

# 基于复杂网络的微博传播溯源方法

王 宁,贾志娟

(郑州师范学院 信息科学与技术学院,河南 郑州 450044)

**摘 要:**针对微博博文信息在传播的过程中容易出现失真、断章取义以及局部抽取等现象,从而导致不良信息的传播扩散,并对社会安定产生影响的问题,提出了一种基于复杂网络的微博传播溯源方法。该方法首先根据处理过后的微博数据还原出单条微博的转发评论关系,进而根据其转发评论关系重构出单条微博的转发评论关系树,并简单描述了重构转发评论树、还原传播路径以及计算用户传播能力的算法流程。然后,在单条微博转发评论关系树的基础上进行拓展,进而重构出一个微博话题的转发评论关系图,并分析了该图中每一个微博用户的传播能力。最后基于该溯源方法设计了一个微博话题关键用户的溯源查找系统,并通过实验证明该溯源方法实现了微博关键用户的查找。

**关键词:**转发评论关系;传播溯源;关键用户;微博话题

**中图分类号:**TP301

**文献标识码:**A

**文章编号:**1673-629X(2019)12-0081-06

**doi:**10.3969/j.issn.1673-629X.2019.12.015

## A Micro-blog Communication Traceability Method Based on Complex Network

WANG Ning, JIA Zhi-juan

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

**Abstract:** In view of the fact that micro-blog information is prone to distortion, out-of-context extraction and local extraction in the process of transmission, which leads to the spread of bad information and influences the social stability, we propose a micro-blog communication traceability method based on complex network. This method firstly restores the forwarding and commenting relationship of a single micro-blog based on the processed micro-blog data, and then reconstructs the forwarding and commenting relationship tree of a single micro-blog based on its forwarding and commenting relationship. The algorithm flow of reconstructing the forwarding and commenting tree, restoring the propagation path and calculating the user's propagation ability is briefly described. Then, on the basis of the relationship tree of forwarding and commenting of a single micro-blog to expand, the forwarding and commenting graph of a micro-blog topic is reconstructed and the propagation ability of each micro-blog user in the graph is analyzed. Finally, a traceability search system for key users of micro-blog topic is designed based on the traceability method, and the experiment proves that this traceability method realizes the search for key users of micro-blog.

**Key words:** forwarding and commenting relationship; communication traceability; key user; micro-blog topic

## 0 引言

随着互联网技术的普及,越来越多的人享受到互联网带来的便利和乐趣。尤其是近几年在线社交网络<sup>[1]</sup>的迅速发展,使人们越来越多地参与到互联网丰富的社交活动中。作为新生网络的应用形式,微博在近年来发展迅猛,并成为目前国内最具影响力的主流网络媒体之一<sup>[2]</sup>。在微博系统中,博文信息在微博博

主之间进行传播,他们之间的传播方式包括两种:一是直接转发;二是对原来的博文信息进行修改,增加一些个人的观点,然后再进行转发。如果要判断微博博文信息在传播过程中是否有失真、断章取义以及局部抽取等情况,就需要借助信息的源头进行查找判断。在微博信息传播过程中,如何快速准确地查找到信息的源头以及理清完微博的传播路径,是微博信息系统分

收稿日期:2018-12-24

修回日期:2019-04-25

网络出版时间:2019-09-24

**基金项目:**国家自然科学基金(U1304614, U1204703);中央高校基本科研业务费资助项目(2012QN087, 2012QN088);河南省重点科技攻关项目(122102310004);郑州市创新型科技人才队伍建设工程资助项目(10LJRC190)

**作者简介:**王 宁(1987-),女,讲师,硕士,研究方向为复杂网络、数据挖掘与量子密码;贾志娟,教授,硕士,研究方向为计算机网络、数据挖掘。

**网络出版地址:**http://kns.cnki.net/kcms/detail/61.1450.TP.20190924.1534.004.html

析中最主要的部分。微博事件的源头发现,就是在这一系列微博中,查找出首发微博。而如何快速查找出微博事件及话题的源头也是近些年来研究者们研究的热点。

Adar E 等提出了隐式结构和博客空间的动态<sup>[3]</sup>,并对博客空间中的流进行了研究<sup>[4]</sup>。J Leskovec 等<sup>[5]</sup>通过创建一个有向图发现了博客空间中信息传播的模式并了解了潜在的社交网络。D Liben-Nowell 等<sup>[6]</sup>提出了一种在全球范围内利用连锁信息数据进行信息溯源的快速方法。该方法利用网络簇以及异步时间的概率模型在社交网络上迅速准确地找到了信息的源头。Jin Xin 等<sup>[7]</sup>提出了话题源头的概念,并基于话题的相关性、文档时间以及文档之间的关系这三个方面提出了 TCL 的话题溯源模型。文卫华等<sup>[8]</sup>以微博事件为研究对象对其传播特征进行研究。G S Bindra 等<sup>[9]</sup>对信息流进行追踪,并分析它在社交媒体中的不完整影响,研究表明,k 树模型是研究级联中缺失数据影响的有效工具。张旸等<sup>[10]</sup>通过对微博上不同特征的重要性进行分析,提出了基于特征加权的预测模型。杨静等提出一种基于话题影响力的微博话题溯源方法<sup>[11]</sup>和一种基于溯源的虚假信息传播控制方法<sup>[12]</sup>,其中传播控制方法利用微博转发关系,结合关系网和信息级联关系网找到微博的真正发起者。郑业鲁等<sup>[13]</sup>提出了蔬菜供应链全程溯源模型并搭建了蔬菜产品质量安全溯源系统。王澍贤等<sup>[14]</sup>对意见领袖参与下微博舆情演化的三方博弈进行了分析。刘荣叁等<sup>[15]</sup>对面向新浪微博的信息溯源技术进行研究。李城等<sup>[16]</sup>提出了一种微博谣言溯源方法,该方法利用改进的最长公共子序列进行比对微博谣言,从而查找源头。

由于微博本质上属于一种复杂的社交网络,它满足复杂网络所具有的一些特性,如网络节点的度、聚类系数和平均最短路径等,故通过对前人的研究成果进行研究和分析,文中提出一种基于复杂网络的微博传播溯源方法。该方法首先从单条微博的传播过程入手,还原出单条微博的转发评论关系,进而重构出单条微博的转发评论关系树;其次,在此基础上进行扩展,重构出一个微博话题的传播过程;然后基于该方法设计了一个微博话题关键用户查找系统,并基于该系统做了一个微博话题传播溯源实验,实验表明该溯源方法实现了关键用户的查找。

## 1 基于复杂网络的微博源头发现方法

在本节的溯源方法中,首先介绍了单条微博的转发评论关系重构,然后在此基础上进行拓展,重构出微博话题的转发评论关系图,最后实现微博关键用户的查找。

### 1.1 重构单条微博的转发评论关系

若干条单独的原创微博信息可以组成一个微博话题,如果想要重新构建出一个微博话题的信息传播关系,那么重新构建出每一条原创微博博文的信息传播关系则是首先需要做的事情。

#### (1) 还原单条微博信息的传播模型。

当微博博主用户发送出一条原创的微博信息后,微博系统将会实时地将该原创微博博文传递给他的微博粉丝用户,粉丝用户可以对原创微博进行评论或转发,而转发则可以把原创微博信息传递到该粉丝用户的粉丝用户,以此类推,最终使得整条微博在整个微博用户关系网中传播开来。实际上,微博博文是按层次进行转发和评论的,通过一层一层进行推进,形成一个类似于树形的结构。这里,如果把原创微博博主用户看作根节点,把每一个转发评论微博用户都看成是一个子节点,而把用户的每一次转发评论行为看成是一条边,那么一条微博的转发评论关系就可以构成一棵微博转发评论关系树,如图 1 所示。

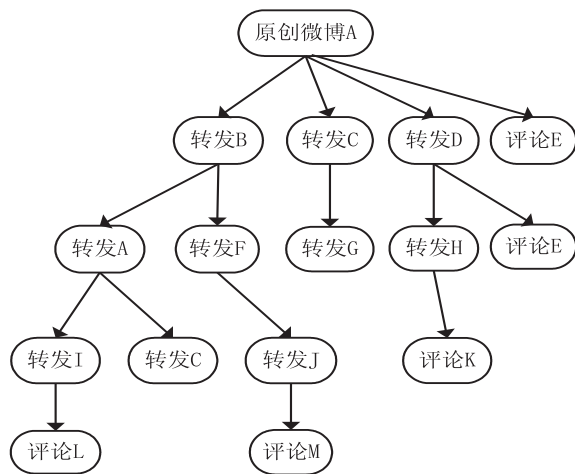


图 1 原创微博的转发评论关系树

由图 1 可知,原创微博博主用户 A 发出原创微博,然后由微博粉丝用户 B、C、D 分别进行转发,并且由微博粉丝用户 E 对该原创微博进行评论,原创微博用户 A 和粉丝用户 F 从粉丝用户 B 处转发。由此能够看出,微博用户 A 虽然是原创微博用户,但其同样能够成为自己所发微博的转发用户,微博用户 B 是微博用户 A 的粉丝,同样微博用户 A 也可以成为用户 B 的粉丝。在微博系统中,所有的转发评论关系图都可以还原成类似图 1 所示的一个树形结构。

#### (2) 实现微博转发评论关系树的重构。

通过对微博系统的分析可以看出,粉丝用户对原创微博博主信息的所有转发评论信息全部都可以归总到原创微博博主用户的微博主页,并且能够在该主页上面查找到所有粉丝的转发微博,而微博评论信息则可以在相应的转发微博和原创微博主页中进行查找,

具体步骤如下:

- (a) 获取转发评论数据。在原创微博的主页获取用户的信息(包括用户 ID、昵称、发送时间等)和所有转发用户的信息(包括 ID、昵称、转发时间、微博 ID),根据转发用户的信息依次访问每一条转发微博的主页,在该主页获取该条转发微博的转发和评论微博信息,并把它们保存在相应的数据库列表中。
- (b) 构造转发评论树。根据第一步得到的转发和评论数据列表,层次地重构转发评论树,其算法流程如图 2 所示。

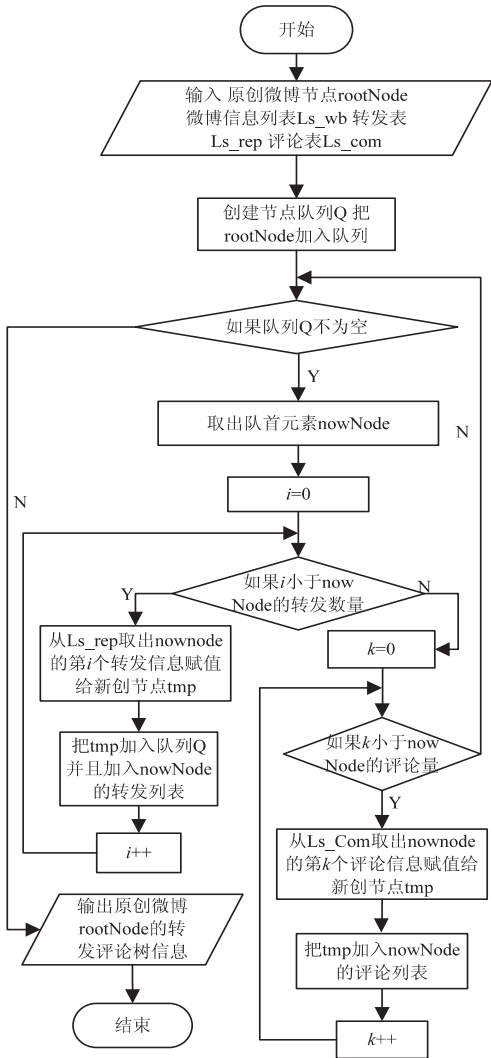


图 2 重构转发评论树的算法流程

- (c) 可视化转发评论树。采用 JavaScript 可视化组件对单条微博的转发评论关系进行可视化,如图 3 所示(这里需要说明的是,由于某些微博的转发量比较大,而整个屏幕能够显示的点是有限的,故选择其中的部分点进行展示)。其中,底部标注 1 所示颜色表示原创微博博主(发布者),2 所示颜色表示转发数量大于 8 的微博用户,3 所示颜色表示转发数量为 1-8 的微博用户,4 所示颜色的点则表示转发量为 0 的用户(末端节点)。如果鼠标被放在图 3 的任一个转发

和评论节点上面,在图 3 中均能够显示出相应微博的微博用户昵称、微博转发或评论的时间以及微博的博文内容等信息。

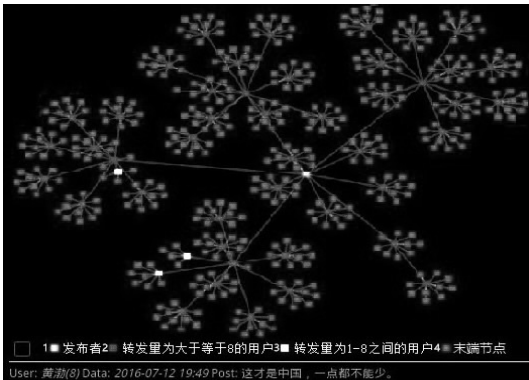


图 3 微博转发评论图

- (3) 微博转发评论关系树的分析。
- 微博转发评论树的分析主要从传播路径和用户传播能力两个方面进行。传播路径包括传播路径的还原及微博信息的传递,而微博用户的信息传播能力则通过信息传播的广度和深度来进行计算。
- 根据图 1 中生成的原创微博的转发评论关系树,还原出微博传播的每一条传播路径,其算法流程如图 4 所示。

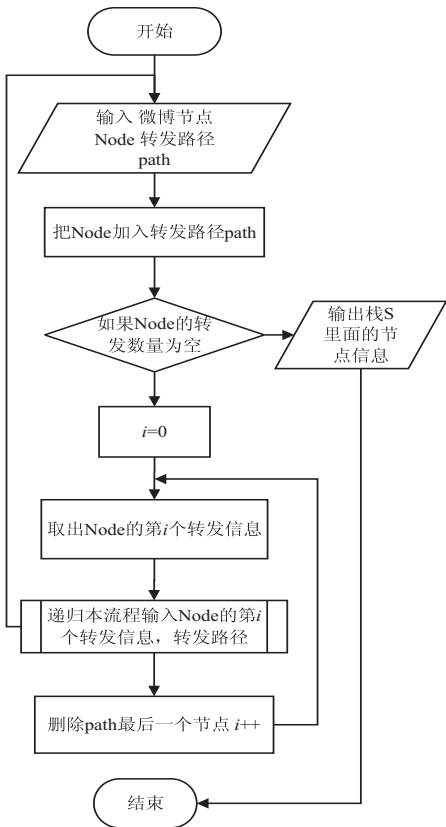


图 4 还原传播路径的算法流程

如果要还原出微博信息在任意时刻被微博用户转发和评论的情况,则需要按图 5 所示的算法流程进行分析。

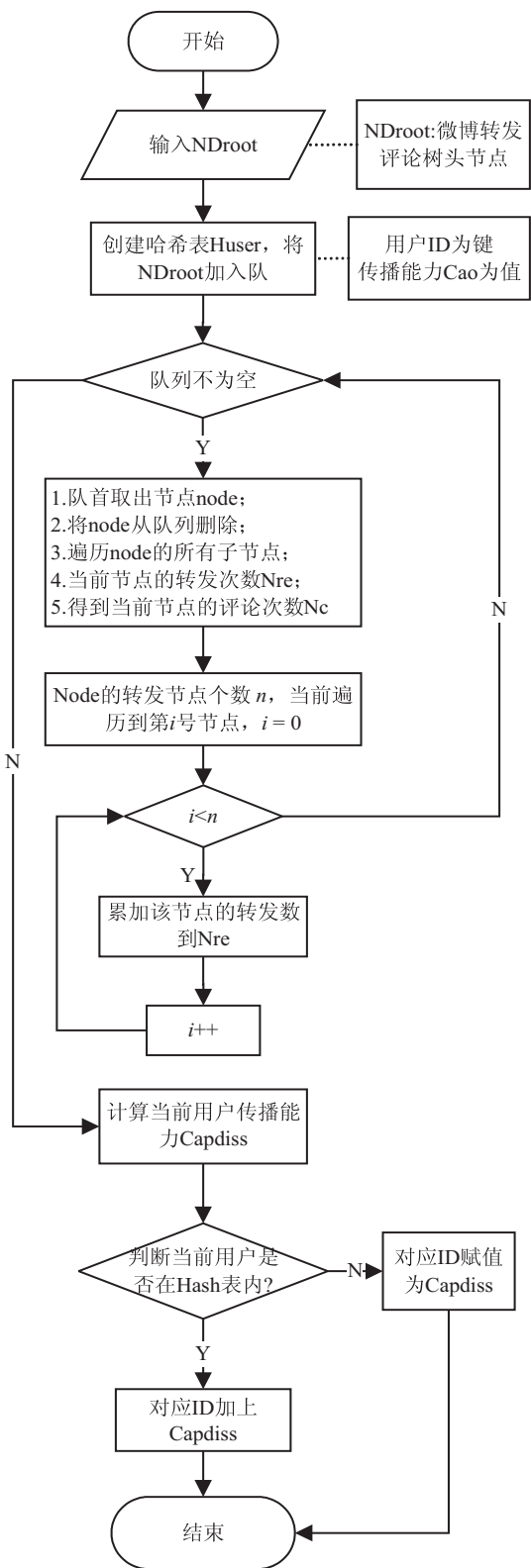


图 5 微博信息传递(原创微博在任意时刻被户转发和评论)的算法流程

信息传播的广度是指该微博用户的微博能被多少微博粉丝用户进行转发或者评论。这里,主要考虑第一层的微博被转发和评论的数量。信息传播的深度是指该微博能被传播几层的深度,这与该微博用户后面用户的粉丝有着相当大的关系,但和当前微博用户却

没有特别大的关系,故这里仅考虑第二层的微博转发数量。

假设  $N_r$  表示用户的微博信息被转发的次数,  $N_c$  表示微博的微博信息被评论的次数,  $N_{re}$  表示第二层的节点一共被转发多少次,那么在微博的传播过程中,某个微博用户的传播能力可以表示为:

$$Cap_{diss} = N_r + \log_2 N_c + \log_2 N_{re} \tag{1}$$

如果一个微博用户曾参与一条微博的多次转发,那么该用户的传播能力则是其对应的每个节点的传播能力的叠加。其算法流程如图 6 所示。

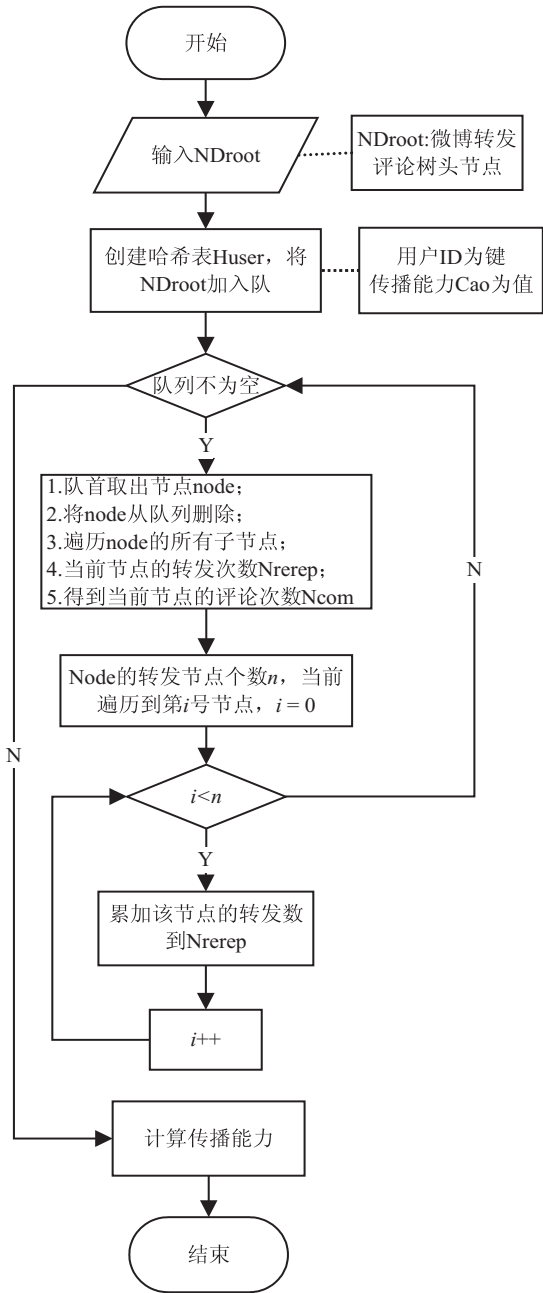


图 6 计算用户传播能力的算法流程

1.2 重构微博话题的转发评论关系

本节就微博话题转发评论关系的构建进行简单介绍。首先需要得到该话题的微博地址的 ID 列表,然后



根据原创微博的地址 (ID) 按顺序构建出每一条微博的转发评论关系树,最终构建出一个微博话题的转发评论关系图,其可视化图如图 7 所示。

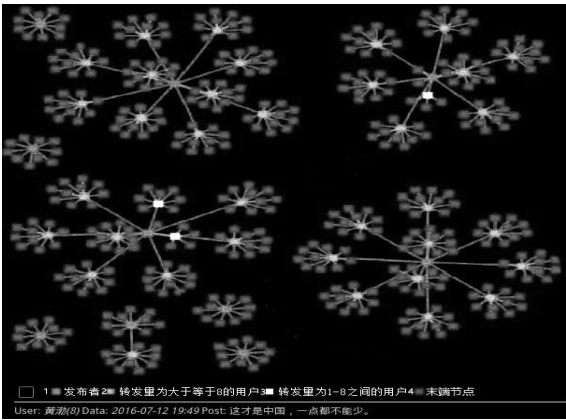


图 7 微博话题的转发评论关系图

图 7 所示是 2016 年 7 月 19 日“中国一点都不能少”话题的微博转发评论关系图,图中的转发评论关系主要由 10 条微博事件构成,图中底部标注 1 所示颜色的点表示原创微博博主(发布者)。

微博话题的信息传播分析和单条微博的信息传播分析相似,而不同的是,微博话题的转发评论分析则是把转发评论关系变成了多棵树的分析,故微博话题用户的信息传播能力为:

T\_i = {M\_1,M\_2,...,M\_n} (2)

另外,一个微博用户或许会参与多条微博信息的转发和评论,因此,对于参与该话题的每一个微博用户 U\_j,他在微博话题中的信息传播能力可以定义为:

Cap\_{U\_j} = \sum\_{i=1}^n cap\_{U\_j} in M\_i (3)

2 系统的设计与实现

基于复杂网络的微博传播溯源方法,本节设计了微博话题关键用户的溯源系统,并通过实验证明该系统实现了关键用户的查找。

2.1 系统设计架构

该系统的设计主要分为微博网络数据的采集及相关处理和微博话题关键用户的溯源,具体步骤如下:

(1) 微博数据的采集。

利用爬虫程序在微博系统的网页上抓取若干个微博种子用户的数据,这里主要抓取原创微博用户的信息、微博地址 (ID) 和发送时间等。

(2) 微博数据的本地处理。

首先对微博进行分词及倒排序索引,构建词典索引库;其次整理出每一条微博的转发评论关系,构建其转发评论树;然后对爬虫获得的微博进行主题提取,得到某些谈论话题,并据此对微博进行聚类;最后对这些话题进行索引,构建话题索引库。

(3) 查找给定话题对应的微博列表。

首先在索引库 (包括词典索引库和话题索引库) 中查找微博话题对应的微博列表 ID,然后根据微博 ID 得到每一条微博的转发评论树,重建该话题的转发评论关系。

(4) 根据重构的转发评论关系,对用户的传播能力进行分析,计算参与该话题的每一个用户的传播能力,并对其进行排序。

(5) 根据微博 ID 和排名靠前的传播能力这两个方面得到微博信息传播过程中的关键用户。

该系统的工作流程如图 8 所示。

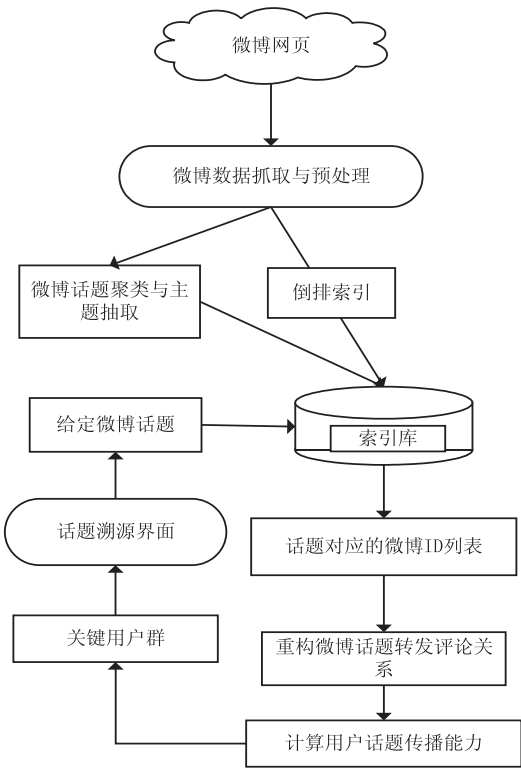


图 8 微博关键用户查找系统

2.2 实验与结果分析

根据前面章节抓取的数据以及对各个算法的处理及实现,最终实现了该微博话题关键用户的查找系统,系统展示如图 9 所示。



图 9 微博话题关键用户查找系统首页

以下通过一个例子分析验证该溯源系统的有效可行性。

首先,假定微博话题为“中国一点都不能少”,再通过分词软件 ICTCLAS 对其进行分词,得到该微博话题的关键词列表为“中国”、“一点”、“都”、“不能少”。

然后,在微博主题索引库中进行关键词匹配,得出该话题与主题“中国”的相似度是最高的,因此,认为所要查的微博话题为“中国”。故当在图 10 的关键词搜索一栏中键入“中国”一词时,会在窗口右侧显示出参与该微博话题的每个微博用户的详细的微博传播情况(由于页面有限,只截取部分微博用户的传播情况)。从图中可以看出,用户的传播能力逐次降低,且传播能力较高的为央视新闻、范冰冰、人民日报、胡歌、Johnny 黄景瑜等。通过对原创微博博主用户以及重点微博用户的分析可知,在传播过程中起关键作用的微博用户群为人民日报和央视新闻,如图 10 左侧窗口所示。

关键词搜索			中国
博主	传播能力	查看详情	
人民日报	19.9658		央视新闻20.5507
央视新闻	20.5507		范冰冰19.9658
			人民日报 19.9658
			胡歌 19.6518
			Johnny黄景瑜 19.5118
			宋茜 16.6439
			王凯kkw 16.6439
			杨洋icon 13.3219
			狮心峰香菜君 13.3219
			馬天宇 13.3219
			任大称 13.3219
			我才不喜欢黄景瑜呢 13.3219
			Johnny黄景瑜全球粉丝后援会 13.3219
			浅水小鱼2222 13.3219
			小黄少 13.3219
			胡歌数据组 13.3219

图 10 微博话题关键用户查找系统查找页面

以上便是从输入待查的微博话题,到最后返回该微博话题在传播过程中起关键作用的用户群体的整个分析过程。

### 3 结束语

基于微博本质上是一个复杂网络的特性,提出了一种基于复杂网络的微博传播溯源方法。该溯源方法首先从微博系统中抓取微博数据并对其进行处理,然后根据处理的数据还原出单条微博的转发评论关系,进而构建出单条微博的转发评论关系树;其次,在单条微博的转发评论关系基础上进行扩展,重构出一个微博话题的转发评论关系树,并对该树中每一个微博用户的传播能力进行分析;然后基于该溯源方法设计一个微博话题关键用户的查找系统,最后基于该系统做了一个微博话题溯源实验,实验表明该方法实现了微博话题关键用户的查找。

#### 参考文献:

[1] BOYD D, ELLISON N B. Social network sites: definition,

history, and scholarship [J]. Journal of Computer Mediated Communication, 2007, 13(1): 210-230.

- [2] 张 晗. 浅析微博时代网络舆论的特点及传播规律——以唐骏学历“造假门”事件为例[J]. 新闻世界, 2011(4): 85-86.
- [3] ADAR E, ZHANG L, ADAMIC L A, et al. Implicit structure and the dynamic of blogspace [C]//Proceedings of the workshop on the weblogging ecosystem. New York: ACM, 2004.
- [4] ADAR E, ADMIC L A. Tracking information epidemics in blogspace [C]//2005 IEEE/WIC/ACM international conference on web intelligence. Compiegne, France: IEEE, 2005: 207-214.
- [5] LESKOVEC J, MCGLOHON M, FALOUTSOS C, et al. Patterns of cascading behavior in large blog graphs [C]//Proceedings of the 2007 SIAM international conference on data mining. [s. l.]: [s. n.], 2007: 551-556.
- [6] LIBEN-NOWELL D, KLEINBERG J. Tracing Information flow on a global scale using internet chain-letter data [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(12): 4633-4638.
- [7] JIN Xin, SPANGLER S, MA Rui, et al. Topic initiator detection on the world wide web [C]//Proceedings of the 19th international conference on world wide web. Raleigh, North Carolina, USA: ACM, 2010: 481-490.
- [8] 文卫华, 张 杰. “微博事件”及其传播特征研究 [J]. 新闻爱好者: 下半月, 2011(10): 8-9.
- [9] BINDRA G S, KANDWAL K K, SINGH P K, et al. Tracing information flow and analyzing the effects of incomplete data in social media [C]//Fourth international conference on computational intelligence, communication systems and networks. Phuket, Thailand: IEEE, 2012: 235-240.
- [10] 张 畅, 路 荣, 杨 青. 微博客中转发行为的预测研究 [J]. 中文信息学报, 2012, 26(4): 109-114.
- [11] 杨 静, 董 圆, 张健沛. 一种基于话题影响力的微博话题溯源方法 [J]. 小型微型计算机系统, 2015, 36(9): 1939-1942.
- [12] 杨 静, 周雪妍, 林泽鸿, 等. 基于溯源的虚假信息传播控制方法 [J]. 哈尔滨工程大学学报, 2016, 37(12): 1691-1697.
- [13] 郑业鲁, 刘晓珂, 郭洛先, 等. 基于供应链的蔬菜安全溯源系统的设计与实现 [J]. 广东农业科学, 2016, 43(1): 145-150.
- [14] 王澍贤, 陈福集. 意见领袖参与下微博舆情演化的三方博弈分析 [J]. 图书馆学研究, 2016(1): 19-25.
- [15] 刘荣叁, 张 宇, 王 星. 面向新浪微博的信息溯源技术研究 [J]. 智能计算机与应用, 2017, 7(2): 94-98.
- [16] 李 城, 沙俊淞, 武 文. 基于最长公共子序列的微博谣言溯源研究 [J]. 计算机与现代化, 2018(1): 107-112.