

广告

搜狐 公众平台 >IT

他一出家就成中国最帅和尚
眼眸深邃、轮廓分明、身材颀长，活生生的一幅画。

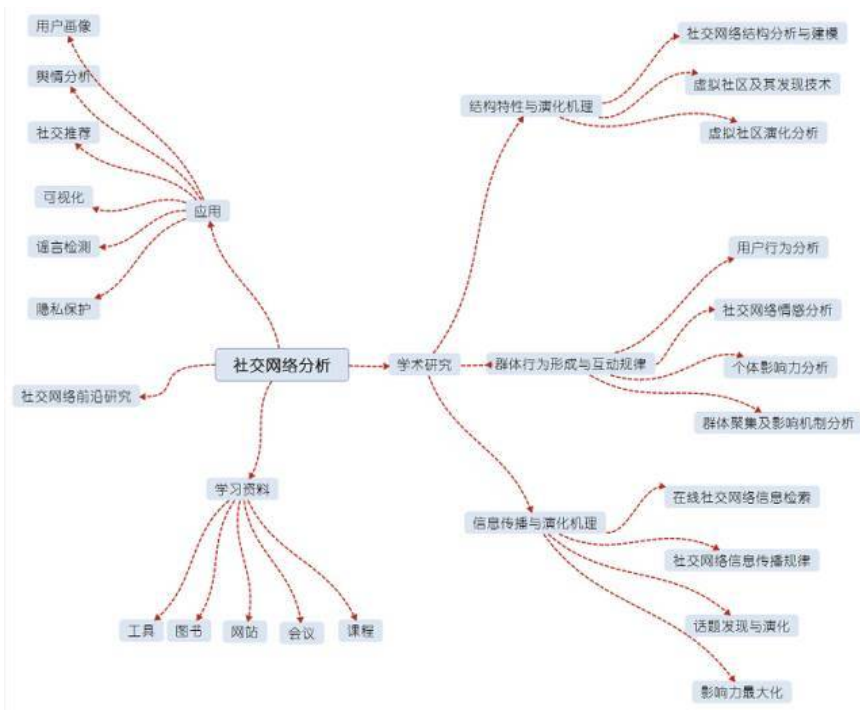
大学副教授与在押服刑女结婚
这在监狱民警看来，那么令人不可思议。

一文读懂社交网络分析（附应用、前沿、学习资源）

PPV课大数据 2017-10-10 08:00:00 社交 阅读(0) 评论(0)

声明：本文由入驻搜狐公众平台的作者撰写，除搜狐官方账号外，观点仅代表作者本人，不代表搜狐立场。

举报



本文主要阐述：

- 社交网络的结构特性与演化机理
- 社交网络群体行为形成与互动规律
- 社交网络信息传播与演化机理
- 社交网络分析的应用
- 社交网络前沿研究
- 学习资源
- 参考资料

前言

社交网络在维基百科的定义是“由许多节点构成的一种社会结构。节点通常是指个人或组织，而社交网络代表着各种社会关系。”在互联网诞生前，社交网络分析是社会学和人类学重要的研究分支。早期的社交网络的主要指通过合作关系建立起来的职业网络，如科研合作网络、演员合作网络等。

本文所指的社交网络分析特指在线社交网络分析（Online Social Network Analysis），该门科学的发

PPV课大数据

0

阅读量

387

文章数

0

评论数

一文读懂社交网络分析（ 搜狗搜索

热词： 死亡航班 饲养蜘蛛侠 夺命房间 引力双眼皮



热门图片

大谈床事不害臊的女星

女星遭有钱人玩弄抛弃

女星因怀孕不得已结婚

清纯玉女突变豪放欲女

热点视频 影视剧 综艺 原创

看黄片副局长已被停职

小学开课教撒蛋合适吗

空姐招聘现场美腿抢镜

准妈妈堕胎捐骨髓救妹

辣条生产场地环境肮脏

百岁老太高空跳伞庆生

<屌丝男士4>爆笑开机

刘欢回应二当冠军导师

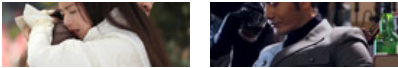
立刻说两句吧！

扎克伯格为脸书诞

搜狐 首页

登录 注册 我的搜狐 邮件

Facebook、Google+ 等），微博类应用（新浪微博、腾讯微博、Twitter 等），共享空间类应用（论坛、博客、视频分享、评价分享等）。



神雕侠侣-大结局

同步热播-锦绣缘

在线社交网络（下文统称社交网络）有着迅捷性、蔓延性、平等性与自组织性等四大特点。正因为这些特性，其在互联网出现的短短数十年内已经拥有数十亿用户并对现实社会的方方面面产生着影响。在2016年的美国总统大选中，当选总统特朗普就很好地利用了推特作为宣传工具；而在国内，从魏则西事件到和颐酒店事件再到最近的“刺死辱母者”事件，无一不是在社交网络上迅速发酵，并最终对现实社会产生影响。而且这种线上影响线下的趋势越来越明显。

除了社交网络给社会和经济带来许多正面影响之外，也带来了不少负面影响。从Facebook 和YouTube上的暴力恐怖信息传播到微博微信上大量谣言和假新闻，这些有害信息借助社交网络的特点迅速传播并且往往产生不可控的后果。

为了利用好社交网络的特性，产生价值，消除危害，所以产生了社交网络分析这门科学。它是一种基于信息学、数学、社会学、管理学和心理学等科学的交叉科学。根据社交网络的特性，其主要研究三大内容：结构与演化，群体与互动，信息与传播。

本文简要概述了社交网络分析领域各个研究方向，对于细节性的内容我只列出参考文献，在文章最后提供了一些学习资源。希望通过阅读本文，对这个领域感兴趣的读者可以对社交网络分析有一个宏观理解并且找到学习的方向。笔者作为社交网络分析的初学者，对某些概念和事实的解释和陈述不免有错误之处，还望各位读者能及时指正，大家共同交流进步。

客服热线：86-10-58511234

客服邮箱：kf@vip.sohu.com

一． 社交网络的结构特性与演化机理

1. 社交网络结构分析与建模

1.1 统计特性

社交网络模型许多概念来自于图论，因为社交网络模型本质上是一个由节点（人）和边（社交关系）组成的图。笔者将简要介绍社交网络模型中常用的统计概念。

- **度（Degree）**：节点的度定义为与该节点相连的边的数目。在有向图中，所有指向某节点的边的数量叫作该节点的入度，所有从该节点出发指向别的节点的边的数量叫作该节点的出度。网络平均度反应了网络的疏密程度，而通过度分布则可以刻画不同节点的重要性。
- **网络密度（Density）**：网络密度可以用于刻画节点间相互连边的密集程度，定义为网络中实际存在边数与可容纳边数上限的比值，常用来测量社交网络中社交关系的密集程度及演化趋势。
- **聚类系数（Clustering Coefficient）**：用于描述网络中与同一节点相连的节点间也互为相邻节点的程度。其用于刻画社交网络中一个人朋友们之间也互相是朋友的概率，反应了社交网络中的聚集性。
- **介数（Betweenness）**：为图中某节点承载整个图所有最短路径的数量，通常用来评价节点的重要程度，比如在连接不同社群之间的中介节点的介数相对于其他节点来说会非常大，也体现了其在社交网络信息传递中的重要程度。

1.2 网络特性

- **小世界现象**：小世界现象是指地理位置相距遥远的人可能具有较短的社会关系间隔。早在1967年，哈佛大学心理学教授 Stanley Milgram 通过一个信件投递实验，归纳并提出了“六度分割理论（Six Degrees of Separation）”，即任意两个都可通过平均五个人熟人相关联起来。1998年，Duncan Watts 和 Steven Strogatz 在《自然》杂志上发表了里程碑式的文章《Collective Dynamics of “Small-World” Networks》，该文章正式提出了小世界网络的概念并建立了小世界模型。

小世界现象在在线社交网络中得到了很好地验证，根据2011年 Facebook 数据分析小组的报告，[Facebook 的7.2亿用户中任意两个用户间的平均路径长度仅为4.74，而这一指标在推特中为4.67，可](#)

立刻说两句吧！

扎克伯格为脸书道

- **无标度特性：**大多数真实的大规模社交网络都存在着大多数节点有少量边，少数节点有大量边的特点，其网络缺乏一个统一的衡量尺度而呈现出异质性，我们将这种节点度分布不存在有限衡量分布范围的性质称为无标度。无标度网络表现出来的度分布特征为幂律分布，这就是此类网络的无标度特性。

1.3 网络模型

- **WS 模型：**WS 模型即小世界模型，通过小世界模型生成的小世界网络是从规则网络向随机网络过渡的中间形态。
- **BA 模型：**BA模型考虑到现实网络中节点的幂律分布特性，生成无标度网络。
- **其他模型：**森林火灾模型，Kronecker 模型，生产模型。



2. 虚拟社区（社团）及发现技术

2.1 定义

虚拟社区基于子图局部性的定义：社区结构是复杂网络节点集合的若干子集，每个子集内部的节点之间的连接相对非常紧密，而不同子集节点之间的连边相对稀疏。

在社交网络中发现虚拟社区有助于理解网络拓扑结构特点，揭示复杂系统内在功能特性，理解社区内个体关系。为信息检索、信息推荐、信息传播控制和公共事件管控提供有力支撑。虚拟社区发现存在着许多经典的算法，这些算法用于挖掘不同规模的虚拟社区，算法在追求高精度的同时力求提高效率（降低时间复杂度）。

2.2 社区发现算法评价指标

以下评价指标可通过搜索引擎获得详细的介绍：

- **模块度(Modularity)：**通过比较现有网络与基准网络在相同社区划分下的连接密度差来衡量网络社区的优劣。
- **NMI (Normalized Mutual Information)：**利用信息熵来衡量预测社区结构一直社区结构的差异，该值越大，则说明社区结构划分越好，最大值为1时，说明算法划分出的社区结构和一直社区结构一致，算法效果最好。
- **Rand Index：**表示在两个划分中都属于同一社区或者都属于不同社区的节点对的数量的比值。
- **Jaccard Index：**Jaccard 系数用来衡量样本之间的差异性，是经典的衡量指标。

Mark Newman 提出了针对模块度的最大化的贪心算法FN。可参考文献：Newman,Mark EJ. "Fast algorithm for detecting community structure in networks." Physical review E 69.6 (2004): 066133.

- 多目标优化算法

Zhao, Yuxin, et al. "A cellular learning automata based algorithm for detecting community structure in complex networks." Neurocomputing 151 (2015): 1216–1226.

Du, Jingfei, Jianyang Lai, and Chuan Shi. "Multi-Objective Optimization for Overlapping Community Detection." International Conference on Advanced Data Mining and Applications. Springer, Berlin, Heidelberg, 2013.

- 基于概率模型的算法

Newman, Mark EJ, and Elizabeth A. Leicht. "Mixture models and exploratory analysis in networks." Proceedings of the National Academy of Sciences 104.23 (2007): 9564–9569.

Ren, Wei, et al. "Simple probabilistic algorithm for detecting community structure." Physical Review E 79.3 (2009): 036111.

- 信息编码算法

Rosvall, Martin, and Carl T. Bergstrom. "Maps of random walks on complex networks reveal community structure." Proceedings of the National Academy of Sciences 105.4 (2008): 1118–1123.

Kim, Youngdo, and Hawoong Jeong. "Map equation for link communities." Physical Review E 84.2 (2011): 026110.

2.4 社区动态发现算法

- 派系过滤算法

Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society." arXiv preprint physics/0506133 (2005).

Kumpula, Jussi M., et al. "Sequential algorithm for fast clique percolation." Physical Review E 78.2 (2008): 026109.

- 基于相似度的聚合算法

Shen, Huawei, et al. "Detect overlapping and hierarchical community structure in networks." Physica A: Statistical Mechanics and its Applications 388.8 (2009): 1706–1712.

Huang, Jianbin, et al. "Density-based shrinkage for revealing hierarchical and overlapping community structure in networks." Physica A: Statistical Mechanics and its Applications 390.11 (2011): 2160–2171.

- 标签传播算法

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. "Near linear time algorithm to detect community structures in large-scale networks." Physical review E 76.3(2007): 036106.

Gregory, Steve. "Finding overlapping communities in networks by label propagation." New Journal of Physics 12.10(2010): 103018.

- 局部扩展优化算法

Lancichinetti, Andrea, and Santo Fortunato. "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities." Physical Review E 80.1 (2009): 016118.

Lee, Conrad, et al. "Detecting highly overlapping community structure by greedy clique expansion." arXiv preprint arXiv:1002.1827 (2010).



3. 虚拟社区演化分析

在线社交网络中存在着大量显性或者隐性的虚拟社区结构，这些虚拟社区结构并不是永恒不变的，随着事件变化，社区结构也在不断演变。分析动态的虚拟社区结构演化有助于理解整个社交网络的演化过程，所以有着重要的研究价值。

3.1 虚拟社区的涌现

虚拟社区涌现即在社交网络中虚拟社区从无到有的过程，其最重要的特征是网络聚集现象。

- **周期闭包**：所谓周期闭包，是指网络节点倾向于和自己在网络中邻居的邻居建立连接关系而形成的结构，该机制是导致虚拟社区形成的主要因素。实验表明三元闭包的出现概率随着两个节点之间测地距离的增减呈指数递减。相反地，焦点闭包和测地距离无关，其生成原因是两个节点之间有共同的兴趣或参与共同的活动。
- **偏好连接**：在很多真实网络中，新增加的边并不是随机连接的，而是倾向于和具有较大度数的连接。

3.2 虚拟社区的演化

在线社交网络虚拟社区演化过程非常复杂，影响因素很多。如何挖掘虚拟社区演化中的关键性因素成为社交网络研究中一个重要而有挑战性的课题，用户个体的累积效应、结构多样性和结构平衡性三个基本因素对虚拟社区演化都存在影响。

3.3 演化虚拟社区的发现

演化虚拟社区发现目前已有大量的研究资料，以下五种是比较成熟的算法模型，具体细节和根据参考文献进一步了解。

- 基于相邻时刻相似度直接比较的演化虚拟社区发现

Hopcroft, John, et al. "Tracking evolving communities in large linked networks." Proceedings of the National Academy of Sciences 101.suppl 1 (2004): 5249–5253.

Greene, Derek, Donal Doyle, and Pdraig Cunningham. "Tracking the evolution of communities in dynamic social networks." Advances in social networks analysis and mining (ASONAM), 2010 international conference on. IEEE, 2010.

- 基于演化聚类分析的演化虚拟社区发现

Chakrabarti, Deepayan, Ravi Kumar, and Andrew Tomkins. "Evolutionary clustering." Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006.

Lin, Yu-Ru, et al. "Facetnet: a framework for analyzing communities and their evolutions in dynamic networks." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

- 基于拉普拉斯动力学方法的演化虚拟社区发现

Lambiotte, Renaud, J-C. Delvenne, and Mauricio Barahona. "Laplacian dynamics and multiscale modular structure in networks." arXiv preprint arXiv:0812.1770 (2008).

- 基于派系过滤算法的演化虚拟社区发现

Palla, Gergely, Albert-Laszlo Barabasi, and Tamas Vicsek. "Quantifying social group evolution." Nature 446. arXiv: 0704.0744 (2007): 664.

- 基于节点行为趋势分析的演化虚拟社区发现

Hopcroft, John, et al. "Tracking evolving communities in large linked networks." Proceedings of the National Academy of Sciences 101.suppl 1 (2004): 5249–5253.

二. 社交网络群体行为形成与互动规律

1. 用户行为分析

社交网络用户行为是用户对自身需求，社会影响和社交网络技术进行综合评估的基础上做出的使用社交网络服务的意愿，以及由此引起的各种使用活动的总和。用户行为是在线社交网络研究的重要内容。现有研究主要基于如下两种思路展开，一是将在线社交网络作为一种特定的信息技术，研究用户对在线社交网络技术的采纳行为、拒绝行为和用户忠诚；二是将在线社交网络视为提供各种服务和应用的平台，研究用户使用各种服务和应用所表现出的特征与规律。

1.1 用户采纳与忠诚

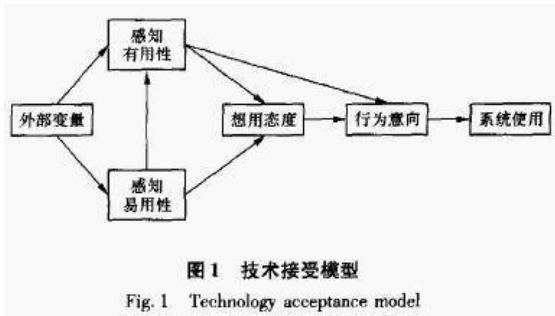
在线社交网络用户采纳是指用户在对自身需求、社会影响和在线社交网络技术进行综合评估的基础上做出的使用在线社交网络服务的意愿或行为，在线社交网络再出现初期能否被尽可能多的用户采纳和试用对于其后续的扩散至关重要。目前已有多种理论被用于揭示在线社交网络用户采纳行为机理。其中，技术接受模型和计划行为理论是研究者们应用最多的两种理论。

在线社交网络用户忠诚是指用户在使用社交网络服务之后，能够继续保持使用的习惯。各种层出不穷的新型网络服务所带来的竞争压力让保持在线社交网络用户忠诚度愈发困难。目前为止，已经有多种理论被用于在线社交网络的用户忠诚研究。其中，期望确认理论和心流体验理论受到较多研究者青睐。

• 基于技术接受模型的在线社交网络用户采纳模型

David Fred 提出技术接受模型是目前信息系统研究领域最经典的模型之一。对模型详细了解可参考：

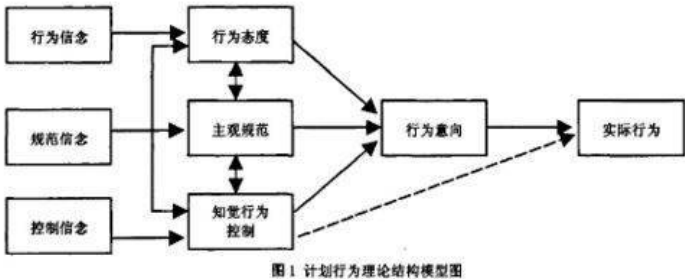
Davis, Fred D. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." MIS quarterly (1989): 319–340.



• 基于计划行为理论的在线社交网络用户采纳模型

Icek Ajzen 提出的计划行为理论已经被广泛用于人类行为研究。对理论详细了解可参考：

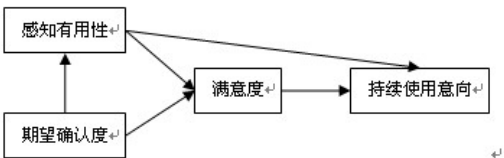
Ajzen, Icek. "From intentions to actions: A theory of planned behavior." Action control. Springer Berlin Heidelberg, 1985. 11–39.



• 基于期望确认理论的在线社交网络用户忠诚模型

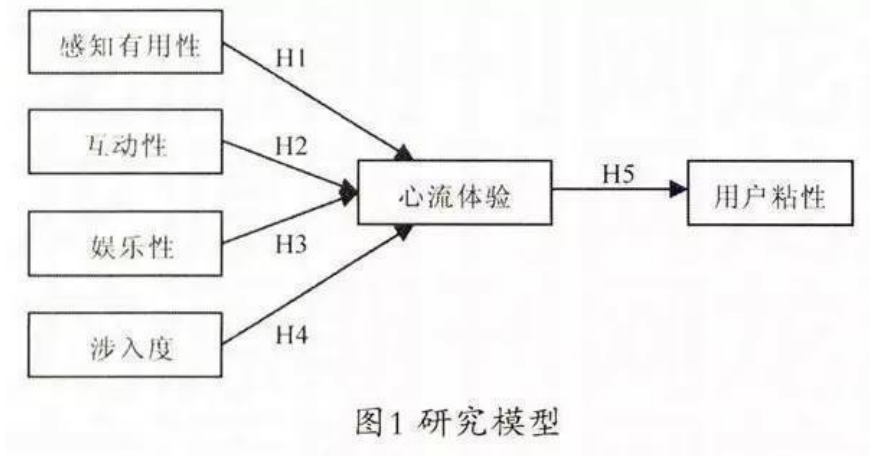
由 Oliver 提出的期望确认理论是研究消费者满意度的基本理论。Anol Bhattacharjee 再该理论的基础上结合信息系统的特点提出了信息系统持续使用的期望确认模型（ECM-ISC）。对模型详细了解可参考：

Bhattacharjee, Anol. "Understanding information systems continuance: an expectation–confirmation model." MIS quarterly (2001): 351–370.



• 基于心流体验理论的在线社交网络用户忠诚模型

Mihaly Csikszentmihalyi等提出的心流体验理论是目前关于用户体验研究的重要理论。对理论详细了解



1.2 用户个体使用行为

- **一般使用行为：**用户可以在社交网络上执行各种各样的行为，例如浏览，点击，分享，点赞，收藏等等。具体的分类可参考：Benevenuto F, Rodrigues T, Cha M, Almeida V. Characterizing User Behavior in Online Social Networks. New York, New York, USA: ACM; 2009:49–62. doi:10.1145/1644893.1644900.
- **内容创建行为：**用户在社交网络通过写博客微博，发帖评论等行为产生内容，对内容创建行为的研究主要研究创建内容的动机、创建内容时的主题选择偏好以及内容创建时的语言表述等。关于主题，可通过搜索引擎搜索 LDA 模型。
- **内容消费行为：**用户在社交网络中通过浏览，分享和评论来满足他们的社交需求，对社交网络内容的消费可分为主动消费和被动消费。被动消费即“浏览”，有研究表明，社交网络中高达92%的行为都是浏览行为。主动消费即社交搜索，例如搜索朋友的信息以及向社交圈内好友提问等等。

1.3 用户群体互动行为

- **群体互动关系选择：**对群体互动关系的研究主要是识别用户之间的关系，通过制定不同的衡量指标，研究用户之间的关系强弱。
- **群体互动的内容选择：**社交网络中用户对内容选择与其社交关系密不可分。例如有研究表明两位维基百科编辑在互动前后产生的编辑内容的相似性有所不同。
- **群体互动的的时间规律：**在线社交网络中人类行为的时间特征研究主要集中于分析行为发生的时间间隔分布。研究发现在线社交网络中用户行为时间间隔分布不同于传统的负指数分布，而是呈现幂律分布，即具有“长尾效应”。对群体互动时间规律的研究可以应用到公共管理和决策等场景中。

2. 社交网络情感分析

随着互联网技术的迅速发展，网络已经成为人们获取信息，发表意见的主要途径，根据文本内容，我们可以将网络中的文本分为两种，一种是客观描述信息，主要针对事件、产品等进行客观描述，另一种是主观性信息，主要产生与用户对人物、事件、产品进行客观性描述；另一种是主观性信息，主要产生于用户对人物、事件、产品等的评价信息。主观性信息表达了人们的各种情感色彩和情感倾向，如“支持”、“反对”、“中立”等。

情感分析，在此等同于意见挖掘，是针对主观性信息进行分析、处理和归纳过程。情感分析最初起源于自然语言处理领域，主要从语法语义规则方面对文本的情感倾向性进行研判。随着社交网络的兴起与发展，情感分析逐渐涉及多个研究领域，如文本挖掘、Web 数据挖掘等，并延伸至管理学及社会科学等学科，并在产品评论、舆情监控、信息预测等多个领域发挥着重要的作用。

2.1 文本情感分析技术

典，这些词语可以代表用户的某种倾向性。基于语义规则的分析技术是计算评价词和情感词典中已经标注倾向性词语的距离，从而达到情感分类的目的。其最经典的算法是 SO-PMI 算法。

- **基于监督学习的情感分析方法：**基于监督学习的方法是首先通过人工标注文本的情感极性，然后将此作为训练集，通过机器学习的方法对目标文本进行情感分类。常用方法：朴素贝叶斯，支持向量机。
- **基于话题模型的情感分析技术：**有两个话题模型，PLSA (Probabilistic Latent Semantic Analysis)和 LDA (Latent Dirichlet Allocation) 模型，网络上有大量的学习资料可供读者进一步了解。

2.2 社交网络情感分析技术

- **面向短文本的情感分析技术：**社交网络产生大量的短文本，例如微博和新闻评论，论坛帖子等等，这些短文本不同于新闻报道，其语法不规则，充斥大量噪声，因此对短文本的分析非常重要。
- **基于群体智能的情感分析技术：**用户在社交网络中表达意见会受到其社交关系的影响，情感会沿着社交关系进行传播，因此可以通过研究社交用户之间的关系来提高情感分析的准确度。
- **社交网络的垃圾意见挖掘技术：**社交网络中的垃圾意见，包括水军与广告等信息，通过对垃圾意见的挖掘，能够有效区分有效信息和垃圾信息，从而提高社交网络使用体验。

3. 个体影响力分析

发现社交网络中的有影响力的个体是社交网络研究中非常重要的研究分支，而且其有着重要的应用价值。例如微博营销，谣言检测，舆情管理等等。

1.1 基于网络结构的个体影响力计算

基于社交网络的图结构特性，有几个指标用来衡量网络中节点的中心度，即节点的影响力。除了以下三种外还有 PageRank 中心度等度量方法。

- **度中心度 (Degree Centrality)：**度中心度是指与该节点直接相连的节点的数量。
- **接近中心度 (Closeness Centrality)：**指某节点与网络中所有其他节点的最短距离之和。
- **介数 (Betweenness Centrality)：**介数用来衡量某节点在社交网络中中介作用大小。网络中某两个节点所有最短路径的数量除以这些路径中经过 A 节点路径的数量便是 A 节点的介数，也叫中间中心度。

1.2 基于行为的个体影响力计算

社交网络中用户的行为决定用户的影响力，以微博为例，用户主要表现的行为是评论、转发、回复、点赞、复制、阅读等等，基于这些行为特征构建多种网络关系图，可通过随机游走等方法发现网络中的影响力个体。

1.3 基于话题的个体影响力计算

在社交网络中用户在不同话题下的影响力不同，可以根据用户的关注网络 and 用户兴趣相似性来计算用户在每个话题上的影响力。

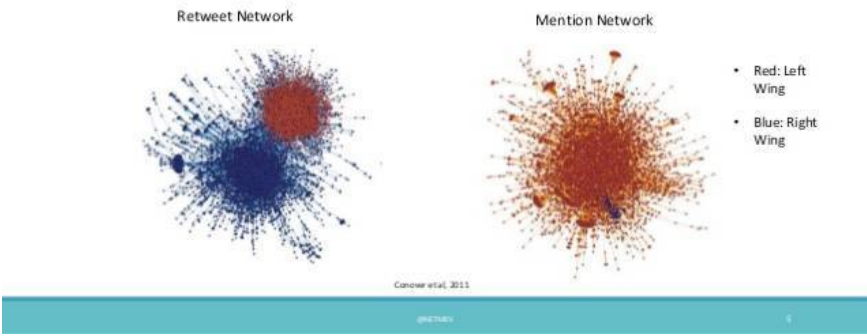
4. 群体聚集及影响机制分析

本部分主要介绍群体极化的概念。群体极化是指在群体决策的情境中，个体意见或决定往往会受到群体间的彼此讨论的影响，而产生一个群体性的结果。群体极化往往表现为群体内的个体不经过个人思考而同意大多数人的观点。群体极化是一个社会心理学概念，在社会学名著《乌合之众》中提到的大众心理状态就是群体极化的体现

群体极化产生的条件可概括为四点：第一，必须有激发事件出现；第二，群体内的个人能看到前人的选择；第三，群体信息缺乏；第四，群体有一定的同质性。

在在线社交网络分析中，人们通过建立分析模型和仿真来研究在线社交网络中的群体极化现象。主要的分析模型有基于博弈论和委托—代理理论的从众行为模型，基于信息瀑的群体一致性模型和基于元胞自动机群决策和行为仿真。

Twitter and Politics: Polarization?



Twitter 中政治观点的极化[4]

三． 社交网络信息传播与演化机理

1. 在线社交网络信息检索

信息检索(Information Retrieval) 是从大规模非结构化数据中获取信息的过程，例如搜索引擎就是典型的信息检索技术的应用。在线社交网络数据结构有其特殊性，以微博的“话题”（#话题名称#）为例，这种新型的信息组织方式是传统信息检索研究没有涉及的，所以对社交网络信息的检索成为了一门研究课题。

1.1 社交网络内容搜索

内容搜索是指给定查询，从大量信息中返回相关信息的过程。例如在微博上搜索相关热点事件名称，能够返回关于热点事件的微博。内容搜索是信息检索最经典的应用形式。经典的信息检索模型有向量空间模型（VSM），概率模型及 BM25检索公式，基于统计建模检索模型及查询拟然模型，基于统计语言建模的检索模型等。

针对微博的内容检索建模，目前有两种主要的方法:

- 时间先验方法：**时间先验是由于语料库中的文档具有不同的重要性，考虑语料库背景定义不同的计算公式，再将计算结果用于检索模型以期得到更好的检索效果的一种检索方法。目前考虑时间信息计算文档先验的研究工作可分为两种：一种定义文档的时间变化关系；另一种为修改 PageRank 的方法，在其中加入时间关系。具体细节可参考：
Li, Xiaoyan, and W. Bruce Croft. "Time-based language models." Proceedings of the twelfth international conference on Information and knowledge management. ACM, 2003.
Yu, Philip S., Xin Li, and Bing Liu. "On the temporal dimension of search." Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. ACM, 2004.
- 多特征组合的方法：**多特征组合方法是通过组合多个微博特性来检索微博内容。下面的参考文献中提到的微博特性有：微博个数，关注数，粉丝数，微博长度，微博是否含有外链。具体细节可参考：

approaches for microblog search." Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on. Vol. 1. IEEE, 2010.

1.2 社交网络内容分类

面向文本的分类称为文本分类。分类包括训练和测试两阶段，简单地说，训练是根据已标注类别的语料来学习分类规则或规律的过程。而测试是将已训练好的分类器用于新文本的过程。不管是训练还是测试，都需要将分类对象进行特征表示，然后利用分类算法进行学习或者分类。以下社交网络中内容主题分类的相关参考文献，读者可自行查阅。

Liu, Zitao, et al. "Short text feature selection for micro-blog mining." Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on. IEEE, 2010.

Yuan, Quan, Gao Cong, and Nadia Magnenat Thalmann. "Enhancing naive bayes with various smoothing methods for short text classification." Proceedings of the 21st International Conference on World Wide Web. ACM, 2012.

Ling, Xiao, et al. "Can chinese web pages be classified with english data source?." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

Zhang, Dan, et al. "Transfer Latent Semantic Learning: Microblog Mining with Less Supervision." AAAI. 2011.

1.3 社交网络推荐

- **协同过滤推荐：**传统的协同过滤根据用户（user）和物品（item）信息构建矩阵，根本的原则是相似用户的选择也相似，例如 a 和 b 都喜欢 m，其中 a 还喜欢 n，那么 b 也有可能喜欢 m。在社会化协同过滤推荐中，我们可以利用用户之间的社交关系，弥补协同过滤矩阵中缺失的内容，从而使协同过滤的结果更加精准。

- **基于模型的推荐：**

- 邻居模型：

Ma, Hao, et al. "Sorec: social recommendation using probabilistic matrix factorization." Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008.

- 矩阵分解模型：

Funk, Simon. "Netflix update: Try this at home." (2006).

- 融入社交网络信息：

Jamali, Mohsen, and Martin Ester. "A matrix factorization technique with trust propagation for recommendation in social networks." Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010.



2. 社交网络信息传播规律

信息传播是人们通过符号、信号、传递、接收与反馈信息的活动，是人们彼此交换意见、思想、情感，已达到互相了解和影响的过程。社交网络信息传播是指以社交网络为媒介进行信息传播的过程。研究社交网络信息传播的规律，有助于我们加深对社交系统的认识，理解社交现象。也有助于模式发现，大影响力节点识别和个性化推荐。下面主要介绍几种社交网络信息传播模型。

2.1 基于网络结构的传播模型

- 线性阈值模型（Linear Threshold）：

Granovetter, Mark. "Threshold models of collective behavior." American journal of sociology 83.6 (1978): 1420–1443.

- 独立级联模型（Independent Cascade）：

Goldenberg, Jacob, Barak Libai, and Eitan Muller. "Talk of the network: A complex systems look at the underlying process of word-of-mouth." Marketing letters 12.3 (2001): 211–223.

2.2 基于群体状态的传播模型

传染病模型（SI, SIS, SIR），传染病模型是经典的信息传播模型，网上有丰富的参考资料。

- 线性影响力模型（Linear Influence Model）：

Yang, Jaewon, and Jure Leskovec. "Modeling information diffusion in implicit networks." Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.

2.3 基于信息特性的传播模型

[在线社交网络中的信息承载着用户网上活动的所有记录，在信息传播分析时起着不可或缺的重要作用](http://mt.sohu.com/20171010/n516921964.shtml)

立刻说两句吧！

扎克伯格为脸书道歉

Myers, Seth A., Chenguang Zhu, and Jure Leskovec. "Information diffusion and external influence in networks." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

Beutel, Alex, et al. "Interacting viruses in networks: can both survive?." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

此外，对社交网络信息传播规律的研究还包括热度预测和信息溯源。感兴趣的读者可自行查阅了解。

3. 话题发现与演化

在话题发现和演化的大部分研究中，话题是指一个引起关注的事件或活动，及其所有相关事件和活动。其中，事件或者活动是指在一个特定的时间和地点，发生的一些事情。社交网络语料库中的数据和传统话题发现语料库的数据区别较大，所以我们必须使用新的方法或对传统方法进行改进来适应社交网络数据特点。

一般社交网络例如 Twitter 的数据有以下特点：数据规模大、内容简短、噪声多、数据特征丰富等。下面介绍几种主要的话题发现和演化模型。

3.1 基于主题模型的话题发现

- 最具有代表性的主题发现模型——LDA

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993–1022.

3.2 基于向量空间模型的话题发现

Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." Communications of the ACM 18.11 (1975): 613–620.

Becker, Hila, Mor Naaman, and Luis Gravano. "Beyond Trending Topics: Real-World Event Identification on Twitter." ICWSM 11.2011 (2011): 438–441.

3.3 基于词项关系图的话题发现

词项共现是自然语言处理技术在信息检索中的成功应用之一。它的核心思想是词项之间的共现频率在某种程度上反映了词项的语义关联。最初学者们利用词项共现来计算文档的相似性，随后学者们利用该方法来完成话题词提取，话题句提取和摘要生成任务。

Sayyadi, Hassan, Matthew Hurst, and Alexey Maykov. "Event detection and tracking in social streams." Icwsm. 2009.

3.4 基于主题模型的话题演化

Yin, Zhijun, et al. "LPTA: A probabilistic model for latent periodic topic analysis." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

Wang, Xiaolong, Chengxiang Zhai, and Dan Roth. "Understanding evolution of research themes: a probabilistic generative model for citations." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

3.5 基于相邻时间片关联的话题演化

Lin, Cindy Xide, et al. "The joint inference of topic diffusion and evolution in social communities." Data Mining (ICDM), 2011 IEEE 11th International Conference on. IEEE, 2011.

Saha, Ankan, and Vikas Sindhwani. "Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization." Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012.

4. 影响力最大化

影响力最大化是在社交网络中选定信息初始传播用户，使得信息的传播范围能达到最大，即影响力最大。影响力最大化算法的目的就是找出一定数量的用户作为影响力传播的初始节点。对影响力最大化的问题的建模是基于社交网络信息传播模型的。其中最经典的模型是线性阈值和独立级联模型。

影响力最大化算法被证明为 NP-hard问题，下面主要介绍两种典型的影响力最大化算法。

4.1 贪心算法

贪心算法从单个节点开始，计算每选一个新节点作为初始节点对每个节点带来的边际收益，取能造成边际收益最大的点加入初始节点集合。贪心算法的缺点是计算时间成本较大，但是计算精度较高。

Kempe, David, Jon Kleinberg, and Éva Tardos. "Maximizing the spread of influence through a social network." Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

Chen, Wei, Yajun Wang, and Siyu Yang. "Efficient influence maximization in social networks." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

4.2 启发式算法

不同于贪心算法选择任何一个点作为初始节点开始计算，启发式算法先通过一定策略选取一定数量的初始节点，然后计算其影响力传播。其优点是速度快，缺点是精度低。

Chen, Wei, Yajun Wang, and Siyu Yang. "Efficient influence maximization in social networks." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2009.

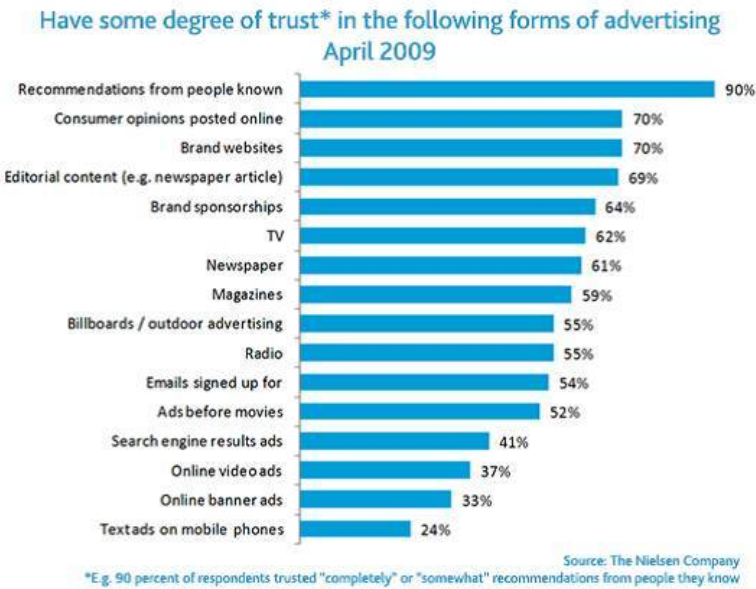
Chen, Wei, Yifei Yuan, and Li Zhang. "Scalable influence maximization in social networks under the linear threshold model." Data Mining (ICDM), 2010 IEEE 10th International Conference on. IEEE, 2010.

四. 社交网络分析的应用

1. 社交推荐

社交推荐顾名思义是利用社交网络或者结合社交行为的推荐，具体表现为推荐 QQ 好友，微博根据好友关系推荐内容等。在线推荐系统最早被亚马逊用来推荐商品，如今，推荐系统在互联网已无处不在，目前大热的概念“流量分发是互联网第一入口”，支撑这个概念有两点核心，其一是内容，另外就是推荐，今日头条在短短几年间的迅速崛起便是最好的证明。

根据推荐系统推荐原理，社交推荐可定义为一种“协同过滤”推荐，即不依赖于用户的个人行为，而是结合用户的好友关系进行推荐。对于互联网上的每一个用户，通过其社交账户能很快定义这个用户众多特点，再加之社交网络用户数之多，使得利用社交关系的推荐近些年备受关注。



人们更愿意接受来自朋友的推荐，来源：尼尔森

笔者所了解到的研究有，根据不同社交网络之间进行信息匹配进而进行推荐，有根据社交关系解决新注册用户的冷启动问题等。总之社交推荐在内容分发、广告宣传等领域有着十分重要的地位。具体应用细节大家可以关注笔者的一篇介绍腾讯社交广告的文章（http://mp.weixin.qq.com/s/mLpNoMdBpDAEb5IZB_A3Rg），如果了解这方面更多信息还可以关注推荐系统领域顶级会议ACM RecSys。

2. 舆情分析

舆情分析在互联网出现之前就被广泛应用在政府公共管理，商业竞争情报搜集等领域。在社交媒体出现之前，舆情分析主要是线下的报纸，还有线上门户网站的新闻稿件，这些信息的特点是相对专业准确，而且易于分析和管理；但随着社交媒体出现，舆情事件第一策源地已经不是人民日报新华社这样的大媒体，而是某一个名不见经传的微博用户，一个个人微信公众号。他们的特点是信息非常新鲜，缺点是真实度较低且传播十分迅速，难以控制。所以在社交网络下的舆情分析是一门新的学问。



“刺死辱母者”微博转发趋势，来源见水印

举几个例子，去年的和颐酒店，今年的北京地铁骂人事件这类急性舆情事件最早就是在微博上爆出，而且在短时间内迅速传播。还有去年的关于快手的“中国农村残酷底层物语”，今年的“北京房价”这类民生话题，也是在微信公众号逐渐发酵。

当然，在新形势下的舆情应对，也已经有新的工具，大家百度“舆情分析平台”或者“舆情分析软件”可以找出一大堆。比较有名的有蚁坊、红麦、清博、知微、新榜等等。一些传统的舆情分析机构开始转型做“大数据”的舆情分析，也有近年来完全基于社交媒体的舆情平台，比如基于微信的新榜和基于微博的知微。除此之外，BAT 等大型平台有自己舆情分析工具，可以私人订制，也有开放的指数（百度指数、微信指数）。

3. 隐私保护

隐私问题在互联网时代已经是老生常谈的问题了。在社交网络中，作为用户，我们可能会留下大量痕迹，这些痕迹有隐性的，也有显性的，好不夸张地，社交服务提供商可以根据你的少量痕迹，挖掘到大量你的个人信息，有些信息是你不愿意别人知道的。

这其中存在一个矛盾，即社交服务提供商处于商业目的想尽可能获取你的个人信息，但是你又担心自己的个人信息被泄露。所以在隐私保护领域，一方面要设计足够安全的机制，技术层面的，法律层面的，在保护个人隐私的前提下最大化商业利益和用户的体验。

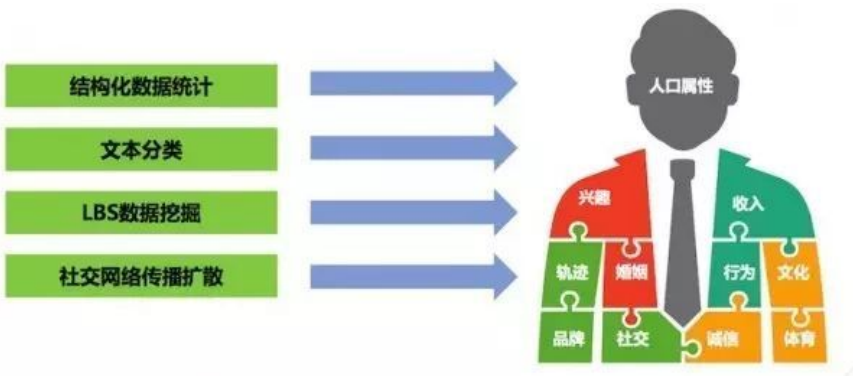


“云端”的隐私，来源：<http://s9.sinaimg.cn>

举一个大家比较熟悉的例子，即许多网站注册账户的时候使用微信、支付宝账户验证，即免去了大家填写个人信息的烦恼，又保护了大家的隐私。同理，蚂蚁金服提供的芝麻信用功能也有隐私保护的功能。

目前学界对隐私保护的研究主要还是从技术层面设计完善的隐私保护机制。

4. 用户画像



一种用户画像流程，来源：<http://www.51callcenter.com>

用户画像，这是个营销术语，即通过研究用户的资料和行为，将其划分为不同的类型，进而采取不同的营销策略。传统的用户画像最常用的手段就是调查问卷，订阅过杂志和报纸的读者都知道，会有各种各样的有奖问卷，一方面用来获得对于产品的反馈，另一方面就是对你进行画像，这些画像资料甚至广泛在黑市流通，这就是你为什么有时候会接到莫名其妙的电话的原因（又扯到了隐私保护问题）。

在社交网络，用户画像方式变得更多了，除了传统的线下问卷变成在线问卷。我们通过用户的行为，一方面通过统计学方法获得一些用户特征（经典的例子是沃尔玛的“啤酒和尿布”，另一方面通过机器学习进行建模和验证获得意外的收获（参见上面提到的腾讯社交广告文章）。

接触过微信公众号后台的读者都知道，公众号后台对微信公众号文章的读者还有公众号粉丝的画像已经做得非常充足了，好像微博会员也有粉丝画像的功能。这些便捷的功能对于媒体运营者和广告投放者都有非常重要的作用。

5. 谣言检测

谣言检测算是舆情分析的一部分，之所以单独提出来是因为这部分非常重要，而且谣言的确定对于舆情管理非常重要。早起微博因为充斥着大量谣言，使得新浪微博不得不推出“微博辟谣”官方账号，到如今微博以及有许多自发和官方的辟谣账号，微信公众号也是如此。



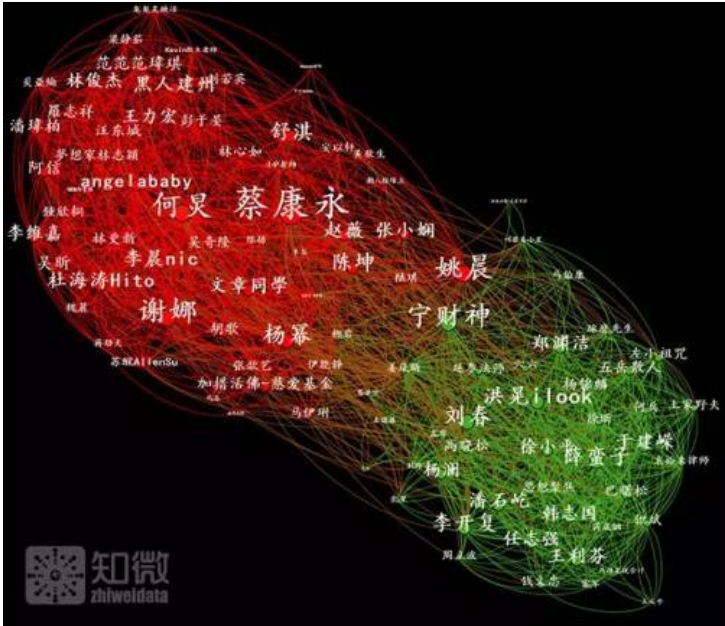
“六小龄童春晚被拒”谣言传播走势，来源见水印

传统辟谣方法无非是进行试试检验，用证据说话，随着现在机器学习技术的迅速发展，我们也可以通过信息传播的轨迹，信息内容等维度自动判断消息是否属于谣言，而且判断地越迅速，对于舆情管理的意义就越大。同理，这种技术也被应用在社交网络有害信息识别。

在国外，有关 Facebook 假新闻的新闻被炒得火热，有兴趣的读者可以关注一下。

6. 可视化

可视化是随着大数据一起成为热门话题的。因为人类对于图像信息的理解速度要大于文字信息数百倍，所以讲一些数据可视化有助于人们更生动地理解某一结论或现象。当然不是所有数据都适合可视化，在社交网络中，我们最常见的有信息传播轨迹还有词云图等。有关这方面的内容可以参考微博账号“社交网络与数据挖掘”。



微博明星好友关系可视化，来源见水印

除了专门可视化的机构，网上也有许多开源的可视化库，百度的 Echarts 就很有名。对于社交网络信息传播以及好友关系等的可视化，使得我们能直观看到一些事实，这对于舆情报告制作以及新闻报道都有很好的辅助作用。

五. 社交网络前沿研究

我在本部分搜集了几篇近两年来在社交网络顶级会议上比较受关注的文章，将文章的摘要翻译并陈列，以供各位读者参考。

1. Negative Link Prediction in Social Media

Tang, Jiliang, et al. "Negative link prediction in social media." Proceedings of the Eighth ACM International Conference on Web Search and Data Mining. ACM, 2015

近年来，符号网络（signed network）越来越受到关注。对于符号网络的研究表明，负关系（negative link）对分析过程有帮助。由于许多网络中用户无法指定这种负关系，这是其被有效利用的主要障碍。话说，负关系的重要性与其在真实数据集之间的应用存在着差距。因此，我们自然而然会探讨是否能通过公开的社交网络数据自动预测用户的负关系。在本文中，我们研究了在社交媒体中仅仅用正关系和内容为中心的交互行为（content-centric interactions）来预测负关系的问题。我们对负关系做了一些列观测并且提出了一个原则性框架 NeLP，该框架可以利用正关系和以内容为中心的交互来预测负关系。我们对在现实社交网络的实验结果表明，NeLP框架可以准确地预测具有正关系和以内容为中心的交互关系的负关系。我们的详细实验还说明了各种因素对NeLP框架有效性的的重要性。

2. Twitter Sentiment Analysis with Deep Convolutional Neural Networks

Severyn, Aliaksei, and Alessandro Moschitti. "Twitter sentiment analysis with deep convolutional neural networks." Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2015

本文介绍了我们用于推特舆情分析的深度学习系统。我们工作主要的贡献是提出了一个初始化卷积神经网络参数权重的模型，这对于准确训练模型至关重要，同时避免增加新的特征。简而言之，我们用无监督神经网络语言模型来训练初始的词嵌入（initial word embeddings），这个词嵌入将被通过我们的基于远程监督语料库（distant supervised corpus）的深度学习模型进一步调整。在最后阶段，预先训练的参数将被用于初始化我们的模型，然后通过由Semeval-2015组织的Twitter情感分析官方系统评价竞赛最近提供的监督

实际价值。

3. Social Recommendation with Strong and Weak Ties

Wang, Xin, et al. "Social Recommendation with Strong and Weak Ties." Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016

随着在线社交网络的爆炸式增长，现在人们普遍了解，社会信息对推荐系统非常有帮助。社会推荐方法能够应对关键的冷启动问题，从而可以大大提高预测精度。主要的原因是，基于信任和影响，人们对其朋友购买过的产品表现出更多的兴趣。尽管在社交推荐领域已经有大量工作，但是很少有人关注社交强关系和弱关系这两个重要的社会学概念之间的区别。在这篇文章中，我们使用邻域重叠来逼近关系强度，并扩展受欢迎的贝叶斯个性化排名（BPR）模型并将其用于区别强弱关系。我们提出了一种基于 EM（EM-based）的算法，它可以根据最优推荐准确度（optimal recommendation accuracy）对强弱关系进行分类并学习所有用户和所有商品的潜在特征向量（latent feature vectors）。我们对四个现实世界数据集进行广泛的实验，并证明我们提出的方法在各种精度指标中显著优于目前最好的成对排名（pairwise ranking）方法。

4. Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior

Althoff, Tim, P. Jindal, and J. Leskovec. "Online Actions with Offline Impact: How Online Social Networks Influence Online and Offline User Behavior." Tenth ACM International Conference on Web Search and Data Mining ACM, 2016:537–546

如今许多应用软件都广泛地利用了社交网络功能并允许用户互相连接、互相关注、分享内容和评价动态。尽管这些功能已经被广泛应用，对于用户在线时和离线后参与还是保留的行为却很少有人理解。本文中，我们通过一个运动记录 APP 研究了社交网络是如何影响用户线下行为的。

我们分析了600万用户五年间的七亿九千一百万条线上和线下活动记录，结果表明社交网络对用户线上和线下的行为有着巨大的影响。具体来讲，我们提出了社交网络影响用户行为的因果关系。我们发现新社交关系的建立能将用户在 APP 中的活跃度提高30%，用户保留率提高17%，线下活跃度提高7%（大约每天多走400步）。通过开展自然实验，我们将新社交关系对用户的影响和用户因为对 APP 的兴趣而走更多步数作了区分。

我们发现社交影响占有所有对用户行为影响因素的55%，剩下的45%可以用用户对 APP 本身的兴趣来解释。此外我们还发现一连串的个人用户之间的社交关系建立对每日步数的增加有显著影响，用户之间每增加一条边都会减弱这种影响，并且这些变化是基于边属性和用户自己的资料属性。最后我们用这些现象设计了一个模型，模型用来判断哪些用户最容易被新建立的社交网络关系影响。

5. Intertwined Viral Marketing in Social Networks

Zhang, Jiawei, et al. "Intertwined viral marketing in social networks." Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on. IEEE, 2016

传统的病毒性营销问题旨在为一个单一产品选择一个种子用户的子集，以最大限度地提高其在社交网络中的知名度。而在实际情况下，许多产品可以同时社交网络中进行推广。从产品层面来看，这些产品之间的关系是互相缠绕的，举个例子，就是竞争、互补且独立的关系。

在这篇文章中，我们将研究“纠缠影响力最大化”问题，它是基于一个目标产品需要在社交网络上进行宣传，而同时有多个竞争/互补/独立的产品在推广这样的场景。纠缠影响力最大化是一个非常具有挑战性的问题，首先是因为很少有模型能模拟多种产品同时宣传时的信息扩散形式；第二是对于目标产品最优种子集的选择可能很大程度上取决于其它产品的营销策略。为了解决此问题，我们提出了一种统一贪心算法框架

6. Who to Invite Next? Predicting Invitees of Social Groups

Yu Han, and Jie Tang. "Who to Invite Next?Predicting Invitees of Social Groups " Proceedings of theTwenty–Sixth International Joint Conference on Artificial Intelligence (IJCAI–17).2017.

WhatsApp、Snapchat 和微信等社交即时通讯工具很大程度上改变了人们工作生活和交流的方式，也受到了多个领域例如计算机科学、心理学、社会学和物理学的关注。在社交即时消息工具中，社交群组在多用户交流中扮演着重要的角色。一个有趣的问题是，社交群组动态演变的机制是什么？更具体来说，在一个群组中，谁将会被邀请加入？这篇文章中，我们研究社交群组潜在加入者这样一个新颖的问题。我们采用微信这个中国最大的社交软件作为实验数据的来源。我们提出了一个概率图模型用来计算影响用户被邀请加入群组概率的因子。我们的实验预测结果表明我们的模型相比目前的其他模型有显著的提高。

7. The Co–Evolution Model for Social NetworkEvolving and Opinion Migration

Gu,Yupeng, Yizhou Sun, and Jianxi Gao. "The Co–Evolution Model for SocialNetwork Evolving and Opinion Migration." Proceedings of the 23rd ACMSIGKDD International Conference on Knowledge Discovery and Data Mining.ACM, 2017.

几乎所有的真实社交网络都是动态且随着时间演化的。新的链路的形成和旧的链路的消失很大程度上取决于社交网络用户的同质性。同时，一些社交网络用户的隐性性质例如用户的观点也随着时间而变化。其中一部分原因是用户从社交网络中接收到影响力，这些改变进而会影响社交网络的结构。社交网络的演化和节点性质的迁移通常被认为是两个独立正交的问题。

在这篇文章中，我们提出一种协演化模型，通过对两种现象的建模形成闭环。模型有两个主要部分：

- 一个已知节点性质的网络生成模型；
- 一个已知社交网络结构的节点性质迁移模型。

通过模拟发现我们的模型有一些不错的特性：

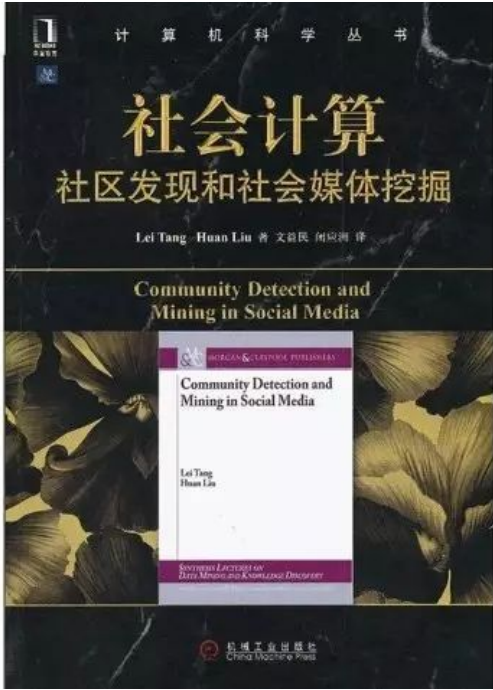
- 它可以模拟一个大范围现象，例如观点的收敛和基于社群的观点差异；
- 它可以通过一系列因子例如社交影响力范围，意见领袖，噪声等级来控制网络的演化。

最后，我们模型的有效性通过在对议会立法议案支持者的预测中得到了验证，并且我们的模型优于一些目前的方法。

六. 学习资料

1. 图书

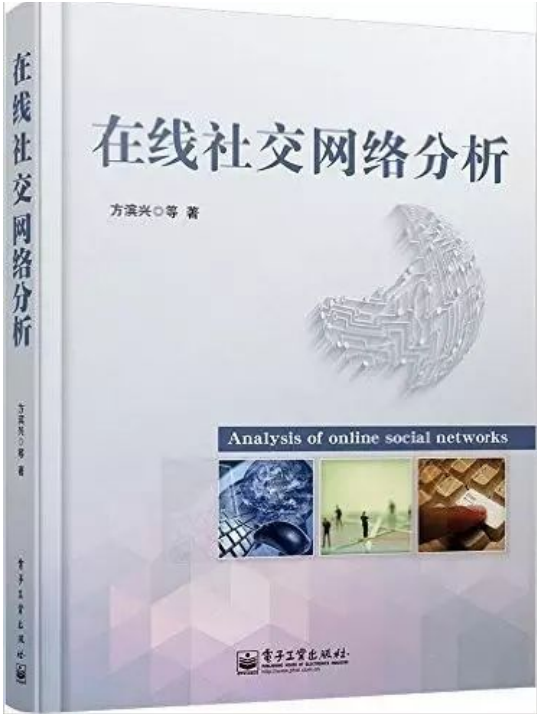
- 《社会计算》 Lei Tang, Huan Liu



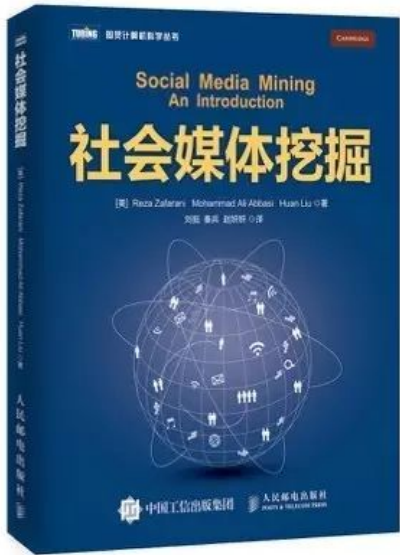
- 《社交网站的数据挖掘与分析》Matthew A. Russell



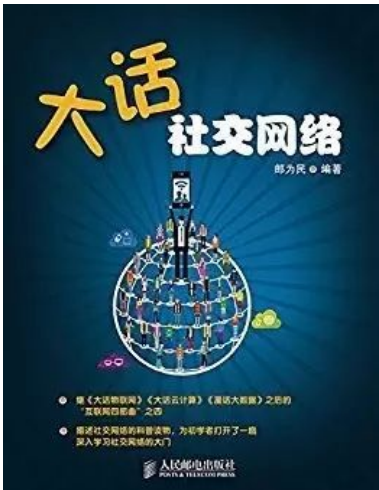
- 《在线社交网络分析》方滨兴等



- 《社交媒体挖掘》Huan Liu等



- 《大话社交网络》郎为民



2. 网站

- 大数据导航（此网站包含很多资源）

<http://hao.199it.com/>

- 斯坦福数据集网站（Jure 男神）

<http://memetracker.org/data/index.html>

- 加州大学欧文分校数据集网站

<http://archive.ics.uci.edu/ml/datasets.html>

- 国内社交网络数据集共享网站

<http://www.socialysis.org/data/project/project>

- 清华大学搭建的学术数据库

<https://cn.aminer.org/>

- 亚马逊商品流行趋势分析平台

<http://132.239.95.211:8080/demowww/index.jsp#>

- 明尼苏达双城分校社会计算实验室

<https://grouplens.org/>

- 新华网信息传播影响力评估

<http://www.xinhuanet.com/xuanzhi/zt/xzyxl/index.html>

- 新榜，微信公众号数据检测平台

<http://www.newrank.cn/>

- 清博新媒体大数据平台

<http://www.gsdata.cn/>

- 百度Echarts数据可视化库

<http://echarts.baidu.com/>

- 阿里云 DataV 数据可视化库

<https://yq.aliyun.com/teams/8>

3. 工具

- Python 及其相关库（scipy, numpy, pandas, scikit, scrapy, twitter）更多请见

<http://blog.csdn.net/hmy1106/article/details/45166261>

- 图分析分析工具 Graphchi, SNAP, Pajek, Echarts

- 可视化工具 Gephi, Graphviz

- 数据挖掘工具 WEKA, AlphaMiner

- 图数据库 Neo4j

4. 会议

笔者仅列出与社交网络相关的部分国际会议，排名不分先后，加粗的会议为专门讨论社交网络话题的会议。

KDD, WWW, ICDM, CIKM, AAAI, SDM, IEEE BigData, **ASONAM**, WSDM, **ICWSM**, ACL, IJCAI,

笔者在这里推荐两个国内的社交网络分析会议，一个是全国社会媒体处理大会（SMP），由中国中文信息学会主办，会议论文 EI 检索。第二个是国际网络空间数据科学会（IEEE ICDSC），会议由中科院，北大，中国网络空间安全协会等机构筹办。

5. 课程

笔者在上一部分提到的国际会议，例如 WWW、KDD 等，每年都有关于社交网络分析方向的 tutorial，其视频和 PPT 都是在网上可获取的，通过 tutorial 能对相关领域有一个宏观了解并且能了解领域前沿动态。

除此之外，在 Coursera 上面密西根大学安娜堡分校开设的一系列 Python 学习课程也值得一看。在网易公开课上面也有中文的 Python 数据挖掘课程可供学习。

万能的淘宝也提供大量廉价的视频和电子学习资料。

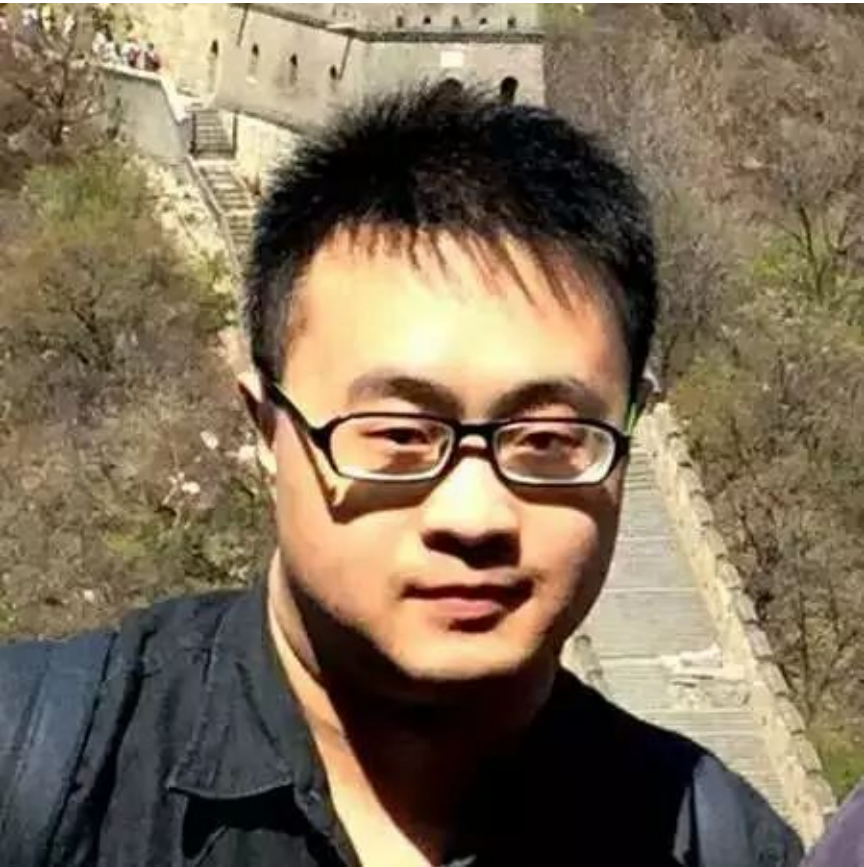
最后，利用好科学上网工具和搜索引擎（不是百度）才是王道。

七. 参考资料

[1] 方滨兴, 许进, 李建华. 在线社交网络分析[M]. 电子工业出版社, 2014.

[2] Reza Zafarani, Mohammad Ali Abbasi, Huan Liu. 社交媒体挖掘[M]. 人民邮电出版社, 2015.

[3] Carlos Castillo, Wei Chen, Laks V.S. Lakshmanan, Information and Influence Spread in Social Networks, KDD 2012 Tutorial



窦英通，伊利诺伊大学芝加哥分校博士生，对社交网络分析，推荐系统感兴趣。希望通过数据派平台在分享交流中成长。

来源：软件定义世界（SDX）

阅读(0)

举报

0 喜欢

0 没劲

分享到

热门关注



搜生活

[+关注](#)

搜狐公众平台官方账号



MAGIC杨梦晶

[+关注](#)

生活时尚&搭配博主 / 生活时尚自媒体 / 时尚...

登录

来说两句吧....

还没有评论，快来抢沙发吧！

搜狐“我来说两句”用户公约

已有 0 人参与，点击查看更多精彩评论

搜狐正在使用畅言