

# 基于社交网络分析的诈骗团体挖掘方法研究

贾志娟 赵 靓 周 娜

( 郑州师范学院 信息科学与技术学院 河南 郑州 450044)

**摘 要:** 微博作为一种重要的社交方式,逐渐融入大众生活,用户在平台上可以随时随地抒发个人情感、分享信息等。微博在给人们带来信息传递之便利的同时,也带来了不少不法分子利用其进行诈骗的问题。诈骗团体利用微博设置语言陷阱,以此骗取他人钱财、夺取他人利益。对此,利用社群图表示微博社会网络,该网络是一有向图,节点表示微博用户,连接线表示微博传播路径,以此连接微博发起节点和微博转发节点。另外,研究社会网络分析的方法和数据挖掘的技术,对诈骗团体进行分析,对该团体应具有的组织结构、特性进行定义,分析出微博中诈骗团体应该具备的特征,并以此寻找微博中潜在的诈骗团体,帮助用户识别诈骗,避免上当受骗。最后用案例验证了该方法的有效性。

**关键词:** 社会网络分析; 数据挖掘; 诈骗团体; 特征向量

中图分类号: TP301

文献标识码: A

文章编号: 1673-629X(2018)05-0090-04

doi: 10.3969/j.issn.1673-629X.2018.05.021

## Research on Mining Fraud Group Based on Social Network Analysis

JIA Zhi-juan ZHAO Liang ZHOU Na

( School of Information Science & Technology Zhengzhou Normal University Zhengzhou 450044 ,China)

**Abstract:** As an important social way ,Weibo has been integrated into public life.Users can express personal feelings and share information anytime and anywhere on the platform.In addition to the convenience of message transmission ,Weibo also brings a lot of criminals to use it for fraud.The fraud groups use Weibo to set the language trap for money and interests of others.For this ,we use the social diagram to express Weibo social network which is a directed graph.Nodes represent Weibo users and the connection lines represent Weibo propagation path to connect Weibo initiated node and Weibo forwarding node.Besides ,we use the method of social network analysis and data mining technology to analyze the fraud group ,thus defining its organizational structure and characteristics ,finding out the potential fraud group in Weibo ,which helps users identify the fraud and avoids being deceived.Finally ,the validity of the method is verified according to a case.

**Key words:** social network analysis; data mining; fraud group; feature vectors

## 0 引 言

根据 2015 年发布的《中国互联网发展状况统计报告》,截至 2016 年 6 月,中国网民规模达 7.10 亿,互联网普及率达到 51.7%,超过全球平均水平 3.1 个百分点。同时,国内微博用户总量从 12 年的 2.74 亿增长到 15 年底的 5.03 亿,可见增长之迅速。作为一种新兴的社交媒体,微博作为信息发布和传播的主流平台,正在逐渐改变着人们的生活方式。网民们热衷于在微博上分享自己的心情,评论当前流行的元素,探讨当今的社会热点,并关注自己的喜好,这给微博的数据挖掘带来了相当大的价值,同时也使得诈骗团体的行动更

加便利<sup>[1-2]</sup>。

社会网络以用户为基础,具有主体繁多、用户影响力差异显著、用户特征与信息资源复杂且事件突发性强等特性。中国社会网络环境比较复杂,尤其是诈骗谣言等信息对社会的影响较大,引导不当极易引发社会矛盾。微博的出现进一步推动了社会网络的发展。而且微博具有较为活跃的用户量,若仅仅依靠传统的统计方法无法高效地提取有价值的信息,这就急需一种更高效的技术能对海量文本数据进行分析 and 挖掘,社会网络分析和数据挖掘技术应运而生。因此利用社会网络分析和数据挖掘技术对微博中诈骗团体的语言

收稿日期: 2017-04-10

修回日期: 2017-08-16

网络出版时间: 2018-02-07

基金项目: 国家自然科学基金( U1304614 ,U1204703 ); 郑州市创新型科技人才队伍建设工程基金项目( 131PCXTD597 ); 河南省科技攻关项目( 162102310238 )

作者简介: 贾志娟( 1973- ),女,教授,在读博士,CCF 会员( 26775M ),研究方向为数据挖掘。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.TP.20180207.1525.008.html>

行为进行研究是可行的而且是很有必要的<sup>[3-4]</sup>。

对此,国内外相关学者做了大量研究。张劲捷等将垂直搜索的技术、文本分析和挖掘的技术应用于微博的舆情分析,分析了网络热点话题的发现模型等,并设计了一个基于微博设计网络的舆情分析系统<sup>[5]</sup>;缪茹一等对微博进行细粒度情感分析,将情感分为七种类别,提出了融合微博显性和隐形特征的情感聚类方法,开发出一个情感分析与监控系统<sup>[6]</sup>。国外微博的发展始于2006年,是由Evan Williams推出的Twitter把人们引入微博的世界,从而一系列关于微博的研究相应出现<sup>[7]</sup>。

鉴于国内外对微博中诈骗团体的挖掘方面的研究较少,因此在前人研究的基础上,通过结合社会网络分析方法和数据挖掘算法,分析出诈骗团体应该具备的特征属性,从而挖掘出微博上潜在的诈骗团体,帮助网民减少受骗。

## 1 相关理论知识

### 1.1 社会网络分析

社会网络是指社会行动者及其关系的集合。一般情况下,社会网络的形式化界定用点和线来表示网络,社会网络可简单地理解为各种社会关系交织成的结构<sup>[8-9]</sup>。社会网络的形式化可表达包括社群图和矩阵两种方法。其中社群图用于表示一个社会群体成员之间的复杂关系,由表示社会成员的和线连成的图构成。举例说明,图1为一个简单的微博传播的社群图,抽象出关系为:用户A发一条微博消息 $M_0$ ,然后B进行转发生成 $M_1$ ,继而C、D、E进行转发 $M_1$ 分别生成 $M_2$ 、 $M_3$ 、 $M_4$ ,然后博主F转发 $M_2$ 生成 $M_5$ ,G转发 $M_5$ 生成 $M_6$ 。

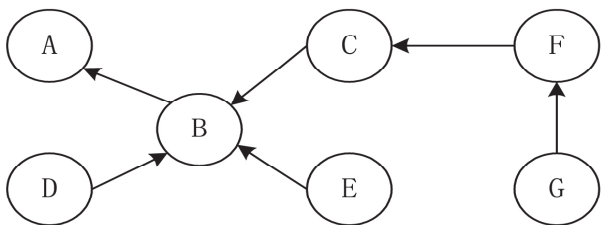


图1 微博的传播社群图

社会网络分析是一个针对社会网络的内部结构和节点之间的关系进行分析和解释的重要工具。通过社会网络分析可以了解社会网络的内部特性和节点之间的角色关系。其中用来表示社会网络内部特征的组件包括社会网络的密度、中间程度、各节点的角色等,以此为基础来分析社会网络的内部结构<sup>[10-12]</sup>。

### 1.2 文本特征选择

特征选择的过程是一个挑选文本特征的过程,首先要查找能够表示训练文本的特质集合,然后按照评

估函数从中挑选出对分类有较高贡献的特征项构成特征子集。最常用的方法有TF\*IDF、互信息/信息增益、期望交叉熵等,文中使用应用较为广泛的TF\*IDF法。

Salton在1988年提出使用TF\*IDF法计算单词权重,其中TF为词频,即特征词在文本中出现的频率,用于表示该词描述文档内容的能力;IDF为反文档频率,即 $\lg(N/n+0.01)$ (其中 $N$ 是文本总数, $n$ 是出现该词的文本数),用于计算该词区分文档的能力。该方法认为:如果某个单词在一个文本中频繁出现,那么它在另一个相同类型的文本中出现的次数也会很多,反之也成立。同时,如果一个单词出现的文本频率越小,则认为该单词的文档区别能力越强,因此引入反文档频率,最终以TF和IDF的乘积来定义特征空间坐标系的值<sup>[13-15]</sup>。

## 2 微博中诈骗团体的模型构建与实现

利用社会网络分析与数据挖掘算法分析出微博中诈骗团体应该具备的特征,并挖掘出新浪微博上潜在的诈骗团体。为了实现这个目的,文中的实证主要分为以下四个步骤:微博数据采集;网络爬虫程序的开发;文本数据清洗;诈骗团体的社会网络特征和文本特征的挖掘;诈骗团体预测及评估。流程如图2所示。

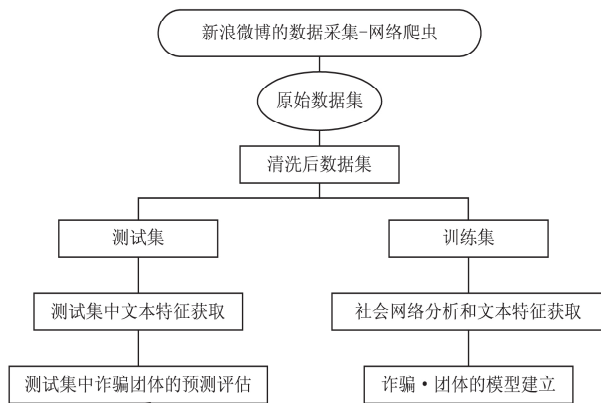


图2 研究流程

### 2.1 微博数据采集: 网络爬虫程序的开发

要分析微博平台上的诈骗团体,首先要对微博上关于诈骗的热点、文章和相关评论数据进行采集,对比多种网络上的爬虫工具。选用C#自己开发出一款爬虫软件,对比C#的网络库采集数据。设定微博为新浪微博,通过关键词“仇恨”字段获取相关的URL,使用C#的httpwebrequest类库访问URL获取返回结果,解析出需要的数据。

### 2.2 数据处理

取到文本信息后,首先要对文本数据进行清洗。文本数据里存在大量的冗余杂乱的数据,好多基本上

是没有任何价值的,如果将这些数据也引入到词频统计里,必然对模型的最终结果产生深远的影响。因此在建模之前需要对文本数据进行预处理,删除掉大量无价值的信息,包括去重、机械压缩去词和短句删除。

数据经过预处理之后,在进行数据挖掘之前还需要对文本数据进行分词处理,即将连续的字序列按照一定的标准重新组合成词的过程。而且不同的分词效果会直接影响到词语在文本中的重要程度,因此采用 Python 中评价较高的中文分词包“jieba”对文本数据进行断词,最后整理成有字词组成的数据集。

生成数据集之后,还需要计算一个词对于一个文本的重要程度,最常用的方法就是 TF-IDF 算法。某个词对文章的贡献度越大,它的 TF-IDF 值就越大,所以按 TF-IDF 值从大到小排序,排在最前面的就是文章的关键词,也就是特征值。其中:

TF = 单词在文章中出现的次数 / 文章的总次数

IDF =  $\log(\text{文章总数} / \text{包含该词的文章数} + 1)$

TF-IDF = TF \* IDF

### 2.3 诈骗团体社会网络分析特征获取和文本特征的挖掘

#### 2.3.1 通过社会网络分析特征获取

首先从整体社会网络的角度对诈骗团体进行社会特征值的挖掘,主要从两个方面进行分析,一是网络密度,二是平均最短路径。

通过网络密度分析可以对诈骗团体的训练集中社会网络之内部节点互动链接的强度进行大致的了解,密度高的社会网络通常代表与中心高度密集或高度相关而且信息传递速度更快。平均最短路径可用于衡量社会网络中,信息在节点与节点之间的传播效率,路径越短,传递信息的效率越高。

其次从网络节点的角度进行分析,主要对节点的连接度和中心性进行分析。通过分析节点与节点之间的连接度,可以了解节点在整个社会网络中的活动范围,而对中心性进行分析,主要是用于衡量单个节点在网络中的重要程度,可以借此来掌控整个社会网络的主要信息流向等,其目的就是为了挖掘出社会网络中的重要节点。

通过对诈骗团体的中心性进行分析,可以挖掘出诈骗团体内部各节点之间的角色担当以及诈骗团体内部各节点的结构,以此来担任诈骗团体的特征值,方便后续挖掘研究。

#### 2.3.2 通过文本挖掘获取特征关键词

在对训练集中的数据进行处理之后,对得到的数据集计算所有字词的 TF-IDF 特征值,然后进行排序。此外,从对诈骗团体的分析观察中发现,诈骗团体的目的在于传播诈骗信息,从而让更多的人上当受骗,其用

词多半强烈且频繁。因此该研究只取微博中关于诈骗的训练样本集中 TF-IDF 排名前 10 的词作为关键词,即此微博的内容特征词。

### 2.4 诈骗团体的预测和评估

该步骤的目的是为了验证上述特征值的确立可否通过对比挖掘出潜在的诈骗团体。主要包括两步:内容特征对比和社会网络特征对比。

#### 2.4.1 内容特征对比

在进行社会网络特征对比之前,需要先比对训练集中得到的内容特征向量与测试集中的内容特征向量的相关程度,判断测试集中的内容特征值与既有的训练集中的内容特征之间的相似度(similarity)。通过上述分析,可以得到测试集中与诈骗内容高度相似的族群,并将此族群列为潜在的诈骗团体。

#### 2.4.2 社会网络特征对比

对于上述分析得到的潜在的诈骗团体,通过分析比对这些潜在的诈骗团体所构成的社会网络特征与训练集中得到的社会网络特征是否存在高度相关性,判断该族群是否真的是诈骗团体。

首先对潜在的诈骗团体进行社会网络构建和分析,建立相同的社会网络特征向量,然后进行特征向量之间的相似度对比,进而判断是否为真的诈骗团体。整体社会网络特征向量  $G_n = [\text{平均连接度}, \text{网络密度}, \text{平均最短路径}]$ ,  $n = 1, 2$ , 其中 1 表示训练集中的社会网络特征向量, 2 表示测试集中的社会网络特征向量。

通过对诈骗团体的中心性进行分析,可以挖掘出诈骗团体内部各节点之间的角色担当以及诈骗团体内部各节点的结构,这些角色在网络中的特征向量可以表示为:  $F_i = [\text{网络中担任领导者角色的比率}, \text{网络中担任中间者角色的比率}]$ ,  $i = 1, 2$ , 其中 1 表示训练集中的社会网络特征向量, 2 表示测试集中的社会网络特征向量。

用向量空间模型中的余弦相似性(cosine similarity)来比较特征向量之间的相似度。对于余弦相似性,可以想象空间中的两条从原点出发指向不同方向的线段,形成一个夹角,如果夹角是  $0^\circ$ ,这就意味着这两条线段方向相同,线段完全重合;如果夹角为  $180^\circ$ ,则说明方向完全相反。因此,可以用夹角的大小来衡量向量的相似性,夹角越小就代表向量越相似。假定  $A = [A_1, A_2, \dots, A_n]$  和  $B = [B_1, B_2, \dots, B_n]$  是两个  $n$  维向量,则  $A$  与  $B$  的夹角  $\theta$  的余弦等于:

$$\cos\theta = \frac{\sum_{i=1}^n (A_i \cdot B_i)}{\sqrt{\sum_{i=1}^n (A_i)^2} \cdot \sqrt{\sum_{i=1}^n (B_i)^2}} = \frac{A \cdot B}{|A| \cdot |B|}$$

通过这个公式可以得到社会网络特征向量  $G_1$  和

$G_2$  和节点角色组成的特征向量  $F_1$  和  $F_2$  之间的余弦, 余弦值越接近 1, 说明夹角越接近 0, 则这两个向量越相似。当这几个特征向量的相似度都较高时, 则可以判定此潜在团体为诈骗团体。

### 3 案例分析

通过一个小案例样本, 分析已经存在的诈骗团体与一般的社会网络团体之间在内容特征和社会网络特征上的差异性。

#### 3.1 实验样本集

通过网络爬虫技术搜集新浪微博平台上已经存在的诈骗团体和讨论女排相关话题的一般社会网络团体的信息, 其中诈骗团体的社会网络包含 15 个独立节点和 17 条连接, 而一般社会网络团体包含 19 个独立节点和 21 条连接。

#### 3.2 内容特征向量获取

通过获取到的诈骗团体和非诈骗团体的信息, 经过上述介绍的数据清洗和处理, 得到两个只包含字和词的数据集, 然后分别计算它们的 TF-IDF 值; 之后再根据 TF-IDF 值从大到小排序, 各取前 10 个关键词作为代表社会网络的内容特征, 其结果如下:

诈骗团体 [骗子、非法、获利、个人信息、曝光、诈骗、隐蔽、电话、短信、拉黑]

一般社会团体 [郎平、中国、女排、冠军、铁榔头、一米八、梦想、夺冠、骄傲、奥运会]

针对不同的热点搜索词所产生的社群获取到的内容特征词便有很大的不同, 通过这种性质便可以作为辨别诈骗团体的依据。

#### 3.3 社会网络特征获取

通过 Pajek32 软件对上面两种社会网络团体进行构建并计算各自的社会网络特征, 由此来观察两者之间的差别。Pajek 是包含上千及至数百万个节点大型网络的分析和可视化操作。

图 2 和图 3 分别表示诈骗团体和一般社会网络团体(女排相关)所呈现的网络图(不带方向)。

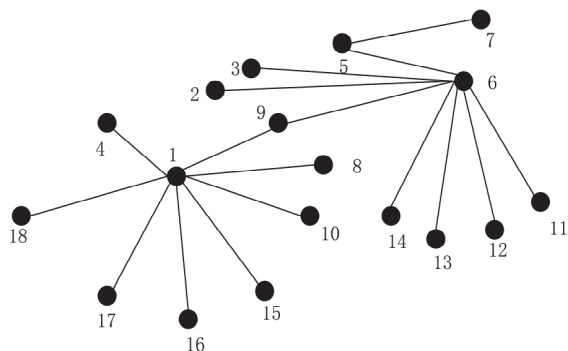


图3 诈骗团体网络图

从图3可以看出, 诈骗团体的网络图拥有两个主

要的领袖, 网络图中其他成员或者节点之间传递信息大多都要经过这两个领袖进行。

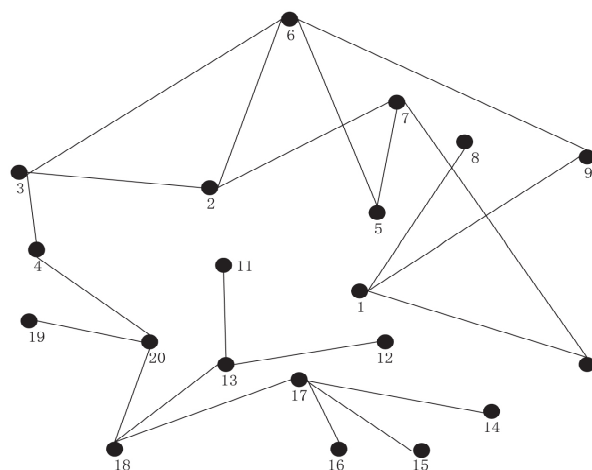


图4 一般社会网络团体网络图

而从一般社会网络团体的网络图来看, 角色大多不是很鲜明, 各节点之间大多直接进行信息传递。

如表1、表2所示, 从这两个团体的社会网络特征值来看, 这两个团体在社会网络角色中存在明显的差异性。诈骗团体存在非常明显的领袖节点, 统一社会网络信息的传播和控制。而一般社会团体(女排相关)各节点直接比较松散, 各节点内部之间大多直接进行交流, 这与诈骗团体的社会网络组成有着显著的差异。

表1 诈骗团体的社会网络特征值

序号	网络中担任领导者角色节点的比率	网络中担任中间者角色的比率
1	0.444	0.611
6	0.444	0.667
9	0.111	0.277
5	0.111	0

表2 一般社会网络团体的社会网络特征值

序号	网络中担任领导者角色节点的比率	网络中担任中间者角色的比率
1	0.213	0.53
2	0.213	0.32
6	0.255	0.24
20	0.213	0.33

### 4 结束语

文中利用社会网络分析法和数据挖掘技术对微博中的诈骗团体进行分析和研究, 挖掘出微博平台上潜在的诈骗团体, 从而帮助人们减少受骗的机会。虽然对微博中的诈骗热点数据进行了分析和挖掘, 但是由于该模型的复杂性, 尚存在一些不足之处: 首先, 由于 (下转第98页)

次使用顺序访问替代随机访问,以增加计算机的存储器访问效率,降低了误判率。仿真结果表明,优化后的哈希查找算法的系统性能更加优越。

#### 参考文献:

- [1] 刘颖,陈煜,林林,等.高性能计算集群中的网络技术研究与实践[J].中国水利水电科学研究院学报,2016,14(2):90-95.
- [2] 岳菲菲,王海军,王新,等.高性能计算通信机制分析与研究[J].计算机工程与科学,2009,31(A1):27-30.
- [3] 李根国,桂亚东,刘欣.浅谈高性能计算的地位及应用[J].计算机应用与软件,2006,23(9):3-4.
- [4] 熊兵,李峰,姜腊林,等.面向高速网络连接记录管理的高效哈希表[J].华中科技大学学报:自然科学版,2011,39(2):19-22.
- [5] CLARK D D, JACOBSON V, ROMKEY J, et al. An analysis of TCP processing overhead[J]. IEEE Communications Magazine, 2002, 40(5): 94-101.
- [6] CHIANG M L, LI Y C. LyraNET: a zero-copy TCP/IP protocol stack for embedded systems[J]. Real-Time Systems, 2006, 34(1): 5-18.
- [7] LI Z, MAKINENI S, JILIKKAL R, et al. Efficient caching techniques for server network acceleration[C]//Advanced networking & communications hardware. [s.l.]: [s.n.], 2004.
- [8] MAKINENI S, BHUYAN L. TCP/IP cache characterization in commercial server workloads[C]//Proceedings of seventh workshop on computer architecture evaluation using commercial workloads. [s.l.]: [s.n.], 2004.
- [9] 马如林,蒋华,张庆霞.一种哈希表快速查找的改进方法[J].计算机工程与科学,2008,30(9):66-68.
- [10] 王果,徐仁佐.结合哈希过滤的一种改进多连接查询优化算法[J].计算机工程,2004,30(7):57-59.
- [11] SONG H, DHARMAPURIKAR S, TURNER J, et al. Fast hash table lookup using extended bloom filter: an aid to network processing[C]//Proceedings of the 2005 conference on applications, technologies, architectures, and protocols for computer communications. New York, NY, USA: ACM, 2005: 181-192.
- [12] KUMAR S, CROWLEY P. Segmented hash: an efficient hash table implementation for high performance networking subsystems[C]//Proceedings of 2005 ACM symposium on architecture for networking and communications systems. New York, NY, USA: ACM, 2005: 91-103.
- [13] HASAN J, CADAMBI S, JAKKULA V, et al. Chisel: a storage efficient, collision-free hash-based network processing architecture[C]//Proceedings of the 33rd annual international symposium on computer architecture. Washington, DC, USA: IEEE Computer Society, 2006: 203-215.
- [14] LIAO G, BHUYAN L N, WU W, et al. A new TCB cache to efficiently manage TCP sessions for web servers[C]//ACM/IEEE symposium on architectures for networking and communications systems. New York, NY, USA: ACM, 2010.
- [15] PONG F. Fast and robust TCP session lookup by digest hash[C]//Proceedings of the 12th IEEE international conference on parallel and distributed systems. [s.l.]: IEEE, 2006.

(上接第 93 页)

数据的局限性,只研究了微博中的诈骗团体,对于其他平台的和沟通工具的诈骗团体有待进一步挖掘;其次,采用结巴分词进行断词,产生了大量的数据集,影响了运行效率,因此提高该算法的效率是后续的研究方向。

#### 参考文献:

- [1] 孙孟.微博营销-新媒体时代的营销宠儿[J].通信企业管理,2011(7):38-39.
- [2] 吴继飞,邓安平.基于互联网时代微博营销的SWOT分析[J].中国集体经济,2011,21:52-53.
- [3] 王利.基于数据挖掘技术的微博营销系统的设计与实现[D].武汉:华中科技大学,2013.
- [4] 邵笑.新媒体诈骗的言语行为研究[D].锦州:渤海大学,2014.
- [5] 张劲捷.基于微博社交网络的舆情分析模型及实现[D].广州:华南理工大学,2011.
- [6] 缪茹一.基于文本数据挖掘的微博情感分析与监控系统[D].杭州:浙江工业大学,2015.
- [7] ZHOU X, CHEN L. Event detection over twitter social media streams[J]. VLDB Journal, 2014, 23(3): 381-400.
- [8] 康泽东,余旌胡,丁义明.微博社交网络的对称程度实证分析[J].计算机应用,2014,34(12):3405-3408.
- [9] FARINE D R, WHITEHEAD H. Constructing, conducting and interpreting animal social network analysis[J]. Journal of Animal Ecology, 2015, 84(5): 1144-1163.
- [10] 孙怡帆,李赛.基于相似度的微博社交网络的社区发现方法[J].计算机研究与发展,2014,51(12):2797-2807.
- [11] 范超然,黄曙光,李永成.微博社交网络社区发现方法研究[J].微型机与应用,2013,31(23):67-70.
- [12] NASON G J, FARDOD O, KELLY M E, et al. The emerging use of Twitter by urological journals[J]. Bju International, 2015, 115(3): 486-490.
- [13] CHEN P, FU X, TENG S, et al. Research on micro-blog sentiment polarity classification based on SVM[C]//International conference on human centered computing. [s.l.]: Springer International Publishing, 2014: 392-404.
- [14] FLEUREN W W M, ALKEMA W. Application of text mining in the biomedical domain[J]. Methods, 2015, 74: 97-106.
- [15] IRFAN R, KING C K, GRAGES D, et al. A survey on text mining in social networks[J]. Knowledge Engineering Review, 2015, 30(2): 157-170.