基于用户行为的微博用户社会影响力分析

毛佳昕 刘奕群 张 敏 马少平
(清华大学智能技术与系统国家重点实验室 北京 100084)
(清华信息科学与技术国家实验室(筹) 北京 100084)
(清华大学计算机科学与技术系 北京 100084)

摘 要 社会影响力分析是当前在线社会网络研究中的热点方向.随着微博成为了一种至关重要的大众媒体,更好的分析和衡量微博用户的社会影响力引起越来越广泛的关注.基于从新浪微博收集的大规模数据集,作者结合社会影响力在微博环境中的传播情况,分析了用户行为因素之间的关系.然后提出了一个通过预测用户传播信息能力大小来分析和度量用户社会影响力的方法.该方法结合了来自社会网络结构和用户行为因素两方面的信息,获得了更好的影响力估计结果.基于大规模数据的实验结果表明,作者提出的方法是较为有效的.

关键词 用户行为分析;社会网络;社会影响力;信息传播;社会计算
中图法分类号 TP391 **DOI 号** 10.3724/SP.J.1016.2014.00000

Social Influence Analysis for Micro-Blog User Based on User Behavior

MAO Jia-Xin LIU Yi-Qun ZHANG Min MA Shao-Ping
(*State Key Laboratory of Intelligent Technology & System, Tsinghua University, Beijing* 100084)
(*Tsinghua National Laboratory for Information Science and Technology, Tsinghua University, Beijing* 100084)
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract Recently, social influence analysis is an emerging topic in the research area of online social networks. As micro-blog becomes a mass media of vital importance, better analysis and measurements of the micro-blogging users' social influence are paid more and more attention. Based a large scale dataset collected from weibo.com, the relationship between user behavior factors are analyzed together with the diffusion of social influence on micro-blogging environment. A learning-based method is then proposed for analyzing and measuring users' social influence via predicting users' capability of propagating information. Both information extracted from social network structures and user behavior factors are combined in the method to gain a better estimation. Experimental results based on the large scale data set show effectiveness of the proposed method.

Keywords user behavior analysis; social network; social influence; information diffusion; social computing

1 引 言

社会影响(Social Influence)指一个人的情绪、意见或者行为被他人影响的现象.社会影响在社会网络(Social Network)中的传播方式和作用结果是社会学中由来已久的研究对象.从上世纪中叶开始,社会学家相继提出了二级传播理论[1]、弱连带优势理论[2]、强连带优势理论[3]和结构洞理论[4]等理论来分析和描述社会网络中的社会影响现象,为该领

收稿日期:2013- - ;最终修改稿收到日期:2013- - .本课题得到国家"八六三"高技术研究发展计划项目基金(2011AA01A205)、国家自然科学基金(61073071)资助.**毛佳昕**,男,1991 年生,博士研究生,主要研究方向为社交网络分析.E-mail:maojiaxin@gmail.com.**刘奕群**,男,1981 年生,博士,副教授,主要研究方向为信息检索.**张 敏**,女,1977 年生,博士,副教授,主要研究方向为机器学习、信息检索.**马少平**,男,1961 年生,教授,博士生导师,主要研究领域为知识工程、信息检索、汉字识别与后处理以及中文古籍数字化.

域奠定了理论基础.

但是,由于真实世界中的社交关系和社会影响往往是难以观测的,社会影响的形式也往往是复杂而多样的,对社会影响的实证研究受到了很大的限制.而这种情况,随着一些在线社会网络的出现和兴起,得到了改善.大型在线社交网站因其用户数量大、用户行为活跃、用户交互方式较为一致、用户行为和关系记录完整等特点,为研究社会网络中的社会影响现象提供了一个较为理想的实验环境.同时,更好的理解社会影响这一现象及衡量在在线社会网络中社会影响力的大小能够加深我们对在线社会网络的认识,并且能够改进好友推荐[5]、专家用户发现[6]、垃圾信息发现、病毒营销和内容排序等应用.所以,通过挖掘在线社交网站的数据,分析社会影响现象,度量社会影响力大小成为了当前的一个研究热点.

具体到国内互联网环境,微博类网站是用户数最多的在线社交网站.根据 CNNIC 在 2013 年 1 月公布的中国互联网络发展状况统计报告[7],微博用户在 2012 年末达到 3.09 亿,年增长率达 23.5%,网民中微博用户的比例达到 54.7%,报告认为微博已经成为中国网民使用的主流应用,处于网络舆论传播中心地位.所以,更好地分析微博类网站中的信息传播和社会影响现象,度量中文微博环境中的用户社会影响力大小,有着尤为重要的意义.

通过比较,我们认为微博类网站与其它在线社交网站的主要区别有两点.首先,微博用户间的关系是单向关注关系,而其它社交网站用户间的关系是双向的好友关系.在一个典型的微博类网站中,系统通常展示给用户一条实时更新的、包含其所有被该用户关注的用户的一段时间内所有微博的时间线①,供其进行浏览和阅读.

其次,与其它社交网站提供了发布状态(类似微博的短消息)、发布日志(有格式的长文章)、分享图片或视频等多媒体内容,以及评论和转发其它用户发布的内容等较为多样的使用方式不同,在微博类网站上,用户通常只能发布一条 140 个字符以内的微博,或者转发一条他所阅读过的微博.而一个用户所有发布或转发的微博,都会以相对统一的形式被展现在关注了该用户的其它用户的时间线上.

以上两个特点使得微博类网站更类似于一个社会化的媒体,而不是普通的社会网络[8].因此,我们认为在微博类网站上,应该以媒体传播效果而非网络拓扑结构中的重要性作为影响力度量的标准.具体来说,可以通过用户所发布的一条微博的平均被阅读次数作为该用户影响力大小的一个可行的度量指标.同时,微博上信息传播扩散的主要机制是转发,而转发一条微博这一行为本身也是受到影响的标志,所以我们还考虑了用户所发布的一条微博的平均被转发次数这一指标.

基于这一研究思路,我们从国内最具影响力的微博类网站——新浪微博(www.weibo.com)上收集了包括 114 万用户,近 7.6 亿条微博以及相关的关注关系信息的数据集.通过对该数据集的分析,我们尝试将用户访问微博的时间、用户阅读微博的方式以及用户转发微博的偏好等行为因素与社会网络结构信息相结合,以更好地估计用户的影响力.

本文第 2 节将主要调研在线社会网络社会影响力研究方面已有的成果和存在的问题;第 3 节将简要介绍论文工作所基于的数据集合,并从时间维度分析微博发布行为与转发行为之间的关联关系;第 4 节将基于数据观察,提出影响力估计模型和相关参数的估计方式;第 5 节介绍实验设置和实验结果;最后给出总结和对未来研究工作的思考.

2 相关工作

正如在第 1 节中提到的,影响力分析是当前在线社会网络研究中的热点方向,研究者们从不同的角度分析了不同形式的社会影响现象.

2003 年,早在微博类社交网站出现前 Kempe 等人[9]就研究了"影响力最大化(Influence Maximization)"模型,并对在给定传播模型参数的条件下,如何寻找最优的种子用户集合,使得最后社会影响的传播扩散规模最大这一 NP 难问题给出了近似解.但他们的研究并没有将重点放在从实际的数据中推测传播模型的参数上,所以不能直接用于分析微博用户的社会影响力大小.2010 年,Goyal 等人[10]基于上述传播模型,从图片分享网站 Flickr 的用户行为记录中推测模型参数.他们基于得到的参数,分析了互为好友的用户对之间的影响力大小,而没有分析用户在社会网络中全局意义上的影响力大小.

在微博类社交网站中,研究者较为集中关注 Twitter 网站上的各种影响力传播现象,并用不同方式开展了针对 Twitter 用户的社会影响力的分析工作.

① 部分微博系统(例如新浪微博)也向用户提供以时间之外的其它因素(如交互程度等)进行微博排序的功能,但由于时间线展示是默认的形式,因此我们仍旧以时间线为例来开展本文的研究工作.

2010 年，Cha 等人[11]比较分析了按照被转发次数、被提及(@mention)次数和关注者数量 3 种衡量用户影响力的方式，并分析了影响力随时间变化的规律. 他们发现，拥有较多关注者的用户并不一定能引发更多的转发和提及行为. 这说明了我们不能简单通过关注者数量等网络拓扑结构特征来衡量用户影响力，但该研究主要着重于衡量用户历史上某个时刻的影响力有多大，并没有提出算法通过历史数据来估计当前或者未来时间内用户的社会影响力.

同年，Weng 等人[5]基于用户间的关注关系和用户发布微博的主题信息，分析了在特定主题下用户影响力的大小，并发现算法给出的排名在用户推荐方面取得了较好的效果. 但根据文献[11]，有更大可能被他人关注的用户并不一定有相对较强的传播信息能力，所以该方法并不一定能较好地估计用户的社会影响力.

2011 年，Bakshy 等人[12]使用回归树的方法估计用户发布的含有短链接的微博在全局的传播规模，并用预计的平均传播规模大小作为用户社会影响力大小的估计指标，取得了较好的效果. 我们认为，这部分工作所达成的目标与用户社会影响力的估计最为接近，因此，我们在本文中也将重点实现并将该方法与我们所提出的方法进行比较.

与上述工作相比，我们的研究在关注关系、历史影响力统计和微博内容等被广泛研究过的因素之外，创新性地将用户的阅读行为因素引入了影响力的计算模型中. 并且，基于中文微博类网站上包含短链接的微博比例较低这一特点，我们估计和预测了所有微博被阅读和被转发的次数，以此作为用户传播信息能力大小的度量指标，分析和衡量用户的社会影响力.

3　微博发布与转发行为的时间关联分析

用户与微博类网站交互的一般过程为：用户首先在某个时间，访问了微博的客户端(如网页客户端、移动客户端等)，然后选择发布原创微博，或者浏览微博系统为其生成的时间线，阅读其中的微博，最后转发一些他感兴趣的微博. 所以，从用户行为分析的角度出发进行用户影响力估计，我们首先尝试回答以下两个问题：

(1) 用户在什么时间访问了微博？

(2) 用户在时间线中会阅读哪些微博？

为了回答这两个问题，我们统计和分析了新浪微博上的微博发布时间和微博转发延迟时间.

3.1　数据集合

为了分析转发行为，我们需要同时获得用户所发布和转发的所有微博，以及与其相关的关注关系信息. 因此，我们首先通过微博搜索接口通过特定关键词(清华大学)获得了 97 个种子用户，再通过微博 API，获取这 97 个用户的所有关注者，总共 12 255 个用户. 再次调用微博 API，获取这 12 255 个用户关注的所有用户，得到总共 1 146 117 个用户和 2 115 949 条用户间的有向关注关系. 最后再抓取这部分用户所发布和转发的所有微博，总数为 759 006 885 条. 时间跨度为从 2009 年 9 月到 2012 年 11 月.

3.2　微博发布行为的时间分布

考虑微博受众的阅读行为，一条在关注者访问微博可能性较大的时刻发布的微博显然更有可能被关注者阅读，并进一步更可能被转发. 虽然从数据集中，我们只能获知用户发布或转发微博的时间信息，但我们可以认为用户发布和转发微博的时间分布可以用于近似地估计用户访问微博的时间分布. 所以，我们先统计了数据集上原创微博和转发微博的时间分布，结果如图 1 所示.

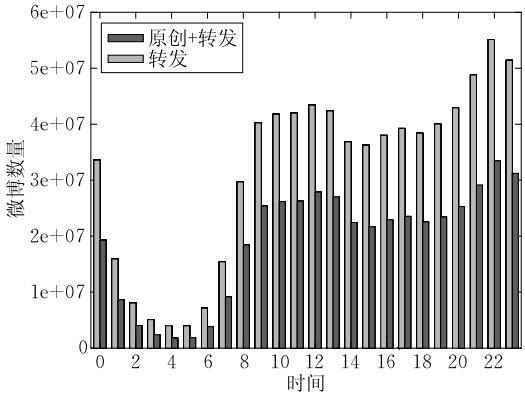


图 1　一天内发布和转发微博的时间分布

从图 1 中我们可以看出，除了每天的 0 点到 8 点的睡眠时间之外，其它时刻的分布都较为平均，并且转发和原创微博的时间分布也较为一致. 但这并不能说明单个用户访问微博的时间分布也较为平均，为了进一步研究不同用户访问微博的时间模式的差异性，我们分别统计了每个用户在在一天中各个时刻(按小时划分)发布(包括转发)微博的频率，并对用户群体进行了 K-means 聚类分析①，相关结果见图 2.

① 经过多个不同参数的测试，选取了 $K=3$ 作为最终聚类结果.

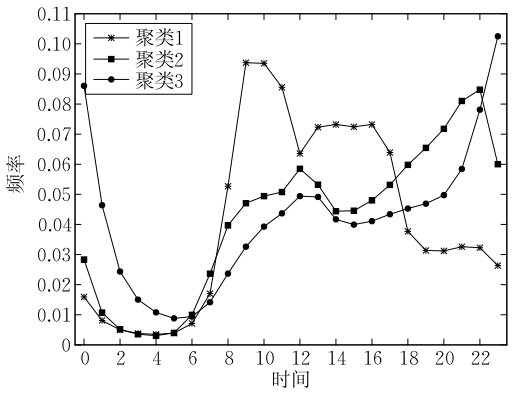


图 2 单个用户发布微博时间分布聚类分析结果

其中,聚类 1 包括 259 732 个用户,聚类 2 包括 513 901 个用户,聚类 3 包括 372 484 个用户.从图中可以明显看出聚类 1 的用户主要在工作时间使用微博,而聚类 2 和聚类 3 的用户则更多在晚间时段使用微博,这说明了不同的用户访问微博的时间模式还是有较大不同的.

3.3 用户关注与微博转发行为的时间维度关联分析

在不同用户访问微博的时间模式不同的基础上,我们进一步探究访问微博的时间分布不同会不会影响用户之间的关注关系和用户之间的转发关系.于是我们设计了两个假设检验来回答以下两个问题:

(1)有关注关系的两个用户的发布微博的时间分布是否更为相似?

(2)若有关注关系的两个用户属于同一聚类,关注者是不是有更大的可能性受到被关注者影响,转发被关注者发布的微博?

对于问题 1,我们随机选取了 10 000 对有关注关系的用户和 10 000 对没有关注关系的用户分别作为实验组与对照组.实验组与对照组内部的平均余弦相似度和相似度方差如表 1 所示.

表 1 实验组和对照组用户发布微博时间分布的余弦相似度统计

	均值 μ	方差 s
实验组(有关注关系)	0.7373	0.0310
对照组(无关注关系)	0.7047	0.0337

我们选取的零假设 H0 是实验组和对照组的均值相等.由于样本较大,均值满足正态分布,实验组和对照组均值的差也满足正态分布.均值的差偏离 H0 假设的理论值 0 多达 12 个标准差,P 值接近 0,在显著性水平 $\alpha=0.01$ 拒绝零假设.这说明,对问题 1 的回答是肯定的.而导致有关注关系的两个用户发布微博的时间更为相似,有一方面可能是关注关系的社会影响带来的结果,使得关注者在行为模式上与被关注者更为相似,另一方面也可能是选择的结果——用户会选择关注在某些方面与自己相似的用户,而访问微博的时间相同,就是用户之间较为相似的表现之一.

对于问题 2,我们随机抽取了 10 000 对之间有关注关系且属于同一聚类的用户对作为实验组,同时随机抽取了 10 000 对之间有关注关系但是不属于同一聚类的用户作为对照组.对于一对关注者 u 和被关注者 v,我们按如下公式计算转发概率:

$$p=\frac{u\text{ 转发 }v\text{ 的微博的次数}}{v\text{ 发布的微博数量}} \tag{1}$$

然后再分别计算实验组和对照组的平均转发概率,结果见表 2.

表 2 实验组和对照组用户转发微博概率统计

	均值 μ	方差 s
实验组(属于同一聚类)	0.001353	3.30e-05
对照组(不属于同一聚类)	0.000993	1.51e-05

选取的零假设 H0 仍然是实验组和对照组均值相等,同样可以认为均值符合正态分布,P 值为 0.0003,在显著性水平 $\alpha=0.01$ 拒绝零假设.这说明,问题 2 的回答依然是肯定的,即与被关注者属于同一个聚类的关注者平均来说有更大的可能性转发该被关注者的微博,所以利用用户访问微博的时间分布信息应该能更好地预测转发行为,进而更准确地估计用户的社会影响力大小.

3.4 微博转发延迟的分布

接下来,我们统计了微博的转发延迟的分布情况.微博的转发延迟指一条转发微博的发布时间和它直接转发的微博的发布时间的差值,即转发树中一个节点与它的父节点的时间差.统计的结果见图 3.

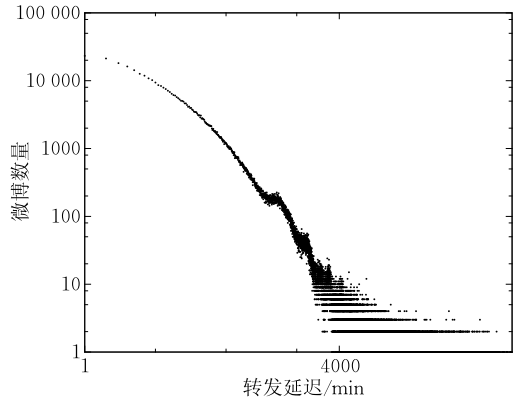


图 3 微博转发延迟的分布

从图 3 可以看出，除了尾部一些少量转发延迟特别长的微博之外，转发延迟的分布近似符合幂律（在对数-对数坐标下近似为一条直线）. 据统计，50％的微博的转发延迟小于 55 min，90％的微博的转发延迟小于 1153 min（约 19.2 个小时）.

这说明，微博上的信息传播的时效性较强，用户通常不会去转发较旧的，例如一天之前发布的微博. 导致这一点的原因一方面可能是用户通常倾向于传播最近的新闻，另一方面可能是由于微博显示在时间线中的顺序是由新到旧，用户可能根本不会看到较早之前发布的微博，更加不可能转发了.

无论是哪种原因，阅读与发布之间的时间间隔无疑会影响一条微博被转发的可能性，所以我们认为利用该信息，同样能更好地估计用户影响力.

4　影响力估计模型

4.1　全局被阅读次数

基于上述对数据的观察和分析，我们提出了一个可以对用户阅读微博的行为进行建模，进而利用不同用户访问微博的时间分布以及微博转发延迟信息来估计用户在某一时刻发布的微博在全局范围内被阅读次数的模型.

首先，我们用 $I_u(t_s,t_e)$ 表示用户 u 在 t_s 时刻发布的微博在 t_e 前在全局范围内被阅读次数的期望值. 然后，我们假设：

假设 1.　用户必须阅读过某条微博后才有可能转发该条微博；

假设 2.　用户会在第一次阅读某条微博时决定是否转发该微博；

假设 3.　用户只会阅读他关注的用户发布或转发的微博；

假设 4.　在已知用户第一次阅读某条微博后，用户的微博转发行为仅受其个人转发偏好影响，而独立于其它条件.

这时，可得以下式(2)：

$$I_v(t_s,t_e)=\sum_{u,u\text{关注}v}\int_{t_s}^{t_e}P_u(t_s,t)[1+q_uI_u(t,t_e)]\mathrm{d}t \tag{2}$$

其中，$P_u(t_s,t)$ 的含义是用户 u 在 t 时刻第一次阅读 t_s 时刻发布的微博的概率. q_u 是用户在阅读某条微博后转发该条微博的条件概率，用来描述用户的转发偏好.

由于假设 3，我们认为某个用户的影响力只由他的所有关注者决定. 并且由于假设 4，总影响力可以被分解为关注者带来的影响力的和.

再单独考虑用户 v 的某个关注者 u 在 t_e 时刻看到 v 在 t_s 时刻发布的微博这一事件. 如果该事件发生，那么 u 一定在 t_s 到 t_e 之间的某个时刻 t 第一次阅读该微博，所以根据全概率公式，上述事件发生的概率可表示为以下的式(3)，注意这也是由用户 u 的直接阅读带来的全局被阅读次数的期望值：

$$\int_{t_s}^{t_e}P_u(t_s,t)\mathrm{d}t \tag{3}$$

由于我们考虑的是全局范围内的被阅读次数，所以还需要考虑转发带来的间接阅读. 根据假设 2，用户 u 在 t 时刻转发用户 v 在 t_s 时刻发布微博的概率为 $P_u(t_s,t)q_u$，再根据全期望公式，由用户 u 的转发行为带来的全局被阅读次数的期望值为

$$\int_{t_s}^{t_e}P_u(t_s,t)q_uI_u(t,t_e)\mathrm{d}t \tag{4}$$

综合式(3)、(4)，我们即可得到式(2).

4.2　全局被转发次数

我们认为，某个用户在某个时刻发布的微博在一定时间段内被阅读的次数，可以直接被用来衡量用户传播信息能力的大小. 然而，由于用户浏览和阅读微博的行为不会被记录下来，我们无法得知一条微博的被阅读次数的真实值，所以无法评价我们的模型是否有效. 为了解决这个问题，我们发现可以通过修改式(2)，得到了一个利用同样的信息，同样的假设，估计用户发布的微博在全局范围内被转发次数的模型：

$$RP_v(t_s,t_e)=\sum_{u,u\text{关注}v}\int_{t_s}^{t_e}P_u(t_s,t)[q_u+q_uRP_u(t,t_e)]\mathrm{d}t \tag{5}$$

其中 $RP_v(t_s,t_e)$ 表示用户 u 在 t_s 时刻发布的微博在 t_e 前在全局范围被转发次数的期望值. 需要注意，由于用户 u 在第一次看到某条微博时会以 q_u 的概率决定是否转发，而只有在选择转发时，才会对用户 v 所发的微博的被转发次数有贡献. 所以，我们在式(5)中，将式(2)中的 1 更改为 q_u.

4.3　模型参数的估计和计算方式

为了计算上述 $RP_v(t_s,t_e)$ 和 $I_v(t_s,t_e)$，我们需要估计 $P_u(t_s,t)$ 和 q_u.

对于 $P_u(t_s,t)$，我们从数据集中只能得到用户发布微博的时间，我们可以通过如下的方式，依据用户发布微博的时间分布计算其在 t 时刻第一次阅读 t_s 时刻发布的微博的概率.

首先进行离散化，对每个用户统计在这每天 24 个

小时组成的区间内发布微博的概率分布.并且,如果用户历史上从未在一小时内发布微博,我们不能简单的认为之后他也不可能在这个小时发布微博,所以,我们对该概率分布进行了拉普拉斯平滑处理,得到平滑后的分布 $\rho_u(t)$,具体计算公式如下:

$$\rho_u(t)=\frac{\text{用户 } u \text{ 在 } t \text{ 小时发布的微博数量}+1}{\text{用户 } u \text{ 发布的微博总数}+24} \quad (6)$$

然后我们需要计算用户在一天内某个小时浏览微博的概率 $p'_u(t)$.注意到该概率和用户的活跃程度相关,并且在这里我们需要估计用户访问微博的可能性,而不仅仅是发布微博的可能性.在这里,我们假设用户每次访问微博网站时平均发布的微博数量是不随时间变化而变化的,那么用户访问微博的时间分布会与 $\rho_u(t)$ 基本一致(在不考虑平滑操作影响的情况下).进一步的,我们可以假设用户每次访问微博时,会按照该时间分布独立的在 24 小时内选择一个时间,那么用户在一天内某个小时访问微博网站,浏览微博的概率 $p'_u(t)$ 可按照以下式(7)计算:

$$p'_u(t)=1-(1-\rho_u(t))^{\alpha n_u} \quad (7)$$

其中 n_u 为用户一天内平均发布微博的次数.同时,我们用 $1/\alpha$ 表示用户每次访问微博平均会发布或者转发多少条微博.所以,αn_u 即为用户平均每天访问微博的次数.由于我们无法获知用户访问微博的情况,α 的值只能作为一个待定参数人为选取,我们将会通过实验选择较为合适的 α.

接着,考虑微博时效性的影响,我们计算用户在 t 时刻阅读一条 t_s 时刻发布的微博的概率 $\lambda_u(t_s,t)$:

$$\lambda_u(t_s,t)=p'_u(t)\cdot f_u(t_s,t) \quad (8)$$

其中 $f_u(t_s,t)$ 用来表示微博时效性对用户的阅读行为和转发行为的影响,也是模型的一个待定参数.在后续实验中,我们尝试并比较了不考虑微博时效性($f_u(t_s,t)$ 恒等于 1)和考虑微博时效性(根据 3.4 小节中的统计,取 $f_u(t_s,t)=\frac{1}{t-t_s+1}$)两种 $f_u(t_s,t)$ 的选取方法.

最后我们计算 $P_u(t_s,t)$,对于时间连续的情况,在已知用户在 t 时刻阅读一条 t_s 时刻发布的微博的概率 $\lambda_u(t_s,t)$ 的情况下,有

$$P_u(t_s,t)=\exp\left(\int_{t_s}^{t}-\lambda_u(t_s,t)\mathrm{d}t\right)\cdot\lambda_u(t_s,t) \quad (9)$$

而在按照小时离散化的情况下,有

$$P_u(t_s,t)=\prod_{\tau=t_s}^{t-1}(1-\lambda_u(t_s,\tau))\cdot\lambda_u(t_s,t) \quad (10)$$

对于 q_u,同样由于我们无法直接从数据集中知道用户阅读过哪些微博,只能人为设定其估计方式,再通过实验验证.我们比较了不考虑转发偏好(q_u 恒等于 0.01)和考虑转发偏好两种方法.对于考虑转发偏好的情况,我们按以下式(11)计算 q_u:

$$q_u=\frac{\text{用户 } u \text{ 历史转发的微博数量}}{\text{用户 } u \text{ 时间线上出现的微博数量}} \quad (11)$$

需要指出,由于本文的主要目标是从用户行为的角度对用户的影响力进行分析,同时为了避免数据稀疏性带来的影响,在这里我们只是简单的考虑了用户 u 转发行为的习惯,而没有对不同的用户 v,分别估计用户 u 转发用户 v 的微博的概率大小.

5 实验结果分析及讨论

5.1 实验设置和评价方式

为了验证第 4 节中提出的方法的有效性,我们利用新浪微博数据进行了实验.由于无法从数据集中获得微博被浏览和阅读的信息,所以实验主要针对估计被转发次数的方法.我们将收集到的 701 921 次直接转发按照直接被转发的微博的发布时间排序,取前 2/3 作为训练集,后 1/3 为测试集.得到划分训练集和测试集的时间分点为 2012 年 6 月 22 日 18 时 15 分 15 秒.然后使用训练集上的数据,按照 4.3 节中的方法,统计各用户平滑后的发布微博时间分布 $\rho_u(t)$、日均发布微博数量 n_u 和转发偏好 q_u 等参数,再基于这些参数和用户之间的关注关系,计算 $RP_u(t_s,t_e)$.在这里,我们均以一小时为最小的时间单位,将式(5)的积分形式转化为离散求和形式进行计算.

虽然我们的模型能够估计每一个用户 v 在 t_s 时刻发布的微博在 t_e 时刻前被转发的次数,但是如果想对于每组 t_s,t_e,v 的选取都在测试集上测试,则会遇到严重的数据稀疏问题.并且,我们的最终目的是估计用户总体的社会影响力大小.所以,我们在训练集上计算 $RP_v(t_s,t_e)$ 完成之后,按以下公式计算用户 u 发布的微博,在接下来的一天时间内,平均全局被转发的次数的期望值 Inf_u:

$$Inf_u=\sum_{t=0}^{23}\rho_u(t)\cdot RP_u(t,t+24 \bmod 24) \quad (12)$$

同时,我们在测试集上,统计用户实际上发布的微博的平均被转发次数.由于我们的目标是衡量用户影响力大小,在这里我们没有采取直接对比 Inf_u 和测试集上的平均被转发次数,计算其误差的方法来进行评价,而是对用户按 Inf_u 从大到小排序得到

的结果，与测试集上将用户按平均被转发次数从大到小排序得到的排序结果进行比较，评价指标选取了 Spearman 秩相关系数 ρ[13].

在实际实验中，我们采用的方法是，从测试集中选择排名最靠前的 N 个用户，得到测试集上的 Top$N$ 排序. 再在要评价的方法计算所得的排序中，过滤得到这 N 个用户的子排序与测试集上的 Top$N$ 排序计算 Spearman 秩相关系数. 我们注意到，测试集中有很多发布微博较少（如只有 1 条）的用户平均每条微博被转发的次数排名会很靠前，我们认为由于其发布微博数量较少，平均被转发次数不具有统计意义. 所以，我们过滤掉了测试集中发布微博数量少于 100 条的用户. 考虑到测试集的时间跨度长达近 5 个月，在分析影响力时，去除这些不活跃用户是合理的.

5.2　模型待定参数的选取

接下来，我们通过实验的方式来选取 4.3 小节中提到的 3 个待定参数：α、$f_u(t_s,t)$和 q_u.

首先，对于 $f_u(t_s,t)$和 q_u，我们总共尝试了 4 种不同的组合方式，详见表 3.

表 3　待定参数 $f_u(t_s,t)$和 q_u的设置方式

实验设置	$f_u(t_s,t)$	q_u
1	$1/(t-t_s+1)$	0.01
2	$1/(t-t_s+1)$	考虑转发偏好*
3	1	0.01
4	1	考虑转发偏好

注：带“*”表示按式(11)计算，下同.

而对于待定参数 α，我们从小到大分别取 1、2、5、10、20、50、100 不同的值进行了实验.

于是，我们总共测试了 28 种不同的参数选取方式，分别计算这些参数选取方式在考虑测试集上 Top100，Top5000 和 Top20000 的用户排序时的相关系数，并绘制相关系数与 α 的关系图. 所得结果分别见图 4、图 5、图 6.

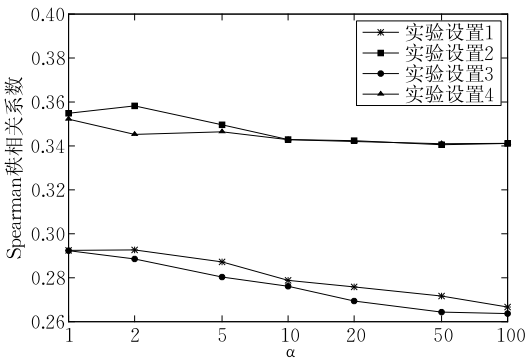


图 4　在表 3 四种实验设置下，考虑测试 Top100 用户时，相关系数与 α 的关系图

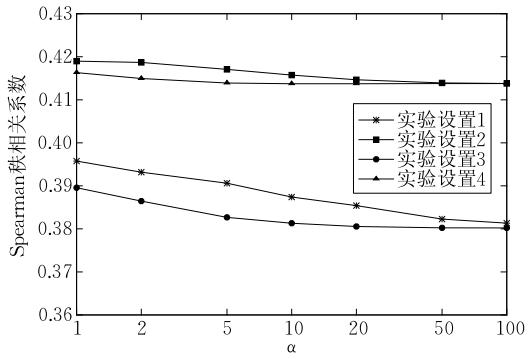


图 5　在表 3 四种实验设置下，考虑测试 Top5000 用户时，相关系数与 α 的关系图

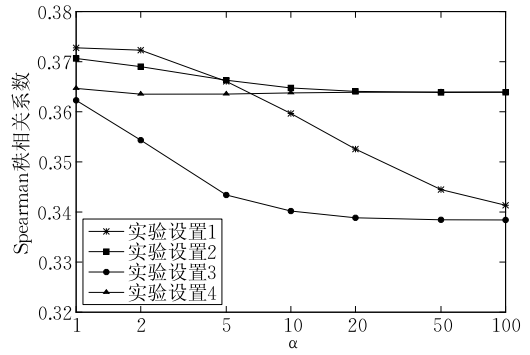


图 6　在表 3 四种实验设置下，考虑测试 Top20000 用户时，相关系数与 α 的关系图

通过分别对比实验设置 1 和实验设置 2 的相关系数以及实验设置 3 和实验设置 4 的相关系数，可以发现考虑用户转发偏好的实验结果通常优于不考虑相应偏好的实验结果（除了在考虑 Top20000 用户，考虑微博时效性，且 α 取值较小时不考虑转发偏好的相关系数更高）.

而通过分别对比实验设置 1 和实验设置 3 的相关系数，以及实验设置 2 和实验设置 4 的相关系数，发现考虑了微博时效性的实验结果均比不考虑微博时效性的实验结果相关性高.

对于 α 值，我们发现相关系数随着 α 的增大而递减. α 取 1（平均而言，用户每访问 1 次微博网站即发布 1 条微博）时的结果通常是最好的（除了在考虑 Top100 用户时，设置 2 下 α 取 2 相关系数最大）. 这也说明了用户与微博的互动还是较为活跃的.

5.3　与其它方法的比较

在找到了一组较好的参数（5.2 小节中的设置 2，并取 $\alpha=1$）的基础上，我们将所提出的方法作为方法 1 与以下 3 种估计平均被转发次数，及衡量用户影响力的方法进行比较，以验证该方法的有效性.

方法 2. 利用训练集上用户一条微博的全局平均被转发次数从大到小排序. 与生成测试集上的排

序类似,我们为了排除发布微博较少用户的干扰,过滤掉了在训练集上发布微博数量少于 200 的用户.这里选取的阈值为测试集上相应过滤阈值的两倍,这是因为训练集的微博数量也为测试集的两倍.

方法 3. 使用类似文献[12]中的方法,在训练集上,选取用户的关注者数量的、用户关注的其它用户数量、发布微博数量、局部(只考虑直接关注者)平均被转发次数、局部最大被转发次数和全局最大被转发次数作为输入特征,将输入取对数后,用回归树拟合用户发布微博的全局平均被转发次数的对数值.再通过回归树的拟合值从大到小排序.

方法 4. 仅仅考虑关注关系网络的信息,使用用户的关注者数量从大到小排序.

所得结果见图 7,从中可以看出,我们提出的方法在估计测试集上用户平均一条微博被转发次数的排序方面,明显优于其它 3 种方法.

值得注意的一点是方法 2,训练集上的用户平均全局被转发次数从大到小得到的排序与测试集上的排序的相关系数大约只有 0.35 左右,这说明用户传播信息能力的大小随着时间的推移是会较为显著改变的.所以,我们不能只简单地根据用户历史上的一些统计数据,如被转发数量、被提及(@)数量等,来估计用户在未来传播信息、造成社会影响、引发其它用户互动的能力.

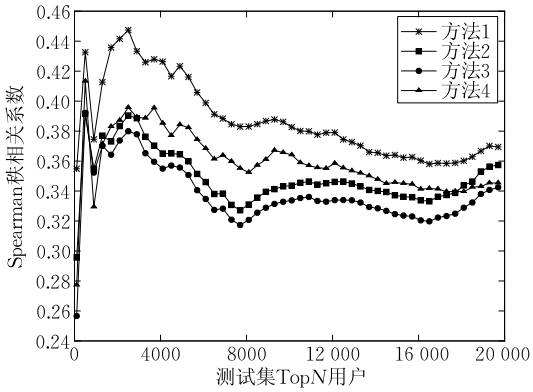


图 7 4 种方法的结果对比(方法 1,表 3 第 4 组设置下本文提出的方法;方法 2,根据训练集上用户微博平均全局被转发次数排序;方法 3,按照回归树在训练集上的拟合结果排序;方法 4,按照关注者数量排序)

与文献[12]中进行的 tweet 级别的回归不同,由于我们的数据集中微博较多,实现方法 3 时,我们采用了用户级别的回归.但需要注意到方法 3 和文献[12]中的特征全都是用户相关的特征,并且回归树选取出的最主要特征——用户关注者数量和训练集上的局部影响力——也与文献[12]中提到的一致.由于方法 3 使用回归树拟合的目标是训练集上用户平均全局被转发次数,所以从结果上看,得到的结果曲线与方法 2 的很相似,但性能上还不及方法 2.

方法 4 仅仅利用了最简单的用户关注关系信息——用户有多少个关注者,但从结果来看,它的性能在很大一个区间内,都比方法 2 和方法 3 要好.这说明了尽管之前文献[5,11]均有提到,在历史数据上,从关注关系信息得到的影响力排序和从转发和提及关系得到的影响力排序相差较大,但考虑到关注关系在时间变化下相对稳定,在预测未来影响力时,关注关系信息还是有较大价值的.

6 结论与未来工作

包括微博在内的在线社会网络的兴起,给社会影响现象的相关研究提供了理想的实验平台.同时,对于社会影响方面的研究又能对改进在线社交网站中的一些关键性的应用起到帮助作用.我们对微博这一在中文环境下占主导地位,而又有其独特之处在在线社会网络的数据进行了分析.发现了用户访问微博的时间分布、微博对用户来说的时效性以及用户转发微博的偏好等用户行为相关的因素会影响用户的转发行为,进而影响用户在微博平台上传播信息的能力.

基于上述发现,我们提出了一个考虑了上述因素的,通过估计用户所发微博在全局范围内被转发的次数这一与影响力的定义较为切合的指标大小,来衡量用户影响力的方法.并通过实验,验证了该方法的有效性.

在接下来的工作中,我们将从两个方面着手,进一步改进我们的方法:一方面,我们希望能够获取更为丰富的用户行为记录.如通过分析浏览器日志、微博客户端日志等记录,来直接地获知用户真实与微博交互的情况,以更好地分析用户行为因素.另一方面,在当前的研究工作中,我们没有考虑微博内容的信息.我们认为,引入主题、情感、主体等语义信息,也将使我们能够更好地分析和度量社会网络上用户的社会影响力.

参 考 文 献

[1] Lazarsfeld P F, Berelson B, Gaudet H. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. New York: Columbia University Press, 1944

[2] Granovetter M. The strength of weak ties. American Journal of Sociology, 1973, 78: 1360-1380

[3] Krackhardt D. The strength of strong ties: The importance of philos in organizations//Nohria N, Eccles R G eds. Networks and Organizations: Structure, Form, and Action. Boston: Harvard Business School Press, 1992: 216-239

[4] Burt R S. The social structure of competition//Nohria N, Eccles R G eds. Networks and Organizations: Structure, Form, and Action. Boston: Harvard Business School Press, 1992: 57-91

[5] Weng Jianshu, Lim Ee-Peng, Jiang Jing, He Qi. Twitterrank: Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 261-270

[6] Pal A, Counts S. Identifying topical authorities in microblogs //Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 45-54

[7] CNNIC. The 30th statistical report on Internet development in China, 2013(in Chinese)
(中国互联网络信息中心. 第 31 次中国互联网络发展状况统计报告. http://www.cnnic.cn/hlwfzyj/hlwxzbg/hlwtjbg/201301/P020130%20122600399530412.pdf, 2013)

[8] Kwak H, Lee C, Park H, Moon S. What is Twitter, a social network or a news media? //Proceedings of the 19th International Conference on World Wide Web. Raleig, USA, 2010: 591-600

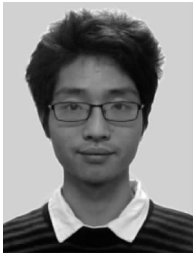
[9] Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 137-146

[10] Goyal A, Bonchi F, Lakshmanan L V S. Learning influence probabilities in social networks//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 241-250

[11] Cha M, Haddadi H, Benevenuto F, Gummadi K P. Measuring user influence in Twitter: The million follower fallacy. International AAAI Conference on Weblogs and Social Media (ICWSM), 2010, 10: 10-17

[12] Bakshy E, Hofman J M, Mason W A, Watts D J. Everyone's an influencer: Quantifying influence on Twitter//Proceedings of the 4th ACM International Conference on Web Search and Data Mining. Hong Kong, China, 2011: 65-74

[13] Myers J L, Well A D. Research Design and Statistical Analysis. 2nd Edition. New Jersey, USA: Lawrence Erlbaum Associates Publishers, 2003



MAO Jia-Xin, born in 1991, Ph. D. candidate. His research interest is social network analysis.

LIU Yi-Qun, born in 1981, Ph. D., associate professor. His research interest is information retrieval.

ZHANG Min, born in 1977, Ph. D., associate professor. Her research interests include machine learning and information retrieval.

MA Shao-Ping, born in 1961, Ph. D., professor, Ph. D. supervisor. His research interests include knowledge engineering, information retrieval, Chinese character recognition and post processing, and digitization of Chinese ancient books.

Background

Social influence occurs when one's emotions, opinions or behaviors are affected by others. The diffusing mechanism and effects of social influence have been studied in sociology for a long time. From the middle of last century, a series of theories, which aim to analyze and describe the social influence phenomenon in social networks, were proposed by sociologists. Their work served as a theoretical foundation of this research area.

However, the social relations and the social influence are usually hard to observe in real world, and the forms of social influence are varied and complex. So the empirical study of the social influence phenomenon had been limited. And this situation was changed recently, as the emergence and fast-growing of the online social networks. For providing a large number of active users, a fairly unified interface for users to communicate, and a complete log of users' actions and relations, the online social websites become ideal laboratories for the study of social influence. At the same time, a better understanding of this phenomenon and a calibration of the measurement of users' social influence will improve the applications like user recommendation, finding expert users, spam detections, viral marketing and content ranking. So the analyzing and measuring social influence, via mining the data produced by online social sites, become a research hotspot in

these days.

Compared to other online social networks, micro-blogging service has two major differences, one-way following relations and an extremely unified and simplified user interface that a user usually can only post or re-post a 140-character "micro-blog", beside reading the posts from others. These two traits make micro-blogging sites more similar to a socialized media, but not to a standard social networks. So the authors tried to use the capabilities of diffusing information, not the importance or status in network structures, as a more accurate indicator of users' social influence. This idea differs our work from most of the existing research on this topic.

Following it, the authors collected a large scale dataset from weibo.com. After carefully analyzing the data, we incorporated some user behavioral factors, including the access time of the website, the user preference of browsing and re-posting, with the social network structures to build a better model that can predict the average re-posted times of a single post of a specified user. And finally, they verified the effectiveness of our model by comparing it with other methods.

Their work shows the value of user behavioral factors for measuring social influence in micro-blogging environment and sheds light on future works, in which they may explore a richer behavioral log of users or take the content of post into account.