

DOI: 10.7652/xjtuxb201704014

众筹项目的社交网络影响力预测与分析

杨扬¹, Chun-Ta LU², 王菲菲³, 许进¹, Philip S. YU²

(1. 北京大学信息科学技术学院, 100871, 北京; 2. 伊利诺伊大学芝加哥分校计算机系, 60607, 美国芝加哥;
3. 北京大学光华管理学院, 100871, 北京)

摘要: 针对众筹项目由于社会影响力不足而成功率较低的问题,提出了面向众筹平台的社交网络影响力预测方法。该方法基于众筹网站和社交网络的实时观测数据,分别提取累积和增量等多类别预测特征,并随着社会推广的进行而渐进地预测项目的社交网络影响力增益,最后采用带 L1 范数约束惩罚的逻辑回归等方法进行预测特征分析。实验结果表明:在整个推广过程中,众筹项目的社交网络影响力可以被精确预测,准确率最高达 88.31%;分类器在采用累积特征时具有比采用增量特征更好的预测效果;项目统计特征和推广者的社会影响力等特征具有更高的重要性和更稳定的显著性。该方法成功地解决了众筹项目的社会影响力预测问题,并为设计更好的社交网络推广策略提供了依据。

关键词: 社交网络;众筹项目;社会影响力;影响力预测

中图分类号: TP391.9 **文献标志码:** A **文章编号:** 0253-987X(2017)04-0091-06

Prediction and Analysis of Social Influence for Crowdfunding Projects

YANG Yang¹, Chun-Ta LU², WANG Feifei³, XU Jin¹, Philip S. YU²

(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;
2. Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA;
3. Guanghua School of Management, Peking University, Beijing 100871, China)

Abstract: A novel influence prediction method is proposed to deal with the social promotion problem on crowdfunding platforms and to investigate the predictability of social influence for fundraising projects. The proposed method gradually predicts future influence growth of each project based on features extracted in different timestamps from data streams of crowdfunding websites and social networks. Experimental results show that features extracted from aggregate data rather than incremental data are better indicators of the social influence growth, and the method has the highest prediction precision 88.31%. Analysis of features also shows that some features such as the statistic of projects and social influence of promoters are consistently significant within the prediction process. These results provide hints for devising better social promotions.

Keywords: social network; crowdfunding project; social influence; influence prediction

与传统的募资方式相比,众筹具有募资速度快和成本低等优势,因而逐渐成为一种重要的募资渠道。然而,由于众筹平台中的大多数项目都以失败告终,众筹项目发起人、投资者和众筹平台的运营者

收稿日期: 2016-12-15。 作者简介: 杨扬(1988—),男,博士生;许进(通信作者),男,教授,博士生导师。 基金项目: 国家重点基础研究发展计划资助项目(2013CB329600);国家自然科学基金资助项目(61372191,61572492,61672050)。

网络出版时间: 2017-03-03 网络出版地址: <http://www.cnki.net/kcms/detail/61.1069.T.20170303.0925.008.html>

都迫切地需要找到能够有效提高项目成功率的方法。

已有研究主要通过两种方式解决募资难题:一种是优化众筹平台中的推荐算法,使得项目能够被更合理地推荐给潜在投资者,其中包括将项目推荐给个体用户的算法^[1-2]和将项目推荐给由个体用户所组成的兴趣社团的算法^[3];另一种方法是在募资的早期阶段对项目成功率进行预测,例如在项目推广周期的前1/4时间内,基于项目在社交网络中的传播状态统计量,预测其最终有多少投资人和能否募资成功^[4]。相关研究还包括采用审查回归模型改进已有算法的预测效果^[5]和针对国内平台上众筹项目成功与否的影响因素分析与预测^[6]等。然而,上述方法仍无法有效解决募资难题:推荐算法只能将项目推荐给正在平台中访问的用户,而众筹项目中的大量投资者是通过社交网络中传播的推广链接导入的,因此推荐算法无法增加整个平台中潜在投资者数以及整个平台的总募资金量;项目成功率预测提供了失败的预警手段,但无法帮助改进项目的推广策略以提高其募资成功概率。具有更高社交网络影响力的众筹项目获得成功概率更高^[7],因此通过研究众筹项目在社交网络中传播的社会影响力预测问题来帮助众筹者制定更合理的推广策略,从而取得更高的社交网络影响力和成功概率。

本文创新性地对众筹项目在社交网络中的影响力传播问题进行了研究,提出同时基于众筹平台数据和社交网络数据进行特征抽取的影响力预测方法,并取得了较好的预测效果;在特征分析过程中,部分预测特征在整个项目推广周期内都具有一致的显著性和较强的相关性,这些研究能够帮助众筹发起者对项目的社会影响力进行实时观测并制定相应推广策略,从而提高项目的成功率。

1 问题定义

从两方面定义众筹项目的社交网络影响力,即社交网络中相关推广微博的被分享次数和相关推广微博的转发者(推广者)的粉丝总数。

定义1 社交网络影响力 对于某个众筹项目,假设其在社交网络中的所有推广者 U 发布了一组推广微博 T ,并且推广者 U 的粉丝集合为 G ,则众筹项目的社会影响力,即分享次数和粉丝总数,可分别被定义为 $|T|$ 和 $|G|$ 。

分享次数和粉丝总数从不同角度反映了项目的社会影响力。直观来看,分享次数反映了网络用户对某一个项目接受和推广的意愿,粉丝总数则反映了某

个项目的推广微博可能被网络用户浏览的总人次。

对于筹资周期长度不同的众筹项目,用相同数目的时间分割点将其分割,使得所有项目被等分为数量相同的时间片段,从而可以被统一处理。

定义2 时间分割点 假设每个众筹项目都可以被等分为 K 个区间,用 $[t_{\text{start}}, t_{\text{end}}]$ 表示某个项目的筹资周期,其中 t_{start} 为起始时间, t_{end} 为终止时间,则该项目在第 k 个时间分割点的时间为 $[k(t_{\text{end}} - t_{\text{start}})/K] + t_{\text{start}}$ 。

设项目 p 在第 k 个时间分割点的社交网络影响力为 $|T_k^p|$ 或 $|G_k^p|$,在终止时间 t_{end} 的社交网络影响力为 $|T_k^p|$ 或 $|G_k^p|$,那么项目 p 在第 k 个时间分割点的未来影响力增益 $Q_k^p = |T_k^p| / |T_k^p|$ 或 $Q_k^p = |G_k^p| / |G_k^p|$ 。对于所有项目 P ,在第 k 个时间分割点存在一个影响力增益的中位数 $M_k = \text{med}(Q_k)$ 。通过判定一个项目在第 k 个时间分割点上的影响力增长率是否能够超过 M_k ,将该预测问题定义为一个均匀分布的二分类问题。根据该定义,依次在每个分割点上预测某个众筹项目的未来影响力增益是否大于中位数,就可以对该项目的社会影响力进行持续观测。

首先,影响力预测问题可被转换为样本均匀分布的二分类问题,而非传统影响力预测中样本分布不均匀的多分类问题^[8];其次,在众筹项目的整个筹资周期内来研究项目社会影响力的可预测性;最后,众筹项目发起人可以实时地持续观测项目的社会推广进程,具有较强的实际操作环境。另外,本文的预测问题比传统影响力预测中的级联传播预测更加复杂:对于每个众筹项目,在整个项目周期内新的级联传播树随着初始推广者的出现而不断出现,因此需要解决的是一组数目和规模不断增长的级联的传播预测问题,而非传统影响力预测中常见的独立级联传播树^[8-9]预测问题。此外,只关注项目筹资周期内的影响力增长,因此众筹项目的影响力增长曲线随着募资周期的结束呈现出截断效应。

2 实验数据集

本文采用已有研究^[4]中的数据,并对其进行了扩充。除了原数据中的KickStarter众筹平台数据和Twitter社交网络中的相应数据外,还扩充了众筹项目的文本描述、社交网络的用户内容、用户行为和关注关系等数据,以丰富用于预测的特征集合。

Kickstarter众筹数据包含2012年12月1日后发布,并且在2013年3月31日之前结束募资的

1 521个项目。其中每个项目包含项目文本介绍、Kickstarter 分类类别、募资目标、募资时间、页面 URL 和投资金额选项等,同时还包括在募资周期内每个项目的日募资金额、投资者人数。

Twitter 数据包含所有与 Kickstarter 项目有关并发表于 2012 年 11 月 1 日至 2013 年 4 月 30 日期间的微博,共 62 473 条。每条微博包含作者、被提及用户、项目页面 URL、分享次数和发表时间。对于每个发布初始推广微博的用户(初始推广者),爬取其个人简介和发表于原始推广微博之前的最多 100 条微博,得到文本内容和被分享次数。同时,还爬取了每个初始推广者的粉丝在当时所发表的最多 100 条微博文本内容。对于所有项目推广者(初始发布或后续转发推广微博的人),爬取其邻居用户和邻居的邻居用户的所有朋友和粉丝,构造出规模为 205 149 431 个节点的好友关注关系网络。

3 预测特征构造

基于 Kickstarter 和 Twitter 数据,可以构造丰富的特征来完成预测任务。对于第 k 个时间分割点,本文首先基于累积数据构造了项目统计特征、内容特征、社会行为特征和社会结构特征等 4 类累积特征,以刻画项目属性的累积变化特性。得到第 $k-1$ 和 k 时间分割点的累积特征值后,通过对累积特征值作差,可以得到第 k 个时间分割点的增量特征值。增量特征可以刻画两个分割点之间项目属性的变化特性。针对预测过程中存在的多重级联现象,本文构造多重级联预测特征进行刻画。

项目统计特征随着募资进程的推进而变化,并且对于募资的最终结果,如总投资者人数和最终募集的资金数等,具有较好的预测效果^[4-5],因此本文构造了相应统计特征,包括已投资人数、已筹集金额、已筹集比例和投资最多/最少人数。

社交网络中存在的大量用户文本内容和众筹平台中丰富的项目文本描述为构造文本内容特征提供了条件。研究表明,社交网络中的用户在话题层面具有不同的相互影响力^[10],因此本文通过推广者与其粉丝之间的文本相似性来刻画相互之间的影响力。类似地,还通过推广者的微博内容与其所推广项目的文本介绍内容之间的文本相似性来刻画推广者在众筹项目所属的领域的专业程度。实验中所有文本都采用 TF/IDF 表示,并用 Pearson 相关性系数度量相似性。构造的特征包括项目简介和推广者的最高内容相似度、项目简介和推广者的平均内容

相似度、初始推广者与其粉丝的内容的最高相似度(特征 R_1)和初始推广者与其粉丝的平均内容相似度(特征 R_2)等。

社会行为是驱动社会影响力增长的重要因素,其中包括发布信息、转发信息和提及他人等。对于每个项目抽取社会行为特征来刻画项目推广者的社会行为和项目推广微博的统计量变化。构造的特征包括推广者人数、推广微博数、平均每条推广微博被分享次数、平均每个推广者发表的推广微博数、平均每个推广者的推广微博被分享次数、微博中被提及人数、独立的级联数目(特征 R_3)、初始推广者占比(特征 R_4)、最大级联中推广者占有所有推广者的比例(特征 R_5)和同时是推广者和投资者的人占有所有推广者比例等。

社交网络结构特征常被用于挖掘具有重大社交网络影响力的用户,而大多数重要的社会结构特征都派生自网络中心性的度量指标。为了充分刻画项目推广者的社交网络影响力,从局部和全局层面抽取中心性度量指标作为预测特征,其中包括初始/所有推广者的粉丝总数、初始/所有推广者的评分粉丝数、被初始/所有推广者关注的总人数、被初始/所有推广者关注的平均人数、所有推广者的平均聚集系数和推广者的平均 PageRank 值等。

上述特征均基于累积数据构造,与累积数据不同,增量数据中包含着累积变量的变化规律,因此与上述累积特征一一对应,本文还构造了增量特征,并研究其预测效果。

在预测任务中每个众筹项目的影响力增长都伴随着多个同时发生的级联传播,因此从文本、结构和行为的角度构造多重级联特征来刻画众筹项目影响力增长过程中的多重级联现象。特征 R_1 、 R_2 、 R_3 、 R_4 和 R_5 均为多重级联特征。

4 影响力增长预测

对于每个众筹项目,在两个影响力定义和不同特征集合下预测社交网络影响力增益。在所有实验中,首先用带 L1 一范数约束惩罚的逻辑回归(L1-penalized logistic regression)算法进行特征维度压缩,进而用 SVM 分类器以获得最好的分类效果。所有实验采用 10 折交叉验证以避免过拟合。

首先用累积特征进行预测,并分析其预测效果。实验采用社会结构特征作为基特征组,进而分别将其与其他 3 个累积特征组组合进行预测。图 1 给出了时间分割点数设为 20 时,4 种特征组合在第 1 个

时间分割点($k=1$)的预测效果。其中分别用 S 、 A 、 P 和 C 代表社会结构特征、社会行为特征、项目统计特征和内容特征。 F 是评价机器学习算法预测效果的综合评价的指标,定义为准确率 D 和召回率 E 的调和平均值

$$F=2DE/(D+E)$$

由图 1a 可以看出,社会结构特征和社会行为特征的组合取得了最好的预测结果,其中 D 为 0.894, F 为 0.914。

社会结构特征与项目统计特征的组合取得了最差的预测效果, D 为 0.727, F 为 0.735。该结果甚至比基础特征组的独立预测效果更差。由图 1b 可以看出,社会结构特征对粉丝数的独立预测效果获得了最高的预测 F (0.795),而社会结构特征与社会行为特征的组合取得了最高的 D (0.771)。对比图 1a 和图 1b 可以看出,分享次数的可预测性比粉丝总数更强,而相同的特征组合在两个不同的预测任务中表现出截然不同的效果。具体来说,与其他组合相比,特征 S 与特征 A 的组合、特征 S 与特征 C 的组合在分享次数预测中取得了更好的预测效果,而在粉丝总数的预测中结果相反。

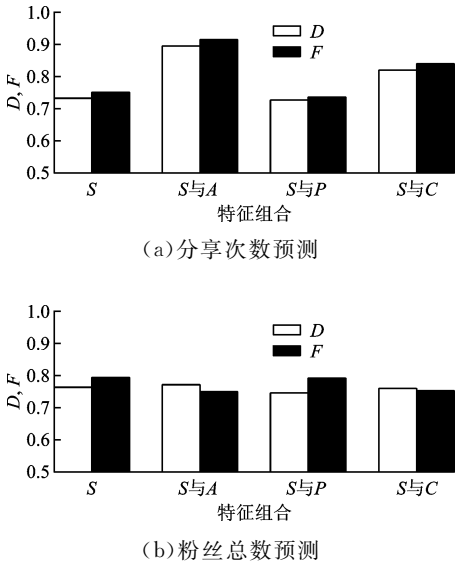
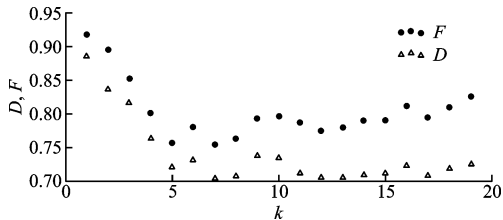


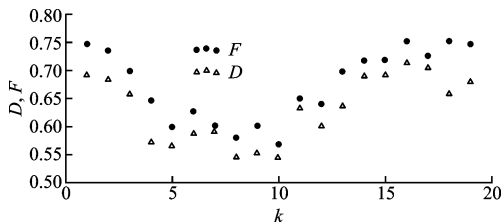
图 1 不同特征组合的影响力增益预测

接下来给出众筹项目在整个募资周期内影响力增长的预测结果,对分享次数和粉丝总数的预测分别如图 2 和图 3 所示。实验结果表明,两项任务都取得了较好的预测效果,并且在募资周期的起始时期和结束时期的预测效果更好。具体地,在图 2a、图 2b、图 3a 和图 3b 中,当时间分割点 k 由 0 不断增大时, F 首先呈现出下降趋势,并在 $k=5$ 附近取得全局最小值;接下来的预测效果增长期持续至募资

结束。该预测效果的变化规律与文献[4]中指出的募资和推广行为吻合,即对于所有众筹项目,投资行为在募资周期的前 1/4 阶段出现起始爆炸性增长,在募资周期的后 1/10 阶段出现末尾爆炸性增长,而在两个增长阶段之间存在一个稳定增长阶段。这个有趣的预测现象可以解释为:在爆炸性增长阶段,项目的募资和推广活动更加频繁,因此预测特征能够捕捉到更多的有效信息,从而使预测效果比稳定增长阶段更好。

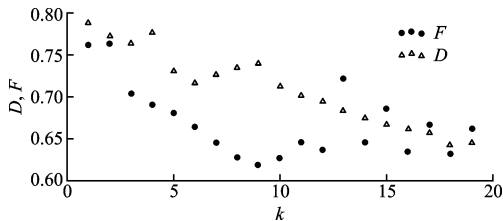


(a) 累积特征

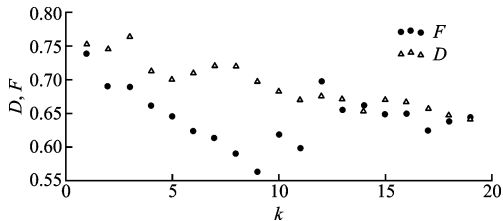


(b) 增量特征

图 2 不同特征的分享次数预测



(a) 累积特征



(b) 增量特征

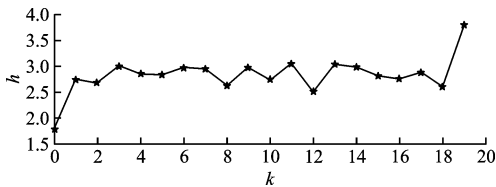
图 3 不同特征的粉丝总数预测

通过对图 2 和图 3 中相应子图中的预测结果可以看出,对分享次数进行预测的 D 和 F 普遍高于对粉丝总数的预测。在分享次数的预测结果中, D 与 F 的变化趋势相近,而在粉丝数的预测中差异很大,因此在本文给出的两个预测对象中,分享次数比

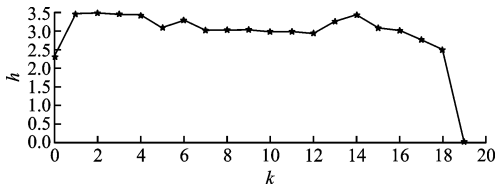
粉丝总数的可预测性更强。对比图 2a 和图 2b 可以看出,图 2a 中最差的预测结果 F 也普遍优于图 2b 中的预测结果,类似现象也出现在图 3a 和图 3b 中,由此可以总结出应用累积特征进行预测比应用增量特征效果更好。

5 特征重要性分析

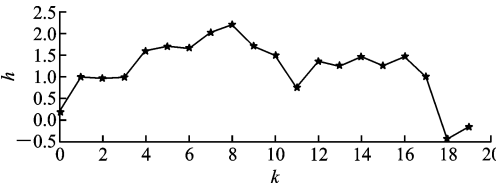
接下来分析特征的显著性在整个预测周期中的变化。本节实验首先用带 L1 一范数约束惩罚的逻辑回归算法进行特征筛选,进而用筛选出的变量再次通过逻辑回归算法进行预测,并在预测过程中记录回归系数及其显著性,最后将持续显著的特征筛选出来进行展示。图 4 和图 5 分别给出了在粉丝总数预测和分享次数预测中表现显著的一部分特征其逻辑回归系数 h 的变化趋势。



(a) 已投资人数



(b) 已筹集金额



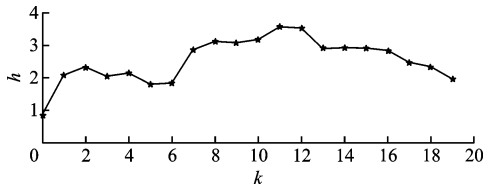
(c) 所有推广者关注的平均人数

图 4 粉丝总数预测中的特征重要性

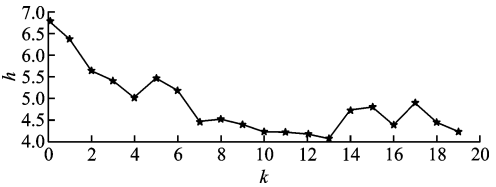
图 4 给出了特征与粉丝总数的关系,可以看出 3 个特征在整个募资过程中基本与粉丝数的增益呈正相关。

预测特征在分享次数预测中的重要性如图 5 所示。其中特征已筹集金额、初始推广者占比和所有推广者的平均粉丝数与预测目标呈现出稳定的正相关性。因此,在社交网络中,若某个众筹项目具有更高初始推广者占比(外部激活推广者占所有推广者的比例),并且推广者的粉丝数更多时,该项目的分

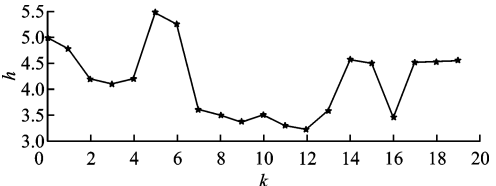
享次数具有更高的增长预期。



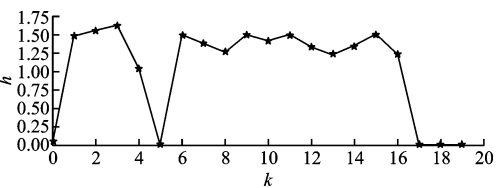
(a) 已筹集金额



(b) 初始推广者占比



(c) 所有推广者的平均粉丝数



(d) 所有推广者关注的平均人数

图 5 分享次数预测中的特征重要性

在两个不同影响力定义下,特征分析都显示当前时间分割点上的项目统计特征能够显著地影响未来的影响力增益。当项目在当前时刻具有更高的投资人数和投资金额时,其在未来的影响力增益更有可能超越所有项目的中位数值。此外,推广者的社交网络影响力,如初始推广者占比、所有推广者的平均粉丝数和所有推广者关注的平均人数,都是驱动项目影响力增长的重要因素。

6 结 论

本文对众筹项目在社交网络中影响力增长的可预测性问题进行了研究,通过将预测任务定义为渐进的影响力增益预测问题,使得项目最终获得的社交网络影响力可以在整个募资周期内被实时监测。本文搜集了众筹平台 Kickstarter 上 4 个月的数据,以及相应的 Twitter 社交网络数据,并从中构造累积、增量等预测特征。实验结果显示,众筹项目社交

网络影响力可以在整个募资周期内被准确预测,其中分享次数比粉丝总数更容易预测,而累积特征的预测效果比增量特征更好。在整个预测周期内,多重级联特征中的初始推广者占比表现出稳定的显著性和较高的重要性。此外,特征重要性分析还显示,部分特征如已投资人数、初始推广者占比和所有推广者的平均粉丝数,与影响力增益具有稳定的正相关性,这些发现为设计更合理的众筹项目推广策略提供了有益的启示。

参考文献:

- [1] AN Jisun, QUERCIA D, CROWCROFT J. Recommending investors for crowdfunding projects [C] // Proceedings of the 23rd International Conference on World Wide Web. New York, USA: ACM, 2014: 261-270.
- [2] RAKESH V, CHOO J, REDDY C K. Project recommendation using heterogeneous traits in crowdfunding [C] // Proceedings of the 9th International AAAI Conference on Web and Social Media. Palo Alto, CA, USA: AAAI Press, 2015: 337-346.
- [3] RAKESH V, LEE W C, REDDY C K. Probabilistic group recommendation model for crowdfunding domains [C] // Proceedings of the 9th ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2016: 257-266.
- [4] LU C T, XIE Sihong, KONG Xiangnan, et al. Inferring the impacts of social media on crowdfunding [C] // Proceedings of the 7th ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2014: 573-582.
- [5] LI Yan, RAKESH V, REDDY C K. Project success prediction in crowdfunding environments [C] // Proceedings of the 9th ACM International Conference on Web Search and Data Mining. New York, USA: ACM, 2016: 247-256.
- [6] 李雅婷, 左文明. 众筹项目筹资结果预测模型[J]. 统计与决策, 2016(2): 86-89.
LI Yating, ZUO Wenming. Results prediction of crowdfunding projects [J]. Statistics and Decision, 2016(2): 86-89.
- [7] HUI J S, GERBER E M, GERGLE D. Understanding and leveraging social networks for crowdfunding: opportunities and challenges [C] // Proceedings of the 2014 Conference on Designing Interactive Systems. New York, USA: ACM, 2014: 677-680.
- [8] CHENG J, ADAMIC L, DOW P A, et al. Can cas-

cases be predicted? [C] // Proceedings of the 23rd International Conference on World Wide Web. New York, USA: ACM, 2014: 925-936.

- [9] WANG Senzhang, YAN Zhao, HU Xia, et al. Burst time prediction in cascades [C] // Proceedings of the 9th International AAAI Conference on Web and Social Media. Palo Alto, CA, USA: AAAI Press, 2015: 325-331.
- [10] TANG Jie, SUN Jimeng, WANG Chi, et al. Social influence analysis in large-scale networks [C] // Proceedings of the 15th ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2009: 807-816.

[本刊相关文献链接]

张磊, 马春光, 杨松涛, 等. 面向路网环境速度预测攻击的隐私保护. 2017, 51(2): 27-32. [doi:10.7652/xjtuxb201702005]

李晓娟, 孙文磊, 袁逸萍, 等. 扰动环境下作业车间网络多瓶颈识别方法研究. 2016, 50(12): 64-72. [doi:10.7652/xjtuxb201612011]

谈博, 蔡远利, 史旻. 采用变时域监督预测控制的机炉协调控制方法. 2016, 50(12): 92-98. [doi:10.7652/xjtuxb201612015]

杨沫, 由磊, 李冰, 等. 采用传递概率与社会网络分析的延迟容忍网络路由. 2016, 50(12): 136-141. [doi:10.7652/xjtuxb201612021]

李扬, 潘泉, 杨涛. 基于短文本情感分析的敏感信息识别. 2016, 50(9): 80-84. [doi:10.7652/xjtuxb201609013]

王菲菲, 杨扬, 蒋飞, 等. 面向用户话题相似性特征的链路预测方法. 2016, 50(8): 103-109. [doi:10.7652/xjtuxb201608017]

魏倩, 蔡远利. 一种基于神经网络的中制导改进算法. 2016, 50(7): 125-130. [doi:10.7652/xjtuxb201607019]

赵辉, 郑庆华, 张未展. 多版本视频点播流媒体服务器集群资源分配方法. 2016, 50(6): 30-35. [doi:10.7652/xjtuxb201606005]

刘兆丽, 秦涛, 管晓宏, 等. 采用用户名相似度传播模型的线上用户身份. 2016, 50(4): 1-6. [doi:10.7652/xjtuxb201604001]

孟宪佳, 马建峰, 王一川. 面向社交网络中多背景的信任评估模型. 2015, 49(4): 73-77. [doi:10.7652/xjtuxb201504012]

叶娜, 赵银亮, 边根庆, 等. 模式无关的社交网络用户识别算法. 2013, 47(12): 19-25. [doi:10.7652/xjtuxb201312004]

张赛, 徐格, 李海涛. 微博类社交网络中信息传播的测量与分析. 2013, 47(2): 124-130. [doi:10.7652/xjtuxb201302021]

莫同, 褚伟杰, 李伟平, 等. 采用超图的微博群落感知方法. 2012, 46(11): 120-126. [doi:10.7652/xjtuxb201211022]

(编辑 武红江)