DOI:10.11992/tis.201603341
网络出版地址:http://www.cnki.net/kcms/detail/23.1538.TP.20160513.0919.014.html

面向用户兴趣与社区关系的微博话题检测方法

刘志雄[1,2],贾彩燕[1,2]
(1.北京交通大学 计算机与信息技术学院,北京 100044; 2.北京交通大学 交通数据分析与挖掘北京市重点实验室,北京 100044)

摘 要:微博话题检测是一种特殊形式的话题检测,传统的话题检测方法并不能取得很好的效果。提出了一种面向微博用户社区的话题检测方法。该方法首先在用户发表的微博文本上,利用 LDA 主题模型分析用户的兴趣分布。接着,结合微博用户关系网络与用户兴趣对用户进行社区划分,使得同一社区的用户不仅具有较稠密的链接关系,还具有相似的兴趣。然后,面向用户社区,在每个社区内部检测用户关心的话题,给出了一种面向用户社区的、融合词重要度与 ε 近邻图的微博话题发现方法。该算法能够有效地去除微博噪声、快速准确检测出每个用户社区内关心的话题并对话题进行热度排行。

关键词:微博;社区;网络;文本;话题;兴趣;噪声;主题
中图分类号:TP393 **文献标志码**:A **文章编号**:1673-4785(2016)03-0294-06

中文引用格式:刘志雄,贾彩燕.面向用户兴趣与社区关系的微博话题检测方法[J]. 智能系统学报, 2016, 11(3): 294-300.
英文引用格式:LIU Zhixiong,JIA Caiyan.Micro-blog topic detection based on users' interests and communities[J]. CAAI transactions on intelligent systems, 2016,11(3): 294-300.

Micro-blog topic detection based on users' interests and communities

LIU Zhixiong[1,2], JIA Caiyan[1,2]
(1. School of Computer and Information Technology, University of Beijing Jiaotong, Beijing 100044, China; 2.University of Beijing Jiaotong Beijing Key Lab of Traffic Data Analysis and Mining, Beijing 100044, China)

Abstract: Microblog topic detection is a special type of topic detection. The traditional topic detection algorithms do not work well in special situations for Chinese microblogs. In this paper, a topic detection method cater to the user community of microblogs is proposed. Firstly, the users´ interests were analyzed by using the LDA(Latent Dirichlet Allocation) topic model on the text of microblogs generated by users/bloggers. Then the user/follower network associated with users´ interests was created and partitioned into different communities so that the users in the same group were not only densely connected but also shared similar interests. Then, the topics of interest in each community were detected. Together, this provides a microblog topic finding method that faces a user´s community and combines the importance of words as well as an ε neighboring graph. The experimental tests show that the method can effectively eliminate microblog noise, compute the importance of words, and rapidly and accurately obtain the topics of interest of each community.

Keywords: microblog; community; network; text; topic; interest; noise; theme

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,这种基于用户关系与短文本特性的信息分享、传播以及获取的平台也随之兴起。微博用户可以通过 PC 端、手机端以及其他客户端组建个人社区,以 140 字左右的文字更新信息,并实现即时分享。微博成为典型的 Web2.0 应用之一。

在现实世界中,有很多系统都可以抽象为网络,这些网络中包含着一些潜在的社区结构,具有社区

收稿日期:2016-03-19. **网络出版日期**:2016-05-13.
基金项目:国家自然科学基金面上项目(61473030)、中央高校基本科研业务专项基金项目(2014JBM031).
通信作者:刘志雄. E-mail:523129791@qq.com.

内部节点链接稠密、社区之间节点链接稀疏的特点。通常,社区内部的节点具有相似的特性,在网络中扮演着相似的角色。对于微博用户关系网而言:同一社区内的用户往往具有相同或者相似的兴趣与爱好。

目前对于微博的研究大多是对用户关系的分析或者微博内容的分析。在用户关系研究领域,主要研究其社区特性。大体思路是:以用户 ID 为节点,用户关注关系为边构建用户关系网络图,然后采用社区划分算法将其划分为若干社区。往往同一社区内的用户拥有共同的兴趣与爱好。在微博内容分析方面,致力于研究微博话题发现方法。大体思路是:以词为特征使用 VSM[3] 模型将微博文本转化为空间向量,并且使用 TF-IDF 算法计算每一维的权重,然后使用聚类方法将相同话题下的微博文本聚集成一个个微博话题簇。例如:周刚等[4] 提出了一种基于组合相似度的微博话题发现方法 MB-SinglePass 来提升聚类效果,他们将余弦相似度、雅各比相似度、语义相似度以一定的权值融合,改进了微博相似度的计算方法;郑斐然等[5] 提出了一种基于词聚类的新闻话题发现方法;方然等[6] 提出了一种基于情感的微博话题检测方法,他们认为倾向消极的词更加具有话题表现力,从而依据词的情感分数改善了话题检测效果。然而微博文本被严格限制在 140 字以内,单纯地使用 VSM[3] 空间向量模型对微博文本进行建模,存在严重的特征稀疏和维度过高问题。更严重的是聚类结果还受到微博噪声的影响,导致话题检测的效果不理想。

本文提出了一种面向用户兴趣与社区关系的微博话题检测方法,首先应用 LDA[1] 主题模型对微博文本进行降维,以用户微博在主题上的分布来表征用户的兴趣与爱好;然后,结合用户兴趣特征对用户关系网进行社区划分,使得同一社区内的用户不仅具有稠密链接的社区关系,还具有相似的兴趣;最后,使用了一种融合词重要度与 ε 近邻图[2] 的微博话题检测方法得出每个社区(主题)对应的话题,并实现相关社区内的话题热度排行。实验结果显示,该算法有效地对微博特征空间进行了降维、微博去噪,使得相似度的计算更加容易;实现了社区内的微博话题检测,以挖掘出社区内的用户共同关心的话题,话题检测结果更加迎合社区内的用户兴趣与爱好,便于进行面向社区兴趣的话题推荐和排行。

1　基于用户社区兴趣的话题发现方法

本文提出的微博话题检测方法以中文微博为处理对象,分为如下 4 个步骤:数据预处理、网络建模、用户社区发现、微博话题检测(流程如图 1)。其中,数据预处理主要对微博数据进行筛选和切词,并且过滤掉停用词以及微博平台常见的噪声。例如:“转发微博”、“分享图片”、“视频”等,然后采用基于吉布斯采样[7] 的 LDA[1] 主题模型对用户微博进行降维处理,以得到用户的兴趣分布。网络建模是以用户 ID 为节点,用户关注关系为边,构建网络模型。用户社区发现主要结合 LDA 模型提取的用户兴趣特征,对用户关系网络进行社区划分,使得找到的社区内的用户对相似的话题感兴趣。话题发现:利用社区划分结果,对社区内微博进行话题检测,挖掘出社区内关心的话题,并对社区内的话题进行热度排行。

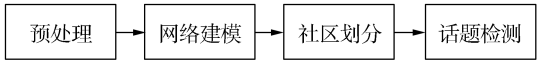


图 1　算法流程

Fig.1　flow of algorithm

1.1　数据预处理

微博是一种非结构化数据,携带信息具有碎片化的特征。并且,携带着大量的垃圾信息(噪声),使得对微博数据的预处理是微博数据分析的重要前提。主要分为以下 2 个方面:1)针对微博用户的处理规则,2)针对微博文本内容的处理规则。

1)针对微博用户

由于某些用户发表微博数目较少,并不能很好地反映用户的兴趣,故选取发表微博总长度大于 5 000的用户及其关注关系作为我们的数据集。

2)针对微博内容

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空格加以区分。因此分词是微博内容处理的关键一步,分词的方法有多种,如基于字符串匹配的分词方法、基于统计的分词方法等。本文采用一种基于最大匹配算法的中文单词识别系统(a word identification system for mandarin chinese text based on two variants of the maximum matching algorithm,MMSEG)进行分词,MMSEG 算法是一种简单、高效的基于词典的中文分词算法。

去停用词:停用词是指在自然语言中具有一定功能但又没有什么实际意义的词。这些词往往以较高的频率出现,会对文本处理造成一定干扰。另外,微博文本中常会出现一些高频词,如:“转发”、“微博”、“分享”、“图片”等,这些高频词会对话题检测

产生较强的干扰,也需要和停用词一起加以过滤。

经过以上预处理步骤,我们过滤掉了一部分噪声。但即便如此,以词来表征微博文本的特征向量的维度也是巨大的,会严重影响微博文本相似度计算的效率以及有效性。

3)基于微博文本的用户兴趣特征抽取

为了学习用户的兴趣特征,如果以用户发表的微博文本上的词为特征,则会面临维数灾难,我们将一个用户发表的所有微博合并为一个长的文本,用以表征用户的兴趣,采用基于吉布斯采样法[7]的 LDA[1]主题模型进行降维。将用户的兴趣表示为其在有限个主题上的分布向量。

1.2 网络建模

1.2.1 建模

本文使用有向无权图表示用户关系网。每一个用户作为图中的一个节点,为每一个节点都分配一个 ID,ID 值从 1~n,用户之间的关注关系作为图的边。如果用户 i(ID 为 i 的用户)关注了用户 j,则有一条由节点 i 指向节点 j 的有向边。

1.2.2 相似度构造方式

1) 链接属性相似度度量

文献[19]提出了一种采用信号传递方法将网络的拓扑结构转换成一个 N 维欧式空间上的几何向量结构,N 是网络中的节点数。我们以该几何向量作为节点的链接属性向量。

2) 内容属性相似度度量

用户微博通过 LDA[1]主题模型降维后,可以得到一个该用户对应微博文档在主题上的分布向量,以该向量表示节点的内容特征向量。

3) 联合相似度

本文采用余弦相似度计算两个节点的链接和内容相似度,公式为

$$\mathrm{sim}(i,j)=\frac{\sum_{k=1}^{n} v(i,k)\times v(j,k)}{\sqrt{\sum_{k=1}^{n} v(i,k)\times v(i,k)}\times\sqrt{\sum_{k=1}^{n} v(j,k)\times v(j,k)}} \tag{1}$$

如果将链接相似度表示为 sim^t(sim^t由链接属性向量采用式(1)求得),将内容相似度表示为 sim^c(sim^c由内容特征向量采用式(1)求得),那么链接与内容相结合的联合相似度可表示为 sim^u,sim^u计算公式为

$$\mathrm{sim}^u=\alpha\,\mathrm{sim}^t+(1-\alpha)\,\mathrm{sim}^c \tag{2}$$

式中 $\alpha\in[0,1]$ 表示链接相似度在联合相似度中占的比例。由于参数 α 的选取通常很困难,故在社区划分过程中采用投票机制来规避这一缺陷,详情见文献[8]。

1.3 用户社区划分

以用户 ID 为节点构建的用户关系网中,同一社区内的用户,通常具有相同或相似的爱好。因此,结合用户的链接关系和用户的兴趣分布,对用户进行聚类,也称为用户社区划分。

本文延用我们设计的社区划分方法 KRLC[8]对微博用户进行社区划分。具体过程如下:

1)选取中心节点

采用 K-rank[9]算法选取初始中心节点,即中心节点不但要具有大的 PageRank 值,中心节点间的相似度要尽可能小。

2)社区划分

采用 K-means[11]算法进行社区划分,过程如下:

输入 用户网络 G,用户微博长文本集 LD,社区数 K;

输出 划分好的社区列表 CommunityList。

①运行 Signal[19]方法将网络的拓扑结构转换成一个 N 维欧式空间上的几何向量。

②运行 Gibbs-sampling-LDA[1]方法将节点的微博文档映射到 K 维特征空间(表示用户在 K 个主题上的兴趣分布);

③采用 K-means[11]算法进行社区划分,将每个用户节点分配得离它最近的中心所属的类中,用户间节点的相似性计算方法用式(2)的联合相似性测度。

对于已经划分的社区,我们根据社区内用户所发表微博在主题上的兴趣分布向量,可以求出该社区关心的主题,如下:

以 $\boldsymbol{t}(i)=(t(i,1),\cdots,t(i,j),\cdots,t(i,k))$ 表示社区 i 在各个主题上的兴趣分布向量,其中 $t(i,j)$ 表示社区 i 在第 j 个主题上的分布值,则

$$t(i,j)=\sum_{u(k)\in c(i)} t(k,j) \tag{3}$$

式中:$c(i)$ 表示社区 i,$u(k)$ 表示 ID 为 k 的用户,$t(k,j)$ 表示用户 k 在第 j 个主题上的分布值。最后,根据 $\boldsymbol{t}(i)$ 取主题分布值最大的 3 个分量对应的主题作为社区 i 关心的主题。

1.4 话题检测

话题是讨论、谈话的中心,在整个微博上,用户经常会针对某一事件、观点展开讨论。对于有大量用户参与讨论的事件和话题,我们称之为热点话题。

本文提出了一种融合词重要度与 ε 近邻图[2]的微博话题检测方法来检测话题。具体步骤如图 2 所示。

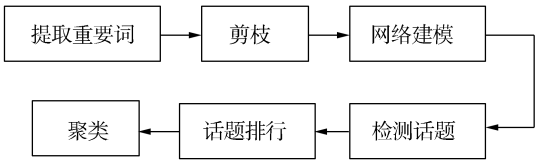


图 2　话题检测流程图

Fig.2　flow of topic detection

1.4.1　*提取重要词*

由话题的定义可知,与话题相关的词语通常会具有更高的重要性。显然,重要性过低的词语,尽管能够表达一定的含义,但并不能构成话题,会对我们话题检测造成一定影响。因此需计算词的重要性。

TextRank[12]算法是在 Google 的 PageRank[10]算法启发下,针对文本里的句子设计的权重算法。最初的目标是对文章提取摘要,目前多用于给词语打分,即计算词语的重要度。本文采用 TextRank[12]算法计算词语重要度并过滤掉重要度过低的词语,步骤如下:

1)将同一社区内所有微博(已切词)做拼接,构成微博文档 D。

2) 采用 TextRank[12]算法对微博文档 D 求词语重要度分数并逆序排序。

3)剔除重要度低于阈值 θ 的词语。

经过如上步骤,得到了微博文档 D 对应的重要词库,记为精英词集 elite。

1.4.2　*剪枝*

将微博特征向量中不属于重要词汇库 elite 的词语剔除。若剔除后微博向量长度过短,则将该微博从该社区剔除,本文设置长度阈值为 3。去除了微博内与话题相关度很低的词语,保留了与话题相关度较高的词语。

1.4.3　*微博文本 ε 近邻图构建*

传统的微博相似度计算方法主要是对微博集合中每一条微博的词进行 TF-IDF 的计算,并将微博中各个词表示成 VSM[3]空间向量,然后采用余弦相似度计算两条微博之间的相似度。但考虑到微博具有短文本高维、稀疏的特点,采用传统的 TF-IDF 向量表示法计算得到的相似性(趋于 0)不能反映两个微博文本的真实相似性。故本文以词语的重要度代替 TF-IDF 值作为词的特征权重。由于经过社区划分以及微博剪枝之后,社区内微博特征已相对稠密,故可采用基于 VSM[3]空间向量模型的余弦相似度计算方法来计算两条微博之间的相似度,公式为

$$\mathrm{sim}(A,B)=\frac{\sum_{w_i\in A\cap B}\mathrm{score}(w_i)\times\mathrm{score}(w_i)}{\sqrt{\sum_{w_i\in A}\mathrm{score}(w_i)\times\mathrm{score}(w_i)}\times\sqrt{\sum_{w_i\in B}\mathrm{score}(w_i)\times\mathrm{score}(w_i)}} \tag{4}$$

式中:$\mathrm{sim}(A,B)$ 表示微博 A 与微博 B 之间的相似度,$\mathrm{score}(w_i)$ 表示词 w_i 的重要度分数。

我们给每一条微博分配一个 ID,ID 从 1 到 n',然后以微博为节点,微博之间的相似度为边,构建一张 ε 近邻图[2]。若微博 i 与微博 j 的相似度大于阈值 ε,则微博 i 与微博 j 之间存在一条边,且该边权重为 $\mathrm{sim}(i,j)$。

1.4.4　*微博聚类*

本文采用社区划分的方法对微博文本进行聚类。由于社区具有社区内部节点连接稠密、社区之间节点连接稀疏的特点,故社区(话题簇)内微博相似度更大,社区(话题簇)间微博相似度更小。故对微博 ε 近邻图进行社区划分,并选取社区节点数最多的 T 个社区作为社区内关心的话题。本文采用经典社区划分算法 BGLL[13]对微博 ε 近邻图进行社区划分。

1.4.5　*话题检测*

本文以主题词来描述话题,提出了一种以主题度来选取主题词的方法。本方法以 $\mathrm{topic}(w_i,j)$ 表示词 w_i 在话题簇 j 内的主题度,计算公式如下:

$$\mathrm{topic}(w_i,j)=\mathrm{fre}(w_i,j)\times\mathrm{score}(w_i)/\mathrm{num}(j) \tag{5}$$

式中:$\mathrm{fre}(w_i,j)$ 表示词 wi 在话题簇 j 内的词频,$\mathrm{score}(w_i)$ 表示词 w_i 的重要度,$\mathrm{num}(j)$ 表示话题簇 j 包含的微博数目,则主题选取过程如下:

1)对于所有话题簇,在话题簇内计算所有词的主题度;

2)在话题簇内按主题度对词进行逆序排序,并保留主题度最大的 15 个词;

3)将所有话题簇内所保留的词加入集合 s;

4) 遍历集合 s,对于词 w_i,遍历所有社区,若 w_i 在社区 t 内的主题度最高,则 $w_i\in tw(t)$。$tw(t)$ 表示话题簇 t 对应的主题词集合。

1.4.6　*话题热度排行*

话题的热度表现在多个方面,本文以主题度来表征话题的热度。计算公式为

$$\mathrm{heat}(j)=\frac{\sum_{w_i\in\mathrm{tw}(j)}\mathrm{topic}(w_i,j)}{m(j)} \tag{6}$$

式中:$\mathrm{heat}(j)$ 表示话题簇 j 对应话题热度,$m(j)$ 表示话题簇 j 对应主题词集合包含词语个数。

最后按话题热度对话题进行逆序排序。

2 实验结果与分析

2.1 实验数据

本实验数据采用自主抓取的新浪微博数据,该数据集于2013 年 9 月—2013 年 12 月采用自主开发的面向新浪微博的网络爬虫爬取。数据集包括用户基本信息、用户关系信息、用户发表微博等 3 部分。

2.2 实验过程与结果

2.2.1 用户社区划分实验与结果

根据新浪微博首页热门微博分类版块,选取 10 个类别作为主题,分别为亲子、体育、公益、娱乐、文艺、时尚、时政、生活、科技、财经。然后将每个用户发表的微博拼接成微博文档,选取微博文档长度大于 5 000 字的 3 490 个用户作为实验数据,并进行网络建模。采用信号传递算法[19]对用户关系网求链接属性向量,并采用该向量求节点链接相似度;采用 LDA[1]主题模型对微博文档求主题分布向量(内容特征向量),并采用该向量求节点内容相似度。然后采用 KRLC 算法[8]对用户进行社区划分,最后采用式(3)求出社区对应兴趣分布,具体结果如表 1。

表 1 使用 KRLC 划分的社区兴趣分布

Table 1 The interest distribution of community by KRLC %

节点数	亲子	娱乐	生活	时政	财经	文艺	时尚	公益	体育	科技	类别
798	58.06	2.11	19.11	1.53	3.58	0.17	8.77	4.64	0.23	1.76	亲子
536	6.63	9.25	21.31	0.96	2.69	0.01	0.16	8.08	48.15	2.71	体育
204	11.14	5.16	12.27	6.92	6.81	0.70	1.70	50.71	1.60	2.92	公益
169	7.81	60.09	19.51	0.35	0.65	0.05	0.77	5.09	1.62	4.01	娱乐
122	10.89	5.00	4.46	7.25	6.27	45.64	2.50	15.07	0	2.86	文艺
70	11.85	2.71	15.34	0.10	2.28	0.49	56.22	4.00	0	6.95	时尚
567	10.52	1.27	3.14	46.45	9.66	0.50	0.51	27.27	0.16	0.48	时政
257	13.46	9.80	58.16	0.54	0.74	0.18	2.60	7.79	1.39	5.29	生活
609	9.38	2.71	8.37	0.14	4.31	0.07	6.97	1.38	0.08	66.53	科技
158	15.64	3.73	4.94	6.41	55.93	2.52	0.83	9.41	0.09	2.52	财经

2.2.2 社区内话题检测结果

根据划分的 10 个社区,在社区内检测话题,检测算法如 2.4 所示。本文选取了 2013.11.10—2013.11.12 共 3 天的微博作为话题检测数据。其中词语重要度阈值 θ 设为 40%,即保留重要度最高的 40% 词语,相似度阈值 ε 设为0.15。由于亲子、文艺、时尚等 3 个主题出现话题几率较小,故本文没有在这 3 个社区内检测话题。部分主题对应社区内话题检测结果如表 2 所示。

表 2 部分社区内微博话题检测结果

Table 2 Part of micro-blog topic detection result within community

主题	编号	主题词	热度	对应话题
体育	1	男篮 亚锦赛 第一 比赛 易建联 朱芳 王治郅	2.13	男篮亚锦赛名单出炉
	2	广州 恒大 2013 亚冠 联赛 冠军 足球 夺冠	2.04	恒大亚冠联赛夺冠
	3	中国 大奖赛 车迷 活动 潘涌涌 现场	1.95	车迷大奖赛潘涌涌现场解说
	4	北京 首钢 支持 陈磊 吉哲 球迷 国安	1.70	明日 CBA 北京 VS 辽宁
	5	央视 羽毛球 世锦赛 林丹 谌龙 世界 锦标赛 男子 单打	1.37	羽毛球世锦赛林丹战胜谌龙
公益	1	河南 大学生 北京 2013 儿童 家人 电话 平安	4.70	河南大学生失联
	2	汇聚 长沙 银行 支行 慈善 拍卖会 拍品 展示 作品	2.99	长沙银行娄底支行慈善拍卖会
	3	头条 汪峰 离婚 瞬间 章子 发布 凌晨 吴奇隆 刘诗诗	1.89	汪峰悲催头条又被抢
娱乐	1	白举纲 丝带 北京 小白 白菜 节目 录制	1.35	白举纲爱的绿丝带北京演唱会
	2	吴亦凡 1106 生日 快乐 祝福 凡凡	0.92	11 月 06 日吴亦凡生日快乐
	3	非同 凡响 131102 首尔 青少年 庆典 颁奖	0.84	非同凡响青少年庆典颁奖

根据表 2 我们可以看出,面向用户社区的话题检测方法,可以针对社区内的用户兴趣找到用户感兴趣的话题,使得话题推荐和排行具有社区兴趣个性化。

2.3　局部算法对照试验

由于本文提出的面向用户兴趣与社区关系的微博话题检测与已有研究不同,一是方法不同,二是研究数据不同,因此我们没有和已发表方法进行对比。但我们对已选取的社区划分方法的差异而造成的结果差异,进行了一些分析。

除了使用 BGLL 算法[13]对微博 ε 近邻图[2]进行话题分割,我们采用被广泛使用的图聚类方法 metis[17]、经典社区划分算法 infomap[20]、基于模块度的快速社区划分算法 fastnewman[16]对微博 ε 近邻图进行话题分割。为了更全面地分析我们的实验结果,选取了 CV[18]值作为评价指标(表 3 中 CV 值为该社区内所有话题 CV 值的平均值),该评价指标由 Mimno[18]基于评估话题质量而提出。

给定一个话题 t 和它的描述主题词 $V(t) = (v_1(t), v_2(t), \cdots, v_M(t))$,则 CV 值定义为

$$C(t, v^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \quad (7)$$

式中:$D(v)$为包含词 v 的文档频次,$D(v, v')$为同时包含词 v 和 v'的文档频次。CV 值基于描述同一话题的词往往同时出现于同一文档中。CV 值越小,所得话题簇的一致性越好。

实验结果如表 3 所示。由表 3 可以知道,选择不同的方法对微博 ε 近邻图进行话题聚类,会得到不同的结果。在本实验中,metis 方法的效果总体上好于 BGLL 方法,但本文的方法只是面向用户兴趣和社区关系的话题检测框架的一个尝试,这类方法都可以找到用户群兴趣个性化的话题。

表 3　BGLL 算法与 metis、infomap、fastnewman 算法对照试验结果

Table 3　The controlled Trials result of BGLL with metis、infomap、fastnewman

划分算法	体育	公益	娱乐	时政	生活	科技	财经
bgll	−15.8	−6.3	−16.3	−22.3	−14.3	−15.3	−14.0
metis	−17.3	−0.5	−17.3	−8.8	−17.0	−8.0	−7.0
infomap	−4.0	−4.2	−8.0	−3.2	−8.4	−6.0	−3.2
fastnewman	−7.2	−5.2	−6.2	−5.2	−5.8	−5.4	−4.4

综上所述,本文提出的算法面向用户兴趣检测话题,基于词重要度的词过滤方法使得社区内的特征向量维度更低、更稠密,有效地解决了微博话题检测过程中出现的特征稀疏问题。与普通话题检测方法相比,该算法所检测话题更有可能被社区内用户所关注,提高用户活跃度。并且,本文采用主题度计算话题热度并排序,使话题展示顺序更加合理。

3　结束语

本文提出了一种基于用户兴趣与社区关系的微博话题检测方法,该方法能够快速准确地在社区内部检测话题,并对话题按热度进行排行。并且,该方法巧妙融合了新浪微博的社区特性与文本特性,检测的话题更加迎合用户的兴趣。

本文以主题词的形式来表现微博话题,但是本文对主题词采用硬划分,导致同一主题词只能属于唯一主题。但在真实情况下,可能多个话题含有同一主题词,如何实现将主题词划入多个话题,有待进一步研究。另外,以主题词表现话题并不是特别直观,如何实现以词组或句子表达主题,也有待进一步研究。

参考文献:

[1] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation[J]. The journal of machine learning research, 2003, 3(4-5): 993-1002.

[2] VON LUXBURG U. A tutorial on spectral clustering[J]. Statistics and computing, 2007, 17(4): 395-416.

[3] 郭庆琳, 李艳梅, 唐琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用研究, 2008, 25(11): 3256-3258.
GUO Qinglin, Li Yanmei, TANG Qi. Similarity computing of documents based on VSM[J]. Application research of computers, 2008, 25(11): 3256-3258.

[4] 周刚, 邹鸿程, 熊小兵, 等. MB-SinglePass: 基于组合相似度的微博话题检测[J]. 计算机科学, 2012, 39(10): 198-202.
ZHOU Gang, ZOU Hongcheng, XIONG Xiaobing, et al. MB-SinglePass: microblog topic detection based on combined similarity[J]. Computer science, 2012, 39(10): 198-202.

[5] 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): 138-141.
ZHENG Feiran, MIAO Duoqian, ZHANG Zhifei, et al. News topic detection approach on Chinese microblog[J]. Computer science, 2012, 39(1): 138-141.

[6] 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.
FANG Ran, MIAO Duoqian, ZHANG Zhifei, et al. An emotion-based method of topic detection from Chinese microblogs[J]. CAAI transactions on intelligent systems,

2013, 8(3): 004: 208-213.

[7] Heinrich G. Parameter estimation for text analysis[R]. Technical report, Darmstadt, Germany: Fraunhofer IGD, 2004.

[8] 乔健. 面向新浪微博的链接和内容相结合的社区划分方法[D]. 北京: 北京交通大学, 2015.
QIAO Jian. Community detection by using link and content and it's application in sina microblog[D]. Beijing: Beijing Jiaotong University, 2015.

[9] JIANG Yawen, JIA Caiyan, YU Jian. An efficient community detection method based on rank centrality[J]. Physica A: statistical mechanics and its applications, 2013, 392(9): 2182-2194.

[10] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the Web[R]. Stanford InfoLab, 1999: 189-194.

[11] KOJIMA K. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability[J]. American journal of human genetics, 1969, 21(4): 407-408.

[12] MIHALCEA R, TARAU P. TextRank: bringing order into texts[C]//Proceedings of EMNLP 2004: association for computational linguistics. Barcelona, Spain, 2004.

[13] CHATURVEDI P, DHARA M, ARORA D. community detection in complex network via BGLL algorithm[J]. International journal of computer applications, 2012, 48(1): 32-42.

[14] ZANGHI H, VOLANT S, AMBROISE C. Clustering based on random graph model embedding vertex features[J]. Pattern recognition letters, 2010, 31(9): 830-836.

[15] XU Zhiqiang, KE Yiping, WANG Yi, et al. A model-based approach to attributed graph clustering[C]//Proceedings of the 2012 ACM SIGMOD international conference on management of data. New York, NY, USA, 2012: 505-516.

[16] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physical review E, 2004, 69(6): 066133.

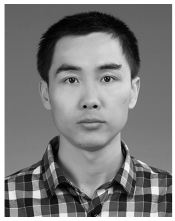
[17] KARYPIS G, KUMAR V. Metis-unstructured graph partitioning and sparse matrix ordering system, version 2.0[Z]. Minnesota: University of Minnesota, Department of Computer, 1995: 202-205.

[18] MIMNO D, WALLACH H M, TALLEY E, et al. Optimizing semantic coherence in topic models[C]//Proceedings of the conference on empirical methods in natural language processing. Stroudsburg, PA, USA, 2011: 262-272.

[19] HU Yanqing, LI Menghui, ZHANG Peng, et al. Community detection by signaling on complex networks[J]. Physical review E, 2008, 78(1): 016115.

[20] BURK C F, HORTON F W. Infomap: a complete guide to discovering corporate information resources[J]. Lincoln: Prentice Hall, 1988.

作者简介:



刘志雄,1990 年生,男,硕士研究生,主要研究领域为数据挖掘、机器学习、复杂网络。

贾彩燕,1976 年生,女,副教授,博士生导师,中国人工智能学会粗糙集与软计算专业委员会委员,主要研究方向为数据挖掘、社会计算、文本挖掘及生物信息学。近年来主持国家自然科学基金面上项目、青年基金面上项目各 1 项;参加国家自然科学基金重点项目、科技重大专项、北京市自然科学基金各 1 项;获湖南省科学技术进步二等奖 1 项。