步骤 1: 设定一个合适的阀值 T;

步骤 2: 计算各微博主题在某一段时间内的关注度 $Y(T_f, T_s)$:

$$Y(T_f, T_s) = N(T_s) - N(T_f)。$$

式中，$N(T_x)$ 表示在时间 T_x 关于该微博主题的有关微博数目，$x = \{s, f\}$。则在 T_f 到 T_s 时间段该微博主题的关注度可以用这段时间内微博数目的差值表示;

步骤 3: 计算各微博主题的关注度，然后与设定阀值 T 比较，如果该微博主题的关注度大于 T，那么该微博主题为热点主题;

步骤 4: 对各热点微博主题按照关注度进行降序排序，从而了解当前最热的微博主题。

以 2014 年 3 月 1 ~ 7 日作为统计周期，将上述各舆情主题的关注度按降序排列，得到舆情热点排行榜如图 5 所示。从图 5 可以看出，在 2014 年 3 月 1 ~ 7 日期间，文章、姚笛出轨事件备受关注，成为了当时的舆情焦点; 乌克兰事件和马航客机失联事件虽然关注度下降，但仍有较多的关注; 而春晚则基本已经没有了关注，这和从舆情趋向性分析得到的结论基本相似。因此，决策者就可以根据民众的不同主题倾向提出相关的改进措施。由此可见，基于主题发现的舆情分析模型能够针对一些热点事件分析出网民的观点、看法，识别出其主题倾向，进而为政府、企业或其他机构的决策提供重要的依据。

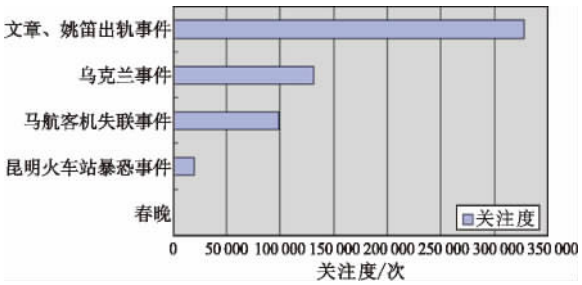

图 5 2014 年 3 月 1 ~ 7 日部分舆情关注度情况

Fig. 5 The public opinion case in March 1 – 7, 2014

5 结束语

本文对微博网络中基于主题发现的舆情分析系统进行了进一步的分析和研究，在提取主题的基础上进行舆情分析，并作出相应的舆情发展趋向性分析，对网络舆情进行预测。同时，本文对微博网络中的舆情分析模型给出了实验结果和分析，将来还将继续对多层社交网络中舆情分析模型的设计和优化等工作进一步研究。

参考文献:

[1] 李岩，韩斌，赵剑. 基于短文本及情感分析的微博舆情分析[J]. 计算机应用与软件，2013，30(12): 240 – 243.
LI Yan，HAN Bin，ZHAO Jian. Analyzing microblog public opinions based on short text and sentiment analysis [J]. Computer Applications and Software，2013，30(12): 240 – 243. (in Chinese)

[2] 唐晓波，宋承伟. 基于复杂网络的微博舆情分析[J]. 情报学报，2012，31(11): 1153 – 1162.
TANG Xiaobo，SONG Chengwei. Analysis of micro – blog public opinion based on complex network [J]. Journal of The China Society for Scientific and Technical Information，2012，31(11): 1153 – 1162. (in Chinese)

[3] Yu M，Yang W，Wang W，et al. Information Diffusion and Influence Measurement Based on Interaction in Microblogging [M]//Social Media Processing. Heidelberg，Berlin: Springer Berlin Heidelberg，2014: 129 – 140. (in Chinese)

[4] 曾振东. 基于灰色支持向量机的网络舆情预测模型[J]. 计算机应用与软件，2014，31(2): 300 – 302.
ZENG Zhendong. The network public opinion prediction models based on grey support vector machine [J]. Computer Applications and Software，2014，31(2): 300 – 302.

(in Chinese)

[5] 殷俊,何芳. 微博在我国的传播现状及传播特征分析[J]. 河南大学学报(社会科学版),2011(3):124 - 129.
YIN Jun,HE Fang. The analysis of current situation and characteristics transmission of micro - blog in China [J]. Journal of Henan University(Social Science Edition),2011(3):124 - 129. (in Chinese)

[6] 吴建军. 网络舆情的云计算监测模式分析与实现[J]. 电讯技术,2013,53(4):476 - 481.
WU Jianjun. The analysis and implementation of the cloud monitoring model of network public opinion [J]. Telecommunication Engineering,2013,53(4):476 - 481. (in Chinese)

[7] 许鑫,章成志. 互联网舆情分析及应用研究[J]. 情报科学,2008(8):1195 - 1204.
XU Xin,ZHANG Chengzhi. Internet public opinion analysis and its application [J]. Information Science,2008(8):1195 - 1204. (in Chinese)

[8] 周俊,郑中华,张炜. 基于改进最大匹配算法的中文分词粗分方法[J]. 计算机工程与应用,2014,50(2):124 - 128.
ZHOU Jun, ZHENG Zhonghua, ZHANG Wei. Chinese word rough segmentation method based on improved maximum matching algorithm [J]. Computer Engineering and Applications,2014,50(2):124 - 128. (in Chinese)

[9] Heffner C C,Dilley L C,McAuley J D,et al. When cues combine: how distal and proximal acoustic cues are integrated in wordsegmentation [J]. Language and Cognitive Processes,2013,28(9):1275 - 1302.

[10] 夏火松,陶敏,王一,等. 停用词表对基于 SVM 的中文文本情感分类的影响[J]. 情报学报,2011,30(4):347 - 352.
XIA Huosong,TAO Min,WANG Yi,et al. The effect of stop list of Chinese text sentiment classification based on SVM [J]. Journal of The China Society for Scientific and Technical Information, 2011, 30 (4): 347 - 352. (in Chinese)

[11] 王旭仁,李娜,何发镁,等. 基于改进聚类算法的网络舆情分析系统研究[J]. 情报学报,2014,33(5):530 - 537.
WANG Xuren,LI Na,HE Famei,et al. The research on analyzing system of network public opinion based on improved clustering algorithm [J]. Journal of The China Society for Scientific and Technical Information,2014,33(5):530 - 537. (in Chinese)

[12] Hong T P,Lin C W,Yang K T,et al. Using TF - IDF to hide sensitive itemsets [J]. Applied Intelligence,2013,38(4):502 - 510.

[13] 何洪华,徐敬德,计哲,等. 基于二阶隐马尔可夫模型的清浊音恢复算法[J]. 电讯技术,2011,51(6):56 - 60.
HE Honghua,XU Jingde,JI Zhe,et al. The algorithm speech recovery algorithm based on Two order hidden Markov model [J]. Telecommunication Engineering,2011,51(6):56 - 60. (in Chinese)

[14] Wang L,Zhao Y,Liang S H,et al. Microblog Social Network Analysis Based on Network Group Behavior [J]. Advanced Materials Research,2013,798 - 799:435 - 438.

[15] 王松,吴亚东,李秋生,等. 基于时空分析的微博演化可视化[J]. 西南科技大学学报,2014(3):68 - 75.
WANG Song,WU Yadong,LI Qiusheng,et al. The visualization of micro - blog evolution based on temporal and spatial analysis [J]. Journal of Southwest University of Science and Technology,2014(3):68 - 75. (in Chinese)

作者简介:

彭　浩(1982—),男,江苏人,2012 年于上海交通大学获博士学位,现为讲师,主要研究方向为计算机通信网、复杂网络安全;

PENG Hao was born in Jiangsu Province,in 1982. He received the Ph. D. degree from Shanghai Jiaotong University in 2012. He is now a lecturer. His research concerns computer communication networks and complex network security.

周　杰(1993—),男,浙江人,主要研究方向为社交网络舆情分析;

ZHOU Jie was born in Zhejiang Province,in 1993. His research concerns analysis of social network public opinion.

周　豪(1993—),男,浙江人,主要研究方向为数据挖掘;

ZHOU Hao was born in Zhejiang Province,in 1993. His research concerns data mining.

赵丹丹(1981—),女,浙江人,2012 年于上海交通大学获博士学位,现为讲师,主要研究方向为网络与信息安全、社会网络。

ZHAO Dandan was born in Zhejiang Province,in 1981. She received the Ph. D. degree from Shanghai Jiaotong University in 2012. She is now a lecturer. Her research concerns network and information security,social network.

Email: ddzhao@ zjnu. edu. cn