嘀咕网用户领域影响力研究

李 敏 肖 盛 刘正捷 张 军

(大连海事大学信息科学技术学院 中国欧盟可用性研究中心 大连 116026)

摘 要 社交媒体的快速发展使得人们越来越关注有影响力的用户的行为及其对他人的影响作用。有一些研究也致力于解决社交媒体用户社会影响力的度量问题。但是选取的度量标准一般都涉及微博数、粉丝数等全局性指标,而没有考虑到用户在不同的领域范围内所具有的影响力大小是不同的,所以对用户影响力的度量比较笼统、不具体。以嘀咕网在线用户数据为对象,对用户发布的信息内容进行领域分类,并提出领域影响力的概念及度量方法。经过验证,该方法可以很好地度量用户在不同领域的影响力。研究发现粉丝数等这类度量维度与用户领域影响力不成正相关的关系。

关键词 社交媒体,嘀咕网,领域分类,领域影响力

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.9.014

Research on Field Influence of Digu Users

LI Min XIAO Sheng LIU Zheng-jie ZHANG Jun

(Sino European Usability Center, Information Science and Technology College, Dalian Maritime University, Dalian 116026, China)

Abstract Social media develops rapidly, which makes people pay more attention to the behaviors of influential social media users and the effects to others. Some studies dealt with the measurement of social influence of social media users. However, they usually chose global metrics, such as number of posts and number of fans, rather than other metrics that might consider varied social influences within different fields. So the measurement metrics are general and unspecific. This research chose online data of Digu users as object to study the classifications of users' posts, and proposed the concept of field influence and the measurement method. At last, the method was verified by a sample study. The results show that it can be well used to measure users' social influence within different fields. It was also found that the measurement metrics such as the number of fans have no positive correlation with user field influence.

Keywords Social media, Digu website, Field classification, Field influence

1 引言

作为信息收集和分享的平台,社交媒体在世界范围内快速的发展,在大多数人们的生活中也起到了越来越大的作用。基于位置的服务(Location-based Services, LBS)类应用是一类考虑了实际地理位置元素的社交媒体应用,拥有大量的用户群,Foursquare是国外最典型的应用,而嘀咕网是国内典型的这类应用之一。嘀咕网于2009年2月上线,提供微博服务,并在2010年7月至2011年11月开始转向提供基于位置类社交服务。截至2011年2月,嘀咕手机客户端用户已达到200万。在此期间,嘀咕网与诺基亚、麦当劳等多家品牌合作推出线上活动与线下实体店联合的服务,鼓励用户线上进行签到等活动,然后到实体可以享有奖品或优惠等服务。

嘀咕网上产生了大量的与实际地理位置有关的用户分享信息,具有高影响力的用户传播的信息对其他用户的行为和消费决策会产生较大的影响。用户的影响力取决于其现实世界中已积攒的声誉、社交关系的数量与质量,以及他们所关注的领域主题等。嘀咕网用户分享的信息内容涉猎不同领域,在不同的领域主题下嘀咕用户所具有的影响力也不同。具有领域影响力的人可以成为该领域的"达人"。人们会在不同领域内受到该领域达人的影响而改变自己的思想和行为。例如,人们广为熟知的姚明因在篮球方面的经历而影响球迷,他就是篮球方面的达人。在消费购物方面,流行的美丽说、蘑菇街这类购物推荐网站上存在各种达人,利用这些达人的影响力来向消费者传播消费相关信息。同时,达人也会涉猎其他领域的信息,但是相对来说,其在擅长的领域内的影响力更大。例如,体育明星发表的关于体育这一领域内的观点会比IT这一领域的观点更有权威性。而现有的研究较多地关注用户整体的影响力,而缺少对用户在不同领域内的影响力进行考虑。我们对嘀咕网用户分享的信息按照领域进行了分类,以美食领域为例研究用户的影响力,并进行了案例研究。通过做上述的研究,可以更好地识别用户在不同领域内具有

到稿日期:2014-03-20 返修日期:2014-06-20 本文受中央高校基本科研业务费专项资金(3132013041)资助。

李 敏(1982-),女,博士生,主要研究领域为人机交互设计、社会网络分析等,E-mail:minleely@gmail.com;肖 盛(1986-),男,硕士,主要研究领域为可用性与人机交互设计;刘正捷(1958-),男,硕士,教授,CCF会员,主要研究领域为可用性工程、用户体验、人机交互和信息无障碍化,E-mail:liuzhj@dlmu.edu.cn(通信作者);张 军(1977-),男,博士,副教授,CCF会员,主要研究领域为可用性与人机交互设计、社交网络分析。

的影响力。产品设计者以及营销方案决策者可以更好地了解他们的目标用户、用户分享行为特点、平台提供的交互功能的局限性及用户使用体验等，由此可以有针对性地进行产品设计、推广，提升用户的体验。而对于用户而言，他们可以获得更好的社交产品使用体验，得到真正感兴趣、需要的信息，避免了信息过载对其造成的影响。

本文首先介绍研究的相关工作；然后描述研究方法，包括数据获取及处理过程；接下来展示了领域影响力的度量方法及案例说明；最后总结全文。

2 相关工作

关于影响力的研究，早期一般都是直接用粉丝数这个维度来度量影响力的大小。Kwak 等人把粉丝数和 PageRank 两种度量方法进行了比较，得出这两个指标有相似性的结论[1]。Ye 等人对 Twitter 用户的影响力进行分析，采用粉丝数、回复、转发这几个维度来度量影响力；并且对这几个维度进行了相关性分析，发现粉丝数与其他几个维度相关性较低[2]。Cha 等人通过关注用户行为，考察微博用户的粉丝数（入度）、用户被提及数（@）、微博转发数这几个用户影响力度量维度，发现粉丝数多少与转发和提及不成正比关系，而用户在单个主题上的活跃度越大越能获得高影响力[3]。还有学者开始研究校园范围内用户信息传播网络的构建和测量，提出了 InfluenceRank 算法来计算微博用户区域影响力[4]，以及利用其他的算法来评估用户影响力。如文献[5]引进 HIT 算法来计算在线社会网络用户的影响力。文献[6]提出 TwitterRank 算法，指出在给定主题下用户社会影响力是所有粉丝影响力的和。文献[7]提出在给定领域下用发布的微博和被转发的微博两个指标来度量该领域下的影响力。另外，还有研究人员考虑更多的维度来度量影响力，如 Liu 等人以新浪微博用户为研究对象，提出粉丝数、关注数、微博数、被转发微博数、评论数、被评论数和被转发评论数 7 个维度，采用主成分分析法获得各维度权重，建立了影响力评价模型[8]。除了可以采用全局性指标来度量直接联系的用户影响力外，还有隐含在非直接联系的用户之间的间接影响力，文献[9]利用 Twitter 用户的转发微博的可能性对间接影响力进行研究。

鉴于高影响力用户背后存在的巨大的商业价值，涌现出了一批网站及应用来计算用户的影响力。Klout 网站可以综合用户在 Facebook、Twitter、LinkedIn 等社交媒体上的表现（在线时长、互动、原创性、领域性），给用户一个评分，最后划分 8 种不同等级的角色。评分高的用户可以享用合作的航空公司提供的贵宾休息室[10,11]。国内的微博风云和微数据也采用多个维度对用户影响力进行排名，考察的维度包括活跃度、影响力、转发评论比、微博价值、微博数粉丝数比、关注率、传播力、覆盖度等[12,13]。

上述研究对用户影响力的考察是存在合理性的，但是维度的选择存在一定的局限性，即多数是考察全局性（如选用粉丝数、发布微博数）指标，没有在特定的领域范围内考量。而针对特定话题的研究采用按照关键字提取热门事件相关内容的方法获得数据，其可理解性以及代表性也有待提高。我们认为影响力是具有领域性的，而且用户间存在间接的相互影响，所以不能泛泛地评价一个用户的影响力，要在特定的领域范围内，考量多种影响因素的评价才具有合理性[7,9]。所以，本研究要判断用户的兴趣领域并在各个领域下判断用户的影响力。

3 研究方法

社交媒体用户的影响力没有固定的度量方法，应该考虑不同的背景环境、领域特性等。本研究认为用户在不同领域范围内所具有的影响力是不同的。以嘀咕网用户为例，通过获取、分析用户在网上留下的行为数据来研究他们在各个领域内所具有的影响力。因为嘀咕网是一个典型的基于位置的社交媒体服务网站，拥有大量的用户群，并且是免费开放的，可以使用工具对嘀咕网在线用户数据进行抓取，所以我们首先使用爬虫工具抓取嘀咕网用户数据，然后对抓取的数据进行结构化处理，用作后续研究使用，最后对嘀咕网用户的领域影响力进行度量。本节将介绍数据获取和数据处理工作，领域影响力的研究将在下一节介绍。

3.1 数据获取

我们使用免费的网页抓取工具 GooSeeker（www.gooseeker.com）来抓取嘀咕网站的用户数据。以一个嘀咕网活跃用户为起始点，抓取该用户的信息，包括个人基本信息（用户名、所在地、性别等）、发布的嘀咕内容、好友（他/她跟随的人）和粉丝（跟随他/她的人）。以上数据作为第一层数据。接下来，抓取第一层用户所有的好友及粉丝的上述信息，这些数据作为第二层数据。以此类推，继续抓取第三、四层数据。所有获得的数据以 XML 文件保存。整个数据获取过程从 2011 年 7 月持续到 12 月，共计 6 个月时间。

3.2 数据处理

以上数据以 XML 格式存储，我们对数据进行了整理提取，把用户基本信息、发布的嘀咕内容、好友、粉丝等信息存储到 MySQL 数据库中，以备后续研究使用。用户信息共有 4 部分内容，如表 1 所列。

表 1 用户信息表

用户基本信息表					
User ID 用户 ID	name 用户名	sex 性别	city 所在城市	count_digu 发布嘀咕数	count_friends 好友数
发布嘀咕内容表					
UserID 用户 ID	scraptime 抓取时间		posttime 发布时间	check in 签到	digu_content 嘀咕内容
好友关系表					
	UserID 用户 ID			friendID 好友的 ID	
粉丝关系表					
	UserID 用户 ID			fansID 粉丝 ID	

因为抓取工具及抓取规则的限制，获得的数据存在冗余及缺失，需要对这些数据进行去噪处理，删除冗余部分的数据以及明显错误的数据。例如，删除用户 ID 为空的数据、删除用户 ID 完全重复的数据等。最后共获得个人基本信息 57000 多条、嘀咕内容 75 万多条、好友关系 234 万多条和粉丝关系 350 多万条。

4 领域影响力

度量用户的领域影响力，首先需要划分领域，即把用户发布的内容进行领域分类；然后依据影响力评价维度，分别在各

领域内度量用户的影响力大小。我们以嘀咕网的实际用户为例，进行案例研究，评价提出的领域影响力。

4.1 领域划分

因为内容的分类没有统一标准，我们借鉴现有的门户网站内容分类标准，并且结合嘀咕用户数据自身的特点，把用户发布的嘀咕内容进行了领域划分。

通过卡片分类的方法将搜狐、新浪、凤凰、新华等网站自己的分类内容进行重新划分。4 名研究人员分别对各网站分类标签进行分类，最后进行统一。共得到包括娱乐、经济、新闻、情感、生活、科技、校园、公益等 8 个领域标签。将这 8 个标签作为参考，结合嘀咕内容的特点，将最终得到领域划分结果。

根据嘀咕用户发布内容的特点，可以将所有内容划分为 7 种形式，包括转载@、话题、@、签到评论、勋章、地盘老大和其他。转载@类的信息数量最少，也即抓取到的嘀咕用户进行转载的行为发生最少。我们认为能进行转载的人，相对来说也会进行其他的行为。所以从进行了转载的用户中随机抽取 15%，把这些用户发布的所有信息内容（上述 7 种形式的内容）抽取出来作为样本进行研究。借鉴文献[14]的微博信息分类方法来对嘀咕用户发布的内容进行分类。由 3 名研究人员分别对嘀咕内容进行分类，然后 3 人对分类进行讨论，原则是必须达到两人以上同意才能确定一个分类。总计确定了 48 个小类。然后结合门户网站得到的 8 个分类，将 48 个小类进行聚类。最终得到关于嘀咕网的 8 个领域分类，即嘀咕相关、科技、交友、美食、服装、音乐、体育、其他。

上述得到的领域分类可以用来划分嘀咕用户发布内容所属领域。一个用户由于发布不同领域的内容，可以被贴上多个领域标签。接下来利用支持向量机（SVM）对用户发布的内容进行领域划分。因为数据样本量比较大，进行了随机取样，选取了 1%的数据大概 76000 多条用于研究。根据关键字对所有数据进行分类。从得到的数据中发现，个别类的样本量级过大，综合考虑各个类的特点，将量级定在 1000 左右。对量级比较大的类别进行随机取样，如其他类，取样 1053 条。各类数据分布如表 2 所列。

表 2 领域类别数据分布

	嘀咕相关	科技	交友	美食	服装	音乐	体育	其他
样本数	1024	1000	382	1020	1004	842	933	1053

从上述样本中随机选取 75%作为训练集，剩余的作为测试集，利用 SVMCLS2.0 工具进行分类。得到的分类模型的查全率和查准率如表 3 所列。基于该模型对剩余的所有嘀咕内容进行分类。最终对于每名用户都可以得到若干领域标签。

表 3 查全率与查准率

查全/准率	嘀咕相关	科技	交友	美食	服装	音乐	体育	其他	总体
查全(%)	98.1	81.2	97.9	80.8	87.7	91.9	76.9	75	85.1
查准(%)	91.3	82.2	85.5	81.1	87.0	88.2	85.7	79.8	85.1

4.2 影响力度量

与全局影响力不同，领域影响力是指用户在某特定领域内所具有的影响力，即把影响力细化到专业领域内，也就是常说的术业有专攻，不能泛泛地说一个人影响力的大小。所以，基于上一部分的领域分类来度量用户的领域影响力的大小。

领域影响力的基本思想是用户发布的关于该领域的信息越多，用户的粉丝越多，用户粉丝关注该领域越多，则该用户的领域影响力就越大。选取用户领域信息质量和领域关系质量两方面的数据来度量用户的领域影响力，如图 1 所示。领域信息质量与用户发布的关于该领域的信息的数量和曝光度有关。领域关系质量与领域内粉丝数量和质量（粉丝的领域信息数量和领域深度）有关。

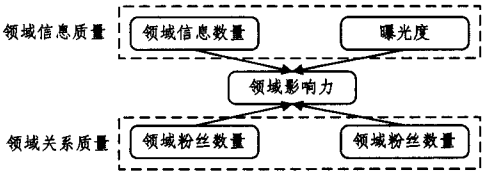

图 1 领域影响力评估模型

领域深度是用户在某领域的贡献的程度。计算方法是：

$$D_{i,a}=\frac{P_{i,a}}{\sum_{1}^{n}P_{j,a}},i=1,2,\cdots,n,a=1,2,\cdots,m \tag{1}$$

其中，$D_{i,a}$ 表示用户 a 在 i 领域的领域深度；而 $P_{i,a}$ 为用户 a 在 i 领域发布的信息数量；n 为领域总数；m 为用户总数。

用户 a 如果能影响某个粉丝 j，那么首先用户 a 的粉丝 j 要关注到 a 发布的信息，即从粉丝 j 接收到的众多好友发布的信息中查看到。所以用户 a 发布的某领域信息的曝光度与该用户的粉丝的好友数目有关。所以，曝光度的计算方法是：

$$E_{i,a}=D_{i,a}\frac{FAN_{i,a}}{\sum_{1}^{r}FRI\_FAN_{i,a,j}},j=1,2,\cdots,r \tag{2}$$

其中，$E_{i,a}$ 表示用户 a 在 i 领域信息的曝光度。$D_{i,a}$ 为用户 a 在 i 领域的领域深度，$FAN_{i,a}$ 代表用户 a 在领域 i 的粉丝数。$FRI\_FAN_{i,a,j}$ 代表用户 a 在领域 i 的第 j 个粉丝的好友数。

领域粉丝数量可以表示成：

$$F_{i,a}=\sum_{1}^{r}FAN\_FAN_{i,a,j},j=1,2,\cdots,r \tag{3}$$

其中，$F_{i,a}$ 表示用户 a 在 i 领域的粉丝数量。$FAN\_FAN_{i,a,j}$ 表示用户 a 在领域 i 的第 j 个粉丝的粉丝数。粉丝数量越多，传播越广，影响力越大。

领域粉丝质量可以表示成：

$$Q\_FAN_{i,a}=\frac{\sum_{1}^{r}FAN\_CONTENT_{i,a,j}\sum_{1}^{r}D\_FAN_{i,a,j}}{FAN_{i,a}^{2}} \tag{4}$$

其中，$j=1,2,\cdots,r$，$Q\_FAN_{i,a}$ 表示用户 a 在领域 i 的粉丝质量。$D\_FAN_{i,a,j}$ 表示用户 a 在领域 i 的第 j 个粉丝的领域深度。$FAN_{i,a}$ 表示用户 a 在领域 i 的粉丝数。$FAN\_CONTENT_{i,a,j}$ 表示用户 a 在领域 i 的第 j 个粉丝发布的信息数量。

综上，可以得到用户领域影响力的计算方法：

$$FI_{i,a}=P_{i,a}*E_{i,a}+p*F_{i,a}*Q\_FAN_{i,a} \tag{5}$$

领域影响力是领域信息质量与领域关系质量之和。其中 p 是阻尼系数，是信息被转发的概率，我们选定为 0.2。

为了验证上述方法，选取抓取的第二层嘀咕用户，共 516 人进行了研究。为了保护用户的隐私，以代号 1 到 516 来代表这一层用户，依次计算了样本用户粉丝的领域深度。图 2 截取了部分中间结果。设定阈值为 0.05，只要大于或等于这个阈值则说明这些粉丝对该领域感兴趣，易受到影响。然后计算粉丝的数量和质量，以及相应的领域曝光度。

fansID	总嘀咕数	领域嘀咕数	领域深度	粉丝好友数	粉丝粉丝数
1	42	2	0.047619048	42	15
2	9	1	0.111111111	16	5
3	17	1	0.058823529	14	2
4	16	1	0.0625	22	6
5	7	1	0.142857143	15	3
6	6	1	0.166666667	119	20
7	20	1	0.05	20	4
8	17	1	0.058823529	29	5
9	22	2	0.090909091	64	23
10	27	2	0.074074074	23	10
11	7	1	0.142857143	16	2
12	31	4	0.129032258	251	53
13	13	3	0.230769231	25	3
14	16	1	0.0625	32	6
15	14	1	0.071428571	28	3
16	43	4	0.093023256	31	13
17	10	1	0.1	58	2
18	27	2	0.074074074	30	8
19	55	47	0.854545455	82	28
20	32	1	0.03125	25	7
21	17	2	0.117647059	51	10
22	10	1	0.1	4	5
23	46	2	0.043478261	55	20
24	43	13	0.302325581	109	8
25	47	6	0.127659574	62	11

图 2　美食领域部分用户的领域深度

以 27 号用户和 42 号用户为例，经过计算，在美食领域的领域影响力 $FI_{美食,27}$ 为 10.26，$FI_{美食,42}$ 为 0.57。如表 4 所列，第一行数据是 27 号用户的，第二行数据是 42 号用户的。

表 4　27 号用户与 42 号用户的领域影响力数据

深度	领域嘀咕数/总	信息曝光度	领域粉丝数	粉丝质量	影响力
0.12	283/524	0.00048	168	9.46	10.26
0.20	267/1000	0.0021	149	0.35	0.57

也就是说在美食领域，27 号的影响力比 42 号大，他发布的关于美食的信息传播得更广泛、更具有权威性。而且可以看出：粉丝数与领域影响力并不成正比关系，也进一步验证了我们的设想，即影响力是分领域的，不能笼统地定义用户影响力的大小。通过计算用户在其他领域的领域影响力，发现用户在不同领域的影响力排名是不一致的，即用户在不同的领域具有不相同影响力，验证了我们的设想。

结束语　本研究主要对嘀咕网的用户领域影响力进行了度量。在获取嘀咕网在线用户数据的基础上，采用 SVM 等方法对用户发布的信息内容进行领域分类，建立领域影响力度量模型，提出领域信息质量和领域关系质量两方面的维度；并且采用样本数据来验证了领域影响力模型，度量了用户在不同领域的影响力。发现用户在不同领域的影响力是有差别的，领域影响力与粉丝数量不成正比关系，而是与领域内粉丝数量等因素有关。未来工作将与其他研究工作提出的度量维度进行对比分析，研究提出的领域影响力的稳定性及与其他全局性维度的相关性等问题。本研究提出的方法可以很好地识别用户在不同领域内具有的影响力。对于关注社交媒体分享行为的研究人员、社交媒体产品设计者以及社会化营销方案决策者来说，分领域影响力的提出可以更好地了解他们的目标用户，比如了解意见领袖或达人分享行为特点及受他们影响的用户的特征。本研究还可以用于研究意见领袖与普通用户进行交流的平台所提供的交互功能的局限性及用户使用体验等问题。而最被熟悉的应用还是社区意见领袖的发现、针对性的信息推荐及改善用户获取准确信息的体验。

参 考 文 献

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

[2] Ye Shao-zhi, Wu Fei-ix. Measuring message propagation and social influence on twitter. com [J]. Social Informatics Lecture Notes in Computer Science, 2010, 6430: 216-231

[3] Cha M Y, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter[C]// the Milloin Follower Fallacy Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM' 10). Washington, Menlo Park: The AAAI Press, 2010: 10-17

[4] 肖宇，许炜，商召玺. 微博用户区域影响力识别算法及分析[J]. 计算机科学，2012，39(9)：38-42
Xiao Yu, Xu Wei, Shang Zhao-xi. Analysis on Algorithms of Identifying Regional Influential Users in Micro-blogging [J]. Computer Science, 2012, 39(9): 38-42

[5] 王菲. 一种改进的 HITS 算法在 SNS 类网站用户影响力评估系统中的应用[D]. 吉林：吉林大学，2012
Wang Fei. An Application of Improved Hits Algorithm in User Influence Valuation System of SNS Websites [D]. Jilin: Jilin University, 2012

[6] Weng Jian-shu, Lim Ee-peng, Jiang Jing, et al. TwitterRank: finding topic-sensitive influential twitterers[C]// Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10). New York: ACM Press, 2010: 261-270

[7] Cataldi M, Mittal N, Aufaure M A. Estimating Donain-based User Influence in Social Networks[C]// Proceedings of SAC' 13. Coimbra, Portugal, 2013: 1957-1962

[8] Liu Qing, Peng Geng, Wang Ping. PCA-Based evaluation system of micro-blog influence and an empirical analysis of Sina microblog[C]// Conference on Web Based Business Management (WBM). Shanghai, China, 2012: 697-700

[9] Shuai Xin, Ding Ying, Jerome B, et al. Modeling Indirect Influence on Twitter [J]. International Journal on Semantic Web and Information System (IJSWIS), 2013, 8(4): 20-36

[10] http://www.36kr.com/tag/klout

[11] http://www.leiphone.com/tag/klout

[12] http://www.tfengyun.com/

[13] http://data.weibo.com/

[14] Yan Q, Chen H, Zhang P Y, et al. Microblogging After a Major Disaster in China-a case study of the 2010 Yushu earthquake[C]// CSCW2011. Hangzhou, China, 2011: 19-23

(上接第 49 页)

[14] Deng Ke-feng, Song Jun-qiang, Ren Kai-jun, et al. Graph-Cut Based Coscheduling Strategy Towards Efficient Execution of Scientific Workflows in Collaborative Cloud Environments[C]// GRID. 2011: 34-41

[15] Dhok J, Varma V. Using pattern classification for task assignment in mapreduce[C]// International Institute of Information Technology, Hyderabad, India, 2005

[16] Polo J, Castillo C, Carrera D, et al. Resource-aware Adaptive Scheduling for MapReduce Clusters[C]// ACM/IFIP/USENIX 12th International Middleware Conference (Middleware 2011). 2011