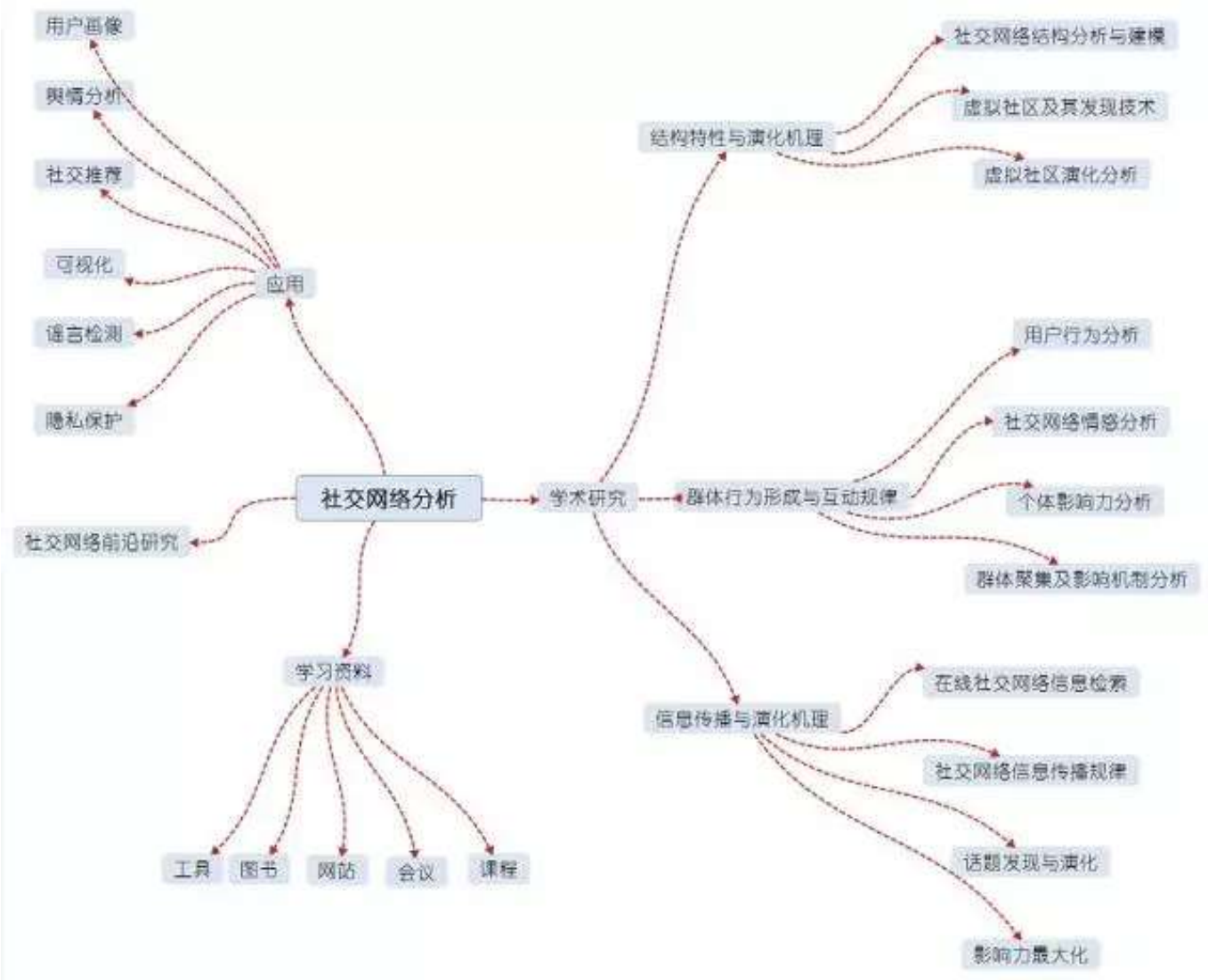
独家 | 一文读懂社交网络分析-下（应用、前沿、学习资源）

原创：窦英通 数据派THU 2017-09-26

DataPi THU, Share and Study

（点击可查看大图）



本文主要阐述：

- **社交网络分析的应用**
- **社交网络前沿研究**
- **学习资料**
- **参考资料**

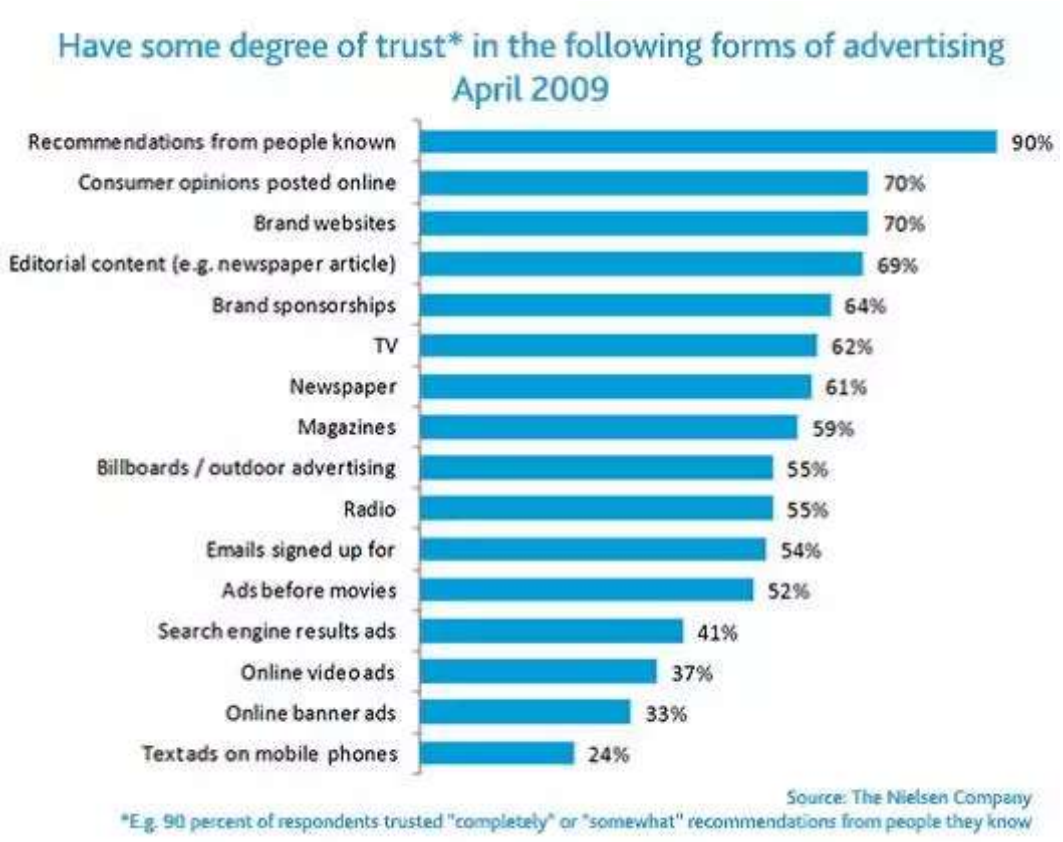
浏览前三章的内容请见上篇（2017年9月26日头条）。

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），如果想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

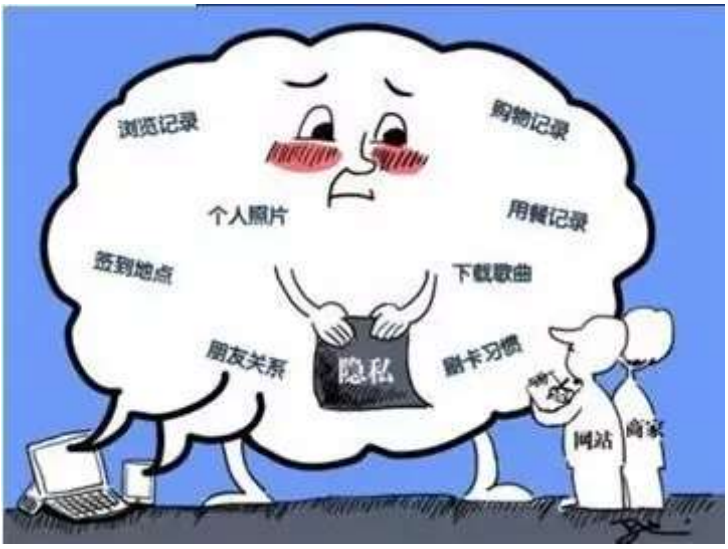
举几个例子，去年的和颐酒店，今年的北京地铁骂人事件这类急性舆情事件最早就是在微博上爆出，而且在短时间内迅速传播。还有去年的关于快手的“中国农村残酷底层物语”，今年的“北京房价”等这类民生话题，也是在微信公众号逐渐发酵。

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。

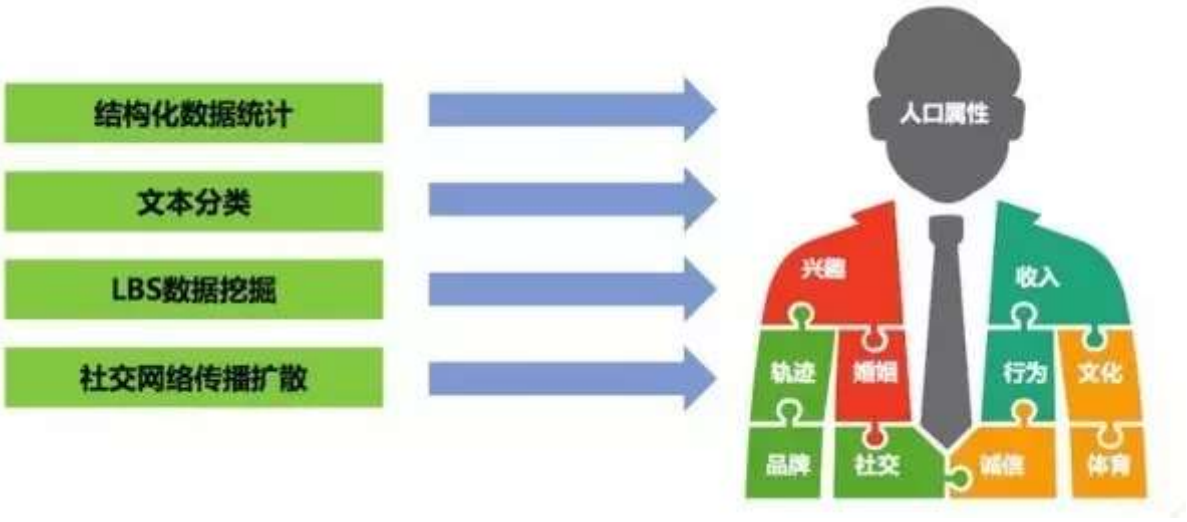


“云端”的隐私，来源：http://s9.sinaimg.cn

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

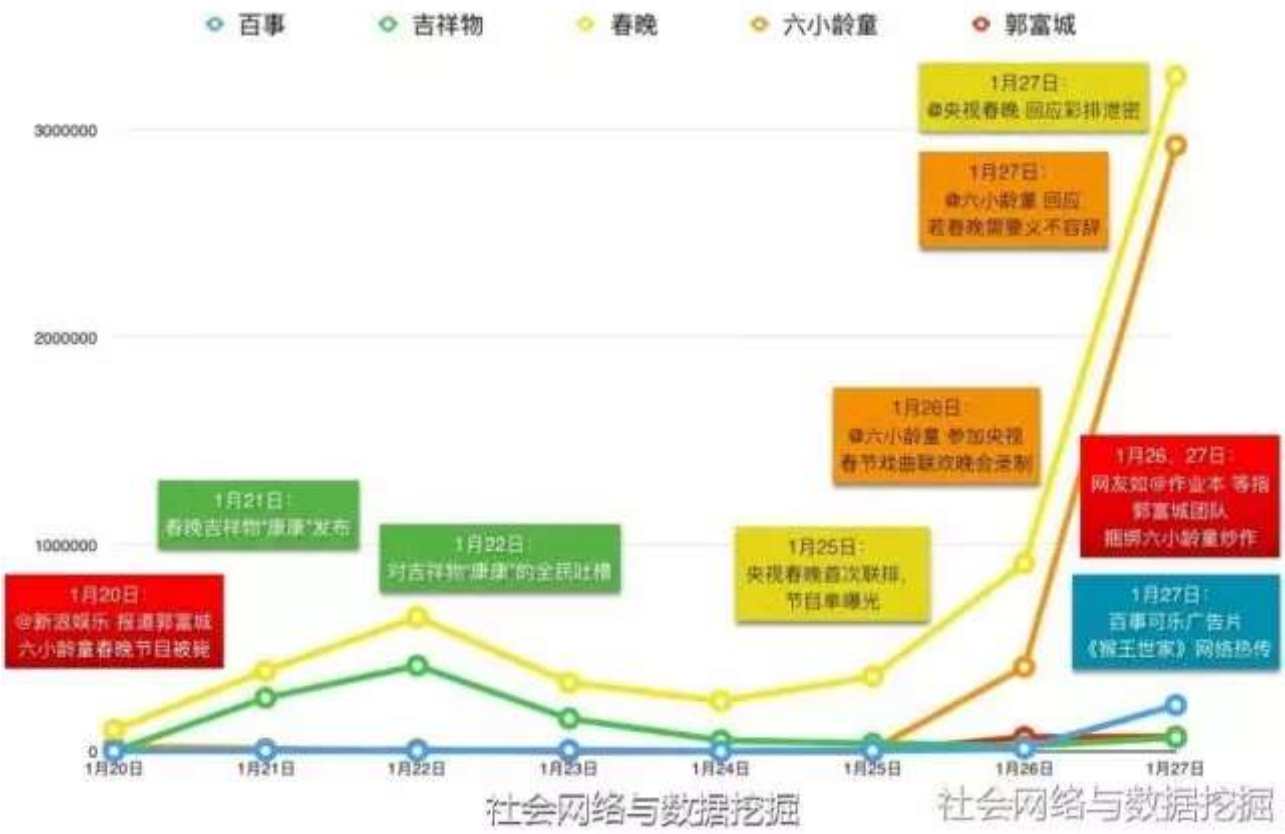
用户画像，这是个营销术语，即通过研究用户的资料和行为，将其划分为不同的类型，进而采取不同的营销策略。传统的用户画像最常用的手段就是调查问卷，订阅过杂志和报纸的读者都知道，会有各种各样的有奖问卷，一方面用来获得对于产品的反馈，另一方面就是对你进行画像，这些画像资料甚至广泛在黑市流通，这就是你为什么有时候会接到莫名其妙的电话的原因（又扯到了隐私保护问题）。

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。



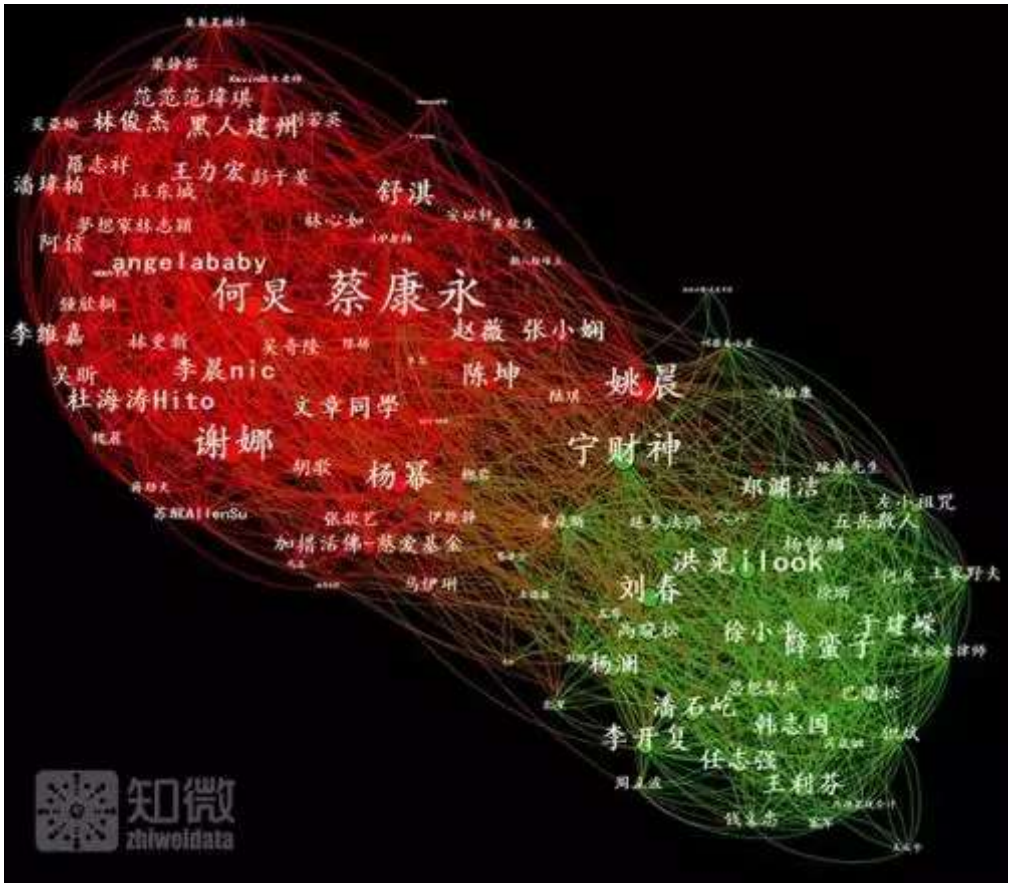
“六小龄童春晚被拒”谣言传播走势，来源见水印

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。话句话说，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，然后我们通过由Semeval-2015组织的Twitter情绪分析官方系统评价竞赛最近提供的监督训练集对后者进行培训。我们的方法得到的结果和参与竞赛的系统的结果之间的比较表明，我们的模型可以分别排在短语级别子任务A（11个团队）和消息级子任务B（40个团队）前两位。这证明了我们解决方案的实际价值。

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%，线下活跃率提高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对会减弱这种影响，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

传统的病毒式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。而然在实际情况下，许多产品可以同时在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（interTwined Influence EstimatoR， TIER），在四种不同类型现实社交网络数据集的实验表明TIER 优于所有的比较方法，在解决纠缠影响力最大化问题上有着显著优势。

6. Who to Invite Next? Predicting Invitees ofSocial 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：

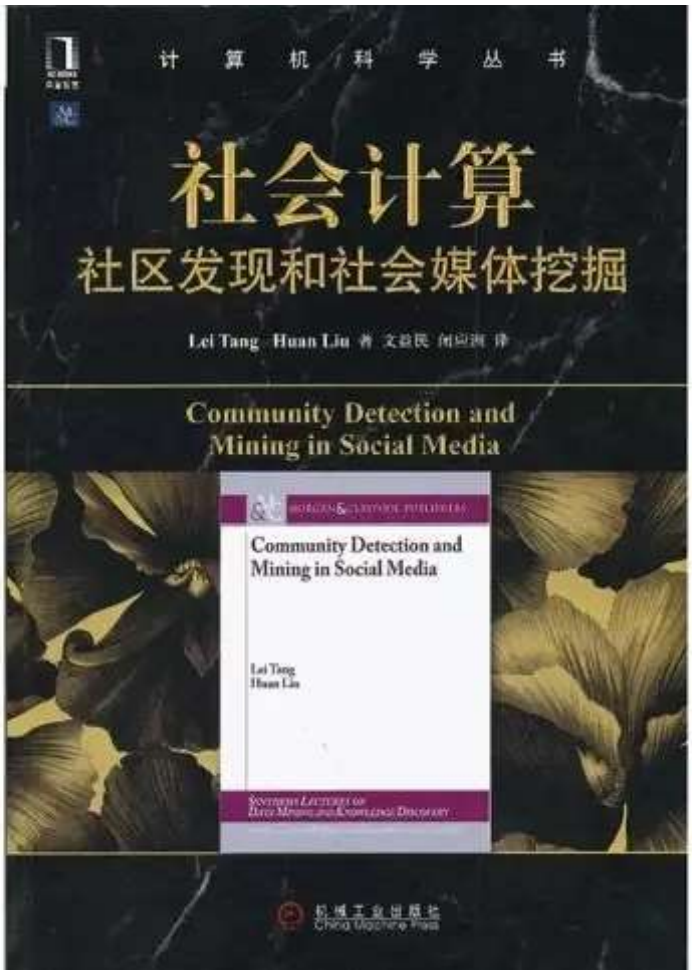
- 它可以模拟一个大范围现象，例如观点的收敛和基于社群的观点差异；
- 它可以通过一系列因子例如社交影响力范围，意见领袖，噪声等级来控制网络的演化。

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

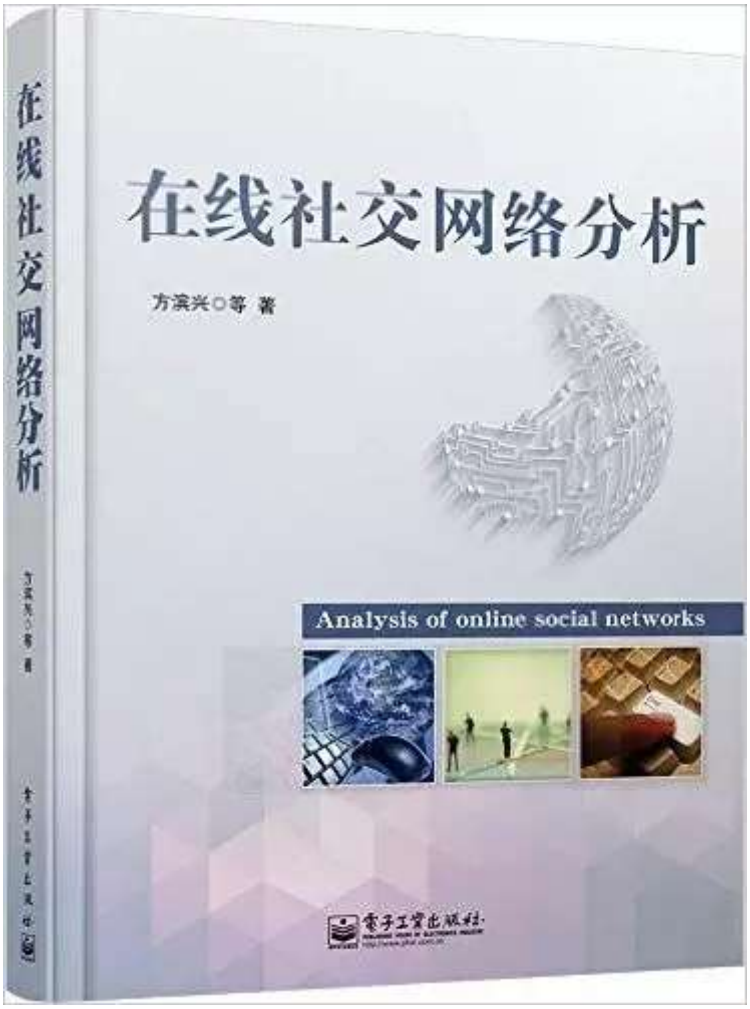
1. 图书

- **《社会计算》Lei Tang, Huan Liu**

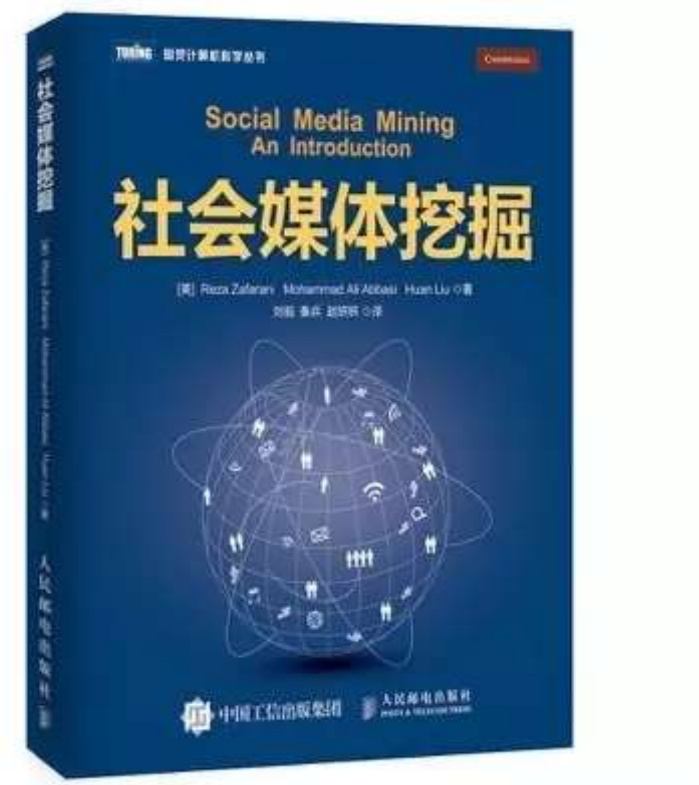


- **《社交网站的数据挖掘与分析》Matthew A. Russell**

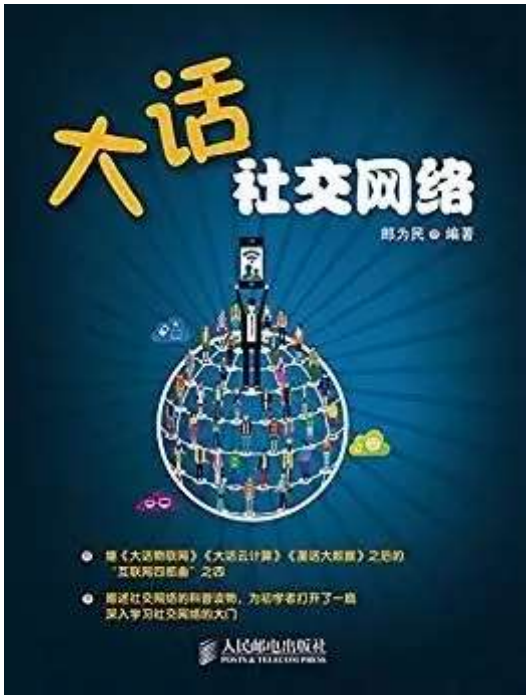
- **《在线社交网络分析》 方滨兴等**



- **《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, NIPS, ICML, ECML-PKDD, VLDB, SIGIR, PAKDD, RecSys, ACM HT, **SBP**, ICWE, PyData

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 PTT 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. 社会媒体挖掘[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。

【一文读懂】系列往期回顾：

独家 | 一文读懂优化算法
独家 | 一文读懂Adaboost
独家 | 一文读懂Apache Kudu
独家 | 一文读懂TensorFlow基础
独家 | 一文读懂Hadoop（一）：综述
独家 | 一文读懂Hadoop（二）HDFS（上）
独家 | 一文读懂Hadoop（二）HDFS（下）
独家 | 一文读懂Hadoop（三）：Mapreduce
独家 | 一文读懂Hadoop（四）：YARN
独家 | 一文读懂语音识别（附学习资源）
独家 | 一文读懂深度学习（附学习资源）

独家 | 一文读懂迁移学习（附学习工具包）
独家 | 一文读懂大数据处理框架
独家 | 一文读懂特征工程
独家 | 一文读懂数据可视化
独家 | 一文读懂聚类算法
独家 | 一文读懂关联分析
独家 | 一文读懂大数据计算框架与平台
独家 | 一文读懂文字识别（OCR）
独家 | 一文读懂回归分析
独家 | 一文读懂非关系型数据库（NoSQL）

数据派研究部介绍

数据派研究部成立于2017年初，以**兴趣为核心**划分多个组别，各组既遵循研究部整体的**知识分享**和**实践项目规划**，又各具特色：

算法模型组：积极组队参加kaggle等比赛，原创手把手教系列文章；
调研分析组：通过专访等方式调研大数据的应用，探索数据产品之美；
系统平台组：追踪大数据&人工智能系统平台技术前沿，对话专家；
自然语言处理组：重于实践，积极参加比赛及策划各类文本分析项目；
制造业大数据组：秉工业强国之梦，产学研政结合，挖掘数据价值；
数据可视化组：将信息与艺术融合，探索数据之美，学用可视化讲故事；
网络爬虫组：爬取网络信息，配合其他各组开发创意项目。

点击文末“阅读原文”，报名数据派研究部志愿者，总有一组适合你~

阅读原文