文章编号：2095-2163(2019)02-0162-07 中图分类号：TP305 文献标志码：A

社交网络中用户行为及影响力评估算法研究

魏杰明，何 慧
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：伴随着网络技术应运而生的社交网络平台打破了固有的信息传播方式，成为日常生活中重要的信息渠道，信息传播的方式发生了翻天覆地的变化。在社交网络平台中，为用户提供了多种交互功能，用户可以进行多种行为。发布行为产生数据信息，这些基本数据信息是社交网络信息的重要组成部分。用户的转发行为促使该基本数据信息在社交网络中有效传播。对原创贴文的点赞和评论增添了原贴文的信息承载量，可以有效地增强原贴文的影响力。本文从用户行为方式和互动规律的角度出发，系统研究了社交网络中用户行为和贴文特征。基于 PCA 主成分分析算法，将各组成因素进行相关性研究，得到社交网络节点影响力函数表达式。

关键词：社交网络；用户行为；影响力算法

Research on user behavior and influence algorithm in social network

WEI Jieming, HE Hui
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 Along with the network technology arises at the historic moment of social networking platform to break the inherent way of information dissemination, which has become the important channels of information in daily life, great changes have taken place in information dissemination way. In the social network platform, it provides a variety of interactive functions for users, and users can conduct a variety of behaviors. Publishing behavior generates data information, which is an important part of social network information. Users´ forwarding behavior makes this basic data information spread effectively in social networks. Thumb up and comments on original posts increase the information load of original posts, which can effectively enhance the influence of original posts. From the perspective of user behavior pattern and interaction rule, this paper systematically studies user behavior and post text characteristics in social network. Based on PCA principal component analysis algorithm, the correlation of each component factor is studied, and the expression of influence function of social network nodes is obtained.

【Key words】 social network; user behavior; influence algorithm

0 引 言

社交网络中的信息传播形式已经随着传播媒介的改变，发生了本质性的变化。在社交网络中，社会事件的传播过程存在着复杂的组成原因。社交网络中用户的行为模式是复杂而多样化的[1]。交互的行为使得信息在网络空间快速传播。用户行为规律分析，已成为当下最热门并且亟待解决的研究课题。

同时，社交网络中的影响力可以改变用户行为[2]，这种影响力可通过信息传播过程得以发挥作用，并因节点影响力强弱而呈现不同的作用效果。社交网络中影响力包含 2 种能力，即：信息传播能力和传播效能。有着高影响力的节点[3-4]可以快速地传播信息，并且依靠接收者进行多层次的迭代传播。同时，接收者获取到该信息后，会因信息传递的内容改变自己的用户行为，配合信息调整活动。

用户的行为方式大致可分为 2 类：发布行为和回复行为。其中，发布行为是用户进行信息传递的起点，其中涵盖了发布时间特征、贴文类型特征、活跃度特征以及发布内容特征等 4 个方面。而回复行为是用户接收信息并进一步传递的过程[5]，主要包括回复时间特征、回复内容特征、回复活跃度特征等 3 个方面。通过对以上 2 种行为的研究，可以分析出节点影响力的组成因素[6-7]。

研究可知，国际上影响力模型的建立，大都是基于 3 个关键属性：追随者数量、话题讨论度和转发传播效能。通过比对各标定对象的属性值，继而进行计算和比较。由用户的行为参数度，判断出该节点的影响力强弱。

针对社交网络中影响力模型提出了系统性的结

基金项目：国家自然科学基金(61472108)；国家重点研发计划(2017YB0801801,2017YFB0803300)。

作者简介：魏杰明(1994-)，男，硕士研究生，主要研究方向：社交网络信息传播；何 慧(1974-)，女，博士，教授，博士生导师，主要研究方向：社交网络信息传播、移动网络安全。

收稿日期：2018-06-08

论。将各种因素汇总，归纳总结出了 4 类影响力模型[8-9]，分别是：基于 PageRank 算法的影响力计算模型、基于用户行为的影响力计算模型、综合前两种算法的影响力计算模型、以及基于社交网络贴文地址的影响力模型。这 4 类影响力模型，涵盖了当今学术界对于社交网络中该课题的所有计算方案，都在应用层面取得了非常好的实验效果[10]。

本文的研究重点在于社交网络中用户行为与节点影响力分析，在此基础上实现对影响力模型的建立。通过从爬取的社交网络数据集中进行特征提取，使用统计分析的方法，进而对用户行为和贴文特征开展深入研究。从 2 类节点的特征中，归纳总结出节点的影响力组成规律，最终形成针对社交网络中节点影响力模型。在定量计算中，需要确定节点的影响力权值。通过对节点影响力的全面分析，确定其影响力组成因素。使用 PCA 主成分分析算法，求得节点影响力函数表达式。

1　社交网络的数据采集

本文以 Facebook 为研究对象，采用基于爬虫技术的社交网络数据自动获取程序。对待分析样本，进行数据采集，通过设定不同情况下，不同数据集下的数据清洗、数据融合以及数据归一化处理的方式，将数据按照一定的规则进行组织，并最终通过使用归一化处理后的数据构建社交网络信息数据集。对此研究，可做阐释论述如下。

1.1　数据爬虫的设计与实现

在数据采集上，采用编写抓取程序和调用 Facebook 提供的 API 以及开放数据等方式相结合，尽量使数据足够完整和具有代表性。

Facebook 面向外部开发者推出了一套比较成熟的 API 接口，开发者通过实名认证可以获得相应的开发者口令 Access token。

首先，建立一个 token 池。该 token 池保存从 Facebook 上申请得到的开发者 Access token。其次，建立一个目标用户池。该目标池保存要爬取的目标用户在 Facebook 上的账号 ID。调取 API 需要用到一个请求信息，该请求信息由接口地址和请求数据段组成。根据所需要爬取的数据组合出目标请求，目标请求和 token 参数组合成一个爬虫任务存放到请求信息池中。考虑到要爬取的目标用户很多，因此需要使用多线程来提高爬取速率。从目标用户池中取出一个用户账户 ID 和请求信息池中的一条请求信息组成一个爬取线程，调用对应的 Post API 接口，即可获取贴文信息。

1.2　数据预处理

从 Facebook 上爬取的数据样本集中含有大量的噪声信息。例如贴文文本中的表情符号属于社交平台的自定义符号干扰数据库的保存，需要进行过滤；伊朗文、拉丁文等非英文的语种需要单独处理，否则数据库会产生异常。因采集的数据来源于全世界各地，时区的不同将直接影响对于时间类型行为的研究分析。这些噪声信息会对开发中的后续研究工作造成麻烦。因此，需要将爬取到的数据进行预处理，清除掉数据中的噪声。

1.3　数据的持久化

对采集到的数据进行数据持久化。建立 2 张数据表，详见表 1、表 2，可用于分别存储贴文数据以及贴文的评论数据。

表 1　贴文数据

Tab. 1　Posts data

字段名	数据类型	字段描述
id	int	数据序号
uid	varchar	用户的唯一标识
nickname	varchar	用户的名字
fid	varchar	贴文的唯一标识
content	text	贴文的内容
type	varchar	贴文的类型
posttime	datetime	贴文的发布时间
repostnum	int	贴文的转发数量
likenum	int	贴文的点赞数量
commentnum	int	贴文的评论数量

表 2　评论数据

Tab. 2　Comments data

字段名	数据类型	字段描述
id	int	数据序号
uid	varchar	用户的唯一标识
nickname	varchar	用户的名字
fid	varchar	贴文的唯一标识
content	text	贴文的内容
ffid	varchar	评论贴文的对应父贴文的 fid
posttime	datetime	贴文的发布时间
likenum	int	贴文的点赞数量
commentnum	int	贴文的评论数量

将获得的数据分别存储在 2 张数据表中，本文的数据集中共获取贴文数据三百多万，评论数据五百多万。数据足够完整和具有代表性，符合实验的要求。

2 用户行为分析

社交网络中用户的行为是复杂而多样的,包括了建立好友关系、发布原创的贴文信息、评论其它人的贴文内容、使用其它的应用等等。这里将社交网络的用户行为分为 3 类,具体如图 1 所示。由图 1 分析可知,社交网络中的用户行为在交互过程中将产生巨大的信息量。为了有针对性地分析用户行为规律,需要系统、全面地研究社交网络中的行为细节,从中挑选有关键特征的信息。而在信息的传播过程中,重要的用户行为可归纳为典型的 5 种,内容解析参见表 3。

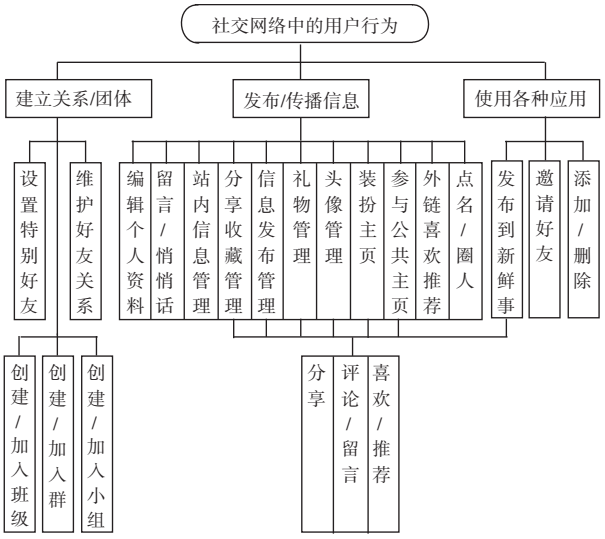

图 1 社交网络中用户行为

Fig. 1 User behavior in social networks

表 3 用户行为

Tab. 3 User behavior

用户行为	行为描述
发布	用户发布一条信息的行为,包括文字、图片、视频等
转发	当用户在浏览信息时,对自己感兴趣的内容进行分享转发的行为
评论	用户对相关信息有自己的见解,对他人发布内容进行评论的行为
点赞	用户对其它用户所发布内容表示赞同时,所做出的点赞行为
关注	用户对该用户发布的内容感兴趣,所做的关注跟随行为

社交网络中用户的内容创作行为和用户之间的各种交互行为构成了社交网络的主体,其中展现的各种规律将会客观反映社交网络的主要特征。基于此,拟展开如下研究论述。

2.1 用户发布贴文的时间特征

用户发布贴文的时间点可以反映用户的日常行为习惯,不同的人会选择在不同的时间点来发布信息。因此,发布贴文的时间特征是一个重要的用户行为特征。这里即以日分布时间为周期进行数据统计,并由此来探寻时间特征规律。研究得到的日分布特征折线结果如图 2 所示。

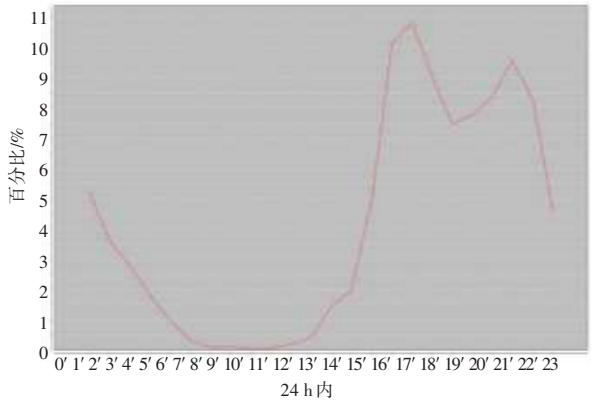

图 2 日分布特征统计图

Fig. 2 Statistical chart of daily distribution characteristics

由图 2 分析可知,凌晨 5 点到早上 9 点的时间段内,用户发布贴文的比例最低,接近于零点。从早上九点开始,用户发贴行为开始逐渐增加。在下午三点到达用户发帖量第一个高峰,仅在三点到四点这一个小时内,发布的贴文数量就超过了全天发布数量的 10%。随后用户发帖量从下午三点的峰值开始下降,在下午六点附近逐渐反弹,形成一个拐点。一直到晚上九点,用户发帖量到达第二个高峰。随着时间的不停前行,发帖量开始逐渐下降,直至凌晨五点时基本接近零点。

在这一过程中,可以发现 2 个关键信息。对此表述如下。

(1)半夜和凌晨是发贴行为比例最低的时候。考虑到此时大部分用户都处于休息状态,没有时间和精力发布贴文。

(2)午后和傍晚是发贴行为比例最高的时候。考虑到午后和晚饭后的时光,用户普遍比较闲适,有充足的时间浏览和发布贴文。

该分析结果与用户的日常行为相吻合。因此,在午后和傍晚发布贴文有较大概率会获得良好的信息传播效果。

2.2 用户发布贴文的类型特征

社交网络平台给用户提供了 4 种待选择的贴文类型:链接类、状态类、图片类、视频类。每种贴文类型传递信息的方式各不相同,产生的效果也不同。

因此，用户选择贴文类型的行为是一个重要的用户行为特征。下面即从贴文类型的分布进行数据统计，并由此来探寻用户发布的贴文类型的特征规律。研究得到的贴文类型分布统计结果如图 3 所示。

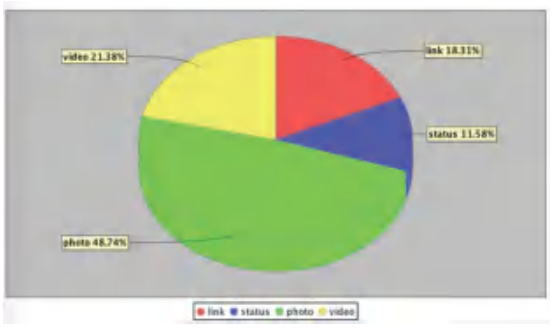

图 3　贴文类型分布统计图

Fig. 3　Post text type distribution statistics

对图 3 讨论分析后，可以发现 2 个关键信息。对此可描述如下。

(1)图片类贴文的占比高，说明了现实生活中，用户更倾向于图文并茂地表达自己的想法，这样更容易让其它用户理解自己的表达。

(2)其它贴文类型发布信息的效率不如图片类效率高。发布视频信息编辑加工的时间成本高、专业性强、操作难度大。链接类和文本类的发布，需要编写大量的文字信息，不如发布图片快捷，需要的思考时间更长。

2.3　用户发布贴文的内容量特征

用户发布贴文的内容长度可以反映用户的日常行为习惯，不同的人发布信息的内容长度是不同的。因此，发布贴文的内容长度特征是一个重要的用户行为特征。下面主要对用户发布信息内容的外部特征进行统计计算，并由此探寻发布贴文的内容长度特征规律。研究得到的贴文内容量分布统计结果如图 4 所示。

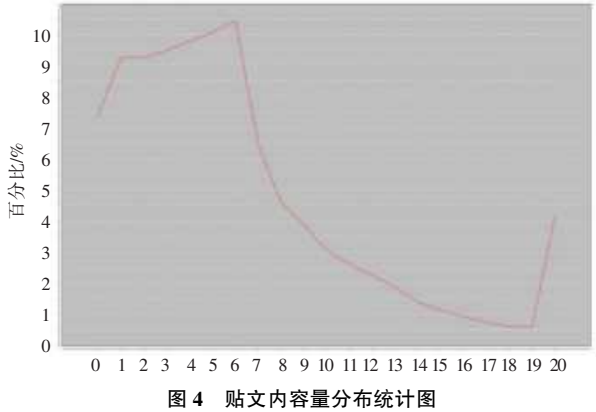

图 4　贴文内容量分布统计图

Fig. 4　The capacity distribution statistics in the post text

对图 4 讨论分析后可以发现，可以发现 3 个关键信息。对此可表述如下。

(1)中少字数的贴文占有比例高。当今的社交网络环境中，简洁明了的信息表达特征已更趋明显。

(2)长篇内容的贴文部分仍然占有很重要的比例。专业化程度高，信息量充足的贴文数量开始增加。

(3)信息内容长度的两极分化程度加剧，信息传递的目的性、针对性日渐增强。

2.4　贴文发布时间与互动量的关系分析

贴文发布后，开始通过社交网络进行信息传递，传递过程中产生的互动数据反映了贴文本身的特性。这种贴文特性由发布时间、贴文类型、发布者等因素共同决定。下面详细地计算出贴文中各项特征的统计结果，由此探寻出其中的规律。研究得到的发布时间与互动量统计曲线如图 5 所示。

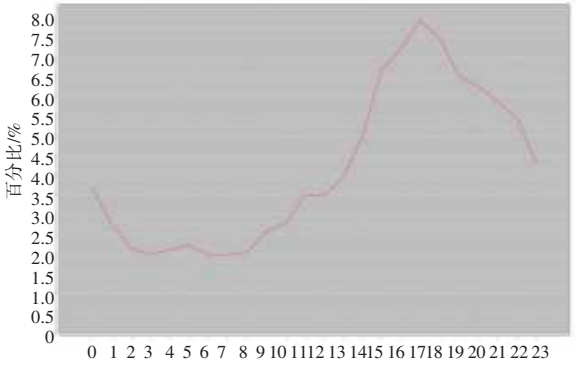

图 5　发布时间与互动量统计图

Fig. 5　Release time and interaction statistics

由图 5 分析可知，从整体趋势上看，从早上七点到下午五点，互动量均呈上升趋势。在下午五点到下午六点的时间段内，发布的贴文获得了最高的用户互动量。在这一小时内发布的贴文获得的互动量占到全部样本的接近 8%。在此之后的互动量比例逐渐下降，到凌晨 3 点附近到达谷底。

在这一过程中，可以发现 2 个关键信息。对此可表述如下。

(1)下午五点附近发布的贴文最容易获得其它用户的关注，并产生互动行为。这个时间点发布的贴文，正值其它用户的使用高峰期，被浏览的概率最高。因而会产生最佳的信息传播效果。

(2)在半夜或凌晨发布的贴文信息传递效果不好，很难与其它用户产生交互。这个时期正是人们休息的时间，是整个社交网络的使用低谷期，活跃度低。而且信息还具有时效性，经过数个小时后，其它发布的新贴文会将该信息覆盖掉，故而更难展示在

其它用户面前。因而造成了信息传播效果不佳,互动量低的现象。

2.5 贴文类型与互动量的关系分析

在社交网络平台中,每种贴文类型传递信息的方式各不相同,产生的效果也不同,有着明显的热度区分。贴文选择以不同的类型进行发布,信息的传递速度、范围以及传递效果也将表现出较大的差别。这些差别可以从贴文产生的互动数据中获得,下面将详细地计算出贴文类型与互动量的统计结果,由此来探寻其中的规律。研究得到的贴文类型与互动量统计结果如图 6 所示。

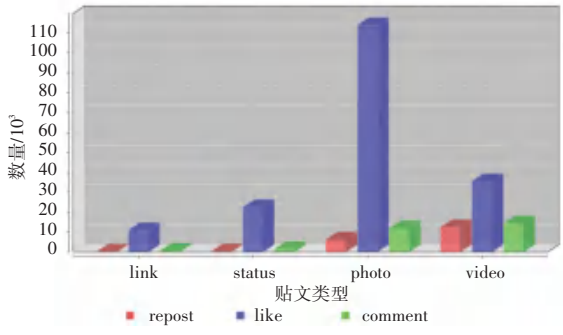

图 6 贴文类型与互动量统计图

Fig. 6 Post type and interaction statistics

对图 6 分析讨论后可以发现,图片类贴文获得了最多的关注,互动量遥遥领先于其它 3 类贴文。图片相较于文字更加直观、感染力更加强大,更能调动用户参与讨论的热情。图片类在获得用户互动行为上都更加明显地优于其它类别。一定程度上,也说明了用户行为的兴趣点中,更倾向于直观的体验,而不是枯燥无味文字的堆积和叠加。纯文字的贴文更容易让用户失去兴趣,进而不参与直接的互动。而视频类的耗时较长,需要一定的观看时间才能让用户真正理解其意图,虽然不及简单直观的图片传播,但是效果也很好,互动的积极性也很高。纯文本的占比很少,这在一定程度上说明了纯文本类贴文在现实生活中的吸引力已经趋于弱化。

3 影响力评估算法

3.1 用户影响力的组成因素

从用户行为分析中,用户发布原创贴文的频率越高,该用户越容易获得高关注度。拥有高粉丝数量的用户,发布的贴文可以赢得更多的话题度和讨论度。发布贴文的转发量增加,会正向带动贴文的阅读量和评论量。结合以上统计特征,去对比社交网络中节点影响力的 2 个关键因素。可以得出以下的分析结果:一个用户节点的信息传播能力主要由其粉丝数、关注数、发帖量、发贴频率、转发量五种因素组成;而该用户节点的传播效果主要由点赞量和评论量这 2 种因素组成。

因此,本文对用户节点的影响力分析,将着重从 6 个方面做出研究,可将其表述为:粉丝量,该用户的粉丝数;发帖量,该用户发布贴文的数量;活跃度,该用户发布原创贴文的频率;点赞量,用户贴文的点赞数量;评论量,用户贴文的评论数量;转发量,贴文节点的转发数量。

3.2 PCA 主成分分析算法

在众多研究中,为了分析目标对象的特征规律,经常需要对此对象进行数据建模。在建模过程中,为了描述一类对象,要使用多种变量从各个角度,对其予以解释说明。在这种情况下,PCA 主成分分析法很好地解决了变量多维度的难题。在对各个指标进行全面分析的同时,将多维度空间进行降维。既保证信息研究的准确性,又降低分析算法的复杂度。

PCA 算法是一种重要的机器学习算法,属于无监督学习。PCA 算法的基本思想是将研究对象的多维特征在尽可能不丢失信息的情况下,在经降维操作后,转换为一组新向量。这组新向量是正交的,也就是说在原数据的基础上做正交变换,生成在低维度空间上的正交映射。目的是将原始基转换为互不相关的新基,并以新基来替代原始基,从而简化复杂问题的一种分析方法。

3.3 影响力模型的建立

构建初始数据矩阵,将样本数据标准化。首先,根据用户节点的影响力组成因素来设计构建数据样本集。本次研究中部分原始数据如图 7 所示。

图 7 部分原始数据

Fig. 7 Partial raw data

其次,将数据样本全部转换为用户节点影响力向量进行保存。同时,建立初始的数据矩阵 $\boldsymbol{M}$,将 Facebook 贴文数据传入 SPSS 函数中,通过 SPSS 函数模块进行数据初始化。分别计算各向量指标的均值和方差,将采集到的社交网络中的原始数据在经过数据标准化的加工环节后,随即转入指标之间的

相关性判定研究，此时需建立系数矩阵即如图 8 所示。

相关矩阵

		粉丝量	发帖量	活跃度	转发量	点赞量	评论量
相关	粉丝量	1.000	0.981	0.988	0.969	0.784	0.738
	发帖量	0.981	1.000	0.961	0.963	0.773	0.729
	活跃度	0.988	0.961	1.000	0.953	0.755	0.697
	转发量	0.969	0.963	0.953	1.000	0.835	0.783
	点赞量	0.784	0.773	0.755	0.835	1.000	0.957
	评论量	0.738	0.729	0.697	0.783	0.957	1.000

图 8　相关性系数矩阵

Fig. 8　Correlation coefficient matrix

根据各个指标的相关性值，分析出各个指标之间的关系。从相关性系数矩阵中，最终可分析探得规律如下：活跃量与粉丝量有很强的相关性，从这个数据可以反映出活跃度高的用户会更容易吸引粉丝关注。发帖量与活跃度有很强的相关性，从这个数据可以反映出发帖量大的用户，发贴频率也很高。粉丝量与转发量有很高的相关性，从这个数据可以反映出，粉丝量高的用户发布的贴文获得的互动量也越高。

经过主成分分析后，从中取出累计贡献率超过 80%的特征值，组成主成分分量。研究中根据主成分的贡献率和累积贡献率，统计出总方差的统计结果如图 9 所示。

解释的总方差

成分	初始特征值			提取平方和载入		
	合计	方差的%	累积%	合计	方差的%	累积%
1	5.299	88.312	88.312	5.299	88.312	88.312
2	0.580	9.665	97.977			
3	0.046	0.761	98.738			
4	0.042	0.703	99.442			
5	0.027	0.456	99.898			
6	0.006	0.102	100.000			

图 9　总方差的解释

Fig. 9　The interpretation of the total variance

由图 9 中可以看到，第一个子成分的累积贡献率已经超过 80%，已经符合 PCA 算法对于主成分计算的要求。因此，可以将用户影响力组成因素合成为一个主成分，并以此表示用户节点的影响力。

主成分分量由一个子分量构成，该子分量由初始数据中 6 个指标的系数组成。各个系数如图 10 所示。至此，研究根据影响力组成因素求得了社交网络中用户影响力评估函数表达式。建立了社交网络中用户影响力评估的重要模型。

成分得分系数矩阵

	成份
	1
粉丝量	0.184
发帖量	0.182
活跃度	0.180
转发量	0.185
点赞量	0.170
评论量	0.163

图 10　成分分量系数组成图

Fig. 10　Composition diagram of component coefficients

4　结束语

在本次研究中，系统地分析了社交网络中用户的行为模式以及表现出的行为规律。提出了一种针对社交网络中用户影响力评估算法。通过从社交网络数据集中有效地提取关键信息，而且从时间特征、贴文类型、内容量等方面，充分探讨、并研究了数据特征。根据用户行为数据，通过使用统计学方法进行规律总结，深入研究影响力的组成因素。同时，又根据数据的分布情况，进一步推导出了各个组成因素的权值。

采用 PCA 主成分分析算法，既保证信息研究的准确性，又降低分析算法的复杂度。结合 6 种用户影响力组成因素，生成一套影响力函数表达式，可以定量计算出用户的影响力大小。建立了社交网络中用户影响力评估的重要模型。

在未来的工作中，研究将考虑把影响力评估算法应用到其它方面，比如信息传播、广告营销、舆情分析等等。将用户的影响力作用发挥到最大程度。

参考文献

[1] MISNER I R. The world´s best known marketing secret: Building your business with word - of - mouth marketing [M]. 4th ed. Austin, TX: Bard Press, 2003.

[2] KIM J, KIM S K, YU H. Scalabe and parallelizable processing of influence maximization for large - scale social networks? [C]// Proceedings of the 29th International Conference on Data Engineering (ICDE). Brisbane, Australia: dblp, 2013: 266-277.

[3] CHENG Suqi, SHEN Huawei, HUANG Junming, et al. IM Rank: Influence maximization via finding self-consistent ranking [C]// Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Gold Coast, Australia: ACM, 2014: 475-484.

[4] HE Jianming, CHU W W. A social network-based recommender system (SNRS) [M]//MEMON N, XU J, HICKS D, et al. Data Mining for Social Network Data. Annals of Information Systems, Boston, MA: Springer, 2010, 12: 47-74.

（下转第 171 页）