ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 33,No. 1,Jan. 2018,pp. 186－194
DOI:10. 16337/j. 1004-9037. 2018. 01. 021

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax: ＋86-025-84892742

基于词共现网络的微博话题发现方法

李　伟[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. 交通数据分析与挖掘北京市重点实验室,北京,100044)

摘　要: 微博作为一个重要的信息平台,每天都有大量用户访问,重要的舆论事件在微博上会形成热门话题。本文提出了一种新的微博话题发现方法:基于词共现网络的话题发现方法(Topic detection in freqent word network,TDFWN),来挖掘微博语料中蕴含的热点话题。该方法首先对微博文本中的k频繁词集($k\geqslant 3$)进行挖掘,利用频繁词集的共现关系构建词共现网络。对该网络进行社区划分,同一社区内的词通常描述同一微博话题,即话题以社区的形式出现。实验结果表明 TDFWN 算法能够快速、全面地发现微博中的热门话题,并且可以实现微博文本的自动聚类。

关键词: 微博;话题发现;短文本;社区划分
中图分类号: TP391　　**文献标志码**:A

Micro-blog Topic Detection in Frequent Word Networks

Li Wei[1,2], Jia Caiyan[1,2]

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

Abstract: As an important information platform, micro-blog has a large number of user visits every day, and important public opinion events will form a hot topic on micro-blog. In this study, we propose a novel micro-blog topic detection method, named TDFWN (Topic detection in frequent word networks), to excavate hot topics in micro-blog corpus. First, frequent k-item sets ($k\geqslant 3$) in Microblog text data are mined. Second, a word co-occurrence network is build based on these mined frequent k-item sets. Third, the network is partitioned into different communities by using a community detection method, where each community represents a micro-blog hot topic. At last, the micro-blog text data are clustered into different groups by computing similarity of each micro-blog text with the found topics. The empirical study shows that the TDFWN method is able to find hot topics in micro-blog text data and cluster the micro-blog text data by the found topics simultaneously.

Key words: micro-blog; topic detection; short text; community detection

引　　言

微博作为一种重要的信息交流平台,得到了广泛的关注和使用。由于微博是一个公平、开放的交流

基金项目:国家自然科学基金(61473030)面上资助项目;中央高校基本科研业务专项资金(2014JBM031)资助项目;数字出版国家重点实验室专项课题。
收稿日期:2015-05-25;**修订日期**:2015-07-08

平台,越来越多的组织、机构通过官方微博发布信息或对突发事件做出回应,明星和企业也将微博当作自我展示的一个重要渠道。同时,一些事件通过微博用户的爆料后成为热点话题被大众所关注。微博颠覆了传统的信息传播方式,以媒体为主导的新闻传播方式正在面临挑战。微博平台使得人人都能成为记者,人人都可以报道新闻。正是因为微博信息传递过程中的方便和自由,对微博平台的管理和微博信息分析变得更具挑战性。微博平台拥有着庞大的数据量,如何挖掘和处理这些数据已成为海内外学者的研究热点,其中微博平台的话题发现方法是微博研究领域的一个重点课题。

传统文本处理领域中使用的话题检测与追踪(Topic detection and tracking, TDT)[1-2]技术已经日趋成熟。在进行微博话题发现的时候借鉴了传统话题检测与跟踪中使用的方法。其大体思路是:以词为特征使用向量空间模型(Vector space model,VSM)[3]将微博文本转化到空间向量,并且使用词频-逆向文档频率(Term frequency-inverse document frequency,TF-IDF)方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划分成一个微博话题簇。周刚等人[4]提出了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;郑斐然等人[5]提出了一种基于词聚类的新闻话题发现方法;Li 等人[6]提出了一种改进的基于增量聚类的微博话题发现方法。然而由于微博文本被限制在了 140 个字以内,使用空间向量对微博文本进行建模存在严重的数据稀疏和维度过高问题。为了解决这个问题,Huang 等人[7]提出了一种基于 LDA 主题模型和潜在语义分析的微博话题发现方法。由于微博主题分散、更新速度快且数据量大,使用 LDA 主题模型进行微博话题发现时存在计算量大的问题。赵文清等人[8]提出了一种基于网络图的微博新闻话题发现方法,该方法先找到微博文本中出现的高频词,然后计算高频词之间的共现度,使用共现度高于阈值共现词,构建共现图呈现潜在的话题。

词共现关系常被用在分析各个学科研究领域的研究主题[8-9],是指几个词在同一文章或者句子、段落而构成的共现关系。当几个词频繁地出现在一起的时候,它们之间很可能存在语义上的关系。随着复杂网络研究的深入,复杂网络中的一些方法在多个领域得到了应用。本文将复杂网络中的社区发现方法应用到微博主题识别领域,提出了一种基于词共现网络的微博话题发现方法:基于词共现网络的话题发现方法(Topic detection in freqent word network,TDFWN)。该方法首先挖掘微博中的 k-频繁词集($k\geqslant 3$),对微博 k-频繁项集中的词构建词共现网络,然后利用社区划分的方法对其进行社区划分。每一个社区对应一个微博热点话题。最后,以这些话题社区为聚类中心点,找到微博话题簇。实验结果显示 TDFWN 方法进行微博话题发现时,能够准确快速地找出微博中的热门话题,并且实现了对大量短文本的快速聚类。

1　TDFWN 方法

TDFWN 方法基于词共现模型基本假设:在大规模的语料中,如果某些词经常共同出现在同一窗口(如一句话,一条微博),则他们在语义上是关联的。共同出现的一组词,通常会被用来表述同一个主题。基于词共现社区的微博话题发现方法 TDFWN 可分为 4 个步骤:数据预处理、频繁词集挖掘、词共现社区发现及微博话题簇获取(算法流程见图 1)。数据预处理对微博数据进行筛选和切词,并且过滤掉一些微博平台常见的无主题高频词组,例如:“转发微博”“美图秀秀手机端发送”等。频繁词集挖掘使用 FP-Growth 算法挖掘微博文本中频繁出现的词集,然后利用频繁词集中的共现关系构建共现网络。词共现社区发现对词共现网络进行社区划分,划分出的每一个社区都是一个微博话题。微博话题簇获取利用计算微博话题与微博文本相似度的方法找到微博话题簇。

1.1　数据预处理

微博数据的预处理包括对微博的筛选、微博文本切词以及特殊词过滤。微博数据是一系列相互独立的短文本,通常微博文本不超过 140 个字。微博文本中存在大量的非正式用语以及表情符号。其中,

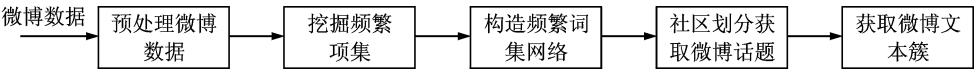

图 1 TDFWN 方法流程

Fig. 1 Flow chart of TDFWN method

“@用户名”这种表达方式用来提到某个用户,“#主题#”用来参与某个主题的讨论。微博的内容大部分是无主题的,为了提高话题发现的效果,本文采用以下方法对微博数据进行过滤。

(1) 忽略粉丝数量和关注数量小于阈值的用户所发布的微博,粉丝和关注较少的用户可能是僵尸用户或者是不活跃用户。

(2) 忽略微博数量少于阈值的用户所发布的微博,发布微博少的用户,在微博平台上不活跃,他们的微博很少涉及热点话题。

(3) 忽略文本长度小于阈值的微博,文字数量较少的微博通常没有明确的主题。

(4) 忽略某些特殊用户所发布的微博,微博平台上存在着一些特别用途的账号,例如:发广告、推销产品、发布笑话等。这些用户发布的微博有特定的目的,对话题发现有一定的干扰。

中文切词工具有很多种[10],如中科院的 NLPIR/ICTCLAS 汉语分词系统和 Java 开源工具 Jcseg。Jcseg 是一款开源的中文切词工具,Jcseg 在进行中文切词时有 3 种模式:①简单模式-FMM 算法,适合速度要求场合。②复杂模式-MMSEG 算法,能够有效地去除歧义,分词准确率达到了 98.41%。③检测模式,只返回词库中已有的词条,很适合某些应用场合。另外,Jcseg 较好地支持了地名、人名等专有名词,还支持自定义词典,可以按照自己的需求在词库里加词。Jcseg 还提供了过滤功能,被加入到过滤列表的词,在分词过程中会被自动过滤掉。

停用词是指在自然语言中具有一定功能但又没什么实际意义的词。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,在自然语言处理过程中往往会将其去掉。本文在切词过程中将出现的停用词去掉。微博系统中有一些词和词组,例如:“转发”“微博”“手机微博客户端发送”等,会以较高的频率出现,但没有实际含义,也和停用词一起去掉。

1.2 频繁词集挖掘

词共现模型是统计自然语言处理领域的重要模型之一[11]。词共现模型假设在大规模的语料中,如果某些词经常共同出现在同一窗口(如一句话,一条微博),则他们在语义上是关联的。共同出现的一组词,通常会被用来表述同一个主题。例如:在同一时间段的微博文本中,“冰桶”“挑战”“ASL”这组词以较高频率共同出现在同一条微博中。这 3 个词的组合描述了“ASL 冰桶挑战赛”这一话题,包含这 3 个词的微博很可能和这一话题相关。

定义 1 频繁词集:设 $W=\{w_1,w_2,w_3,\cdots,w_n\}$ 为 n 条微博文本,微博文本 $w_i(i=1,2,\cdots,n)$ 的词集合为 C_i,$C_i=\{c_{i1},c_{i2},c_{i3},\cdots,c_{im}\}$,其中 c_{ij} 是 w_i 的第 j 个词。T 为 W 中所有微博词的集合,$T=\{t_1,t_2,t_3,\cdots,t_N\}$。词集 U 为 T 的子集,定义 U 的支持度 $\mathrm{SUP}(U)=W(U)$,其中 $W(U)$ 为 W 中包含词集 U 的微博的条数。当 $\mathrm{SUP}(U)$ 大于阈值 θ 的时候,称 U 为一个频繁词集。

本文使用 FP-GROWTH 算法[12]挖掘 k 频繁词(k 指频繁项集项的个数)。将经过数据预处理的微博文本作为频繁词集挖掘的输入,最小支持度 SUP 取 3%。k 为 1 和 k 为 2 的频繁词集包含大量的噪声,因此将其舍弃。

1.3 词共现网络构建

使用从微博网络中挖掘出的 k 频繁词集($k\geqslant 3$)中的词汇作为网络中的点,词和词之间的共现关系作为边,构建词共现网络。使用复杂网络中常用的数据格式 NET 来描述词共现网络,常用的复杂网络

分析工具,例如:Pajek(Pajek 是大型复杂网络分析工具,是用于研究目前所存在的各种复杂非线性网络的有力工具)[13]、Gephi(Gephi 是一款开源免费跨平台基于 JVM 的复杂网络分析软件)[14]等都支持该格式的网络描述文件。NET 文件分为 * Vertices 和 * Edges 两部分内容,Vertices 描述了网络中存在的节点,Vertices 描述了节点之间的关系。假设{A, B, C}是从微博语料中挖掘出的一个频繁词集,将该集合以 NET 格式进行描述,结果如下所示。

```
* Vertices
1 "A"
2 "B"
3 "C"
* Edges
1 2
1 3
2 3
```

将从微博文本中挖掘到的全部频繁项集整合成节点的集合 Vertices Set 和边的集合 Edges Set,两个集合依次输出到 NET 文件中。

1.4　微博话题簇获取

社区是复杂网络中的常见现象,它由一群高度聚集、紧密联系的节点聚集而成。社区是一种介于宏观和微观之间的网络特征。在真实网络中,同一个社区的节点往往具有相似的性质或者相近的功能。在以频繁词集为基础构建的词共现网络中,同一社区内的词通常描述同一话题,即话题以社区的结构出现。当两个话题有具有较多共同的特征词时,将会出现重叠的社区结构。这种重叠很可能是两个话题语义上相关造成的,也可能仅仅是因为有相同的特征词。

本文在检查频繁词集共现网络中存在的社区结构时,使用的是经典的社区发现算法 GN 算法[13]。GN 算法是一种采用分裂思想的算法,在执行社区发现任务时通过不断地移除边介数最高的边来对网络进行分类。边介数是网络中的边所具有的一种属性,是指在一个网络中全部经过了这个边的两个点的最短路径的个数与网络中所有经过了这个边的路径的个数的比值。GN 算法是一种层次化的社区发现算法,最后能得到不同层次的社区结构,GN 算法的执行流程如下所示:

(1) 依次地算出待挖掘的网络中每一条边的边介数;

(2) 找到网络中边介数最大的一条边然后将它删除;

(3) 重新计算剩下的所有边的边介数;

(4) 重复上述几个步骤,只到所有的边都删除为止。

为了使话题发现结果更加直观,TDFWN 算法在得到微博话题社区后,以同一社区内的词(微博话题词集)作为聚类中心点,对 W 中的 n 条微博文本进行聚类,以找到同一微博话题下的微博话题簇。在进行聚类时使用单遍聚类方法,利用式(1) 计算微博与微博话题词集之间的相似度 S,当微博与微博话题词集的相似度 S 大于阈值时,认为该微博是这个话题下的微博。

设 C,H 为两个词集 $C=\{c_1,c_2,c_3,\cdots,c_t\}$,$H=\{h_1,h_2,h_3,\cdots,h_m\}$。计算两个词集相似度的时候,引入函数 $R_{(C,H)}$ 表示词集 C 相对于 H 的相似度,表达式为

$$R_{(C,H)}=\frac{C\cap H}{C} \tag{1}$$

进而,定义 C 与 H 相似度 $S_{(C,H)}$ 为

$$S_{(C,H)}=\frac{R_{(C,H)}+R_{(H,C)}}{2} \tag{2}$$

当 H 与 C 相似度 $S_{(C,H)}$ 大于某阈值时认为 H 与 C 是相似的。

传统的文本聚类方法处理微博数据时，因为微博数据中存在大量噪声，导致大量的微博文本无主题。因此会得到大量的无主题微博簇，聚类结果不理想。TDFWN 算法首先找到微博话题的词集，确定了聚类的中心(即微博话题)，可以快速准确地对微博文本聚类。

2 实　　验

基于共词网络的微博话题发现方法较少，目前只有赵文清等[8]基于词共现图的中文微博新闻话题识别方法。因此，本实验采用赵文清等人基于词共现图的中文微博新闻话题识别作为对照实验。但本文与对照实验的方法之间的差别主要在于：(1)构建词共现网络时所使用的关系不同，本文使用微博文本中词的 k 频繁项集($k \geqslant 3$)构建词共现网络，而赵文清等人的方法利用高频词之间的共现关系构建网络。因为本文过滤掉了 $k \leqslant 2$ 时的共现关系，因此能够有效消除大量噪声。(2)对照实验以图的方式呈现微博话题发现结果，当节点和边过多时无法直观地看出话题，不能自动区分微博话题。本文采用了社区发现的方法对词共现网络进行社区划分，当节点和边较多时也能以社区的方式将微博话题清晰地呈现。(3)本文在发现微博话题的同时，实现了微博文本的聚类。在发现微博话题社区后，以话题社区内的关键词为聚类中心，采取单遍聚类的方法，将相似度大于阈值的微博文本分配到同一微博话题簇内。

2.1 实验环境

电脑型号：联想 ThinkPad X230 笔记本电脑。

操作系统：Windows 7 专业版 64 位 SP1。

处理器：英特尔 第三代酷睿 i5-3320M @ 2.60 GHz 双核。

主板：联想 23255NC。

内存：4 GB (三星 DDR3L 1 600 MHz)。

主硬盘：三星 MZ7TD128HAFV-000L1(128 GB/固态硬盘)。

2.2 实验数据

实验数据集是自然语言处理与信息检索共享平台公开的 NLPIR 微博内容语料库(见表 1)。从新浪微博获取实验数据有两种方式：调用微博 API，网络爬虫抓取。调用微博 API 会受到新浪微博系统 API 调用规则的限制，无法大规模获取微博数据。使用网络爬虫抓取微博数据可以获得大量的微博数据，但由于微博系统反爬虫措施，在技术上较难实现。

2.3 FSWCN 实验过程与结果

对 NLPIR 微博内容语料库 2012-02-01 的 1 586 条微博数据进行实验。首先进行频繁词集的挖掘，选取最小支持度 MINSUP 为 0.01，频繁词集挖掘部分结果如表 2 所示。

表 1　微博数据

Tab. 1　Micro-blog data

实验数据集	条数/万条	获取方式
NLPIR 微博内容语料库	23	NLPIR

表 2　频繁词集部分结果

Tab. 2　Frequent words set

编号	频繁词集	SUP	编号	频繁词集	SUP
1	不住　摊　撤	24	5	易　吴英　中天	28
2	摊　撤　摊主	24	6	请　韩寒　吴英	24
3	经营　摊　鱼	24	7	请　子　韩寒　方舟	27
4	城管　围观　鱼	24	8	执法　城管　南京　围观　小贩	24

频繁词共现网络如图 2 所示,对该网络进行社区划分,结果如图 3 所示:本次试验共发现 5 个微博话题,分别对应深蓝、浅蓝、红色、绿色和紫色 5 个共现网络社区。这 5 个社区对应的微博话题如表 3 所示。

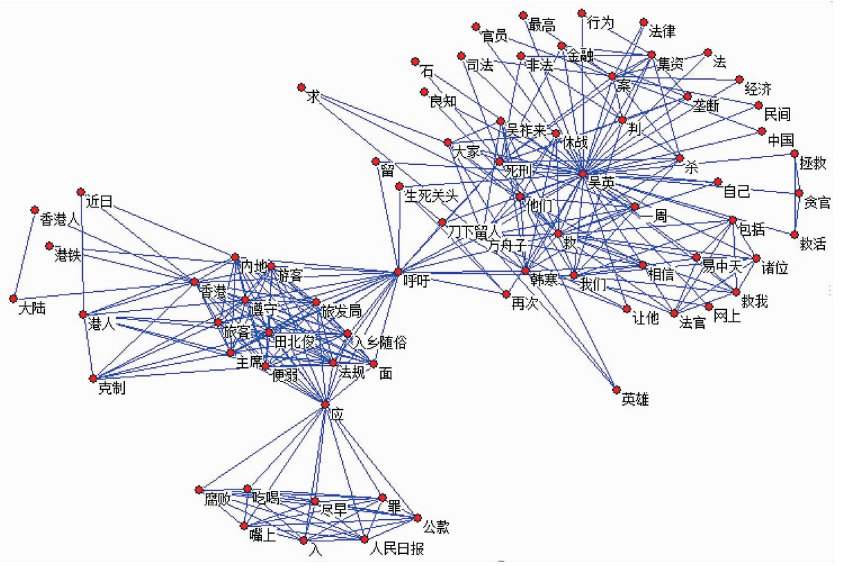

图 2　频繁词共词网络

Fig. 2　Frequent co-word net

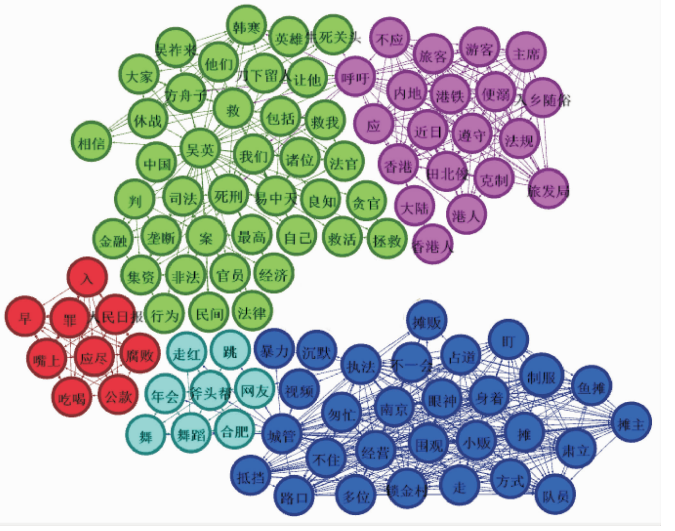

图 3　社区划分结果

Fig. 3　Result of community partition

表 3　微博话题

Tab. 3　Microblog topic

编号	颜色	对应微博话题	编号	颜色	对应微博话题
1	深蓝	南京金锁村城管眼神盯走小贩	4	绿色	吴英案网民呼吁刀下留人
2	浅蓝	合肥城管年会跳斧头帮舞蹈	5	紫色	大陆游客香港扰乱秩序引发冲突
3	红色	人民日报报道公款吃喝腐败			

使用式(2),设定相似度 S 阈值为 0.2,进行单遍聚类得到微博话题簇,部分结果为:

(1) 深蓝,南京金锁村城管眼神盯走小贩。

南京锁金村一处路口,十多位身着制服的城管队员肃立围观占道经营的鱼摊,不一会,摊主便抵挡不住,匆忙撤摊。城管队伍能够在媒体的不断曝光下在执法方式上求变,这起码算是一种进步。

(2) 浅蓝,合肥城管年会跳斧头帮舞蹈。

身着黑衣、手拿板斧、群魔乱舞……传说中的斧头帮来了? 这是合肥高新区城管局年会上的舞蹈串烧,整段舞蹈由斧头舞、甩葱舞、草裙舞组成。网友质疑尺度略大,城管回应只为自娱。——城管跳起斧头帮舞蹈,土匪形象和抢葱归来欢庆场面,创意反变成妖魔化自己,弄巧成拙呀。

(3) 红色,人民日报报道公款吃喝腐败。

"贪污和浪费是极大的犯罪",这是句众人皆知的话。实际上,公款大吃大喝既是贪污,也是浪费。我国也因此成为泔水大国,以至于地沟油泛滥成灾,连政府机关食堂也未能幸免。"嘴上腐败"应尽早入刑治罪。

(4) 绿色,吴英案网民呼吁刀下留人

老易始终是明白人。以他和韩寒的交情,现在却更关注吴英! 李庄,易中天:请最高院的法官大人刀下留人,最好能够重审! 至少,不要马上签署死刑命令。救人一命,胜造七级浮屠。今天救下吴英,明天就会有更多的人来救我们,包括诸位法官。大人勾决的朱笔只要现在停住,就是为法治积德,也是为自己积德!

(5) 紫色,大陆游客扰乱秩序引发冲突。

香港特区旅发局主席田北俊关注近日港人与内地旅客有争拗,他呼吁双方克制。他认为内地旅客到访香港时,应入乡随俗,遵守香港法规,不应在地铁上吃东西或者随处便溺。

2.4 对照实验

按照文献[14]中的参数设置进行共现词挖掘,得到共现度结果(见表 4),使用共现度结果构建共现网络,得到词共现图(见图 4)。从图 4 中可以得出微博语料中存在的话题(见表 5)。

表 4 共现度结果

Tab. 4 Degree of co-word

共现词	共现度	共现词	共现度	共现词	共现度
死刑-吴英	0.623	制服-城管	0.539	执法-城管	0.602
集资-吴英	0.523	南京-围观	0.702	小贩-城管	0.560
韩寒-方舟	0.634	大陆-香港	0.562	舞蹈-斧头	0.780
围观-城管	0.543	队员-南京	0.677	摊主-队员	0.757

2.5 实验分析

TDFWN 算法没有对词性进行过滤,因此 TDFWN 算法结果图中包含的节点和边较多,使得每一个话题下的词集元素较多。较多的词汇有助于提升聚类效果。对照实验只保留了动词和名词,所得结果图较为简洁直观。TDFWN 算法因为使用了社区划分,将不同热门话题下的关键词分成不同的社区,并且着以不同颜色,能够直观生动地看出微博话题。对照实验在边和节点较少时能够直观地看出微博话题,为了达到能够直观看出微博话题这一效果,设置了较高的共现度阈值并过滤了词性,信息有所缺失,使得微博话题发现结果有所缺失,部分话题没有找到。对照实验最后需人来识别出图中蕴含的话题,当微博数据中同属含有多个话题,形成的网络将难以识别。TDFWN 算法能够找到各个话题的特征词集,对数据敏感性低,同时可以实现聚类。

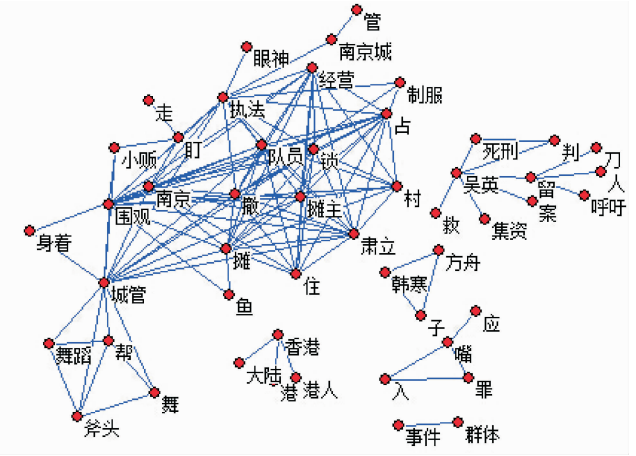

图 4　对照实验结果

Fig. 4　Result of contrast experiment

表 5　热点话题

Tab. 5　Hot topic

编号	热点话题表示	编号	热点话题表示
1	吴英、死刑、呼吁、集资	3	南京、小贩、围观、舞蹈
2	香港、大陆、港人	4	大陆、香港、港人

3　结束语

本文提出了一种微博话题发现方法 TDFWN,该方法能够快速准确地发现微博数据中潜在的话题,并实现微博文本的自动聚类。TDFWN 将已有的复杂网络社区划分算法应用到短文本处理中,以词社区的形式确定了微博文本聚类时的中心,利用聚类中心可以快速地实现聚类。目前关于频繁词集网络的研究较少,除 TDFWN 算法外暂未发现使用频繁词集网络中存在的社区结构进行话题发现的算法。频繁词集共现网络是本文提出的一种全新的短文本模型,基于该模型除了识别微博话题外还能用作短文本聚类。使用高频词共现网络(Frequent word network,FWN)对文本进行聚类的方法在本文中已作介绍,FWN 网络的一些其他应用正在研究中。

随着网络科学的发展,复杂网络的理论和科研成果在许多方面得到了应用。本文将复杂网络中的方法应用到微博话题发现中,结果表明微博词集间的共现网络结构能够有效地发现微博中的热点话题。微博是一个融合了社交网络、兴趣网络、主题网络等多个因素的大型复杂网络系统,如何将复杂网络的方法及微博系统中的这些多模态信息相融合,进一步提高对微博数据的信息抽取效用,还有待进一步研究。

参考文献:

[1]　洪宇, 张宇, 刘挺,等. 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):71-87.
Hong Yu , Zhang Yu, Liu Ting,et al. Topic detection and tracking review[J]. Journal of Chinese Information Processing, 2007, 21(6):71-87.

[2]　宗慧, 刘金岭. 基于短文本信息流的热点话题检测[J]. 数据采集与处理, 2015,30(2):464-468.
Zong Hui, Liu Jinling. Hot topic detection based on short text information flow[J]. Journal of Data Acquisition and Processing, 2015,30(2):464-468.

[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 base 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 Zheifei, et al. News topic detection approach on chinese microblog[J]. Computer Science, 2012, 39(1):138-141.

[6] Li G, Meng K, Xie J. An improved topic detection method for Chinese microblog based on incremental clustering[J]. Journal of Software, 2013, 8(9): 2313-2320.

[7] Huang B, Yang Y, Mahmood A, et al. Microblog topic detection based on LDA model and single-pass clustering[C]// Rough Sets and Current Trends in Computing. [S. l.]:Springer Berlin Heidelberg, 2012: 166-171.

[8] 赵文清，侯小可. 基于词共现图的中文微博新闻话题识别[J]. 智能系统学报，2012，7(5):444-449.
Zhao Wenqing, Hou Xiaoke. News topic recognition of Chinese microblog based on work co-occurrence graph[J]. CAAI Transactions on Intelligent Systems, 2012, 7(5):444-449.

[9] 刘则渊，尹丽春. 国际科学学主题共词网络的可视化研究[J]. 情报学报，2006(5):634-640.
Liu Zeyuan, Yin Lichun. Visualization of international science of science co-word network[J]. Journal of the China Society for Scientific and Technical Information, 2006(5):634-640.

[10] 黄昌宁，赵海. 中文分词十年回顾[J]. 中文信息学报，2007,21(3): 8-19.
Huang Changning, Zhao Hai. Chinese word segmentation:A decade review[J]. Journal of Chinese Information Processing, 2007,21(3):8-19.

[11] Hankerson D, Hernandez J L, Menezes A. Software implementation of elliptic curve cryptography over binary fields[C]// Cryptographic Hardware and Embedded Systems—CHES 2000. Berlin Heidelberg:Springer, 2000:1-24.

[12] Han J, Pei J, Yin Y, et al. Mining frequent patterns without candidate generation:A frequent-pattern tree approach[J]. Data Mining and Knowledge Discovery, 2004(8):53-87.

[13] Heymann S. Gephi[M]. New York:Springer, 2014.

[14] Batagelj V, Mrvar A. Pajek[J]. Encyclopedia of Social Network Analysis & Mining, 2014, 39(6):114-115.

作者简介：

李伟(1989-)，男，硕士，研究方向：数据挖掘与机器学习，E-mail:623267658@qq.com。

贾彩燕(1976-)，女，副教授，研究方向：数据挖掘、复杂网络聚类分析、生物序列分析，E-mail: cyjia@bjtu.edu.cn。

（编辑：夏道家）