中国科学: 信息科学 2017 年 第 47 卷 第 8 期: 967–979
SCIENTIA SINICA Informationis
从大数据到大知识工程专刊 · 评述

社会影响力分析综述

张静[1], 唐杰[2*]

1. 人民大学信息学院计算机系, 北京 100872
2. 清华大学计算机系, 北京 100084
* 通信作者. E-mail: jietang@tsinghua.edu.cn

收稿日期: 2017–06–21; 接受日期: 2017–08–11; 网络出版日期: 2017–08–23

摘要 社会影响力是个人由于社会地位、社会联系以及社会财富等因素, 改变他人思想或行为的能力. 研究社会影响力, 特别是在大规模社会网络数据上对人与人之间的影响力进行建模与度量, 对于社交应用中的商品推广、好友推荐、专家发现以及用户行为预测等都具有非常重要的意义, 对于虚拟商业市场以及国家舆情监控等都具有重要的促进作用. 从计算学的观点来看, 社会影响力的研究包括: 社会影响力检测、建模、度量以及应用等方面. 本文主要从这几个方面介绍国内外的最新研究成果, 并展望未来可能的研究方向.

关键词 社会影响力, 影响力传播模型, 影响力检测, 影响力度量, 社会网络

1 引言

在线社交网络飞速发展, 物理社交世界和虚拟网络加速融合, 国际上最大的社交网站脸书 (Facebook) 的月活用户数已经超过 19 亿 (截止 2017 年 3 月), 稳稳占据世界第一人口 “大国”; 推特 (Twitter) 的活跃用户数也超过 3 亿; 国内方面, 近年崛起的微信已有 9 亿多活跃用户, 新浪最新公布的数据表明新浪微博的注册用户数也已经超过 5 亿. 人们在社交网络上停留的时间已经超过了人们使用传统搜索引擎的时间. 无处不在的社交网络不仅产生了海量的网络大数据, 还为科学研究和互联网应用开辟了一个全新的研究领域. 尤其值得关注的是和传统数据不同的事, 社交网络存在大量人和人之间的交互信息, 这些交互使得我们能够研究人与人之间的相互影响, 以及信息在人与人之间的传播模式.

社会影响是指人的行为、观点、情感受到朋友或者群体的影响从而发生改变. 社会影响在社会科学等领域已有一定的研究 [1~3]. 在计算机领域, 随着在线社会网络平台的蓬勃发展, 社会网络环境下的社会影响力研究于十多年前开始兴起. 其中主要研究内容包括: 社会影响力的检测、建模与度量问题, 以及所有这些研究成果的实际应用. 影响力检测的目标是确定某种用户行为之间是否存在互相影响的因果关系. 只有切实存在这种因果关系, 才能够假设影响力传播模型, 用以刻画用户行为在影响力作用下进行传播的现象. 而影响力度量则是为了求解用户之间影响力的程度大小, 即求解影响力传播模型中的参数, 使得影响力传播模型能够切实发挥预测的作用. 图 1 给出了社会影响力的研究框架.

引用格式: 张静, 唐杰. 社会影响力分析综述. 中国科学: 信息科学, 2017, 47: 967–979, doi: 10.1360/N112017-00137
Zhang J, Tang J. Survey of social influence analysis and modeling (in Chinese). Sci Sin Inform, 2017, 47: 967–979, doi: 10.1360/N112017-00137

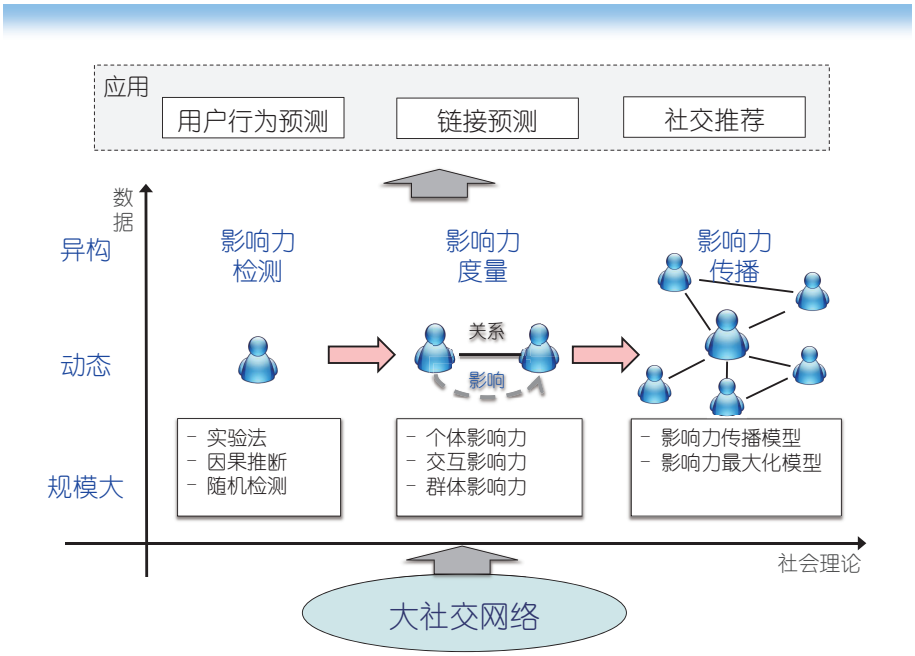


图 1　(网络版彩图) 社会影响力研究框架

Figure 1　(Color online) Architecture of social influence investigation

2　社会影响力检测

社会影响力检测的目标是给定某种行为, 例如发帖或购物等行为, 验证用户与用户相同的行为 (例如, 转发同样的帖子或购买同样的商品等) 之间是否存在互相影响的因果关系. 直观地, 如果两个用户在较短的时间间隔内发生了相同的行为, 则认为先发生行为的用户对后发生行为的用户具有一定的影响作用. 然而事实上, 也有可能是因为用户同时对同一事物 (例如帖子或商品) 产生兴趣而在相近时间内发生了相同的行为, 即由同质性 (homophily) 造成. 如何真正检测到影响力的驱动原因具有一定的挑战性. 这方面的研究一般分为两种模式. 一种是以真实实验的手段来验证影响力的存在性, 另一种是基于收集数据进行统计因果关系推断的方法来验证影响力的存在性.

2.1　实验法

传统的在检测社会影响力的真实实验中, 一般通过招募志愿者进行实验 [4, 5]. 随着社交网络平台的兴起, 可以将社交网络平台当做真实的实验环境. 例如基于 Facebook 平台, Bakshy 等 [6] 设计随机控制实验来检测社会影响力在消费者对推送广告的反馈上所起的作用. 他们采用广告点击行为作为研究目标, 在一些广告下面显示点击过该广告的朋友信息, 发现这些提示能够显著增加用户的广告点击率. 此外, 显示强社会关系的广告点击行为对于促进广告点击率具有更加显著的提升效果. 同样基于 Facebook 平台, Bond 等 [7] 采用随机控制实验来检测社会影响力对于政治投票行为的影响. 实验发现将朋友的投票结果展示给用户, 可以直接或间接地带来额外 34 万投票, 即朋友的投票行为会显著影响个人的投票行为. 在此基础上, 强社会关系的投票行为能够进一步带来 0.224% 的提升. 然而遗憾的是, 此类实验对于一般研究者是不可行的, 只能通过实际操控在线平台的公司来做这类实验.

2.2 因果推断法

与真实实验法相比, 基于收集数据进行统计因果关系推断更加可行. 例如, Anagnostopoulos 等 [8] 提出一种随机洗牌测试 (shuffle test) 的检测方法. 洗牌测试的假设是, 如果影响力不起作用, 那么一个节点的激活时间应该与邻居节点的激活时间独立. 在此假设前提下, 对节点的激活时间进行洗牌, 如果数据分布特征没有发生变化, 则假设成立, 否则推翻假设前提. Anagnostopoulos 等人在检测时, 假设社会网络是静态的, Fond 等 [9] 将情况扩展到动态的社会网络, 并在此基础上, 提出一个自回归模型来刻画影响与选择 (selection) 的交互过程. 所谓选择, 指的是用户由于同质性做出建立相互关系的选择. 模型的假设是, t–1 时刻具有相似属性或相似行为特征的用户会在 t 时刻建立相互关系 (选择过程), t–1 时刻建立关系的用户会在 t 时间趋于相似属性或相似行为特征 (影响过程). Fond 等人也采用随机洗牌测试的方法来验证这两种过程之间因果关系的显著性. Arala 等 [10] 采用癖好分数匹配法 (propensity score matching) 在排除同质性的提前下, 检测影响力的存在性. 其基本思想是, 在构造控制实验时, 让对照组的癖好分数与控制组尽可能相匹配, 做到最大限度地排除选择偏差带来的检测不准确问题, 即排除同质性的原因. 检测结果认为由同质性引起的信息传播受节点的网络结构分布特征所支配, 而由影响力引起的信息传播则是自我驱动的, 呈现出快速、指数级的增长特点. 然而, 由于此类方法是基于收集数据进行分析得出结论, 因此严格来讲, 只能从一定程度上反映影响力与用户行为的因果关系, 不可能排除所有其他因素的影响.

3 社会影响力建模

社会影响力建模的目标是构建社会网络中的影响力传播模型, 用于刻画网络中用户的行为状态影响相邻用户的行为状态, 并造成某一行为状态在网络中扩散传播的机制. 用户行为状态分为激活态与未激活态两种, 一般只关注激活态的传播机制. 例如对于某个帖子或商品, 只关注用户的转发或购买状态的传播机制. 研究者针对不同情境提出多种类型的影响力传播模型. 例如确定性与随机性模型、离散与连续模型、递进式与非递进式模型等, 以及基于传播事物的语义 (如帖子讨论的话题) 、节点的语义 (如用户的角色) 以及关系的语义 (如好友关系与敌对关系) 扩充的模型等.

3.1 影响力传播模型

早期的传染病模型属于确定性模型. SIR 模型 [11] 以及 SEIR 模型 [12] 是两个流行的传染病模型, 描述了疾病在人群中的传播趋势. 模型的基本思路是首先将人群分为给定类别: 易感染者 (susceptible)、感染病者 (infective) 以及恢复者 (recovered) 等, 然后依据目前情况下各类别的整体观察量, 估计类别之间的转移速率, 从而可以预测疾病未来的发展趋势. 信息传播模型大多属于随机模型, 一般是通过刻画人与人之间的影响力来描述整个信息的传播过程. Kempe 等 [13] 提出两个经典的信息传播模型, 线性阈值模型 (LT) 和独立级联模型 (IC).

在独立级联模型中, 网络中每条关系 e_{ij} 会对应一个概率值 $p \in [0,1]$, 表示节点 v_i 独立激活节点 v_j 的概率, 即个体间的影响力. 独立级联模型是刻画影响力在离散时间下的传播机制, 其具体的传播过程如下: 在 t_0 时刻, 一个预先选好的初始集合首先被激活, 将该初始集合称之为种子集合. 在之后的每个时刻 $t \geqslant 1$, t–1 时刻新激活的节点 v_i 会对其尚未激活的节点 v_j 以概率 p_{ij} 尝试激活一次, 且这次尝试与其他尝试事件相独立. 如果尝试成功, 则 v_j 转为激活态. 当某个时刻不再有新的节点被激活时, 传播过程结束. 独立级联模型的随机性由激活概率 p_{ij} 决定. 因为该模型过程是随机的, 最后产

生的激活节点集合也是个随机集合. 因此基于该模型, 一般关心的是给定初始种子节点集, 传播结束后被激活节点个数的期望值, 也叫作影响力的延展度 (influence spread). 独立级联模型以概率来描述个体之间相互影响的强弱, 并假设个体之间独立影响的行为特征. 这与消息或病毒等诸多实体的传播特点相符合 [14], 因此独立级联模型是目前研究最广泛的模型.

在线性阈值模型中, 网络中每条关系 e_{ij} 会对应一个权重 $w_{ij} \in [0,1]$, 表示节点 v_i 在节点 v_j 的所有邻居中重要性的比例, 且满足 $\sum_{v_i \in N(v_j)} w_{ij} \leqslant 1$. 此外, 每个节点 v_i 还关联一个被影响的阈值 $\theta_i \in [0,1]$. 该阈值在一次传播开始前, 在 0 到 1 的范围内随机等概率选取. 与独立级联模型类似, 在 t_0 时刻有且仅有种子集合中的节点被激活. 在之后的每个时刻 $t \geqslant 1$, 每个尚未激活的节点 v_i 根据其激活邻居节点的线性加权值是否超过该节点的被影响阈值来决定其是否被激活, 即判断是否满足 $\sum_{v_i \in N(v_j) \cap y_{jt}=1} \geqslant \theta_i$. 若满足, 则 v_i 转为激活态. 当某个时刻不再有新的节点被激活时, 传播过程结束. 线性阈值模型以概率 θ_i 来描述个体的被影响力, 并假设邻居对节点的激活行为不是独立的, 而是联合发生的. 这与人类做复杂决策 (例如购买商品、政治投票等) 时的从众行为 [15] 相符合. 线性阈值模型的随机性完全由个体的被影响阈值 θ_i 来决定. 然而实际中, 个体的被影响阈值虽有随机性, 却波动不大. 但如果限制其在小范围内变化, 会给模型的计算增加难度 [13]. 因此这也是线性阈值模型不及独立级联模型应用广泛的一个原因.

Kempe 等人 [13] 在独立级联模型与线性阈值模型的基础上, 提出泛化的级联模型与阈值模型, 使其应用场景更加灵活. 上述独立级联模型与线性阈值模型都是离散型模型, 也有研究提出连续型模型, 将节点之间影响的传播延迟用一个概率密度函数来表示 [16], 避免了对连续时间的离散化. 当考虑用户状态的变化情况时, 倘若用户只能一次性地从未激活态切换到激活态, 即状态切换不可逆, 则为递进式 (progressive) 模型, 例如独立级联模型与线性阈值模型都属于递进式模型. 反之如果状态之间可以随意切换, 则为非递进式 (non-progressive) 模型 [17,18]. 递进式模型一般描述信息与商品等的传播, 而非递进式模型一般刻画观点与情绪等的传播. 此外, 一些考虑不同因素或者语义性的信息传播模型也相继提出, 例如时间衰减的独立级联传播模型 [19,20], 话题敏感的独立级联传播模型和线性阈值模型 [21], 同时考虑正面和负面意见的传播模型 [22], 同时考虑朋友与敌人关系的传播模型 [23], 多个竞争实体并发传播的传播模型 [24], 区分意见领袖、结构洞和普通用户角色的传播模型 [25], 社会关系形成之间的影响传播模型 [26].

3.2 影响力最大化

研究者在影响力传播模型的基础上定义了影响力最大化 (influence maximization) 的研究问题, 目的是优化影响力的传播. 其中的一个重要应用场景是在虚拟市场中寻找影响力最大的初始客户, 对其提供免费试用商品的机会, 期望其使用后能够主动在朋友圈中对商品进行推广, 以影响其朋友、朋友的朋友接受并购买该商品 [1,3,27~30].

影响力最大化问题的形式化描述是, 在社会网络中寻找 k 个初始种子节点将其激活, 使其通过社会网络关系的影响传播最终所产生的影响力延展度最大. 之前该问题一直是市场决策与商业管理的研究范畴. 后来, Domingos 和 Richardson 采用 Markov 随机场模型将该问题形式化为一个排序问题, 并采用高效的算法求解该排序问题 [28]. 但该方法将市场决策过程刻画成一个 “黑盒子”, 并没有描述用户之间互相影响的机理. Kempe 等人 [13] 首次将影响力最大化问题形式化为一个离散优化问题. 他们从理论上证明了该优化问题是一个 NP 难优化问题, 并基于子模函数 (submodular function) 给出近似的贪心求解算法. 函数的子模性是指一个元素在集合基础上的增量效应随着集合的增大而递减, 这也反应了经济学中经常提及的边界效用递减现象. 贪心算法用到子模函数的性质进行计算. 具体地, 每次

迭代选择一个节点, 使得其在已激活集合的基础上, 能够触发的新增激活节点最多. 理论证明该贪心算法可以达到 (1–1/e) 的近似程度. 然而上述贪心算法中的一个核心计算问题, 即求解一个种子集合的延展度, 又是一个 NP 难问题. Kempe 等人 [13] 在论文中提出用 Monte-Carlo 方法来随机模拟影响力的传播过程, 从而估计出延展度的大小. 然而该方法的计算效率很低, 因此之后很多研究者提出加速的影响力最大化求解算法. 一种是改进的 Monte-Carlo 贪心算法, 例如 Leskovec 等人 [31] 基于子模性质利用偷懒估计法来减少对延展度估计的次数. 另一种是启发式算法, 例如 Chen 等人 [32] 基于网络中心度的指标, 对激活邻居对应的中心度打一定的折扣, 提出简单高效的启发式方法来求解影响力最大化问题人. 也有工作研究基于扩展的传播模型如何加速影响力最大化的计算 [21~23]. 给定传播模型, 影响力最大化应用的假设前提是网络中用户之间的影响力大小是已知的.

4 社会影响力度量

上一节介绍了社会影响力建模. 然而, 要让影响力传播模型在真实数据中发挥更大的作用, 就必须基于真实数据对节点之间的影响力进行分析与度量. 社会影响力度量一般基于两种类型的数据: 一是给定社会网络结构数据, 由用户之间的社会关系可以推测出影响力模式及其强弱; 二是给定用户行为数据 (网络结构数据可有可无) , 由用户行为发生的时间先后次序推测出其影响力模式及其强弱. 社会影响力度量的目标是从真实数据中挖掘出用户之间的影响模式并估计其对应的影响力强弱. 最简单的影响模式是个体的影响力与被影响力, 即针对一个用户, 度量其影响他人的能力与被他人影响力的倾向; 略复杂一些的模式是个体之间的影响力, 即针对两个用户, 度量一个对另外一个的影响力; 更加复杂的模式是群体影响力, 即针对多个用户组成的群体, 度量群体对一个用户的影响力.

4.1 个体的影响力与被影响力

个体本身有影响他人的能力, 也有被他人影响的倾向. 其度量方法一般采用基于图的迭代算法. 其中图可以是原始的社会网络结构图, 也可以是基于用户交互行为构建的交互网络图. Weng 等人 [33] 构造话题敏感的 Twitter 关系网络, 并提出改进的页面排名算法 (PageRank) 来计算用户在每个话题上对他人的影响力. Heidemann 等人 [34] 首先根据用户的行为记录构造行为网络, 在此基础上采用类似页面排名的算法计算每个用户的中心度来表征影响力. 社会学上将个体被影响力称之为从众度 [15, 35~37]. Tang 等人 [38] 基于用户行为数据量化个体被他人影响的程度, 并提出因子图模型来估计个体被影响力在行为预测中所起的作用. Li 等人 [39, 40] 为每个用户赋予一个影响他人的分数与被他人影响的分数, 并提出同时考虑影响力与被影响力的图迭代算法对其进行度量.

4.2 个体之间的影响力

个体之间的影响力是影响力传播模型的重要参数, 例如独立级联模型中的激活概率 p_{uv} 表征了个体之间的影响力大小. 在对个体之间的影响力进行度量时, 值得注意的是, 个体之间的影响力与个体之间的同质性 (或相似性) 密切相关. 它们之间的关系是: 两个个体之间的相似性越强, 则它们越容易做出建立相互之间关系的选择 (selection); 反之, 有关系的两个个体, 互相之间的影响力会导致他们的相似性越来越强. 这两者永远是交织在一起, 不太好区分开来. Holme 和 Newman [41] 提出一个生成模型来平衡选择与影响过程. 基本思想是每次迭代时要么为两个节点之间建立关系 (选择过程), 要么将一个节点的属性变换为与其某个邻居节点相同的属性 (影响过程). Crandall 等人 [42] 提出一个更全面

的生成模型来刻画一个人的行为. 其基本思想是一个人的行为决策既取决于其本身的历史行为分布, 又取决于邻居的行为分布, 还取决于大众的行为分布. Scripps 等人 [43] 明确给出了相似性与影响力的数学定义. 由于相似性与影响力紧密交织在一起, 区分的难度很大, 因此除了以检测影响力为根本目标的研究, 一般对这两者不做特别区分.

当同时给定社会网络结构与用户行为数据时, 个体之间影响力估计的基本思想是, 两个历史上经常在较短的时间间隔内发生相同行为的用户与那些很少在较短时间内发生相同行为的用户相比, 互相之间的影响力更强. 个体之间影响力度量的主要方法包括频度统计法与最大似然法. 最大似然法基于一个给定的影响力传播模型得到一次传播结果的似然度, 然后通过最大化似然度来求解影响力传播模型中的参数 —— 即个体之间的影响力大小 [20, 44, 45]. 然而, 最大似然法一般很难得到一个精确解, 需要对似然函数做变换并采用近似迭代方法来求解, 其增加了计算的复杂度. 因此, 一些研究者直接对给定数据进行频度统计来度量个体之间的影响力. 例如, Goyal 等人 [46] 提出基于信用分布的频度统计法. 基本思想是对于一个激活节点, 认为其所有激活的邻居都有均等的贡献. 该方法虽然是启发式的, 但是与最大似然法相比, 计算效率非常高. Konstantin 等人 [47] 进一步将用户行为数据扩展到流数据情境下, 提出基于最小哈希值 (MinHash) 的近似算法来估计个体之间的影响力.

某些情况下仅能得到社会网络结构数据, 缺失了用户行为数据. 这种情况下, 个体之间影响力估计的基本思想是两个个体越相似, 则他们互相之间的影响力越强. 基于网络结构的相似度指标大致可以分为两种. 一种是近邻相似度, 其衡量的思想是两个节点之间的直接或间接邻居越多, 则这两个节点越相似. 这种衡量方法导致相似的节点在网络上的距离比较近. 早期的近邻相似度的度量指标, 包括在学术研究网络中的参考文献耦合度 (bibliographical coupling) [48] 和共同引用数 (co-citation) [49], 基本假设是两个节点的共同邻居越多, 则它们越相似. 这种方法不能估计那些没有共同邻居的节点之间的相似度. 之后提出的一些方法可以避免该问题. 例如, Katz [50] 认为两个节点之间可达的短路径越多, 则这两个节点越相似. Tsourakakis 等人 [51] 依据邻接矩阵为每个节点学习一个低维度的向量, 来表征节点的社区分布, 然后通过计算向量相似度来表征两个节点之间的相似度. Jeh 和 Widom [52] 提出了 SimRank 算法. 该算法是个递归的算法, 其基本思想是两个节点相似当且仅当它们的邻居节点相似. SimRank 只能计算路径长度为偶数的节点之间的相似度. VertexSim 扩展了 SimRank, 允许计算路径长度为奇数的节点之间的相似度 [53]. 因为 SimRank 的相关算法都是递归算法, 因此其算法复杂度比较高. 此外, 还有一些随机游走相关的图相似度计算方法 [54, 55].

还有一种是结构相似度, 其衡量的思想是两个节点的网络拓扑结构越接近, 则这两个节点越相似. 这种衡量方法导致相似的节点与其在网络上的距离无关, 可以很近, 也可以很远, 甚至不连通. Blondel 等人 [56] 提出基于 HITS 的迭代方法来度量两个不连通图上节点之间的相似度. RoleSim [57] 扩展了 SimRank 算法, 允许计算两个不连通节点之间的相似度. 此外, 不同于 SimRank 对任意两两邻居的相似度求和, RoleSim 仅对匹配上的邻居对之间的相似度求和. 与 SimRank 类似, 这两种方法的时间复杂度都比较高. 另外有一类方法给每个节点构造一个特征向量, 用向量之间的相似度当做是节点之间的相似度. 例如, Burt [58] 用一个节点局部中心网络中 36 种三角形形状的个数来代表一个节点的结构特征. 类似地, 节点的各种衡量指标, 诸如度数中心性 (degree centrality)、亲近中心性 (closeness centrality) 以及中介中心性 (betweenness centrality) 等也可以代表节点的结构特性 [59]. ReFex 方法为一个节点定义了 3 种结构特征, 包括中心度、局部中心网络中的边数、局部中心网络连接到外部的边数, 以及基于所有邻居对这 3 个指标递归做平均或者求和 [60, 61]. 更多基于特征向量计算相似度的方法介绍参见综述 [62].

4.3 群体影响力

群体影响力是指多个个体共同组成一个群体所释放的整体影响力. Tang 等人从不同粒度上区分了个体本身、个体之间以及群组影响力, 并提出因子图模型来求解不同影响力的大小 [38]. Belak 等人 [63] 进一步对群组与群组之间的影响力进行度量. Myers 等人 [64] 以及 Lin 等人 [65] 将整个社会网络的外部因素看作一种隐式的群体影响力, 并提出相应的概率模型来解释这种隐式影响力的大小.

部分研究考虑了群体影响力的结构特性. 例如 Ugander 等人 [66] 首先提出了影响力的结构多样性特征. 他们发现一个用户加入 Facebook 的概率与其已经加入 Facebook 的朋友之间结构的多样性成正相关关系. 之后有一系列工作在不同情境下对这一特性进行了探索 [67~70]. 例如, Fang 等人 [68] 对在线游戏中用户的付费行为研究影响力的结构多样性, 得出用户的付费可能性与其已付费朋友的结构多样性成正相关关系. Zhang 等人 [71] 对结构影响力给出了具体的数学定义, 并在大规模网络与用户行为流数据下, 提出对结构影响力的快速采样计算方法.

5 社会影响力应用

社会影响力的主要应用场景包括广告推荐、链接预测与用户行为预测等. 例如, 给定影响力传播模型以及网络中用户之间的影响力大小, 便可以从整个网络中挑选出影响力最大的初始用户, 为其提供费商品体验的机会, 使其将体验感受传播出去, 影响最多的用户购买该商品, 这也是影响力最大化的目标所在. 另一方面, 倘若在链接预测以及用户行为预测中考虑其他用户的影响效应, 则有可能达到更精确的预测效果.

5.1 链接预测

已有链接预测的主要方法包括监督与非监督的学习方法. Liben-Nowell 和 Kleinberg [72] 调研了大部分非监督学习的方法, 包括择优链接法 (preferential attachment) [73], 带跳回的随机游走法 (random walk with restart, RWR) [74], SimRank [52] 以及 Katz [50] 等. 其主要思想是两个节点越相似, 则它们之间越有可能形成链接关系. 监督学习的方法包括 Markov 随机场 [75], 逻辑回归模型 (logistic regression) [76] 以及监督随机游走法 (supervised random walk) [77] 等. Lichtenwalter 等人 [78] 提出了一种监督学习的框架, 将现有的非监督学习方法计算得出的分数都当做特征, 结果表明其效果远远好于纯粹的非监督学习方法.

此外, 网络生成模型也可以用来做链接预测. 网络生成模型刻画了整个网络中关系的生成过程 [79~82]. Barabasi 等人 [80] 提出择优链接的方法生成大规模无尺度网络. Leskovec 等人 [82] 发现了网络密度的幂率分布 (densification powerlaw) 以及直径萎缩 (shrinking diameters) 的网络性质, 并提出森林大火模型 (forest fire models) 使其能够符合这些性质. Leskovec 等人 [81] 提出三角闭合模型, Romero 等人 [79] 提出变型的择优链接模型来拟合上述这些性质. 网络生成模型着眼于拟合宏观网络特性, 譬如长尾效应以及短直径等.

已有链接预测的方法重点考察链接之间的静态结构因素, 并没有过多考虑社会影响力所起的作用.

5.2 用户行为预测

社会影响力是影响用户行为的一个至关重要的因素. 已经存在很多基于社会因素进行推荐的研究. 例如, Ma 等人 [83] 提出一个概率矩阵分解模型, 在分解用户对事物的打分关系矩阵的基础上, 加入了

对用户与用户之间好友关系矩阵的分解. Jiang 等人 [84] 除了对用户之间的静态好友关系矩阵进行分解, 还考虑了用户与用户之间动态的交互关系, 例如转发与评论等行为. Fang 等人 [68] 将影响力结构特性用于游戏玩家付费行为预测, Qiu 等人 [67] 将影响力结构特性用于加入微信群组的行为预测, Zhang 等人 [70] 将这种影响力结构特性应用于微博数据上对用户转发行为进行预测. 具体到转发预测, 已有大量工作研究人们转发微博的原因与机理. 例如 Boyd 等人 [85] 对转发原因做了深入的分析. 该研究主要使用调查问卷的方式, 因此其结果有待于在大规模真实数据上做验证. 不同的研究从不同的角色对转发原因进行解释, 例如, 一些研究重点分析了帖子内容对于转发概率的影响. Naveed 等人 [86] 训练了一个机器学习模型来学习从帖子内容中抽取特征的权重. 他们发现那些包含 hashtag, URL 以及用户名的帖子更容易被转发. Macskassy 等人 [87] 给每个帖子打一个标签, 并将一个用户发表或转发所有帖子的标签组合在一起当做该用户的兴趣. 其中标签空间来源是维基百科的目录. 他们尝试了 4 种不同的模型, 发现两个用户兴趣越相似, 越有可能互相转发对方的帖子. 还有一些研究从话题的流行度、社会关系的强度, 以及发布者的社会地位等角度来研究转发的概率 [88~91].

除此之外, 影响力在广告推荐 [92,93] 以及权威用户发现 [33,92,94~96] 等方面也有广泛的应用.

6 研究工作展望

本文主要从社会影响力的检测、建模与度量方面对社会影响力方面的研究进行了阐述. 总结了一些常见的影响力模型的刻画方式, 以及个体影响力、交互影响力和群体影响力的结构特点等. 作者认为, 随着社会网络的飞速发展, 社会网络显著呈现出动态性与大规模的特征, 未来对社会影响力的研究需要着重从这些方面展开:

融合结构影响力的传播模型. 已有研究对群体影响力的结构多样性进行了分析与度量 [70], 并进一步提出结构影响力的形式化定义与度量方法 [71]. 然而, 并没有研究如何将该结构特性融合进影响力传播模型中. 已有的独立级联影响力传播模型假设激活邻居对目标用户进行尝试激活的事件与其他尝试激活的事件相独立, 基于此假设形式化传播机制, 其数学模型简洁美观. 然而一旦认为激活邻居的尝试激活事件之间互不独立, 且激活邻居共同组成的网络结构特征会带来群体影响力的差异, 那么整个传播模型的数学表达就会变得复杂. 如何结合社会影响力的结构特性并提出一个简洁的数学模型来刻画影响力传播过程是一个尚未解决的挑战.

基于动态网络结构数据与用户行为数据对影响力进行建模与度量. 当网络动态性与行为传播动态性交织在一起时, 影响力的建模与度量变得非常复杂. 用户行为, 例如转发帖子、购买商品等, 与网络结构变化并没有直接的关系, 但潜在地, 行为发生后由于同质性等原因会促使网络结构动态演化. 因此, 在动态社会网络中, 研究用户行为之间的影响力对网络动态演化的作用, 是第二个潜在研究问题.

大规模动态网络数据与用户行为数据中的影响力快速度量方法. 已有研究在大规模网络数据与用户行为流数据中, 对个体之间的影响力与结构影响力进行快速度量 [71]. 其前提假设是网络结构是相对静止的, 可以全部加载进内存, 只有用户行为数据是动态的, 需要实时获取并处理. 然而, 现实情况是, 虽然网络结构数据与用户行为数据相比, 变化比较缓慢, 但其仍然是动态演化的. 在动态的网络结构数据与动态用户行为数据中, 如何快速且实时地度量影响力大小是值得研究的第三个问题.

参考文献

1 Brown J J, Reingen P H. Social ties and word-of-mouth referral behavior. J Consum Res, 1987, 14: 350–362

2 Fowler J H, Christakis N A. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. British Medic J, 2008, 337: a2338

3 Mahajan V, Muller E, Bass F M. New product diffusion models in marketing: a review and directions for research. In: Diffusion of Technologies and Social Behavior. Heidelberg: Springer-Verlag, 1991. 125–177

4 Lorenz J, Rauhut H, Schweitzer F, et al. How social influence can undermine the wisdom of crowd effect. Proc Nation Academy Sci, 2011, 108: 9020–9025

5 Zhu H, Huberman B A. To switch or not to switch: understanding social influence in online choices. Am Behav Sci, 2014, 58: 1329–1344

6 Bakshy E, Eckles D, Yan R, et al. Social influence in social advertising: evidence from field experiments. In: Proceedings of the 13th ACM Conference on Electronic Commerce, Valencia, 2012. 146–161

7 Bond R M, Fariss C J, Jones J J, et al. A 61-million-person experiment in social influence and political mobilization. Nature, 2012, 489: 295–298

8 Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 7–15

9 La Fond T, Neville J. Randomization tests for distinguishing social influence and homophily effects. In: Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010. 601–610

10 Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc Nation Academy Sci, 2009, 106: 21544–21549

11 Kermack W O, McKendrick A G. A contribution to the mathematical theory of epidemics. Proc Royal Soc London A: Math, Phys Engineer Sci, 1927, 115: 700–721

12 Li M Y, Graef J R, Wang L, et al. Global dynamics of a SEIR model with varying total population size. Math Biosci, 1999, 160: 191–213

13 Kempe D, Kleinberg J, Tardos É. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2003. 137–146

14 Ghasemiesfeh G, Ebrahimi R, Gao J. Complex contagion and the weakness of long ties in social networks: revisited. In: Proceedings of the 14th ACM Conference on Electronic Commerce. New York: ACM, 2013. 507–524

15 Bernheim B D. A theory of conformity. J Polit Econ, 1994, 102: 841–877

16 Rodriguez M G, Balduzzi D, Schölkopf B. Uncovering the temporal dynamics of diffusion networks. arXiv: 1105.0697, 2003

17 Rosa D, Giua A. A non-progressive model of innovation diffusion in social networks. In: Proceedings of IEEE 52nd Annual Conference on Decision and Control (CDC). Washington: IEEE, 2013. 6202–6207

18 Yang Z, Tang J, Xu B, et al. Active learning for networked data based on non-progressive diffusion model. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York: ACM, 2014. 363–372

19 Chen W, Lu W, Zhang N. Time-critical influence maximization in social networks with time-delayed diffusion process. AAAI, 2012, 2012: 1–5

20 Kimura M, Saito K, Ohara K, et al. Learning information diffusion model in a social network for predicting influence of nodes. Intell Data Analys, 2011, 15: 633–652

21 Barbieri N, Bonchi F, Manco G. Topic-aware social influence propagation models. Knowl Inform Syst, 2013, 37: 555–584

22 Chen W, Collins A, Cummings R, et al. Influence maximization in social networks when negative opinions may emerge and propagate. In: Proceedings of the SIAM International Conference on Data Mining, Mesa, 2011. 379–390

23 Li Y, Chen W, Wang Y, et al. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In: Proceedings of the 6th ACM International Conference on Web Search and Data Mining. New York: ACM, 2013. 657–666

24 He X, Song G, Chen W, et al. Influence blocking maximization in social networks under the competitive linear threshold model. In: Proceedings of SIAM International Conference on Data Mining, Anaheim, 2012. 463–474

25 Yang Y, Tang J, Leung C W K, et al. RAIN: social role-aware information diffusion. AAAI, 2012, 2012: 367–373

26 Zhang J, Fang Z, Chen W, et al. Diffusion of "following" links in microblogging networks. IEEE Trans Knowl Data

Eng, 2015, 27: 2093–2106
27 Bass F M. A new product growth for model consumer durables. Manage Sci, 1969, 15: 215–227
28 Domingos P, Richardson M. Mining the network value of customers. In: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011. 57–66
29 Goldenberg J, Libai B, Muller E. Talk of the network: a complex systems look at the underlying process of word-of-mouth. Market Lett, 2001, 12: 211–223
30 Richardson M, Domingos P. Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002. 61–70
31 Leskovec J, Krause A, Guestrin C, et al. Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2007. 420–429
32 Chen W, Wang Y, Yang S. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. 199–208
33 Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010. 261–270
34 Heidemann J, Klier M, Probst F. Identifying key users in online social networks: a pagerank based approach. In: Proceedings of the International Conference on Information Systems, Saint Louis, 2010. 12–15
35 Asch S E. Opinions and social pressure. In: Readings About the Social Animal. New York: Worth Publishers, 1955. 17–26
36 Kelman H C. Compliance, identification, and internalization three processes of attitude change. J Conflict Resolut, 1958, 2: 51–60
37 Cialdini R B, Goldstein N J. Social influence: complicance and conformity. Annual Rev Psychology, 2004, 55: 591–621
38 Tang J, Wu S, Sun J. Confluence: conformity influence in large social networks. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013. 347–355
39 Li H, Bhowmick S S, Sun A. Casino: towards conformity-aware social influence analysis in online social networks. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. New York: ACM, 2011. 1007–1012
40 Li H, Bhowmick S S, Sun A. Cinema: conformity-aware greedy algorithm for influence maximization in online social networks. In: Proceedings of the 16th International Conference on Extending Database Technology. New York: ACM, 2013. 323–334
41 Holme P, Newman M E. Nonequilibrium phase transition in the coevolution of networks and opinions. Physic Rev E, 2006, 74: 056108
42 Crandall D, Cosley D, Huttenlocher D, et al. Feedback effects between similarity and social influence in online communities. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 160–168
43 Scripps J, Tan P N, Esfahanian A H. Measuring the effects of preprocessing decisions and network forces in dynamic network analysis. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. 747–756
44 Gruhl D, Guha R, Liben-Nowell D, et al. Information diffusion through blogspace. In: Proceedings of the 13th International Conference on World Wide Web. New York: ACM, 2004. 491–501
45 Saito K, Nakano R, Kimura M. Prediction of information diffusion probabilities for independent cascade model. In: Knowledge-based Intelligent Information and Engineering Systems. Heidelberg: Springer, 2008. 67–75
46 Goyal A, Bonchi F, Lakshmanan L V. Learning influence probabilities in social networks. In: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York: ACM, 2010. 241–250
47 Kutzkov K, Bifet A, Bonchi F, et al. Strip: stream learning of influence probabilities. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013. 275–283
48 Kesslere M M. Bibliographic coupling between scientific papers. J Assoc Inform Sci Tech, 1963, 14: 10–25
49 Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. J Assoc Inform Sci Tech, 1973, 24: 265–269
50 Katz L. A new status index derived from sociometric analysis. Psychometrika, 1953, 18: 39–43
51 Tsourakakis C E. Toward quantifying vertex similarity in networks. Internet Math, 2014, 10: 263–286

52 Jeh G, Widom J. SimRank: a measure of structural-context similarity. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2002. 538–543

53 Newman M E. Finding community structure in networks using the eigenvectors of matrices. Physic Rev E, 2006, 74: 036104

54 Fujiwara Y, Nakatsuji M, Shiokawa H, et al. Efficient ad-hoc search for personalized pagerank. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. New York: ACM, 2013. 445–456

55 Sarkar P, Moore A W. Fast nearest-neighbor search in disk-resident graphs. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010. 513–522

56 Blondel V D, Gajardo A, Heymans M, et al. A measure of similarity between graph vertices: applications to synonym extraction and web searching. SIAM Rev, 2004, 46: 647–666

57 Jin R, Lee V E, Hong H. Axiomatic ranking of network role similarity. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011. 922–930

58 Burt R S. Detecting role equivalence. Social Network, 1990, 12: 83–97

59 Freeman L C. A set of measures of centrality based on betweenness. Sociometry, 1977, 35–41

60 Henderson K, Gallagher B, Li L, et al. It's who you know: graph mining using recursive structural features. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2011. 663–671

61 Henderson K, Gallagher B, Eliassi-Rad T, et al. Rolx: structural role extraction mining in large graphs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012. 1231–1239

62 Rossi R A, Ahmed N K. Role discovery in networks. IEEE Trans Knowl Data Eng, 2015, 27: 1112–1131

63 Belák V, Lam S, Hayes C. Cross-community influence in discussion fora. In: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, Dublin, 2012

64 Myers S A, Zhu C, Leskovec J. Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012. 33–41

65 Lin S, Wang F, Hu Q, et al. Extracting social events for learning better information diffusion models. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2013. 365–373

66 Ugander J, Backstrom L, Marlow C, et al. Structural diversity in social contagion. Proc Nation Academy Sci, 2012, 109: 5962–5966

67 Qiu J, Li Y, Tang J, et al. The lifecycle and cascade of WeChat social messaging groups. In: Proceedings of the 25th International Conference on World Wide Web. New York: ACM, 2016. 311–320

68 Fang Z, Zhou X, Tang J, et al. Modeling paying behavior in game social networks. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York: ACM, 2014. 411–420

69 Kloumann I, Adamic L, Kleinberg J, et al. The lifecycles of apps in a social ecosystem. In: Proceedings of the 24th International Conference on World Wide Web. New York: ACM, 2015. 581–591

70 Zhang J, Liu B, Tang J, et al. Social Influence Locality for Modeling Retweeting Behaviors. In: Proceedings of International Joint Conference on Artificial Intelligence, Beijing, 2013. 2761–2767

71 Zhang J, Tang J, Zhong Y, et al. StructInf: mining structural influence from social streams. AAAI, 2017, 73–80

72 Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks. J Assoc Inform Sci Tech, 2007, 58: 1019–1031

73 Newman M E. Clustering and preferential attachment in growing networks. Physic Rev E, 2001, 64: 025102

74 Tong H, Faloutsos C, Pan J Y. Fast random walk with restart and its applications. In: Proceedings of the Sixth International Conference on Data Mining. Washington: IEEE, 2006. 613–622

75 Wang C, Satuluri V, Parthasarathy S. Local probabilistic models for link prediction. In: Proceedings of Seventh IEEE International Conference on Data Mining. Washington: IEEE, 2007. 322–331

76 Leskovec J, Huttenlocher D, Kleinberg J. Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web. New York: ACM, 2010. 641–650

77 Backstrom L, Leskovec J. Supervised random walks: predicting and recommending links in social networks. In: Proceedings of the 4th ACM International Conference on Web Search and Data Mining. New York: ACM, 2011.

635–644
78 Lichtenwalter R N, Lussier J T, Chawla N V. New perspectives and methods in link prediction. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2010. 243–252
79 Romero D M, Kleinberg J M. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In: Proceedings of 4th International AAAI Conference on Weblogs and Social Media, Washington, 2010
80 Barabási A L, Albert R. Emergence of scaling in random networks. Science, 1999, 286: 509–512
81 Leskovec J, Backstrom L, Kumar R, et al. Microscopic evolution of social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2008. 462–470
82 Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. New York: ACM, 2005. 177–187
83 Ma H, Yang H, Lyu M R, et al. Sorec: social recommendation using probabilistic matrix factorization. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM, 2008. 931–940
84 Jiang M, Cui P, Liu R, et al. Social contextual recommendation. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. New York: ACM, 2012. 45–54
85 Boyd D, Golder S, Lotan G. Tweet, tweet, retweet: conversational aspects of retweeting on twitter. In: Proceedings of the 43rd International Conference on System Sciences, Hawaii, 2010. 1–10
86 Naveed N, Gottron T, Kunegis J, et al. Bad news travel fast: a content-based analysis of interestingness on twitter. In: Proceedings of the 3rd International Web Science Conference. New York: ACM, 2011
87 Macskassy S A, Michelson M. Why do people retweet? anti-homophily wins the day! In: Proceedings of Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, 2011. 209–216
88 Chen K, Chen T, Zheng G, et al. Collaborative personalized tweet recommendation. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, 2012. 661–670
89 Duan Y, Jiang L, Qin T, et al. An empirical study on learning to rank of tweets. In: Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, 2010. 295–303
90 Suh B, Hong L, Pirolli P, et al. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Proceedings of IEEE 2nd International Conference on Social Computing. Washington: IEEE, 2010. 177–184
91 Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. New York: ACM, 2010. 1633–1636
92 Goyal A, Bonchi F, Lakshmanan L V. Discovering leaders from community actions. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management. New York: ACM, 2008. 499–508
93 Provost F, Dalessandro B, Hook R, et al. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2009. 707–716
94 Agarwal N, Liu H, Tang L, et al. Identifying the influential bloggers in a community. In: Proceedings of the 2008 International Conference on Web Search and Data Mining. New York: ACM, 2008. 207–218
95 Song X, Chi Y, Hino K, et al. Identifying opinion leaders in the blogosphere. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management. New York: ACM, 2007. 971–974
96 Probst F, Grosswiele D K L, Pfleger D K R. Who will lead and who will follow: identifying influential users in online social networks. Bus Inform Syst Eng, 2013, 5: 179–193

Survey of social influence analysis and modeling

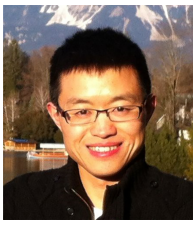
Jing ZHANG[1] & Jie TANG[2*]

1. *Computer Science Department, Information School, Renmin University of China, Beijing* 100872, *China*;
2. *Computer Science Department, Tsinghua University, Beijing* 100084, *China*
* Corresponding author. E-mail: jietang@tsinghua.edu.cn

Abstract Social influence occurs when one's opinions or behaviors are affected by others. It forms a prevalent, complex, and subtle force that governs the dynamics of social networks. With the rapid proliferation of online social networks such as Twitter, Facebook, Yelp, and Amazon, modeling the influence diffusion mechanism and quantitatively measuring social influence between people become more and more critical for algorithms behind features such as friend recommendations, expert finding, and behavior prediction. In addition, they can also benefit the development of virtual marketing and supervision by public opinions. Social influence has been extensively studied and has recently been attracting great attention from different communities. This paper introduces the study and the future work of social influence, in particular, influence tests, modeling, and measurement.

Keywords social influence, diffusion model, influence test, influence measurement, social network

Jing ZHANG was born in 1984. She received her Ph.D. degree from Computer Science Department, Tsinghua University, in 2016. Currently, she is an assistant professor at Computer Science Department, Information School, Renmin University of China. Her research interests include information diffusion, social influence and social network mining.



Jie TANG was born in 1977. He received his Ph.D. degree from Computer Science Department, Tsinghua University, in 2006. Currently, he is an associate professor at Computer Science Department, Tsinghua University. His research interests include social network theories, data mining methodologies, machine learning algorithms, and semantic web technologies.