

文章编号:1001-9081(2012)07-2027-03

doi:10.3724/SP.J.1087.2012.02027

基于信息数据分析的微博研究综述

王晶*, 朱珂, 汪斌强

(国家数字交换系统工程技术研究中心, 郑州 450002)

(*通信作者电子邮箱 wangjingniu_2003@sina.com)

摘要:近年来随着微博信息传播力和组织能力的突显,微博吸引了各类学者的关注。对当前基于信息数据分析的微博研究进行系统梳理,提出微博信息传播三大构件的概念,归纳了此类研究的主要研究内容及方法,总结了国内外围绕微博信息传播三大构件所取得的主要研究成果。最后探讨了未来在微博网络管控方面相关工作。

关键词:微博; 信息数据分析; 用户关系; 消息传播; 影响力; 网络管控

中图分类号: TP393.4 **文献标志码:**A

Survey on microblog research based on information data analysis

WANG Jing*, ZHU Ke, WANG Bin-qiang

(National Digital Switching System Engineering and Technology Center, Zhengzhou Henan 450002, China)

Abstract: In recent years, with the advances in information communication and organizational ability, microblog has attracted the attention of scholars of all kinds. This paper reviewed the present study of microblog based on the information data analysis, and presented the concept of three components in microblog information transmission. Besides it summarized the main problems and methods in this field, generalized the domestic and foreign achievement. Finally, the trend for future work on the monitoring and management of microblog was discussed.

Key words: microblog; information data analysis; member relationship; message transmission; influence; network monitoring and management

0 引言

微博(Microblog)是近年来新兴的一种网络服务,它是一种基于互联网的交流工具,允许用户之间交换短篇内容,如句子、图像和视频链接等^[1]。它是一个基于用户关系的信息分享、传播以及获取平台。用户可以通过网络、手机以及各种智能联网的客户端发送文字,并实现即时分享。微博具有使用简单便捷、支持开放多平台接入方式、消息更新传播速度快等特点,短短5年内吸引了全球上亿用户,截止2011年上半年,中国的微博用户已经达到1.95亿。微博比传统的社交网络具有更强的信息传播能力和成员组织能力,这一独特优势使其迅速成为当前主要社会媒体之一,作为一种非常重要的消息来源和传播途径,在越来越多的社会事件中起到关键作用。微博正在改变着人类的生活方式。

随着微博的爆炸式发展,它逐步成为国内外学者关注的焦点。各领域的科研人员在社交网络现有研究基础上,开展了大量与微博相关的理论和实践研究工作^[2]。本文基于微博网络信息数据分析研究,对当前计算机网络以及通信领域微博研究的主要问题、主要研究方法以及国内外的主要研究成果进行归纳和梳理,并对未来的研究工作进行展望。

1 微博网络服务概述

1.1 微博服务系统网络架构

微博本质上是一种Web2.0网络服务,系统的工作原理与所有的Web2.0服务相同,也是通过HTTP协议承载网络业

务数据。一个完整微博服务系统由微博服务器、微博数据中心和微博客户端组成,微博服务系统的网络体系结构如图1所示。

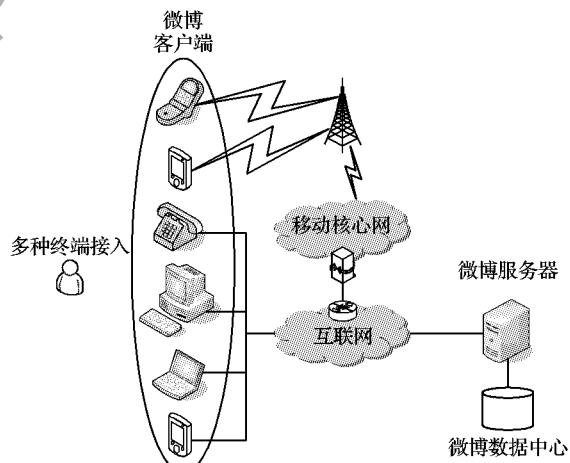


图1 微博服务系统的网络体系结构

在微博服务系统中,其服务器、客户端的功能及实现都基于Web2.0网络服务框架,数据中心是微博服务系统的重点。微博数据中心是一个庞大而复杂的数据库系统,其中存储了所用微博用户的个人信息、用户之间的跟随关系、用户发表的微博消息等。此外,微博数据中心还运行着社交网络服务(Social Network Service, SNS)程序,计算用户之间的关系、统计微博热点话题等。微博消息依靠微博数据中心维护的用户关系进行广泛传播。

收稿日期:2011-12-16;修回日期:2012-02-09。 基金项目:国家863计划项目(2011AA01A103)。

作者简介:王晶(1980-),女,江西黎川人,讲师,博士研究生,主要研究方向:宽带信息网络安全管控; 朱珂(1975-),男,河南开封人,副教授,博士,主要研究方向:网络体系结构、网络安全管控; 汪斌强(1963-),男,安徽安庆人,教授,博士,主要研究方向:网络体系结构、下一代互联网关键技术。

1.2 微博消息传播中的三大构件

强大的消息传播能力是微博服务最显著的特点之一,也是国内外学者关注的焦点问题之一。微博消息的传播主要有两个途径:转发途径和粉丝途径。在微博消息传播中,用户、微博消息和用户关系是直接影响着微博的传播力的三个要素,本文称之为微博消息传播中的三大构件,它们是构成微博消息传播的主体。三大构件之间相互影响作用,使微博消息在用户之间迅速蔓延。构件间的关系如图2所示。

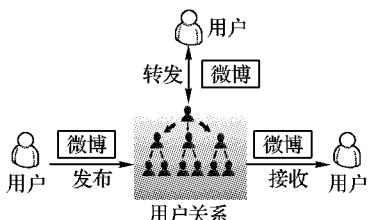


图2 微博消息传播三大构件关系图

微博用户是微博消息的制造者、转发者,也是微博消息的接收者,它是微博消息传播过程的起点、中间点和终点。微博消息是传播的主要内容,直接影响微博消息的受关注程度和转发度。用户关系是传播的主要途径,微博消息根据用户之间的跟随关系迅速传播到大范围用户,它是影响消息传播范围的关键因素。

2 基于微博信息数据分析的研究

2.1 主要研究的问题

在计算机网络和通信等学科领域,微博的信息数据特征是近年来关注的主要问题,很多微博研究都基于网络信息数据测量分析展开^[3-9]。微博信息数据是指微博数据中心中存储的各类数据,主要包括:微博用户档案、微博用户关系、微博消息、热点话题等,它是此类研究方法的基础。

在这类研究中,大多以微博消息传播三大构件为研究对象,以微博消息传播和微博成员组织为主要研究内容,目的在于发现微博中的用户、消息传播、热点话题、用户关系网络等的规律。主要研究的问题如下。

1) 基于微博用户的研究^[3-4,6-7]。主要研究用户的行为特征及用户的影响力。

2) 基于微博用户关系的研究^[3,8]。主要研究用户关系网络的基本属性、关系网络生成和演进、微博人员关系挖掘、微博用户人际关系特点。

3) 基于微博内容的研究^[3,5-7]。主要研究微博消息内容特点、消息活跃时间特点、微博热点话题特点。

4) 基于微博消息传播的研究^[3,7,9]。主要研究微博消息传播的特点,微博消息传播影响力。

此外,由于微博在近期一些群体性事件和突发事件中所起的作用,微博网络管控研究成为一个重要的研究课题,它旨在解决预测和控制突发事件在微博中的出现和扩散,研究内容主要包括微博舆情发现、真伪消息甄别、核心人员关联挖掘等。

2.2 微博信息数据分析的研究方法

基于信息数据分析的微博研究方法,以分析微博信息数据为基础,致力于发现微博中的各种规律和特点。它可以分为两个阶段:信息数据获取和信息数据分析。

在信息获取阶段,主要任务是获取大量微博信息数据,主要采用三种方法:基于微博第三方应用程序接口(Application

Programming Interface, API)编程实现微博信息爬取^[10],利用网络爬虫在微博网页上爬取关键字信息,利用网络数据采集设备直接获取微博服务网络传输数据。在数据分析阶段,主要任务是对微博信息数据进行特征提取和分析,挖掘出微博中的关键特征,采用的主要方法包括统计学数据分析方法、复杂网络分析方法^[11]、数据分类及挖掘方法等。微博信息数据分析方法基本流程如图3所示。

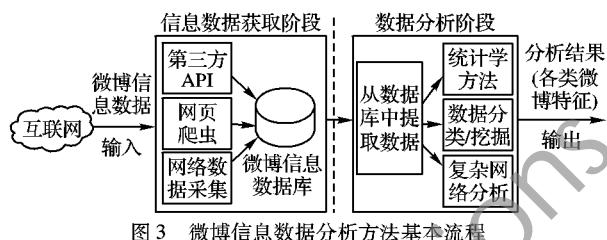


图3 微博信息数据分析方法基本流程

基于微博信息数据分析的研究是微博研究中非常重要的一个方向,它是开展其他微博研究的基础,能够积累大量的微博特征,对微博的理论和实践研究都至关重要。

3 主要研究成果

基于微博信息数据分析的研究近年来在国内外都取得了很多成果,掌握了微博中的大量特征。本章从微博消息传播三大构件的角度,对此领域目前掌握的国内外微博特征进行梳理总结,目的是为以后开展微博其他研究提供依据。

3.1 微博用户特征

微博用户大致可以分为两大类——普通用户和精英用户^[4,360,112],精英用户又分为四类,分别是媒体(media)、名人(celebrities)、博主(bloggers)和组织机构(formal organizations)。分析微博各类用户的特征能够指导微博管控研究的开展。针对用户的微博特征主要包括用户的行为特征和用户的影响力特征。

用户的行为特征主要表现如下。

1) 用户关注倾向方面^[7,363]。精英用户之间的关注关系,具有很强的同质性;普通用户更倾向于关注名人,其次是媒体。

2) 用户关注的话题方面^[3,596,6,7]。在国外新闻类的话题最受关注,在国内关注度高的话题多是一些笑话、话题讨论、图片等,另外不同类型用户关注的话题内容有差异。

3) 用户发布微博的习惯方面。研究表明用户发表微博的数量与其粉丝数量之间存在一定的关系^[3,598],通常当粉丝数量比较少的用户基本不发布消息,粉丝数量在100到1000之间时,用户发布的微博数量的变化平滑,而当粉丝数量大于5000时,发布的微博数量的变化将会是数量级上的。

4) 用户转发习惯方面。国外用户^[7,364]更喜欢自己直接发表消息,而不喜欢转发,倾向于转发精英用户的消息;国内用户^[6,8]更喜欢转发,有50.24%的消息都是转发消息。

5) 用户参与话题讨论方面^[3,597]。一半以上的用户都会参与话题讨论,并且其中有部分用户会参与多个话题的讨论。

对用户影响力的研究也至关重要。传统研究认为,在社会媒体中只有少数的人具有影响力,只要控制住这少数人,就能够用最少的市场投入,获得最大规模的影响力。但在微博这个新型的社会媒体中,大量的实验数据和研究^[4-5]表明,用户的影响力受到多个因素的影响,可以通过多个角度来评价用户的影响力,其中包括粉丝数量、转发微博和通过@符号提及用户。粉丝数量代表用户的受欢迎程度,通过@符号

提及用户代表了用户名字的价值,转发微博代表了用户的微博消息的内容价值。要评价微博用户的影响力,必须综合考虑上述三个因素。此外研究还表明,最具影响力用户的影响能够覆盖到多个话题;普通用户可以通过关注某个话题,并且围绕此话题发布一些有创意或者比较深刻的微博消息来获得影响力。

3.2 微博消息内容特征

微博消息内容包括微博消息和微博热点话题,目前在这个方面的主要研究成果包括微博消息传播特征和微博热点话题特征,为揭示在突发事件中的微博特征提供相关研究基础。

微博消息的传播具有速度快、范围广的基本特征。从已有的文献看,微博在消息传播方面表现出来的主要特征如下:

1)在传播方式上,大部分微博消息的传播方式为两步(two_step)传播^{[7]362},也就是说信息并非直接传播给大众,而是通过媒介(中介)把消息传播大众,中介在消息传播过程中起到重要作用;

2)在传播范围上,微博信息传播范围与微博消息发布者拥有的粉丝数量无关^{[3]598-599},只要消息被转发,则最终消息接收者的数量变化不大,这一点说明转发在消息传播过程中的表现出了巨大的力量;

3)从时间方面分析,微博消息传播的速度很快^{[3]599},有一半的消息会在一个小时之内被转发,75%在一天之内转发,只有10%在1个月之后才被转发;

4)围绕一个话题的微博数量与时间呈线性增加的关系,这是因为微博具有社会媒体的公共传播效应;

5)谣言在微博中的传播与传播节点的度分布和有效传染率相关^[9],该属性对于抑制谣言在微博中的传播具有重要作用。

特点话题(trend topic)是各微博网站通过统计学方法计算出的转发频度高、关注度高的话题^[13],能够反映出网络的舆情态势,现有研究揭示了微博热点话题的部分特征,主要包括如下。

1)在话题参与方面,通常一个长时期话题的参与人数并不会随着时间增加,但围绕话题的微博消息却不断增加,这说明在参与话题讨论的用户中一定存在核心用户,他们决定该话题的消息数量。此外,对于不同类型的话题,各类消息(转发、回复、通过@符号提及用户等)所占的比例不同。

2)在话题出现时间方面,新闻热点话题在微博中出现早于其他社会媒体。

3)话题的活跃期方面,大部分话题只有1个活跃期,只有很少的话题具有3个以上的活跃期,具有多个活跃期的话题主要是由于在不同时区的用户关注的时间存在差异,并且大部分的话题的活跃期的时间都小于1个星期,这说明微博话题具有很强的时效性。

4)在话题内容方面,国外微博的热点话题大多是新闻、时政等;国内则是一些笑话、爱好兴趣讨论等。

5)不同内容的话题具有不同的寿命,长寿的话题大多是与影视、音乐和书籍相关;而寿命短的话题大多是一些新闻类的时政话题。

6)影响话题成为热点的主要因素是转发,通常能够成为热点的话题都是可以引起公众共鸣的话题,粉丝数量和微博数并不起主导作用。

3.3 微博用户关系特征

用户关系网络是微博消息传播的主要途径,它直接影响

微博消息传播的范围。目前业界对微博用户关系网络的研究虽然处于起步阶段,但现有成果已揭示出微博用户关系网络中的很多特征^{[3]592-593,[8]}。

1)在用户粉丝分布方面,关系网络中的节点度分布不再服从标准的幂律分布,用户粉丝数量小于 10^5 时,服从指数为2.267的幂律分布,当大于 10^5 时,幂律分布的特性消失,这说明微博这种社交网络不再是典型的复杂网络;

2)从网络直径这个属性分析,微博用户关系网络中的有向边并没有增加网络的平均直径,在这个方面它与普通社交网络具有相似的特征,平均半径为4.12;

3)在聚类性方面,研究数据显示微博用户具有很强的聚类性,即在微博用户网络中,彼此互相关注的用户在地理位置上很接近,并且他们受到粉丝的关注度相似,这说明微博网络仍然具有复杂网络的属性,并且为用户关系挖掘提供了参考;

4)在微博用户关系中,真正的朋友关系和用户之间的关注关系之间表现出较大的差别,互相关注的用户并不一定是朋友关系,并且真正的朋友数量比用户粉丝数量少得多。

4 结语

基于信息数据分析的微博研究已经取得了大量成果,微博网络中的各类特征已相对清晰。但在当前形势下,微博技术在国内外的众多突发事件中起到了重要的消息传播和人员组织作用,如何有效、健康地使用微博已成为当前国内外学者亟待解决的问题。

为应对上述威胁,未来此领域的工作将主要针对微博网络管控展开,有效防止微博负面影响在国内外的泛滥。下一步的主要工作包括:

1)微博消息内容可信度评估研究,建立评估模型,并深入研究谣言在微博网络中的传播模式。

2)微博网络的演进过程研究,从网络演进的角度分析微博网络特征,力图从网络节点角度深入实施监管。

3)微博网络组织关系特征研究,结合数据挖掘技术,从海量微博数据中提取关键人物信息和组织成员脉络。

4)三网融合下的全方位立体管控方案研究,深入研究微博短信平台的管控技术,包括微博短信识别、危害短信过滤技术,并且利用移动网络中等微博用户的注册信息对核心人员进行位置定位、行动追踪等。

参考文献:

- [1] KAPLAN A M, HAENLEIN M. The early bird catches the news: Nine things you should know about micro-blogging[J]. Business Horizons, 2010, 54(2): 1-9.
- [2] 王红,王鹏.我国微博研究综述[J].消费导刊,2011(9):35-40.
- [3] KWAK H, LEE C, PARK H, et al. What is Twitter, a social network or a news media? [C]// Proceedings of the 2010 International World Wide Web Conference Committee. New York: ACM Press, 2010: 591-600.
- [4] CHA M, HADDADI H, BENEVENUTO F, et al. Measuring user influence in Twitter: The million follower fallacy [C]// Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Washington, DC: AAAI Press, 2010: 10-17.
- [5] MATHIOUDAKIS M, KOUDAS N. TwitterMonitor: Trend detection over the Twitter stream [C]// SIGMOD'10: Proceedings of the 2010 International Conference on Management of Data. New York: ACM press, 2010: 1155-1158.

(下转第2037页)

和 NTUSD 词典三个实验数据集的测试结果表明,该方法弥补了传统基于语素方法对于种子词语数量的依赖,克服了基于图的方法召回率较低的缺点, *MicroF1* 最高可以达到 92.8%, 并且在种子词语数量仅为 100 时, *MicroF1* 依然可以达到 84.1%, 在迭代次数大于 1 时, 算法已经达到收敛。

虽然目前的方法在词语褒贬分类上取得较好的效果,但是依然有很多需要研究和改进之处。本文只考虑了词语之间的语言学关系,但是词语之间的统计关系并未考虑。因此,在未来的工作中,将会从大语料库和互联网上获取词语之间的统计关系融入词语的图模型。其次,词语在具体语言环境中呈现出不同的情感,因此结合具体上下文来对词语的情感进行分类将是研究热点之一。最后,实际应用要求对词语进行更细粒度的情感分类,因此词语的多情感分类也将是研究重点。

参考文献:

- [1] XU GE, MENG XINFAN, WANG HOUFENG. Build Chinese emotion lexicons using a graph-based algorithm and multiple resources [C]// Proceedings of the 23rd International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 1209–1217.
- [2] KIM S M, HOVY E. Identifying and analyzing judgment opinions [C]// Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2006: 200–207.
- [3] KIM S M, HOVY E. Automatic detection of opinion bearing words and sentences [C]// Proceedings of the Second International Joint Conference on Natural Language Processing. Jeju Island: [s. n.], 2005: 61–66.
- [4] HATZIVASSILOGLOU V, MCKEOWN K. Predicting the semantic orientation of adjectives [C]// ACL-97: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Madrid, Spain: [s. n.], 1997: 174–181.
- [5] VELIKOVICH L, BLAIR-GOLDENSOHN S, HANNAN K, et al. The viability of Web-derived polarity lexicons [C]// Proceedings of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 777–785.
- [6] 路斌, 万小军, 杨建武, 等. 基于同义词词林的词汇褒贬计算 [C]// 第七届中文信息处理国际会议论文集. 北京: 电子工业出版社, 2007: 17–23.
- [7] KU L W, HUANG T H, CHEN H H. Using morphological and syntactic structures for Chinese opinion analysis [C]// Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2009: 1260–1269.
- [8] KU L W, LO Y S, CHEN H H. Using polarity scores of words for sentence-level opinion extraction [C]// Proceedings of NTCIR-6 Workshop Meeting. Tokyo, Japan: [s. n.], 2007: 316–322.
- [9] TURNER P, LITTMAN M L. Measuring praise and criticism: Inference of semantic orientation from association [J]. ACM Transactions on Information Systems, 2003, 21(4): 315–346.
- [10] KIM S M, HOVY E. Determining the sentiment of opinions [C]// Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 1367–1373.
- [11] KAMPS J, MARX M J, MOKKEN R J, et al. Using WordNet to measure semantic orientations of adjectives [C]// LREC 2004: Proceedings of the 4th International Conference on Language Resources and Evaluation. Lisbon: [s. n.], 2004: 1115–1118.
- [12] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算 [J]. 中文信息学报, 2006, 20(1): 14–20.
- [13] YUEN R W M, CHAN T Y W, LAI T B Y, et al. Morpheme-based derivation of bipolar semantic orientation of Chinese words [C]// Proceedings of the 20th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 1008–1014.
- [14] KU L W, CHEN H H. Mining opinions from the Web: Beyond relevance retrieval [J]. Journal of the American Society for Information Science and Technology, 2007, 58(12): 1838–1850.
- [15] TAKAMURA H, INUI T, OKUMURA M. Extracting semantic orientations of words using spin model [C]// Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2005: 133–140.
- [16] ESULI A, SEBASTIANI F. PageRanking WordNet Synset: An application to opinion mining [C]// Proceedings of the 45th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2007: 424–431.
- [17] RAO D, RAVICHANDRAN D. Semi-supervised polarity lexicon induction [C]// Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2009: 675–682.
- [18] 杜伟夫. 文本倾向性分析中的情感词典构建技术研究 [D]. 哈尔滨: 哈尔滨工业大学, 2010.
- [19] LU BIN, SONG YAN, ZHANG XING, et al. Learning Chinese polarity lexicons by integration of graph models and morphological features [C]// Proceedings of the 6th Asia Information Retrieval Societies Conference on Information Retrieval Technology, LNCS 6458. Berlin: Springer, 2010: 466–477.

(上接第 2029 页)

- [6] YU L, ASUR S, HUBERMAN B A. What trends in Chinese social media [C]// SNA-KDD'11: Proceedings of the Fifth International Workshop on Social Network Mining and Analysis. SanDiego, CA: KDD Press, 2011: 37.
- [7] WU S, OFMAN J M, MASON W A, et al. Who says what to whom on Twitter [C]// WWW'11: Proceedings of the 20th International Conference on World Wide Web. New York: ACM Press, 2011: 359–367.
- [8] HUBERMAN B A, ROMERO D M, WU F. Social networks that matter: Twitter under the microscope [EB/OL]. [2011-08-10]. <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063>.
- [9] 许晓东, 肖银涛, 朱士瑞. 微博社区的谣言传播仿真研究 [J]. 计算机工程, 2011, 37(10): 272–274.
- [10] 廉捷, 周欣, 曹伟, 等. 新浪微博数据挖掘方案 [J]. 清华大学学报: 自然科学版, 2011, 51(10): 1300–1305.
- [11] 汪小帆, 李翔, 陈关荣. 复杂网络理论及其应用 [M]. 北京: 清华大学出版社, 2006: 49–100.
- [12] WEN E, SUN V. 新浪微博研究报告 [EB/OL]. [2011-05-20]. <http://www.techweb.com.cn/data/2011-02-25/916941.shtml>.
- [13] 李爽. 从微博中挖掘有用信息 [J]. 网络与信息, 2011(6): 98–102.