基于用户行为和影响覆盖的微博价值评价模型

王 涛,周 斌,曲 铭,王忠振
(国防科学技术大学 计算机学院,湖南 长沙 410073)

摘 要:在已有的对单条微博信息价值的研究中,一般都是从构成微博信息价值的多维因素出发,分析每个因素的权重然后综合进行评定,这样就带来了计算复杂的问题。但是,由于中文语义的复杂性,这种先验的分析方法,难免会有“一概而论”的嫌疑。由于转发行为的本质就是对微博信息价值的认同,而转发人占收到此信息总人数的比率则反映了该信息的普遍影响程度,因此从这两个因素出发研究微博信息价值非常值得思考。为此,文中从粉丝行为和影响覆盖率两个角度出发,引入了“单条微博影响力饱和度”概念,对单条微博信息价值进行了研究,在研究思路上进行了积极拓展。

关键词:微博价值;影响力饱和度;转发覆盖人数;转发量
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2015)02-0021-04
doi:10.3969/j.issn.1673-629X.2015.02.005

Microblog Influence Evaluation Model Based on Follower Behavior and Influence Coverage Ratio

WANG Tao, ZHOU Bin, QU Ming, WANG Zhong-zhen
(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Existing research on the information value of a single microblog are generally based on multi-dimensional factors that form the single microblogging information value, analyzing the weight of each factor and then evaluating the overall value, which results in the computational complexity problem. However, due to the Chinese semantic complexity, this transcendental method of analysis will inevitably be in question of "over generalization". Since the nature of the forwarding behavior is the recognition of the microblogging information value, and the ratio between the forwarding number and the total number of people receiving the information reflects the widespread impact of the information, the research based on these two factors are worth considering. In this paper, introduce a novel concept of "single microblogging influence saturation" in perspective of the follower behavior and the influence coverage ratio, study the value of single microblgging information value, which effectively expands the research ideas.

Key words: microblogging value; influence saturation; forwarding coverage; amount of forwarding

0 引 言

微博作为信息化发展的一个重要产物,凭借其开放的平台、便捷的操作、丰富的应用等储多优势,极好地满足了人们获取新闻时事、参与人际交往、进行自我表达的愿望。在极短的时间内,微博便完成了从普及到成为社会公共舆论重要平台的蜕变,并对国家安全和社会发展产生了深刻、巨大影响。据中国互联网络信息中心(CNNIC)2013 年 7 月公布的《第 32 次中国互联网络发展状况统计报告》[1] 显示,截至 2013 年 6 月底,中国微博网民规模已达到 3.31 亿,网民中微博使用率达到了 56.0%。由此产生的信息冗余大、垃圾内容多等储多问题已非常突出。因此,研究单条微博信息价值的评价方法便成了一个非常重要的问题。

1 相关研究

目前,国内外学者对微博的研究主要集中在话题分析[2]、情感分析[3]、信息检索与推荐[4]、关系分析与挖掘[5]、信息传播[6-7]及意见领袖影响力[8-9]等方面,而对单条微博价值研究较少。在已知的对单条微博价值的研究中,韩朝阳[10]等从信度和效度评价,即从可

收稿日期:2014-02-25 **修回日期**:2014-05-28 **网络出版时间**:2014-12-27
基金项目:国家“973”重点基础研究发展计划项目(2013CB329604)
作者简介:王 涛(1979-),男,硕士研究生,研究方向为数据挖掘;周 斌,研究员,研究方向为数据挖掘。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141227.1340.009.html

靠性评价、一致性评价、完备性评价、时效性评价、精确度评价五个方面,对旅游应用的微博信息价值进行了评价。齐娜[11]等认为微博发布的随意性、数量的无限增长、发布的即时性、收集的便利性最终造成了医疗信息领域质量参差不齐、判断标准混乱、科学性和可信度不高等问题。张豪锋[12]等则从教育微博社群中首帖质量的角度,提出了改进社群质量的对策。以上研究关注领域都比较单一,缺乏对全局领域微博信息质量的思考。莫祖英[13]等在总结归纳的基础上,从全局角度出发,对微博信息质量评价模型进行了更全面的分析,提出了以信息量、内容质量、来源质量、信息利用指标为主要内容的评价模型,但实际应用中计算复杂度较高。

2 基于影响力饱和度的评价模型

由于中文语义的复杂性,以权重方法衡量微博信息价值,难免会有"一概而论"的嫌疑。周庆山[14]等研究表明"当意见领袖进行议题设置、发布与个人心灵感受或者个人生活有关的介绍性微博时,大量粉丝会去参与讨论,而当意见领袖发布评论性微博、深刻意见的微博时,用户更多地会参与转发。"可见用户转发行为对微博信息的传播起着关键作用,其本质就是对其价值的认同。但以转发量作为唯一指标,显然忽略了粉丝基数的影响,评价难免失之片面。如,周庆山等在微博中意见领袖甄别与内容特征的实证研究[14]中认为"娱乐明星微博信息的转发量和评论数都高出文化和商业明星很多。究其因,是因为中国网民主要以中低收入、无业、低学历人群为主,使互联网文化带有很强烈的娱乐性。"例如,一条娱乐明星的微博被转发几千次,但用户本身却拥有几千万粉丝,则实际上转发的人是相当少的,此条微博信息是否具备价值也值得怀疑。可见用户转发行为反映了微博信息的价值,而转发人占覆盖总人数的比率则反映了该微博的普遍影响程度。

显而易见,微博用户发布一条微博,经其粉丝转发后,势必也会给粉丝的粉丝造成影响,转发的层级越多,能看到此微博的用户就越多。文中充分考虑了这种影响,在更宽泛的范围内将受此微博影响的用户都纳入了发布信息者的直接或间接粉丝中。如图 1 所示:A、B、C、D 四个节点为微博中的 4 个用户,黑色实线有向边表示关注关系,黑色虚线有向边表示转发关系。当 A(A 有 4 个粉丝)发布一条微博,B 对之进行了转发(B 有 2 个粉丝),C 又对 B 进行了转发(C 有 2 个粉丝),由于转发关系的存在,B、C 二者的粉丝都将受到 A 这条微博的影响,并可视为 A 的间接粉丝。所以,A 的这条微博转发覆盖人数 = A 的粉丝数(4) + B 的粉丝数(2) + C 的粉丝数(2) = 8。

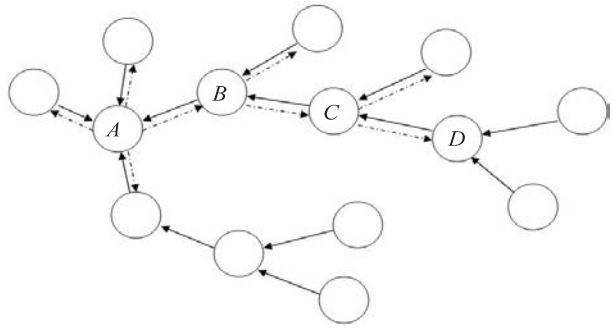

图 1 转发覆盖图

于是给出如下两个定义:

转发覆盖人数:直接粉丝数和间接粉丝数之和。

单条微博影响力饱和度:单条微博转发数与转发覆盖人数之比。

综上,单条微博影响力饱和度计算方法可表示为:

$$\frac{T_i}{F_A + \sum_{j=1}^{M} F_j}$$

其中,T_i 是用户 A 的第 i 条微博的转发数;F_A 是用户 A 的直系粉丝;M 是非直接粉丝转发用户数量;F_j 是第 j 个非直接粉丝转发用户带来的粉丝数。

如上例所示,用户 A 单条微博影响力饱和度等于转发数(2)/单条微博转发覆盖人数(8) = 25%。

3 实验分析

3.1 实验数据获取

文中在新浪微博数据中心的"名人影响力排行榜"各领域随机选取了 3 至 5 名,共 68 名具有代表性的名人于 2013 年 12 月 18 日发布的 177 条微博作为研究对象。以模拟登录自动填写表单的方式,每隔 5 分钟自动抽取了新浪微博的两个应用平台(微博引爆点 newgraph. sinaapp. com、孔明社交管理 www. kmsocial. cn。可以获取粉丝数、转发总量、转发覆盖人数等信息,但不能分析转发量超过 2 000 的微博,不过并不影响本研究)上每条微博的转发数、覆盖人数和爬取时间。截至 2013 年 12 月 20 日 18 时 09 分,共计爬取数据 24 226 条,其中有效数据 20 013 条。为验证数据统计的准确性,文中对两个应用的分析结果进行了对比,两者分析结果基本吻合,具备一定的准确性。

3.2 实验结果分析

表 1 和表 2 分别列出了转发量和影响力饱和度排在前 5 位的微博(对于表 1 至表 4 中内容较长的微博,为便于表述,文中对其内容做了省略处理)。表格中 U 代表用户,CT 代表微博内容,RC 代表转发量,IS 代表影响力饱和度,FD 代表直系粉丝量,FC 代表粉丝覆盖量。

表 1　转发量排在前 5 位的微博

U	CT	RC	IS	FD	FC
英式没品笑话	如果有人天生就是耳聋,那他在心中自言自语时,讲的是哪一国语言	1 990	0.000 84	963 628	2 370 662
思想聚焦	婚姻其实就这么真实! 男人虽爱美人,但不一定会娶她,到最后,男人娶的都是适合做老婆的。女人虽然爱钱,但不一定会嫁富人,到最后嫁的都是待她好的	1 988	0.000 34	5 215 162	5 903 877
延参法师	认识自己才能了然世界,世间万物千变万化,众缘和合,而其中无一为我所有,这便是真实。所以,珍惜,尊重,路过	1 982	0.000 09	21 355 688	22 715 976
Kevin 凯文	原諒別人,就是原諒自己! 大家加油	1 981	0.000 04	48 366 700	49 035 887
新浪娱乐	由@杜淳、@佟丽娅、周一围、@薛佳凝等联袂的都市时尚爱情剧《恋爱的那点事儿》即将于 12 月 23 日登陆浙江卫视	1 960	0.000 04	9 143 991	47 339 636

表 1 是转发量在前五位的微博。从表中可以发现,第 2、第 3 及第 4 条微博转发量都比较大,观察其内容特征,可知这几条微博都是属于生活感悟类的。这类信息一般是发布人在特定的心境或情绪下,对某事物一时的慨叹,既没有强烈的互动性,也没有向外传递太多有价值的信息,近 2 000 次的转发看似较高,但实际上相对于数百万,甚至上千万的粉丝量,被转发的概率是相当小的,即影响力饱和度很小;而第 5 条微博则属于娱乐信息,也有较高转发量,受到一定追捧,但相对于新浪娱乐近一千万的粉丝量来说,实际转发也不多,影响力饱和度也比较小。第 1 条信息在表 2 中排行是比较靠前的,将在下面进行分析。

表 2　影响力饱和度排在前 5 位的微博

U	CT	RC	IS	FD	FC
斯道	不要太伤感,晚安啦各位…对了问一下,大家希望#继承者们#话题继续跟进主演们的动态吗	1 325	0.002 15	302 711	616 279
剑神葡萄	万一中了这笔巨款我该怎么花呢(有图)	1 269	0.001 40	150 648	904 490
海蓝博士	你会发现,当你真的爱上自己,爱你的人也翩然而至,而且不单单局限于爱情	390	0.000 92	232 599	423 913
英式没品笑话	如果有人天生就是耳聋,那他在心中自言自语时,讲的是哪一国语言	1 990	0.000 84	963 628	2 371 871
沈阳铁路	#连客之花#乘务员孙佳慧按标准在边凳值岗,偶尔也会望向车窗外,会在夜深人静的时候思念家……	345	0.000 78	402 430	443 445

相反,表 2 所示第 1、2、4 条微博转发量和表 1 的转发量基本属同一量级,其绝对值虽然不及表 1 中的微博,但影响力饱和度却比较高。从其内容上看,都有较强的互动性,更能激发用户转发兴趣,而且都是大众能广泛参与的。如表 2 中的第 1 条微博,“继承者”正是当前热播的电视剧,有广泛的受众基础,人们也愿意分享心得进行自我表达;而第 2 条微博则既是大众普遍关心的问题,看似自言自语,实际上却具有强烈的互动和暗示,也能引起用户转发兴趣;第 4 条微博所反映的现象广泛存在于生活之中,却又常常被大众疏漏,做为微博发布,具备一定的新鲜性,极好地满足了大众的猎奇心理和求知的欲望;第 5 条微博的博主是“沈阳铁路”,考虑到关注的用户可能都属于铁路领域,属同一体系,且所发布的内容反映了发生在火车上的真人真事,加之此条微博配有实际的图片,因此影响力饱和度相对也比较高。此外,与影响力饱和度相应,这几条微博转发量也是比较可观的。

表 3 和表 4 分别列出了转发量和影响力饱和度排在后 5 位的微博。

表 3　转发量排在后 5 位的微博

U	CT	RC	IS	FD	FC
余英	关键是薪酬能否市场化! ……	5	0.000 007 8	638 303	640 512
瀟湘墨人	前天下午,在网易经济学家年会会场,一个 30 出头的男子在上午论坛结束后追着茅于轼先生谩骂,在拥挤的人群中,茅先生边走边回头递了一张名片给该男子,说这是我的联系方式……	6	0.000 083 6	70 408	71 740
秋叶	【一页纸 PPT 大赛:说说你的 2013 年】抢先奖第 10 号作品,帅得不忍直视,你还憋屈啥	7	0.000 043 7	155 248	160 233
开眼视点	一到冬天人们就易变胖,进而抵抗力减弱,易患疾病。7 个让人冬天发福的原因:……	8	0.000 024 0	330 666	334 090
互联网信徒王冠雄	哈哈,忍不住吐槽	9	0. 000 008 7	1 028 885	1 038 528

从表 3 的第 1、第 2、第 4 及第 5 条微博看,内容都有一定的针对性,受众较少,由此造成转发量较小。如第 1 条反映的是经济政策领域的内容,与其说不重要,不如说专业性太强,其他用户很难参与;第 2 条讲述的是“网易经济学家年会会场”发生的事,受众范围也比较小,只有参会的人才会有兴趣参与转发;而第 4 条微博则更倾向于身体较胖的人,也有特定的人群,且内容都是人们耳熟能详的,基本上没有传递什么有价值的信息;第 5 条微博是对互联网特定产品的吐槽,显得毫无意义;第 3 条发布了一条号召大家制作反映年度生活幻灯片的博文,由于幻灯片短时间并不容易制作,因此转发的人也比较少。此外,从影响力饱和度上看也是比较低的。

表 4 影响力饱和度排在后 5 位的微博

U	CT	RC	IS	FD	FC
上海发布	【身边探宝:明代洪武青花“春寿”云龙纹瓶】#上博珍品#这件青花瓷瓶制作精美……	21	0.000 005 1	4 088 744	4 097 501
新浪娱乐	电视剧《女人帮》正在热播,看剧的各位来打个分吧:http://t.cn/8kxdn4a	52	0.000 005 6	9 143 991	9 259 097
平安北京	【燃气灶具使用安全—不要私装乱改,定期检查】现在,楼房住户主要使用的是管道天然气,……	37	0.000 005 8	6 282 166	6 367 914
赵晓	@施玮:呵呵,我推荐! 这期图文并茂,不仅文章有益有趣,而且美工超赞的,我很喜欢	45	0.000 006 1	7 242 983	7 385 223
汽车之家	【一起去旅行】四姑娘山进击团,508/408/逍客和蓝天的完美融合。与最佳损友的“猥琐”自驾行。从此节操是路人! 传送门:http://t.cn/8k9m7lx	32	0.000 006 4	4 849 916	4 969 153

而表 4 的 5 条微博,同表 3,都针对了特定领域。如“上海发布”的明代洪武青花“春寿”云龙纹瓶属于古玩收藏领域,圈子很小;“平安北京”的燃气灶使用属于安全领域,虽然重要,但并没有传递太多有价值的信息,很难激起大众共鸣;“汽车之家”的四姑娘山进击团属于特定的汽车俱乐部,仅是少数人的游戏。几条微博的受众群体都比较局限,很难大范围得到传播;第 2 条微博要求给影视剧“女人帮”打分,考虑到操作复杂的问题,转发人数和影响力饱和度都是比较低的。

综合以上分析可以很清楚地发现:内容有较高价值、受众范围较广、互动性较强、能满足大众猎奇心理,且图文并茂的微博,影响力饱和度较高,与之相应,也有一定的转发量;相反,对于没有传递有价值信息、缺少特色、内容针对性较强、普遍可参与性较低、仅能在一定领域形成影响的微博,不论是影响力饱和度还是转发量都比较低。由此可以得出以下结论:影响力饱和度较高的微博信息,往往具备不低的转发量,既能反映信息具备较高价值,又能在广泛的范围造成影响;相反,转发量较高的微博信息,影响力饱和度不一定高,这种情况很可能是因为粉丝基数大,或出于对信息发布者的追捧而进行的转发,其价值普遍不高;而转发量比较低的微博,通常影响力饱和度也比较低,这样的信息本身不具备价值,粉丝也并没有因追捧博主而盲目转发;最后,对于影响力饱和度比较低的微博,其转发量也比较低,但由于用户转发行为的随机性和不确定性,其值并没有太多参考价值。因此,利用影响力饱和度评价微博价值显然具备一定的合理性。

4 结束语

微博方兴未艾,未来必有更为广泛的使用,随之而来的信息冗余大、垃圾内容多等问题也会长期存在。为合理衡量微博信息价值,文中以较新的视角对单条微博价值进行了思考,厘清了评价微博价值的核心因素,建立了一个基于影响力饱和度的评价模型,在一定程度上解决了一些微博价值评价“虚高”和计算复杂的问题。但由于微博信息价值评价本身具有一定的主观性,因此文中并不确保评价的精确性,而仅从研究思路上作了一些思考,以期对当前研究有所拓展。

参考文献:

[1] 第 32 次中国互联网络发展状况统计报告[R/OL].(2013-07-17)[2013-10-06].http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201307/t20130717_40664.htm.

[2] Sakaki T, Okazaki M, Matsuo Y. Earthquake shakes Twitter users:real-time event detection by social sensors[C]//Proc of the 19th international conference on World Wide Web. New York:ACM,2010:851-860.

[3] Barbosa L, Feng J. Robust sentiment detection on Twitter from biased and noisy data[C]//Proc of the 23rd international conference on computational linguistics. Stroudsburg, PA:ACL,2010:36-44.

[4] Sarma A D, Sarma A D, Gollapudi S. Ranking mechanisms in Twitter like forums[C]//Proc of the 3rd ACM international conference on web search and data mining. New York:ACM,

(下转第 46 页)

3 结束语

文中提出了一种关系相似性框架下基于大规模文本数据识别蛋白质交互的方法,与广泛采用的基于单句的机器学习方法不同,该方法直接以蛋白质对为研究对象,以大规模文本为依据提取特征建立相似性计算模型,并利用 K 近邻分类器识别给定的两个蛋白质是否存在相互作用关系。识别结果可直接用于 PPI 网络的构建,此方法还能充分利用已有的 PPI 数据而无需额外的人工标注。文中比较了多种相似性度量策略对识别效果的影响,结果可以看出基于余弦距离的相似性计算函数在建立的向量空间模型下较为合理,根据余弦相似性采用近邻的方法自动识别 PPI 取得了较高且较均衡的精确度和召回率。

参考文献:

[1] Prasad T SK, Goel R, Kandasamy K, et al. Human protein reference database – 2009 update[J]. Nucleic Acids Research, 2009, 37: 767–772.

[2] Bader G D, Betel D, Hogue C W V. BIND: the biomolecular interaction network database[J]. Nucleic Acids Research, 2003, 31(1): 248–250.

[3] Salwinski L, Miller C S, Smith A J, et al. The database of interacting proteins: 2004 update[J]. Nucleic Acids Research, 2004, 32: 449–451.

[4] Kerrien S, Alam – Faruque Y, Aranda B, et al. IntAct – open source resource for molecular interaction data[J]. Nucleic Acids Research, 2007, 35: 561–565.

[5] Ceol A, Aryamontri A C, Licata L, et al. MINT, the molecular interaction database: 2009 update[J]. Nucleic Acids Research, 2010, 38: 532–539.

[6] Bunescu R, Mooney R, Ramani A, et al. Integrating co–occurrence statistics with information extraction for robust retrieval of protein interactions from Medline[C]//Proceedings of the workshop on linking natural language processing and biology: towards deeper biological literature analysis. [s. l.]: Association for Computational Linguistics, 2006: 49–56.

[7] Koike A, Kobayashi Y, Takagi T. Kinase pathway database: an integrated protein–Kinase and NLP–based protein–interaction resource[J]. Genome Research, 2003, 13(6A): 1231–1243.

[8] 杨志豪,洪 莉,林鸿飞,等. 基于支持向量机的生物医学文献蛋白质关系抽取[J]. 智能系统学报,2008,3(4):361–369.

[9] 唐 楠,杨志豪,林鸿飞,等. 基于多核学习的医学文献蛋白质关系抽取[J]. 计算机工程,2011,37(10):184–186.

[10] 崔宝今,林鸿飞,张 霄. 基于半监督学习的蛋白质关系抽取研究[J]. 山东大学学报:工学版,2009,39(3):16–21.

[11] Grimes G R, Wen T Q, Mewissen M, et al. PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature[J]. Bioinformatics, 2006, 22(16): 2055–2057.

[12] Ananiadou S, Kell D B, Tsujii J. Text mining and its potential applications in systems biology[J]. Trends in Biotechnology, 2006, 24(12): 571–579.

[13] 陈治纲,何丕廉,孙越恒,等. 基于向量空间模型的文本分类方法的研究与实现[J]. 计算机应用,2004,24(06Z):277–279.

[14] 饶文碧,柯慧燕. Web 文本分类技术研究及其实现[J]. 计算机技术与发展,2006,16(3):116–118.

[15] University of Illinois at Urbana–champaign. Sentence segmentation tool[EB/OL]. [2011–09–23]. http://cogcomp.cs.illinois.edu/page/tools_view/2.

[16] 许 幸,张启蕊. 基于 KNN 算法的医药信息文本分类系统的研究[J]. 计算机技术与发展,2009,19(4):206–209.

[17] 王 煜,白 石,王正欧. 用于 Web 文本分类的快速 KNN 算法[J]. 情报学报,2007,26(1):60–64.

(上接第 24 页)

2010: 21–30.

[5] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks[C]//Proc of the 19th ACM international conference on information and knowledge management. New York: ACM, 2010: 1633–1636.

[6] Romero D M, Meeder B, Kleinberg J. Differences in the mechanics of information diffusion across topics: idioms, plitical hashtags, and complex contagion on Twitter[C]//Proc of the 20th international conference on World Wide Web. New York: ACM, 2011: 695–704.

[7] 吴雨蓉. 微博信息传播模式分析[J]. 渤海大学学报:哲学社会科学版,2012(2):140–143.

[8] Weng J, Lim E P, Jiang J, et al. TwitterRank: finding topic sensitive influential Twitters[C]//Proc of the 3rd ACM international conference on web search and data mining. New York: ACM, 2010: 261–270.

[9] 李 军,陈 震,黄霁崴. 微博影响力评价研究[J]. 信息网络安全,2012(3):10–13.

[10] 韩朝阳,张仁军. 面向旅游应用的微博信息信度和效度评价[J]. 重庆理工大学学报:社会科学版,2011,25(10):37–40.

[11] 齐 娜,宋立荣. 医疗健康领域微博信息传播中的信息质量问题[J]. 科技导报,2012,30(17):60–65.

[12] 张豪锋,杨绪辉. 教育微博社群中首帖质量的分析与对策[J]. 远程教育杂志,2012(2):98–103.

[13] 莫祖英,马费成,罗 毅. 微博信息质量评价模型构建研究[J]. 信息资源管理学报,2013(2):12–18.

[14] 周庆山,梁兴堃,曹雨佳. 微博中意见领袖甄别与内容特征的实证研究[J]. 山东图书馆学刊,2012(1):22–27.

基于用户行为和影响覆盖的微博价值评价模型

作者： 王涛， 周斌， 曲铭， 王忠振， WANG Tao， ZHOU Bin， QU Ming， WANG Zhong-zhen
作者单位： 国防科学技术大学 计算机学院,湖南 长沙,410073
刊名： 计算机技术与发展 ISTIC
英文刊名： Computer Technology and Development
年，卷(期)： 2015(2)