计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130079
51(4):691-706,2014

# 微博数据挖掘研究综述

丁兆云[1,2,3] 贾 焰[3] 周 斌[3]
[1](国防科学技术大学信息系统与管理学院 长沙 410073)
[2](国防科学技术大学信息系统工程重点实验室 长沙 410073)
[3](国防科学技术大学计算机学院 长沙 410073)
(zyding@nudt.edu.cn)

# Survey of Data Mining for Microblogs

Ding Zhaoyun[1,2,3], Jia Yan[3], and Zhou Bin[3]
[1](*College of Information Systems and Management, National University of Defense Technology, Changsha* 410073)
[2](*Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha* 410073)
[3](*School of Computer, National University of Defense Technology, Changsha* 410073)

**Abstract** The past few years the rapid development and popularization of microblogs have already been witnessed. Due to their openness, terminal expansion, content simplicity, low threshold and so on, microblogs deeply affect our daily life by providing an important platform for people to publish comments, transform information and acquire knowledge, to name just a few. Though bearing such advantages, microblogs may cause serious impacts on the national security and social development if they are out of control. Therefore, the research on microblogs is quite valuable from both theoretical and practical perspective, especially in this age of the Internet. Analyzing and mining microblogs also brings great challenges. As can be seen, microblogs can be treated as a generalization and extension of human life in the virtual network world. However, different from traditional information networks, microblogs have their unique characteristics, including noisy data diversity, social media, multi-relations, the rapid spread and evolutionary, nonlinearity, large scalability and etc. Such differences bring forth great challenges in analyzing and mining the microblogs. In this paper, we survey the data mining for microblogs and analyze the dataset of Twitter. Moreover, we summarize the challenges of data mining for microblogs.

**Key words** microblogs; data mining; text mining; social network; social media

**摘 要** 随着近几年微博的快速发展与普及,微博凭借平台的开放性、终端扩展性、内容简洁性和低门槛等特性,在网民中快速渗透,已发展成一个重要的社会化媒体,微博成为网民获取新闻时事、人际交往、自我表达、社会分享以及社会参与的重要媒介以及社会公共舆论的重要平台,对国家安全和社会发展产生了深远的影响.微博是人类在虚拟网络世界生活的抽象概括和延伸,与一般信息网络不同,微博本身具有大规模、噪音数据多样性、快速传播演化性、非线性、社会媒体性以及多关系等特征,因此其在分析方法和挖掘目标上都与传统信息系统具有很大差别,在相关技术的研究上也带来了更大的挑战.针对

收稿日期:2013-01-21;修回日期:2013-08-27
基金项目:国家"九七三"重点基础研究发展计划基金项目(2013CB329601,2013CB329602);国家自然科学基金项目(61372191,71331008,61302144)

微博的新特性,研究了微博近几年的相关研究现状,同时分析了 Twitter 数据集特征,且总结了未来研究面临的挑战.

关键词 微博;数据挖掘;文本挖掘;社会网络;社会媒体

中图法分类号 TP391

互联网正逐步演变为无处不在的计算平台和信息传播平台.微博、在线社交网站、博客、论坛、维基等社交网络应用的出现和迅猛发展,使得人类使用互联网的方式产生了深刻变革——由简单信息搜索和网页浏览转向网上社会关系的构建与维护、基于社会关系的信息创造、交流和共享.特别是近几年微博的快速发展与普及,微博凭借平台的开放性、终端扩展性、内容简洁性和低门槛等特性,在网民中快速渗透,发展成为一个重要的社会化媒体.微博成为网民获取新闻时事、人际交往、自我表达、社会分享以及社会参与的重要媒介及社会公共舆论、企业品牌和产品推广、传统媒体传播的重要平台.

微博正在成为人类社会中社会关系维系和信息传播的重要渠道和载体,对国家安全和社会发展都会产生深远的影响:1)社会个体通过各种连接关系在微博中构成"关系结构",包括以各种复杂关系关联而成的虚拟社区;2)基于微博的关系结构,大量网络个体围绕着某个事件而聚合,并相互影响、作用、依赖,从而形成具有共同行为特征的"网络群体";3)基于微博关系结构和网络群体,各类"网络信息"得以快速发布并传播扩散形成社会化媒体,并反馈到现实社会,从而使得微博与现实社会间形成互动,并对现实世界产生影响.

基于微博用户间关系的单向性,用户可以构建起一个强关系和弱关系并存的网络,从而同时满足了其多层次的社交需求,是人类在虚拟网络世界生活的抽象概括和延伸.与一般信息网络不同,微博本身具有大规模、噪音数据多样性、快速传播演化性、非线性、社会媒体性以及多关系等特征,因此其在分析方法和挖掘目标上都与传统信息系统具有很大差别,在技术上也带来了更大的挑战.

1) 微博发展现状

Twitter 作为全世界最流行的微博服务,由 Dorsey 于 2006 年 3 月创办并在当年 7 月启动的,截至 2012 年 3 月,Twitter 共有 1.4 亿活跃用户,这些用户每天会发表约 3.4 亿条推文,Twitter 每天处理约 16 亿的网络搜索请求.著名流量统计网站 ALEXA 的数据显示,Twitter 在 2012 年 10 月的日均 IP 访问量约为 5000 万,日均 PV 浏览量约为 3 亿次.市场研究机构 Semiocast 对 2012 年 6 月份发布于 Twitter 上的 10.58 亿条博文进行了分析,结果显示,截至 2012-07-01,Twitter 的注册用户数已经达到 5.17 亿,如图 1、图 2 所示,美国本土注册的 Twitter 用户数目最多,约为 1.418 亿,日本和韩国注册用户数目增长缓慢,而 Twitter 上发帖数量最多的三大城市分别是雅加达、东京与伦敦.

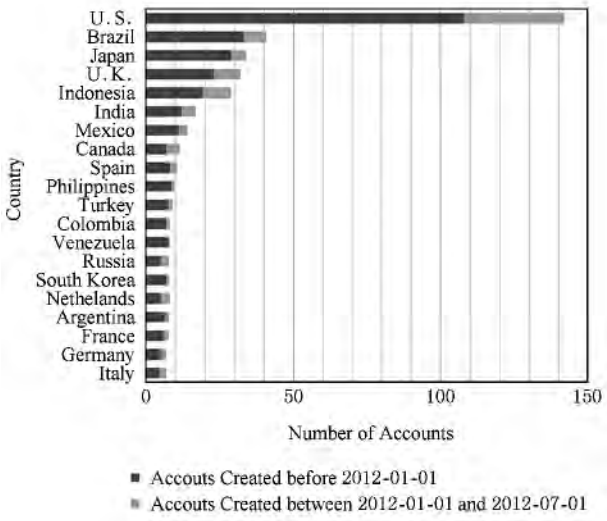


Fig. 1 Top 20 countries in terms of Twitter accounts.

图 1 注册账号数目排名前 20 的国家分布[1]

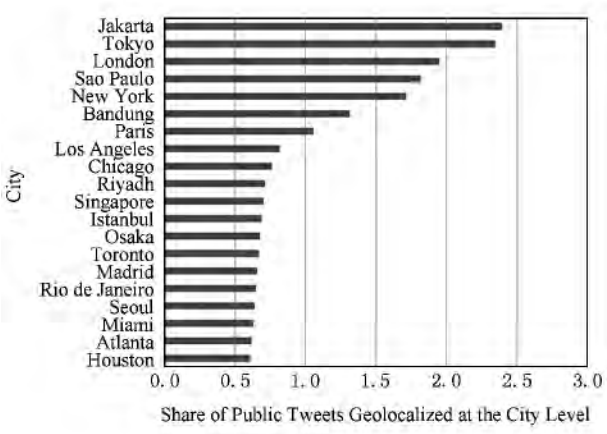


Fig. 2 Top 20 cities in terms of Twitter accounts.

图 2 发布博文数目排名前 20 的城市分布[1]

从 2007 年中国第 1 个具有微博特点的饭否网的创办,到 2009 年 8 月中国最大的门户网站新浪网

推出"新浪微博"内测版,微博正式进入中文上网主流人群视野,2010 年国内微博像雨后春笋般崛起,四大门户网站均开设微博,微博在国内迅速发展.根据中国互联网信息中心(CNNIC)发布的《中国互联网络发展状况统计报告》统计近几年国内微博发展趋势,国内微博用户数目逐年呈上升趋势,且在 2010 年与 2011 年短时间内聚集了大量用户,截至 2011 年 6 月,中国微博用户数量达到 1.95 亿,半年内增长超过 2 倍,增长率为 208.9%.网民使用率从 13.8%迅速提升至 40.2%,成为增长速度最快的互联网应用,截至 2012 年 6 月,超过一半的中国网民使用微博,网民使用率为 50.9%,微博用户数达到 2.74 亿,较 2011 年底增长 9.5%,微博用户规模进入平稳增长期.

2) 微博特点

微博即微博客(microblogs)的简称,是一个基于用户关系的信息分享、传播以及获取平台,用户可以通过 WEB,WAP(手机客户端)以及各种客户端组建个人社区,以不超过 140 个字符更新信息,并实现即时分享.境外微博主要包括 Twitter,plurk 等,境内微博主要包含新浪微博、腾讯微博等.区别于其他类 Facebook 的社交网络应用,微博的社会网络关系为单向的,用户不需要其他用户权限就可以关注它们.例如,Twitter 中社会网络由关注(following)关系形成,用户关注的人称为该用户的好友(friend);关注某用户的人称为该用户的粉丝(follower),用户发布的所有推文(tweets)将出现在公共时间线上(public timeline),该用户所有粉丝时间线上将显示该用户的所有消息.

韩国科学技术院 Kwak 等人[2]研究表明微博不仅具有社交网络(social network)功能,更倾向于具有社会媒体(social media)功能,表现为自媒体性,微博将用户从内容的消费者转换为内容的生产者.微博具有短文本性、终端扩展性、即时性、"裂变式"信息传播等特点.

1) 短文本性.区别于传统博客(blog)的长文本,微博限制用户发布的博文(post)不超过 140 个字符.

2) 终端扩展性.微博平台具有开放性,用户可以通过 WEB,WAP 以及各种客户端方便地使用微博.据美国互联网统计公司 comScore 的统计分析可知,2012 年 3 月份 Twitter 的手机用户数目相对 2011 年同时期增长了约 101%,成为增长速度最快的社交网络应用,如图 3 所示.中国互联网信息中心 CNNIC 发布的《第 30 次中国互联网络发展状况统计报告》指出:微博在手机端的增长幅度明显,用户数量由 2011 年底的 1.37 亿增至 1.70 亿,增速达到 24.2%.

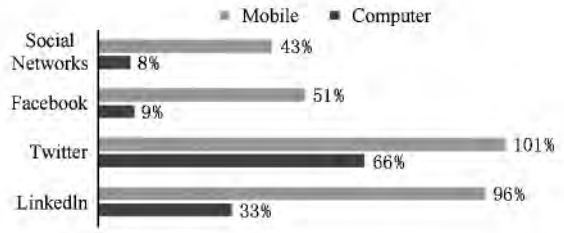


Fig. 3 Percentage growth in unique visitors.

图 3 访问 Twitter 的用户数目增长率分布[3]

3) 即时性.微博的即时性表现为内容发布的即时性和信息传播的即时性.由于微博的短文本性和终端扩展性,用户随时随地都可以不假思索地通过 WEB,WAP 以及各种客户端将所见所闻所感,用简便的语言通过微博迅速发送,这让微博彻底改变了信息传递的模式,成为即时性较强的信息传播平台.另外,微博用户所关注的好友更新了消息之后,系统自动将更新的信息按时间顺序主动推送到好友个人主页中,从而进一步强化了微博信息传播的即时性.

4) "裂变式"信息传播.微博的转发功能("RT @")使得信息无限制地被转发,其信息传播范围呈"核裂变"式的几何级数式扩大,且结合微博的主动推送功能,信息快速地扩散到大量用户中.

本文针对微博新特性,分析了微博近几年的相关研究现状,随后分析了 Twitter 数据集特征,且总结了未来研究面临的挑战.

## 1 微博研究现状

微博数据由于公开应用程序接口(application programming interface, API),数据获取便捷,学者能够在大规模微博数据集上挖掘隐含信息、验证信息理论等.

近年来,对微博数据挖掘以及社交网络中的影响力分析受到了学术界、工业界的广泛关注,微博数据挖掘代表性的研究主要包括话题事件分析、情感分析、信息检索与推荐、网络关系分析、信息传播、影响力分析等.

### 1.1 微博话题事件分析

事件(event)指由某些原因、条件引起,发生在特定时间、地点,并可能伴随某些必然结果的一个特例.

话题(topic)包括一个核心事件或活动,以及所有与之直接相关的事件或活动.微博中的话题事件分析研究主要包括事件检测与跟踪、首事件检测、突发事件检测、话题摘要以及话题模型等.

1) 事件检测与跟踪:事件检测与跟踪的目标为对文本信息流进行新话题的自动识别和已知话题的持续跟踪.事件检测与跟踪的基础方法为计算文档之间的相似性,文档之间相似性常用度量方法为夹角余弦,即

$$sim(D_t, D_s) = \cos(\boldsymbol{v}_t, \boldsymbol{v}_s), \quad (1)$$

其中 $D_t$ 与 $D_s$ 分别表示两篇文档,$\boldsymbol{v}_t$ 和 $\boldsymbol{v}_s$ 分别表示两篇文挡的向量,$sim(D_t, D_s)$表示计算两篇文挡相似性的函数.

Sakaki 等人[4]针对 Twitter 信息的实时性,提出了一种算法来监控博文以及检测目标事件,针对目标事件设计了时空模型,发现事件扩散的地点轨迹. Popescu 等人[5]提出了监督式的机器学习方法,检测 Twitter 中的争议事件. Weng 等人[6]针对微博中大量无意义的噪音数据,利用小波分析法过滤琐细的词,更加准确地检测 Twitter 中的事件. Becker 等人[7]针对 Twitter 中的博文流,提出了在线聚类技术识别真实世界的事件. Ritter 等人[8]针对 Twitter 的短文本以及富含噪音数据等特性,提出了开放领域事件抽取方法,利用潜在变分模型发现重要的事件类别. Lin 等人[9]针对微博流的短文本特性,利用基于复杂的分类器过滤博文流,提出了面向在线语言模型的平滑技术跟踪微博流中的话题. Hong 等人[10]针对 Twitter 中的位置服务,利用统计话题模型和稀疏编码技术,提出了一种稀疏产生式模型发现微博流中地理位置话题.

2) 首事件与突发事件检测:首事件与突发事件检测的目标为对文本信息流中的每篇文档、顺序判断其是否描述了一个新的或者突发的事件,如图 4 所示:

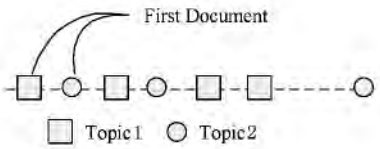


Fig. 4 First story detection.

图 4 首事件与突发事件检测示意图

Petrovic 等人[11]针对微博流,提出了位置敏感的 Hash 方法检测 Twitter 中的首事件. Phuvipadawat 等人[12]针对微博中短文本带来的相似性计算问题,提出了一种方法来收集、分组、排序以及跟踪 Twitter 中的突发新闻. Hong 等人[13]利用 Twitter 中的转发特征预测微博中的流行信息,检测突发新闻.

3) 话题摘要:话题摘要的目标为对一个话题的文档集合自动生成摘要,有助于理解话题的核心语义. Chakrabarti 等人[14]利用隐马尔科夫模型学习 Twitter 中事件的隐状态,摘要话题的所有博文. Yang 等人[15]针对 Twitter 中的流数据,提出了增量式的话题摘要方法,能够以低压缩率、高质量、高效率的方式在内存中摘要博文. Zhao 等人[16]提出了上下文相关的话题 PageRank 方法[17]排序关键词,利用话题关键词摘要博文.

4) 话题模型:常见的话题模型为向量空间模型和潜在的狄利克雷分布(latent Dirichlet allocation, LDA). 模型 LDA 最早由普林斯顿大学的 Blei 于 2003 年提出,近几年内被广泛应用. LDA 模型是一个 3 层贝叶斯概率模型,包含词、主题和文档 3 层结构,将每个文档表示为一个主题混合,每个主题是固定词汇表上的一个多项式分布,LDA 的产生式模型如图 5 所示:

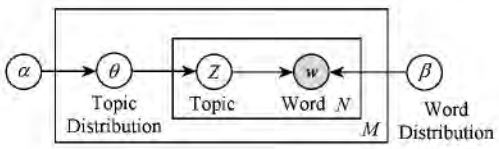


Fig. 5 The model of LDA.

图 5 LDA 模型

因此,给定参数 $\alpha$ 和 $\beta$,话题分布 $\theta$,$N$ 个话题 $z$,以及 $N$ 个词汇 $w$ 的联合概率分布为

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta). \quad (2)$$

Ramage 等人[18]针对 Twitter 中博文的内容特征,利用标签 LDA(labeled LDA)模型将博文内容映射到 4 个维度. Zhao 等人[19]利用话题模型系统地比较了 Twitter 和传统媒体. 国内浙江大学张晨逸等人[20]综合考虑了微博的联系人关联关系和文本关联关系,提出了一种基于 LDA[21]的微博生成模型 MB-LDA 来辅助挖掘微博的主题.

微博数据的实时性、大规模性、短文本特性、以及富含噪音数据等特性为话题事件分析带来了新的挑战.由于微博数据集中含有大量的噪音以及用语不规范、文章短小,使得传统的话题挖掘技术不能直接应用到微博数据中;微博里富含大量现实中的实

时话题,如何克服微博流的高速、海量特性,快速、准确地检测与跟踪实时话题,是微博面临的巨大挑战;微博数据不仅包含大量文本内容,而且富含链接信息,结合内容与链接关系,研究微博的话题模型,这也将是微博亟待解决的问题.

### 1.2 微博情感分析

情感(sentiment)分析又称意见挖掘,是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程.给定文档 $D$,情感分析具有两个目标.首先,情感分析算法将文档 $D$ 分成两类:1)主观的;2)客观的.情感分析的另一个目标为判断文档 $D$ 在对应话题类别下的态度:积极的或者消极的.因此,情感分析能够应用到信息检索领域,判断用户是否喜欢指定文档 $D$.情感分析通常利用分类技术判断文挡 $D$ 的态度,基于相似度的方法为情感分析常用方法.

对于词语 $t$ 和基准词集($S_p$ 代表正向基准词;$S_n$ 代表负向基准词),则词语 $t$ 的语义倾向 $O(t)$ 可以表示为它与正、负基准词之间互信息的差值,即

$$O(t)=\sum_{t_i\in S_p}PMI(t,t_i)-\sum_{t_i\in S_n}PMI(t,t_i),\quad(3)$$

其中,词语的逐点互信息可以通过语料中词语的共现信息等方式得到.

微博平台的开放性使得每个用户都能通过简单的文字表达自己的情感、情绪,微博中的情感分析已经引起相关学者的关注.

Barbosa 等人[22]利用博文特征以及词的元信息检测 Twitter 中的情感. Davidov 等人[23]利用 Twitter 中的 50 个 Hashtag 情感标签和 15 个“笑脸”标签,提出了监督式的分类方法识别情感. Bermingham 等人[24]针对微博的短文本特性,研究了微博中的情感分类技术. Diakopoulos 等人[25]针对 2008 年美国大选时的演讲辩论,依靠 Twitter 数据分析选民情感. Tumasjan 等人[26-27]针对 2009 年德国联邦选举,利用 Twitter 数据分析博文在政治方面的倾向性. Bifet 等人[28]针对微博的流数据特性,提出了基于滑动窗口的 Kappa 统计方法来分析 Twitter 中的情感. Jiang 等人[29]针对 Twitter 中的查询需求,提出了目标依赖的情感分类方法. Thelwall 等人[30]研究了 Twitter 中事件的情感变化趋势,结果表明流行事件通常都具有消极的情感色彩,随着事件发展,情感极性变强,且在事件高潮期具有更强的极性. Bollen 等人[31]通过提取 Twitter 上的关键词分析公众情绪,再将情绪曲线与道琼斯工业指数进行对照,分析股票市场;同时,Bollen 等人[32]利用 Twitter 中约 5 个月的数据,提取了微博中的 6 维情绪:紧张、忧伤、生气、热情、疲劳以及困惑.

微博中用户之间的互动性会导致公众情绪演变、漂移等特征;同时微博的复杂性将导致情感的多维特性,简单地以正、负、中立 3 维已经无法全面衡量用户情绪,这些都将为微博中的情感分析带来新的挑战.

### 1.3 微博信息检索与推荐

信息检索(information retrieval)是从大规模非结构化数据的集合中找出满足用户信息需求的资料的过程.信息推荐(information recommendation)是将满足需求的信息通过某种方式推荐给相关用户.微博数据的海量性、短文本性、富含噪音性等特性为信息检索和信息推荐带来了新的问题.

1) 信息检索:信息检索的目标为将信息按一定的方式组织起来,并根据信息用户的需要找出有关的信息.信息检索主要模型为概率模型,文档 $d_j$ 对于查询串 $q$ 的相关度值定义为

$$Sim(d_j,q)=P(R|d_j)/P(R|d_j).\quad(4)$$

根据贝叶斯原理:

$$Sim(d_j,q)=P(d_j|R)P(R)/P(d_j|R)P(R),\quad(5)$$

其中,$P(d_j|R)$ 代表从相关文档集合 $R$ 中随机选取文档 $d_j$ 的概率,$P(R)$ 表示从整个集合中随机选取一篇文档作为相关文档的概率,依此定义 $P(d_j|R)$ 和 $P(R)$.

概率模型是基于以下基本假设:

① 给定一个用户的查询串 $q$ 和集合中的文档 $d_j$,概率模型估计用户查询串与文档 $d_j$ 相关的概率;

② 概率模型假设这种概率只决定于查询串和文档;

③ 该模型假定在文档集合中存在一个子集,即相对于查询串 $q$ 的结果文档子集,这种理想的集合用 $R$ 表示,集合中的文档是被预料与查询串相关的.

为了解决微博中信息检索问题,著名文本检索会议 TREC 从 2011 年开始,增加了微博检索(microblog track)这一新任务,公布了约 1 600 万条 Twitter 中的博文数据,目的是实现微博的实时检索.信息检索主要任务之一是检索项的排序问题,Sarma 等人[33]利用用户的评论研究了 Twitter 中博文排序方法;Dong 等人[34]针对微博检索中的实时性问题,利用微博流数据实时地检测最新的 URLs,同时根据 URLs 的新颖度与有效特征对 URLs 排序.

信息检索另一项主要任务是索引问题，Yao 等人[35]利用 Twitter 数据集，研究了微博中的索引技术，能够有效地支持微博中的查询检索任务. Teevan 等人[36]系统地对比了微博搜索与 Web 搜索问题. Spina 等人[37]比较了信息检索方法与意见目标识别方法，利用微博流识别微博中的实体，有利于微博中的实体检索.

2) 信息推荐：信息推荐的目标为分析大量用户的行为规律，计算大部分用户的行为偏好，从而自动向用户推荐相关信息.

协同过滤技术是信息推荐中最广泛使用的技术，协同过滤算法的推荐原理就是查找与目标用户相似的近邻用户，通过近邻用户的评价对目标用户产生推荐. 近邻用户的选择方法如下：计算目标用户与推荐系统中其他所有用户的相似性，根据相似性排序从大到小依次选择前面 $K$ 个最相似的用户作为目标用户的近邻集合. 其中，相似性度量方法的选择对于推荐精度有着至关重要的影响，常用的相似性度量方法有皮尔逊相关、余弦相似性、修正的余弦相似性等.

为了解决微博中的信息推荐问题，数据挖掘及知识发现专委会主办的“国际知识发现和数据挖掘竞赛(KDD-CUP)”于 2012 年增加了微博信息推荐这一新项目，会议组提供了腾讯微博(Tencent Weibo)约 1 000 万个用户、50 000 个推荐项、以及 3 亿个推荐记录的数据集，目的是预测用户是否会关注推荐项. Ting 等人[38]利用博文内容与网络关系实现微博中的信息推荐. Brzozowski 等人[39]利用有向网络的结构模式向用户推荐好友. Abel 等人[40]分析了微博中的用户模型，实现了 Twitter 中的个性化新闻推荐.

微博通常反映实时新闻，且信息存在大量的重复性、琐碎性等特点，如何检索实时的、有价值的、有影响力的信息将是微博检索面临的挑战. 微博中用户特性复杂、表现存在差异性，如何克服微博海量性、富含噪音性等特点对每个用户建立个性化模型，实现微博个性化推荐将是微博信息推荐面临的挑战. 所谓微博个性化推荐是指根据不同用户的兴趣特点，对每个用户建立不同的推荐模型，向用户推荐感兴趣的信息.

### 1.4 微博关系分析与挖掘

微博用户之间的交互多样性使得微博网络呈现多关系特性，用户可以根据关注关系构造朋友网络；根据转发关系构造传播网络；根据回复关系构造评论网络. 目前相关研究主要集中在分析关系形成机制、关系预测(prediction)等.

1) 关注关系形成机制研究：微博中，用户可以通过关注关系结交新的好友，比如在新浪微博中用户可以点击“关注”一个用户，而使得自己成为该用户的粉丝，从而形成一条“关注边”. 关注关系形成机制研究有助于了解微博社交网络形成机理.

Romero 等人[41]根据 Twitter 中的关注关系，研究了微博中的三角闭包关系，分析了 Twitter 中关注关系形成机制；同时，Romero 等人[42]研究了微博中用户关注关系的交互机制，验证了关系保持理论：平衡性、交互性、中介性. Yin 等人[43]研究了微博中关注关系形成机制，实验结果表明大约 90% 的新链接由两跳关系形成. Kwak 等人[44]研究了 Twitter 中不关注其他用户的动态行为，分析了影响行为的因素：关系互惠、关系周期性、好友信息以及关系重叠性. Meeder 等人[45]利用一个静态关注网络和用户账号创建时间，推断 Twitter 的网络链接形成时间. Java 等人[46]研究了 Twitter 关注网络的拓扑结构以及地理属性，分析了用户如何以及为什么使用微博.

2) 转发关系形成机制研究：微博中，用户可以点击“转发”将一条博文扩散出去，微博转发关系形成机制研究有助于了解微博中信息扩散的机理.

Yang 等人[47]研究 Twitter 中的转发机制，结果表明用户发表的博文中，约 25.5% 的信息是依靠转发好友博文而产生的. Macskassy 等人[48]研究表明微博中大部分用户并不一定转发他们熟知的话题. Recuero 等人[49]研究了社会资本是如何影响转发行为的. Boyd 等人[50]研究了 Twitter 中用户是如何利用转发而形成对话的. Welch 等人[51]研究了 Twitter 中关注关系和转发关系的语义信息，发现转发关系具有更强的话题关联性.

3) 关系预测：关系预测的目的为通过分析历史数据，预测未来两用户之间是否会形成新的边，传统方法通常依靠两用户间的共同邻居数据来计算用户间的关系强度，从而来预测是否会形成新的边，利用杰卡德相似系数(Jaccard coefficient)计算两节点的共同邻居数据.

$$S(A,B)=\frac{|n_A \bigcap n_B|}{|n_A \bigcup n_B|}. \tag{6}$$

Yin 等人[52]利用 Twitter 中的关注关系，提出了一种基于结构的链接预测方法，预测用户下一步将关注哪些用户. Cheng 等人[53]利用决策树和回归模型预测 Twitter 中的互惠关系. Ediger 等人[54]针

对 Twitter 网络关系的大规模性,提出 GraphCT 方法提取有价值的数据,使得分析能够集中于更小的数据集.

微博中用户数量庞大、交互复杂,使得微博网络呈现大规模性、异构性等特点;同时微博网络是人类在虚拟网络世界生活的抽象概括和延伸,具有社会性.微博的新特点使得传统的链接结构分析技术不能直接应用到微博中,需要进一步地研究微博的结构特性与演化机理,挖掘微博中的社区结构,分析社区形成机制及演化特性.针对微博的社会性,需要研究微博用户之间的博弈关系,利用博弈理论分析微博中的关系形成机制.

### 1.5 微博信息传播

微博的转发功能使得信息无限制地被转发,其信息传播范围呈"核裂变"式的几何级数式扩大,研究微博中信息的传播模式与扩散机制,有助于微博舆情的控制与引导、以及企业品牌和产品的推广.传统的信息传播模型主要包括线形阈值模型和独立级联模型.其中,线形阈值模型中,节点受到影响被激活(感染)的条件为

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geqslant \theta_v, \tag{7}$$

表示所有邻居节点概率 $b_{v,w}$ 之和大于一定阈值.

独立级联模型中,节点受到影响被激活(感染)的条件为

$$b_{v,w} > \theta, \tag{8}$$

表示每个节点都已一定概率 $b_{v,w}$ 激活(感染)其邻居节点.

微博信息传播相关研究主要集中在传播机制理论分析与实验验证、传播预测、传播案例研究等.

1) 传播机制的理论分析与实验验证:微博的转发功能使得信息无限制快速地被转发,从而使得信息快速、大范围地被扩散,微博中传播机制的理论分析与实验验证有助于了解微博中的信息扩散模式与扩散机理.

Romero 等人[55]研究了不同的 HashTag 在 Twitter 中的传播模式,发现有争议性的政治话题传播通常持续更长的时间,但习语和新词持续时间通常较短,且传播路径也存在差异性. Wu 等人[56]研究了 Twitter 中的名人、作家、媒体、组织、和普通用户的信息扩散机制,发现约 50% 的 URLs 仅由两万个名人产生,媒体产生了大多数信息;同时验证了微博中的两阶段传播理论. Dabeer 等人[57]分析了影响微博信息传播的因素:粉丝节点的活跃性、粉丝节点对源信息节点的响应性、粉丝节点出度以及信息本身特点,提出了基于马尔科夫决策处理的框架来度量 Twitter 中信息传播效果. Lehmann 等人[58]跟踪了 Twitter 网络中的 HashTag 的扩散过程,发现流行病传播模式起着重要作用. Yang 等人[59]对比了博客与微博的信息扩散模式,系统地对比了两种媒体的贡献、导航和互动结构模式. Lerman 等人[60]通过实验对比了 Digg 和 Twitter 网络中的信息传播模式,结果表明网络结构通常影响信息的传播模式. Lin 等人[61]综合考虑了微博中的博文内容、影响力和话题演化,提出了概率产生式模式来推理微博中的话题传播与演化模式. Xue 等人[62]研究了微博中信息传播的源头识别技术. Lumezanu 等人[63]研究了 Twitter 中扩散者发帖行为,发现:①短时间内发布大量博文;②仅转发大量博文;③转发速度快;④勾结其他用户共同发布相同信息. Naaman 等人[64]和 Yang 等人[65]分别研究了微博中信息扩散的时序变化模式. An 等人[66]研究表明,Twitter 用户通常直接或者间接与媒体用户关联,且媒体用户提高了信息扩散范围约 60%~98%.

2) 传播预测:微博中传播预测的目标为预测信息是否会被某用户传播以及信息的传播范围等.

Yang 等人[67]预测了微博中信息传播的速度、规模和范围. Tsur 等人[68]结合博文内容与网络拓扑结构,利用线性回归方法预测给定时间内的信息扩散. Petrovic 等人[69]利用一种基于被动进取的机器学习方法预测微博中的博文是否会被其他用户转发. Yang 等人[70]利用微博的网络关系提出了线性影响力模型,预测信息扩散路径,对每个节点的全局扩散能力建模.

3) 传播案例研究:Lampos 等人[71]利用 Twitter 开发了一种能够自动跟踪英国流行病的工具. Ratkiewicz 等人[72]利用 Twitter 跟踪政治信息传播,检测诽谤活动、误导信息等.

微博的自身特点为信息传播带来了良好的滋生环境,使得以前复杂网络中仅能依靠模拟的传播过程得以现实化,但微博的开放性打破了传统研究中的"封闭式"假设,微博中的信息传播不仅依靠网络结构,而且还与外部环境、信息本身等因素相关,从而使得传统的信息传播模型不能直接应用到微博中.同时微博的信息传播也为传播学、社会学、管理学的基础理论提供了实验验证数据,比如是否可以用羊群效应和从众行为解释微博中的信息传播,这些都是微博信息传播面临的问题.

### 1.6 微博中影响力分析

微博中的用户受到网络中其他用户的影响作用而导致个人行为变化,影响力(influence)在微博中是一个普遍存在的现象.传统个体影响力度量技术的相关研究主要包括点度中心度(degree)、接近中心度(closeness)、中间中心度(betweenness)、HITS、PageRank及扩展方法等.

1)点度中心度:指的是该节点的度数,即与该节点直接相连的节点个数.点度中心度用来分析节点直接影响力,即考察个体的直接社会关系.令 $A$ 是网络图的邻接矩阵,$deg(i)$ 为节点 $i$ 的度,则节点 $i$ 的点度中心度 $c_i^{\mathrm{DEG}}$ 即该节点的度数:

$$c_i^{\mathrm{DEG}}=deg(i). \tag{9}$$

点度中心度比较直观地衡量一个节点的影响力,计算开销相对较小,但针对大规模的微博网络,将忽略部分影响力个体.

2)接近中心度:指个体与社交网络中所有其他节点的捷径距离(最短路径)之和.接近中心度用来分析个体通过社交网络对其他个体的间接影响力.节点 $i$ 的接近中心度 $c_i^{\mathrm{CLO}}$ 定义如下:

$$c_i^{\mathrm{CLO}}=\boldsymbol{e}_i^{\mathrm{T}}\boldsymbol{S}\boldsymbol{e},\boldsymbol{e}=(1,1,\cdots,1)^{\mathrm{T}}, \tag{10}$$

其中 $\boldsymbol{S}$ 为一个矩阵,它的第$(i,j)$个元素表示从节点 $i$ 到节点 $j$ 最短路径长度,$\boldsymbol{e}_i^{\mathrm{T}}$ 表示第 $i$ 个元素为 1 的向量.

接近中心度需要计算网络中所有节点对之间的最短路径,计算开销大,优点是能够衡量一个节点的间接影响力.

3)中间中心度:指的是节点处于其他节点最短路径上的能力.中间中心度用来分析节点对信息传播的影响,即个体在多大程度上处于其他个体的中间,是否发挥出"中介"作用.节点 $i$ 的中间中心度 $c_i^{\mathrm{BET}}$ 定义如下:

$$c_i^{\mathrm{BET}}=\sum_{j,k}\frac{b_{jik}}{b_{jk}}, \tag{11}$$

其中,$b_{jk}$ 表示节点 $j$ 与 $k$ 之间最短路径数目;$b_{jik}$ 表示节点 $j$ 与 $k$ 之间,且通过节点 $i$ 的最短路径数目.计算中间中心度的朴素方法为计算所有节点对之间的最短路径,需要 $O(n^3)$ 的时间开销与 $O(n^2)$ 的空间开销.

4)HITS:由康奈尔大学的 Kleinberg 提出,英文全称为 Hypertext Induced Topic Search,最初应用在搜索引擎中,根据一个网页的中心度(hub)和权威度(authority)来衡量网页重要性.对网络图中的每个节点 $v_i$,令 $a(v_i)$ 为该节点的权威度,$h(v_i)$ 为该节点的中心度.则节点权威度与中心度定义如下:

$$a^{(k+1)}(v_i)=\sum_{v_j\in inlink[v_i]}h^{(k)}(v_j),$$
$$h^{(k+1)}(v_i)=\sum_{v_j\in outlink[v_i]}a^{(k+1)}(v_j). \tag{12}$$

HITS 算法综合考虑了节点的权威度与中心度,需要迭代计算,但却忽略了节点影响力的划分.

5)PageRank:由 Google 创始人之一 Page 提出,最初应用在搜索引擎中,根据网页之间的超链接计算网页排名,一个页面的得票数由所有链向其页面的重要性决定,但随后学者将 PageRank 算法应用到社会网络中,为个体影响力度量的基础算法.PageRank 算法用一种基于马尔科夫的随机游走思想来模拟用户浏览网页的行为.令 $\pi$ 为网络中节点影响力得分向量,$\boldsymbol{P}$ 为网络图的转移矩阵,则 PageRank 计算公式如下:

$$\boldsymbol{\pi}=\alpha\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\pi}+(1-\alpha)\frac{1}{n}\boldsymbol{e},\boldsymbol{e}=(1,1,\cdots,1)^{\mathrm{T}}, \tag{13}$$

其中,$\alpha$ 为跳转因子,$\frac{1}{n}\boldsymbol{e}$ 为自重启向量.

PageRank 算法考虑了节点影响力的传播,需要迭代计算,但却忽略了节点自身特征,微博中用户行为表现复杂、且用户规模数量庞大,仅依靠网络结构将忽略更加细粒度的影响力个体,比如无法发现话题层次的影响力个体.相关学者针对这一问题,在 PageRank 算法基础上提出了结合个体特征与网络结构的影响力度量技术.

6)PageRank 算法扩展:社会网络中个体特征主要包括个体发布信息所属话题类别,研究每个话题类别的影响力个体;另外还包括个体发布信息的新颖度、敏感度等,创新能力强的个体通常具有更高的影响力,同时发布敏感信息的个体通常具有更高的影响力.Haveliwala 等人考虑个体用户特征,在 PageRank 算法基础上,提出了 Personalized PageRank 算法,计算公式如下:

$$\boldsymbol{\pi}=\alpha\boldsymbol{P}^{\mathrm{T}}\boldsymbol{\pi}+(1-\alpha)\boldsymbol{r}. \tag{14}$$

Personalized PageRank 算法将均分自重启向量 $\frac{1}{n}\boldsymbol{e}$ 改为个性化向量 $\boldsymbol{r}$,比如元素 $r_i$ 表示个体对话题的偏好程度、个体发布信息的新颖程度与敏感程度等.

微博中影响力分析的目标为利用微博的网络关系以及文本信息,综合衡量每个用户的影响力,挖掘微博中的意见领袖.目前,微博中影响力主要依靠扩散能力、个体特征与网络结构等来衡量.

7) 依靠扩散能力衡量影响力:Bakshy 等人[73]在 Twitter 数据集中,根据每个相同的网页链接 URL 构造传播级联树,用种子节点的扩散范围来衡量每个种子节点的影响力. Lee 等人[74]在 Twitter 数据集上模拟关注网络中的信息传播,通过计算用户的有效读者数来衡量一个用户的影响力. Aggarwal 等人[75]提出了一种随机信息流模型来发现 Twitter 中有代表性的权威节点. Steeg 等人[76]利用转移熵理论刻画用户间的信息流,识别 Twitter 网络中有影响力的链接.

8) 依靠个体特征与网络结构衡量影响力:Cha 等人[77]在 Twitter 数据集中分别利用个体的粉丝数目、被转发数以及被提及数来衡量个体的影响力. Pal 等人[78]在 Twitter 数据集上考虑了个体的发帖数、回复数、被转发数、被提及数(mention)和粉丝数目,分别计算个体的转发影响力、被提及影响力和扩散影响力等. Quercia 等人[79]研究了 Twitter 关注网络结构,发现大多数网络结构洞是意见领袖,且发布各种话题,情感变化丰富. Tunkelang 等人[80]针对 Twitter 中的关注关系构造了一种类似 PageRank 的算法,该算法用粉丝的影响力来衡量个体的影响力,粉丝越重要,且关注其他用户越少,则粉丝对该用户影响力贡献越大. Weng 等人[81]在 Twitter 数据集上提出了 TwitterRank 算法,该算法为 PageRank 算法扩展,根据关注网络和用户兴趣相似性计算个体在每个话题上的影响力. Romero 等人[82]综合考虑了影响力与冷漠性,提出了类 HITS 算法[83]的 IP(influence-passivity)算法,度量 Twitter 中个体影响力. Liu 等人[84]针对 Twitter 网络的异构性,提出了产生式图模型来度量异构网络中的话题影响力. Li 等人[85]依靠微博中的历史消息和社会交互记录,利用统计学习过程构造历史意见和意见影响力,提出了话题级的意见影响力模型,合并话题因素与社会影响力. Ding 等人[86]针对微博交互的多关系特性,研究多关系的影响力个体挖掘方法. Ding 等人[87]综合考虑时间、博文内容以及网络关系研究微博中的影响强度计算方法.

微博的快速发展吸引了大量垃圾用户的加入,垃圾用户依靠机器人程序频繁地交互与发布大量博文,其特征类似影响力个体,垃圾用户的存在降低了传统影响力个体发现方法的准确率;微博的社会媒体性使得用户不仅使用微博交友,更倾向于使用微博接受有价值的信息,使得传统的仅依靠网络结构衡量影响强度的方法不能直接应用到微博中;微博的多关系网络特性也为影响力衡量提供了新的挑战.

### 1.7 其 他

微博的其他相关研究还包括微博自身特性研究、微博地理位置研究、博文价值评价与可信度研究、微博用户分类和垃圾用户发现、以及微博在日常生活的应用等.

1) 微博自身特性研究:Kwak 等人[2]研究了微博网络拓扑与信息共享特征,表明微博不仅具有社交网络功能,更倾向于具有为社会媒体功能. Mislove 等人[88]研究了 Twitter 用户的人口特征,发现 Twitter 中的美国用户大多数为男性. Huang 等人[89]研究了 Twitter 中的 Tag 行为,发现 Twitter 中的 Tag 具有更强地过滤与标记信息的能力. Kıcıman 等人[90]研究了美国实际天气情况与 Twitter 发帖速率的关系,利用天气信息能够解释 Twitter 中超过 40% 的发帖速率变化情况. Perreault 等人[91]研究了移动平台在 Twitter 产生内容过程中的作用. 国内国防科学技术大学樊鹏翼等人[92]对新浪微博的网络拓扑和用户行为特征进行了分析和比较,发现了新浪微博的小世界等特性.

2) 微博地理位置研究:Cheng 等人[93]针对微博中地理位置稀疏特性,利用用户的博文内容,提出了概率框架来评估 Twitter 中用户城市级的地理位置. Kulshrestha 等人[94]分析了 Twitter 中用户行为特性与地理位置关系,发现相同国家的用户通常交互更频繁.

3) 博文价值评价与可信度研究:博文价值评价与可信度研究的目标为评价微博中每篇博文的价值以及可信度,即该博文的真实性.

Castillo 等人[95]利用 Twitter 中用户的发帖和转帖行为,引用的外部链接等特征评价博文信誉度. Morris 等人[96]研究了 Twitter 信誉度评估,发现仅依靠博文内容信誉度评价性能较低,随后提出了结合博文内容与用户行为的信誉度评估方法. Gupta 等人[97]利用事件图优化方法提高 Twitter 中事件信誉度的评估性能.

4) 微博用户分类和垃圾用户发现:垃圾用户发现的研究目标为针对微博用户的行为特征以及交友特性,发现类似依靠程序自动发贴的“发贴机器人”、“转发机器人”等.

Pennacchiotti 等人[98]针对 3 个具体任务:政治派别检测、种族识别、特定业务的紧密性检测,提出了基于机器学习框架的大规模微博用户分类方法. Chu 等人[99]依靠 Twitter 中用户发帖行为、博文内容

和账号属性等特征来区分 Twitter 中的垃圾机器人、类人机器人和正常用户. Thomas 等人[100]利用 Twitter 中暂停账号分析垃圾用户特性,发现约 77%的垃圾用户账号在注册的第 1 天就被暂停使用,超过 91%的垃圾用户的网络关系不同于正常用户,17%的垃圾用户劫持热点信息,52%的垃圾用户恶意地提及(mention)正常用户. Lee 等人[101]利用 Twitter 中 7 个月的数据自动检测微博中的垃圾用户. Yang 等人[102]针对 Twitter 的网络犯罪生态系统,研究垃圾用户的社区结构. Ghosh 等人[103]针对微博网络链接结构,研究了 Twitter 网络中的垃圾链接形成机制. Ding 等人[104]利用双向投票模型发现微博中的垃圾用户.

5) 微博在日常生活的应用:Golbeck 等人[105]分析了美国议员使用 Twitter 的行为,分析表明议员主要使用 Twitter 发布信息、报告日常活动等. Junco 等人[106]研究了 Twitter 在大学生中的作用. Oh 等人[107]研究了 Twitter 在跟踪恐怖袭击中的作用. Vieweg 等人[108]研究了 Twitter 在危机管理中作用. Paul 等人[109]分析了 Twitter 在公共医疗中的作用. Paul 等人[110]研究表明 Twitter 可以用来咨询问题. Conover 等人[111]研究了 Twitter 在政治中的作用. Forte 等人[112]分析了教师是如何使用 Twitter 的.

## 2 Twitter 数据集分析

### 2.1 数据集获取

本文获取了 Twitter 的中文用户,中文用户包括 4 个特征:1)昵称(username)使用中文符号;2)个人简介使用中文符号;3)地理位置为中国城市;4)至少发布一条中文符号的博文.

利用 Twitter 的 API 获取实验数据,对于 Twitter 的每个中文用户,获取最近 200 条博文(如果用户发布的博文数目小于 200,则获取该用户所有博文). 实验数据获取采用"滚雪球"式的获取方法,如图 6 所示,首先人工从 Twitter 中选取粉丝大于 1 万的 100 个用户作为种子用户. 对于每个种子用户,获取最近 200 条博文以及该种子用户所有粉丝和好友,根据中文用户 4 个特征判断粉丝和好友是否为中文用户,如果是中文用户,且未曾获取该用户的信息,则该用户加入到新的种子用户集合,如此反复迭代,直到新加入的种子用户数目与已获取信息用户数目的比例小于一定阈值,则终止迭代过程.

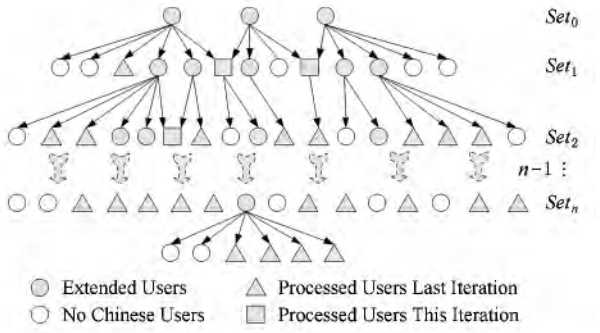


Fig. 6 Data collection.

图 6 数据获取方法

实验总共获取 Twitter 的 261 954 个中文用户,10 091 543 条博文和 17 546 289 条关注关系边,在下一节将分析数据集的特征.

### 2.2 数据集特征分析

本节首先分析了用户粉丝数目、发帖数目分布情况;随后分析了粉丝数目与好友数目的关联性和粉丝数目与发帖数目的关联性;继而分析了中文用户的时区分布情况;最后分析了中文用户数量随时间的变化趋势.

图 7 表示用户粉丝数目和发帖数目皆近似服从

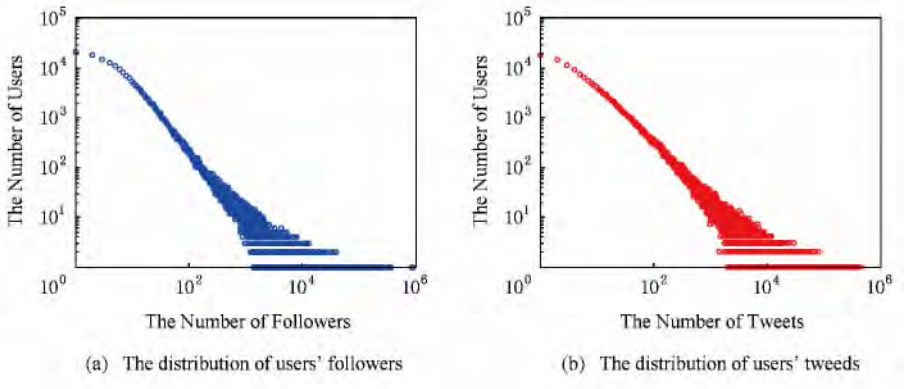


Fig. 7 The distribution of users' followers and tweets.

图 7 粉丝数目和发帖数目分布

幂律分布,表示 Twitter 中大部分中文用户拥有少量的粉丝和发布少量的博文,仅存在少部分中文用户拥有高数量的粉丝或者发布大量的博文.

图 8 表示 Twitter 的中文用户时区分布情况,时区分布前八分别为北京、阿拉斯加、香港、台北、夏威夷、新加坡、美国太平洋时区、美国东部时区. 由于 Twitter 的系统默认时区为阿拉斯加时区,从而导致部分用户使用缺省时区,使得阿拉斯加时区比例偏高. Twitter 的中文用户时区分布表明,Twitter 的大部分中文用户来自中国境内、香港、和台湾地区.

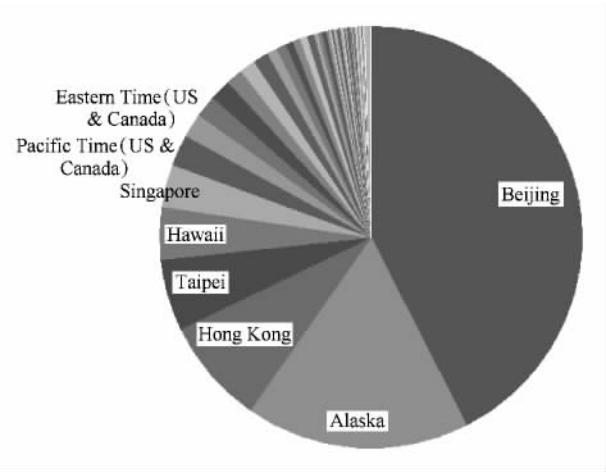


Fig. 8 The distribution of time zone for Twitter.

图 8 Twitter 中文用户时区分布

根据 Twitter 中账号的创建时间分析每个时间段内中文用户数量,如图 9 所示,结果表明在 Twitter 刚创建时,仅有少部分中文用户使用 Twitter,在 2007 年和 2008 年,Twitter 的中文用户数量缓慢上升;2009 年和 2010 年两年内,Twitter 的中文用户数量开始急剧上升;在随后的 2011 年内,Twitter 中文用户数目基本保持稳定,略有下降.

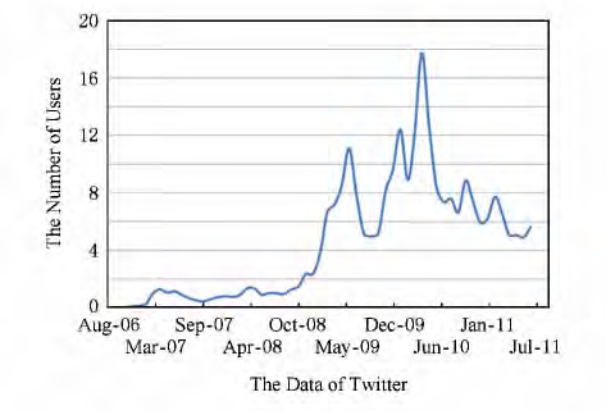


Fig. 9 The trend of users' number for Twitter.

图 9 Twitter 中文用户数量随时间变化

## 3 总 结

本文主要总结了微博的话题事件分析、情感分析、信息检索与推荐、网络关系分析、信息传播、影响力分析等研究现状. 微博的噪音数据多样性、开放性等特性为将来研究带来了新的挑战.

1) 噪音数据多样性:随着微博服务的普及,存在大量以刺探隐私情报、商业推销、推高用户人气、制造与传播舆论等为目的人工垃圾链接和垃圾用户. 垃圾用户存在的目的不同,导致其行为特征的差异性与多样性. 部分垃圾用户为达到特定目的,其行为特征类似于意见领袖. 比如:①连续发布大量的博文吸引正常用户的关注;②连续地提及(mention)正常用户而吸引其他用户的关注;③关注大量正常用户而吸引其他用户的关注. 同时部分"虚假"的意见领袖为了提高自己的人气,利用技术手段或者微博服务漏洞,制造大量"僵死粉",这些"僵死粉"行为特征具有隐蔽性,甚至部分"僵死粉"由特定商业公司操纵,其行为特征更加多样化,因此需要更进一步地研究微博的垃圾用户发现.

2) 微博开放性:微博的开放性,使得用户在一定程度上受到外部环境的支配,打破了传统社会网络分析的封闭式假设. 同时,博文自身携带的语义信息也决定博文的扩散情况,比如用户经常更愿意传播有价值、且富含政治色彩的博文信息,因此需要研究基于开放性的信息传播模型.

## 参 考 文 献

[1] Semiocast. Twitter reaches half a billion accounts more than 140 million in the U. S [EB/OL]. (2012-07-30) [2013-07-23]. http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

[2] Kwak H, Lee C, Park H, et al. What is Twitter, A social network or a news media [C] //Proc of the 19th Int Conf on World Wide Web (WWW'10). New York: ACM, 2010: 591-600

[3] Comscore. Mobile driving majority of growth for leading EU5 social networks [EB/OL]. (2012-05-18) [2013-07-23]. http://www.comscoredatamine.com/2012/05/mobile_driving_majority_f_growth_for_leading_eu5_social_networks

[4] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users: Real-time event detection by social sensors [C] //Proc of the 19th Int Conf on World Wide Web (WWW'10). New York: ACM, 2010: 851-860

[5] Popescu A M, Pennacchiotti M. Detecting controversial events from Twitter [C] //Proc of the 19th ACM Int Conf on Information and Knowledge Management (CIKM'10). New York: ACM, 2010: 1873-1876

[6] Weng J, Lee B S. Event detection in Twitter [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 401-408

[7] Becker H, Naaman M, Gravano L. Beyond trending topics: Real-world event identification on Twitter [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 438-441

[8] Ritter A, Mausam B, Etzioni O, et al. Open domain event extraction from Twitter [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'12). New York: ACM, 2012: 1104-1112

[9] Lin J, Snow R, Morgan W. Smoothing techniques for adaptive online language models: Topic tracking in Tweet streams [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'11). New York: ACM, 2011: 422-429

[10] Hong L, Amr A, Gurumurthy S, et al. Discovering geographical topics in the Twitter stream [C] //Proc of the 21st Int Conf on World Wide Web (WWW'12). New York: ACM, 2012: 769-778

[11] Petrovic S, Osborne M, Lavrenko V. Streaming first story detection with application to Twitter [C] //Proc of the 11th Annual Conf of the North American Chapter of the Association for Computational Linguistics (NAACL'10). Stroudsburg, PA: ACL, 2010: 181-189

[12] Phuvipadawat S, Murata T. Breaking news detection and tracking in Twitter [C] //Proc of the 9th IEEE/WIC/ACM Int Conf on Web Intelligence and Intelligent Agent Technology (WI-IAT'10). New York: ACM, 2010: 120-123

[13] Hong L, Dan O, Davison B D. Predicting popular messages in Twitter [C] //Proc of the 20th Int Conf Companion on World Wide Web (WWW'11). New York: ACM, 2011: 57-58

[14] Chakrabarti D, Punera K. Event summarization using Tweets [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 66-73

[15] Yang X, Ghoting A, Ruan Y. A framework for summarizing and analyzing Twitter feeds [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'12). New York: ACM, 2012: 370-378

[16] Zhao X, Jiang J, He J, et al. Topical keyphrase extraction from Twitter [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11). Stroudsburg, PA: ACL, 2011: 379-388

[17] Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web [OL]. (1999-01-20) [2013-07-23]. http://ilpubs.stanford.edu:8089/422/

[18] Ramage D, Dumais S, Liebling D. Characterizing microblogs with topic models [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 130-137

[19] Zhao X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models [C] //Proc of the 33rd European Conf on Advances in Information Retrieval (ECIR'11). Berlin: Springer, 2011: 338-349

[20] Zhang Chengyi, Sun Jianling, Ding Yiqun. Topic mining for microblog based on MB-LDA model [J]. Journal of Computer Research and Development, 2011, 48 (10): 1795-1802 (in Chinese)
(**张晨逸, 孙建伶, 丁轶群. 基于 MB-LDA 模型的微博主题挖掘**[J]. **计算机研究与发展**, 2011, 48 (10): 1795-1802)

[21] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022

[22] Barbosa L, Feng J. Robust sentiment detection on Twitter from biased and noisy data [C] //Proc of the 23rd Int Conf on Computational Linguistics (COLING'10). Stroudsburg, PA: ACL, 2010: 36-44

[23] Davidov D, Tsur O, Rappoport A. Enhanced sentiment learning using Twitter hashtags and smileys [C] //Proc of the 23rd Int Conf on Computational Linguistics (COLING'10). Stroudsburg, PA: ACL, 2010: 241-249

[24] Bermingham A, Smeaton A. Classifying sentiment in microblogs: Is brevity an advantage [C] //Proc of the 19th ACM Int Conf on Information and Knowledge Management (CIKM'10). New York: ACM, 2010: 1833-1836

[25] Diakopoulos N A, Shamma D A. Characterizing debate performance via aggregated Twitter sentiment [C] //Proc of the 28th ACM Conf on Human Factors in Computing Systems (CHI'10). New York: ACM, 2010: 1195-1198

[26] Tumasjan A, Sprenger T O, Sandner P G, et al. Predicting elections with Twitter: What 140 characters reveal about political sentiment [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 178-185

[27] Tumasjan A, Sprenger A T, Sandner P G, et al. Election forecasts with Twitter: How 140 characters reflect the political landscape [J]. Social Science Computer Review, 2011, 29 (4): 402-418

[28] Bifet A, Frank E. Sentiment knowledge discovery in twitter streaming data [C] //Proc of the 13th Int Conf on Discovery Science (DS'10). Berlin: Springer, 2010: 1-15

[29] Jiang L, Yu M, Zhou M, et al. Target-dependent Twitter sentiment classification [C] //Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11). Stroudsburg, PA: ACL, 2011: 151-160

[30] Thelwall M, Kevan B, Paltoglou G. Sentiment in Twitter events [J]. Journal of the American Society for Information Science and Technology Archive, 2011, 62 (2): 406-418

[31] Bollen J, Mao H, Zeng X. Twitter mood predicts the stock market [J]. Journal of Computational Science, 2011, 2 (1): 1-8

[32] Bollen J, Mao H, Pepe A. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 450-453

[33] Sarma A D, Sarma A D, Gollapudi S, et al. Ranking mechanisms in Twitter-like forums [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining (WSDM'10). New York: ACM, 2010: 21-30

[34] Dong A, Zhang R, Kolari P, et al. Time is of the essence: Improving recency ranking using Twitter data [C] //Proc of the 19th Int Conf on World Wide Web (WWW'10). New York: ACM, 2010: 331-340

[35] Yao J, Cui B, Xue Z, et al. Provenance-based indexing support in micro-blog platforms [C] //Proc of the 28th Int Conf on Data Engineering (ICDE'12). Piscataway, NJ: IEEE, 2012: 558-569

[36] Teevan J, Ramage D, Morris M R. TwitterSearch: A comparison of microblog search and Web search [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 35-44

[37] Spina D, Meij E, Rijke M D, et al. Identifying entity aspects in microblog posts [C] //Proc of the 35th Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'12). New York: ACM, 2012: 1089-1090

[38] Ting I H, Chang P S, Wang S L. Understanding microblog users for social recommendation based on social networks analysis [J]. Journal of Universal Computer Science, 2012, 1(1): 554-576

[39] Brzozowski M J, Romero D M. Who should I follow? Recommending people in directed social networks [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 458-461

[40] Abel F, Gao Q, Houben G J, et al. Analyzing user modeling on Twitter for personalized news recommendations [C] //Proc of the 19th Int Conf on User Modeling, Adaption, and Personalization (UMAP'11). Berlin: Springer, 2011: 1-12

[41] Romero D M, Kleinberg J. The directed closure process in hybrid social information networks, with an analysis of link formation on Twitter [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 138-145

[42] Romero D M, Meeder B, Barash V, et al. Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 606-609

[43] Yin D, Hong L, Xiong X, et al. Link formation analysis in microblogs [C] //Proc of the 34th Int ACM SIGIR Conf on Research and Development in Information Retrieval (SIGIR'11). New York: ACM, 2011: 1235-1236

[44] Kwak H, Chun H, Moon S. Fragile online relationship: A first look at unfollow dynamics in Twitter [C] //Proc of the 29th ACM Conf on Human Factors in Computing Systems (CHI'11). New York: ACM, 2011: 1091-1100

[45] Meeder B, Karrer B, Sayedi A, et al. We know who you followed last summer: Inferring social link creation times in Twitter [C] //Proc of the 20th Int Conf Companion on World Wide Web (WWW'11). New York: ACM, 2011: 517-526

[46] Java A, Song X, Finin T, et al. Why we twitter: Understanding microblogging usage and communities [C] //Proc of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD'07). New York: ACM, 2007: 56-65

[47] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks [C] //Proc of the 19th ACM Int Conf on Information and Knowledge Management (CIKM'10). New York: ACM, 2010: 1633-1636

[48] Macskassy S A, Michelson M. Why do people retweet? Anti-Homophily wins the day [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 209-216

[49] Recuero R, Araujo R, Zago G. How does social capital affect retweets [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 305-312

[50] Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter [C] //Proc of the 43rd Hawaii Int Conf on System Sciences (HICSS'10). Los Alamitos, CA: IEEE Computer Society, 2010: 1-10

[51] Welch M J, Schonfeld U, He D, et al. Topical semantics of twitter links [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 327-336

[52] Yin D, Hong L, Davison B D. Structural link analysis and prediction in microblogs [C] //Proc of the 20th ACM Int Conf on Information and Knowledge Management (CIKM'11). New York: ACM, 2011: 1163-1168

[53] Cheng J, Romero D M, Meeder B, et al. Predicting reciprocity in social networks [C] //Proc of the 3rd Int Conf on Social Computing (SOCIALCOM'11). Piscataway, NJ: IEEE, 2011: 49-56

[54] Ediger D, Jiang K, Corley C, et al. Massive social network analysis: Mining Twitter for social good [C] //Proc of the 39th Int Conf on Parallel Processing (ICPP'10). Piscataway, NJ: IEEE, 2010: 583-593

[55] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter [C] // Proc of the 20th Int Conf on World Wide Web (WWW'11). New York: ACM, 2011: 695-704

[56] Wu S, Hofman J M, Mason W A, et al. Who says what to whom on Twitter [C] //Proc of the 20th Int Conf on World Wide Web (WWW'11). New York: ACM, 2011: 705-714

[57] Dabeer O, Mehendale P, Karnik A, et al. Timing tweets to increase effectiveness of information campaigns [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 105-112

[58] Lehmann J, Gonçalves B, Ramasco J J, et al. Dynamical classes of collective attention in Twitter [C] //Proc of the 21st Int Conf on World Wide Web (WWW'12). New York: ACM, 2012: 251-260

[59] Yang J, Counts S. Comparing information diffusion structure in weblogs and microblogs [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 351-354

[60] Lerman K, Ghosh R. Information contagion: An empirical study of the spread of news on Digg and Twitter social networks [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 90-97

[61] Lin C X, Mei Q, Jiang Y, et al. Inferring the diffusion and evolution of topics in social communities [C] //Proc of the 5th Int Workshop on Social Network Mining and Analysis (SNA-KDD'11). New York: ACM, 2011: 1231-1240

[62] Xue Z, Yao J, Cui B. Temporal provenance discovery in micro-blog message streams [C] //Proc of the 39th ACM SIGMOD Int Conf on Management of Data (SIGMOD'12). New York: ACM, 2012: 864-864

[63] Lumezanu C, Feamster N, Klein H. Bias: Measuring the tweeting behavior of propagandists [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM'12). Menlo Park, CA: AAAI, 2012: 210-217

[64] Naaman M, Zhang A X, Brody S, et al. On the study of diurnal urban routines on Twitter [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM'12). Menlo Park, CA: AAAI, 2012: 258-265

[65] Yang J, Leskovec J. Patterns of temporal variation in online media [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 177-186

[66] An J, Cha M, Gummadi K, et al. Media landscape in Twitter: A world of new conventions and political diversity [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 18-25

[67] Yang J, Counts S. Predicting the speed, scale, and range of information diffusion in Twitter [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 355-358

[68] Tsur O, Rappoport A. What's in a hashtag ?Content based prediction of the spread of ideas in microblogging communities [C] //Proc of the 5th ACM Int Conf on Web Search and Data Mining (WSDM'12). New York: ACM, 2012: 643-652

[69] Petrovic S, Osborne M, Lavrenko V. RT to win ! Predicting message propagation in Twitter [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 586-589

[70] Yang J, Leskovec J. Modeling information diffusion in implicit networks [C] //Proc of the 10th IEEE Int Conf on Data Mining (ICDM'10). Piscataway, NJ: IEEE, 2010: 599-608

[71] Lampos V, Bie T D, Cristianini N. Flu-detector: Tracking epidemics on Twitter [C] //Proc of the 10th European Conf on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD'10). Berlin: Springer, 2010: 599-602

[72] Ratkiewicz J, Conover M, Meiss M, et al. Truthy: Mapping the spread of astroturf in microblog streams [C] // Proc of the 20th Int Conf on World Wide Web (WWW'11). New York: ACM, 2011: 249-252

[73] Bakshy E, Hofman J M, Mason W A, et al. Everyone's an influencer: Quantifying influence on Twitter [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 65-74

[74] Lee C, Kwak H, Park H, et al. Finding influentials based on the temporal order of information adoption in Twitter [C] //Proc of the 19th Int Conf Companion on World Wide Web (WWW'10). New York: ACM, 2010: 1137-1138

[75] Aggarwal C C, Khan A, Yan X. On flow authority discovery in social networks [C] //Proc of the 11th SIAM Int Conf on Data Mining (SDM'11). Philadelphia, PA: SIAM, 2011: 522-533

[76] Steeg G V, Galstyan A. Information transfer in social media [C]. //Proc of the 21st Int Conf on World Wide Web (WWW'12). New York: ACM, 2012: 509-518

[77] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter: The million follower fallacy [C] //Proc of the 4th Int AAAI Conf on Weblogs and Social Media (ICWSM'10). Menlo Park, CA: AAAI, 2010: 10-17

[78] Pal A, Counts S. Identifying topical authorities in microblogs [C] //Proc of the 4th ACM Int Conf on Web Search and Data Mining (WSDM'11). New York: ACM, 2011: 45-54

[79] Quercia D, Capra L, Crowcroft J. The social world of Twitter: Topics, geography, and emotions [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM'12). Menlo Park, CA: AAAI, 2012: 298-305

[80] Tunkelang D. A Twitter analog to PageRank [EB/OL]. (2009-01-13) [2013-07-23]. http://thenoisychannel. com/2009/01/13/a_twitter_analog_to_pagerank

[81] Weng J, Lim E P, Jiang J, et al. TwitterRank: Finding topic-sensitive influential twitterers [C] //Proc of the 3rd ACM Int Conf on Web Search and Data Mining (WSDM'10). New York: ACM, 2010: 261-270

[82] Romero D M, Galuba W, Asur S, et al. Influence and passivity in social media [C] //Proc of the 20th Int Conf Companion on World Wide Web (WWW'11). New York: ACM, 2011: 113-114

[83] Kleinberg J M. Authoritative sources in a hyperlinked environment [J]. Journal of the ACM, 1999, 46 (5): 604-632

[84] Liu L, Tang J, Han J, et al. Mining topic-level influence in heterogeneous networks [C] //Proc of the 19th ACM Int Conf on Information and Knowledge Management (CIKM'10). New York: ACM, 2010: 199-208

[85] Li D, Shuai X, Sun G, et al. Mining topic-level opinion influence in microblog [C] //Proc of the 21st ACM Int Conf on Information and Knowledge Management (CIKM'12). New York: ACM, 2012: 1562-1566

[86] Ding Z Y, Jia Y, Zhou B, et al. Mining topical influencers based on the multi-relational network in micro-blogging sites [J]. China Communications, 2013, 10(1): 93-104

[87] Ding Z Y, Jia Y, Zhou B, et al. An influence strength measurement via time-aware probabilistic generative model for microblogs [C] //Proc of the 15th Asia-Pacific Web Conf. Berlin: Springer, 2013: 372-383

[88] Mislove A, Lehmann S, Ahn Y Y, et al. Understanding the demographics of Twitter users [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 554-557

[89] Huang J, M Thornton K, Efthimiadis E N. Conversational tagging in Twitter [C] //Proc of the 21st ACM Conf on Hypertext and Hypermedia (HT'10). New York: ACM, 2010: 173-178

[90] Kıcıman E. OMG, I have to tweet that ! A study of factors that influence tweet rates [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM'12). Menlo Park, CA: AAAI, 2012: 170-177

[91] Perreault M, Ruths D. The effect of mobile platforms on Twitter content generation [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 289-296

[92] Fan Pengyi, Wang Hui, Jiang Zhihong, et al. Measurement of microblogging network [J]. Journal of Computer Research and Development, 2012, 49(4): 691-699 (in Chinese)
(**樊鹏翼, 王晖, 姜志宏, 等. 微博网络测量研究[J]. 计算机研究与发展**, 2012, 49(4): 691-699)

[93] Cheng Z, Caverlee J, Lee K. You are where you tweet: A content-based approach to geo-locating twitter users [C] //Proc of the 19th ACM Int Conf on Information and Knowledge Management (CIKM'10). New York: ACM, 2010: 759-768

[94] Kulshrestha J, Kooti F, Nikravesh A, et al. Geographic dissection of the Twitter network [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM'12). Menlo Park, CA: AAAI, 2012: 202-209

[95] Castillo C, Mendoza M, Poblete B. Information credibility on Twitter [C] //Proc of the 20th Int Conf on World Wide Web (WWW'11). New York: ACM, 2011: 675-684

[96] Morris M R, Counts S, Roseway A, et al. Tweeting is believing ? Understanding microblog credibility perceptions [C] //Proc of the 15th ACM Conf on Computer Supported Cooperative Work (CSCW'12). New York: ACM, 2012: 441-450

[97] Gupta M, Zhao P, Han J. Evaluating Event Credibility on Twitter [C] //Proc of the 12th SIAM Int Conf on Data Mining (SDM'12). Philadelphia, PA: SIAM, 2012: 153-164

[98] Pennacchiotti M, Popescu A M. Democrats, republicans and starbucks afficionados: User classification in Twitter [C] //Proc of the 17th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining (KDD'11). New York: ACM, 2011: 430-438

[99] Chu Z, Gianvecchio S, Wang H, et al. Who is tweeting on Twitter: Human, bot, or cyborg [C] //Proc of the 26th Annual Computer Security Applications Conf (ACSAC'10). New York: ACM, 2010: 21-30

[100] Thomas K, Grier C, Paxson V, et al. Suspended accounts in retrospect: An analysis of Twitter spam [C] //Proc of the 11th Conf on Internet Measurement Conf (IMC'11). New York: ACM, 2011: 243-258

[101] Lee K, Eoff B D, Caverlee J. Seven months with the devils: A long-term study of content polluters on Twitter [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 185-192

[102] Yang C, Harkreader R, Zhang J, et al. Analyzing spammers' social networks for fun and profit [C] //Proc of the 21st Int Conf on World Wide Web (WWW'12). New York: ACM, 2012: 71-80

[103] Ghosh S, Viswanath B, Kooti F, et al. Understanding and combating link farming in the Twitter social network [C] //Proc of the 21st Int Conf on World Wide Web (WWW'12). New York: ACM, 2012: 61-70

[104] Ding Z Y, Zhang J F, Jia Y, et al. Detecting spammers in microblogs [J]. Journal of Internet Technology, 2013, 14 (2): 289-296

[105] Golbeck J, Grimes J, Rogers A. Twitter use by the U. S. Congress [J]. Journal of the American Society for Information Science and Technology Archive, 2010, 61 (8): 1612-1621

[106] Junco R, Heiberger G, Loken E. The effect of Twitter on college student engagement and grades [J]. Journal of Computer Assisted Learning, 2011, 27(2): 119-132

[107] Oh O, Agrawal M, Rao H R. Information control and terrorism: Tracking the Mumbai terrorist attack through Twitter [J]. Information Systems Frontiers Archive, 2011, 13(1): 33-43

[108] Vieweg S, Hughes A L, Starbird K, et al. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness [C] //Proc of the 28th ACM Conf on Human Factors in Computing Systems (CHI'10). New York: ACM, 2010: 1079-1088

[109] Paul M J, Dredze M. You are what you tweet: Analyzing Twitter for public health [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 265-272

[110] Paul S A, Hong L, Chi E H. Is Twitter a good place for asking questions? A characterization study [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 578-581

[111] Conover M D, Ratkiewicz J, Francisco M, et al. Political polarization on Twitter [C] //Proc of the 5th Int AAAI Conf on Weblogs and Social Media (ICWSM'11). Menlo Park, CA: AAAI, 2011: 89-96

[112] Forte A, Humphreys M, Park T. Grassroots professional development: How teachers use Twitter [C] //Proc of the 6th Int AAAI Conf on Weblogs and Social Media (ICWSM'12). Menlo Park, CA: AAAI, 2012: 106-113

**Ding Zhaoyun**, born in 1982. PhD and lecturer. His main research interests include Web mining and social computing.

**Jia Yan**, born in 1960. Professor and PhD supervisor. Her main research interests include data mining and information security (jiayan@nudt.edu.cn).

**Zhou Bin**, born in 1971. Professor. His main research interests include text mining and information security (binzhou@nudt.edu.cn).