科学通报 2015 年 第 60 卷 第 5-6 期：509 ~ 519
专辑：大数据研究与应用 评 述 www.scichina.com csb.scichina.com
《中国科学》杂志社 SCIENCE CHINA PRESS

面向大社交数据的深度分析与挖掘

唐杰*, 陈文光

清华大学计算机科学与技术系, 北京 100084
* 联系人, E-mail: jietang@tsinghua.edu.cn

2014-10-10 收稿, 2014-12-15 接受, 2015-01-20 网络版发表

摘要 社交网络在线化是大数据时代的典型特点, 也是大数据产生的重要原因之一. 本文从大数据的特点着手, 结合互联网络尤其是在线社交网络的发展趋势, 介绍大数据在提升国家信息产业科学化水平、引领新型互联网经济发展、推动社会学与信息科学交叉发展等方面带来的重大机遇; 分析在线社会网络中存在的关键问题, 阐述网络大数据研究在语义理解与分析、多模态关联与融合、群体行为分析与挖掘、多维分析与可视化、系统的研发与集成等方面面临的巨大技术挑战, 以及当前国内外在大数据分析和在线社交网络领域的主要研究工作; 总结和展望网络大数据研究的未来方向和前景.

关键词
社交网络
社会影响力
社区发现
社交行为预测

随着互联网的迅速发展和广泛普及, 网络数据资源呈爆炸式增长, 各国政府高度重视大数据将带来的时代性变革. 2012年3月, 美国总统奥巴马亲自主持召开会议, 将“大数据研究”上升为国家战略. 同年, 奥巴马竞选团队通过大数据的收集和分析, 帮助奥巴马总统成功连任, 此举标示着政府领域的大数据时代已经到来. 英国政府计划未来两年内投资1.89亿英镑用于大数据建设, 要求不失时机地做好迎接大数据革命的准备. 当今世界, 互联网(特别是移动互联网)、物联网和云计算的快速崛起, 包括文本、图像、音频、视频等各种网络大数据迅速增长, 以大数据为代表的第3次工业革命正在向我们走来, 大数据必将成为全世界下一个创新、竞争和生产率提高的前沿.

另一方面, 在线社交网络取得飞速发展, 众多社交网站如脸谱(Facebook)、推特(Twitter)以及国内的新浪微博、人人网、腾讯网等迅速崛起. 2004年成立的Facebook公司已经有超过13亿的注册用户, 相当于世界人口数第二多的“国家”; 2006年发布的Twitter也有超过6亿的注册用户; 国内的腾讯公司则拥有超过8亿活跃用户; 而新浪最新公布的数据表明新浪微博的注册用户数已经超过5.6亿. 表1列出了国际流行在线社交网站以及国内对应网站用户数的统计情况. 据报道, 在美国, 16%用户的上网停留在Facebook上, 这一数字超过了人们使用传统搜索引擎(如谷歌)的10%. 毫无疑问, 在线社交网络已经成为连接物理社交世界和虚拟网络空间的桥梁. 网络用户和信息的交互以及用户之间的交互在社交网络上留下了各种“足迹”, 直接促成了网络大数据时代的到来. 在线社交网络存储了大量用户资料、用户之间的社交关系以及用户之间的交互, 这些海量社交数据有着巨大的研究价值, 同时也在广告、推荐系统等方面具有广阔的应用前景.

社交网络为用户提供了一个交互和传播信息的平台, 同时为大规模社交网络的研究提供了数据基础. 现有的社交网络研究包括网络结构拓扑分析(如ER模型[1]、small-world模型[2]、Barabási-Albert模型[3]等)、网络演化分析(如网络微观演化[4]), 社交关系和影响分析(如链接预测[5]、影响力分析[6~9]、社交纽带关系推断[10])以及用户行为预测[11,12]等.

1 大数据带来的机遇

一方面, 大数据研究已经成为提高国家信息产

引用格式: 唐杰, 陈文光. 面向大社交数据的深度分析与挖掘. 科学通报, 2015, 60: 509–519
Tang J, Chen W G. Deep analytics and mining for big social data (in Chinese). Chin Sci Bull, 2015, 60: 509–519, doi: 10.1360/ N972014-00954

表1 国际国内流行在线社交网站基本统计信息(截至2014年6月)
Table 1 Statistics of the popular international and domestic online social networking sites (as of June 2014)

在线社交网站(国际)	用户数	在线社交网站(国内对应)	用户数
Facebook	13亿	人人网	2.8亿
Twitter	62亿	新浪微博	5.6亿
Amazon	月活跃2.37亿	阿里巴巴	5亿
Wechat	月活跃2.37亿	腾讯网	8亿活跃

业科学化水平的支撑点. 随着网络媒体技术的日益普及, 公众参与并产生大量网络数据, 其中包含着公众在社会生活、金融服务、医疗卫生等各个方面的需求表达. 准确、及时地获取并理解这些数据而得到的信息, 可成为相关政府部门发现和处理民生问题、制定有效政策提供重要的辅助决策依据.

另一方面, 大数据必将是引领新型互联网经济发展的制高点. 巨量的网络数据中蕴含着丰富的客户行为信息及个性化需求信息, 通过智能服务系统将网络媒体内容转化为政府管理的可用信息, 来提高政府在行业规划、地方经济战略布局等方面的决策和服务水平. 在经济全球化和商业竞争日益激烈的今天, 谁率先掌握人类社会活动的基本规律(包括个体兴趣偏好、群体消费趋势、关系和行为分析等), 谁就可以在市场营销、商业计划、社会规划、经济建设中做到有的放矢. 例如, 通过分析微博中信息传播微观机理, 可以对信息传播的影响规模和速度进行预先判定, 从而为企业的产品推广选择最好的投放目标用户群和种子用户; 通过分析社交网络用户的社会关系和在线商店中用户的浏览、点击、购买行为, 可以为商家提供精准的产品推荐和优质的在线配送, 更好地满足用户的消费需求和提高用户对商家的满意度; 通过对网络空间的用户消费行为感知, 就可以在产品策划、设计和营销过程中做到有的放矢; 通过网络舆情分析了解社会民生, 就可以为国家的经济建设和社会管理科学化提供决策支持.

最后, 大数据研究将是推动社会学与信息科学交叉发展的着力点. 信息技术的发展带来了现实人类社会与虚拟网络空间的深度融合, 人们的工作和生活在信息化环境中留下的数字足迹汇聚形成了可感知、可计算的关系多样的社会关系网络. 大数据完整记录了数以10亿计用户的所言所行、记录了用户间形成的种类繁多的社会关系、记录了用户产生的海量网络信息的传播轨迹, 这些人类社会活动的真实记录为研究社交网络及其上的信息传播规律提供了宝贵的基础数据, 为科学研究带来了很多全新的挑战, 必将极大地促进信息科学与社会科学交叉领域及其相关方向(包括模式识别、数据挖掘、人工智能、信息检索等)的革新与发展, 具有重大的学科发展意义.

2 大数据研究面临的科学挑战

大数据研究在以下几个方面面临着巨大的科学挑战:

(ⅰ) 网络大数据的语义理解与分析. 以文本、图像、音频和视频等为载体的网络大数据已成为一种主要的形式. 谷歌、百度等通用搜索引擎在很大程度上可以帮助用户快速检索图像等信息, YouTube、优酷网等则提供了视频检索机制, 可以搜索网络上的视频数据, Facebook、Twitter、新浪微博、微信等社交媒体网站通过用户共享的形式也包含了海量的图像、视频等数据. 海量的网络大数据带来了存储、检索、管理等多方面的挑战. 百度、谷歌、YouTube 等商用搜索引擎均基于网络数据的文本描述进行检索; 在网络大数据的背景下, 很多数据缺乏文本描述, 需要算法自动分析并理解可视化或者音频内容, 因此现有的完全基于文本的技术将很难应用, 需要网络大数据的语义进行全面的分析.

(ⅱ) 网络大数据的多模态关联与融合. 随着各种模态媒体信息的不断增加, 面临着"信息多但用不了, 有信息但找不到"的重要问题, 为异构媒体的研究与应用带来了新的机遇和挑战. 如何实现异构媒体的关联与模式发现成为了研究和应用的关键问题. 然而, 目前常用的以文搜文、以图搜图等单媒体检索方式返回结果局限于单媒体数据, 异构媒体内容形式多样, 一般包括图像、视频、音频和文本等, 基于内容的单媒体检索忽略了共存的异构网络数据相关性, 不能很好地理解异构网络数据语义. 如何跨越不同媒体, 利用异构媒体之间的语义关系来实现异构网络大数据的模式发现技术, 从而支持异构媒体的关联和大数据模式发现, 是数字媒体行业发展面临的重要问题, 是下一代搜索引擎所需的核心技术. 尽管目前异构媒体的关联与挖掘技术已经有了一些相关研究, 但仍然困难重重, 很多关键问题还没有解决, 包括准确性及可用性差、媒体类型有限、评测数

据集缺乏等, 这严重阻碍了异构媒体的关联与模式发现等技术的研究及应用.

(ⅲ) 社交网络大数据的群体行为分析与挖掘. 社交网络的快速发展构建了网络化、数字化、虚拟化的工作和生活环境, 给人们带来了前所未有的信息自主权, 人类社会的信息化水平进入了一个全新阶段. 社交网络的快速发展在使人们信息交流需求极大释放的同时, 也带来了信息产生社会化、信息内容碎片化和信息传播网络化的问题, 这对网络信息环境的科学管理和合理利用带来了新的挑战. 图灵奖获得者Hopcroft教授提到, 社交网络的不确定性使得传统物理学中的复杂动力学方程不再适用[13]. 社交网络中的群体行为模式尚未得到深刻理解和充分掌握, 导致社交网络在信息的可信性、传播的可预测性、群体行为的可控性等方面仍处于一种无序状态, 造成人们创造大量社会数据却对其知之甚少的现状. 深入分析社交网络结构演化及群体行为的原动力和本质特征, 对于提高社交网络管理的科学化水平、培育文明理性的网络环境都具有广泛的现实意义.

(ⅳ) 网络大数据的多维分析与可视化. 随着网络媒体的发展, 各种新闻、论坛、博客、微博、社交网站等新媒体平台迅猛发展, 大量媒体内容产生. 现有的媒体信息呈现方式一般采用简单罗列方式, 如搜索引擎往往按照相关程度顺序排列结果, 新闻网站采用人工编辑方式按类别分块呈现. 这些媒体信息之间往往是孤立和单一的, 浏览效率比较低下, 无法满足网络大数据呈现的需求. 上述问题导致人们无法快速感知网络热点信息的发生和进行全面准确地了解. 因此, 迫切需要对网络大数据的聚合与呈现技术进行研发, 这将能大大改变网络大数据的分析与理解, 从而在很大程度上提高网络大数据的使用效率和效果.

(ⅴ) 网络大数据系统的研发与集成. 网络数据具有海量、异构、多样复杂等特性, 这些特性给数据的采集、整合、存储管理以及相应的分析挖掘带来诸多的挑战.

基于上网数据、社交数据、网页数据、多媒体数据、海外新闻数据、用户行为数据等网络大数据, 通过对各种信息源的分析, 实现广义通用的民意分析, 提供从事件发现、事件分析、事件处理到事件总结的一条龙政府决策服务. 通过进一步的应用层研究(如: 热点预判、多媒体数据分析预警、热点事件提取、社交网络分析、深度学习大数据的语义分析等高层应用), 使网络大数据的研究应用到具体的政府工作中, 实现网络大数据智能服务的落地.

接下来本文将主要从社交网络大数据的角度详细阐述大数据分析相关技术的在国内外的研究现状及发展趋势.

3 国内外研究现状和发展趋势

面向社交网络大数据的内容深度分析与理解的关键技术涉及到网络媒体数据本身的概念表示和多角度语义描述, 以及面向社交网络的用户社会关系分析, 海量异构信息的协同关联挖掘, 以及多层次信息的聚合与呈现, 下面从这几个方面介绍国内外研究现状.

3.1 网络大数据的语义理解与分析技术

大数据语义分析技术将为基于网络大数据理解提供关键支持, 是众多大数据应用的基础. 这里面主要的问题是随着网络异构数据快速产生, 数据本身以不同媒体形式存在, 如何从异构媒体中识别对应的概念成为研究热点. Yang等人[14]设计了领域自适应算法, 即在目标领域视频数量有限的情况下, 如何从已有领域的模型获得目标领域的模型; 同时, 如何再从已有领域的模型挑选最适合目标领域的模型. 此外, 对于高维海量的网络数据, 主要问题之一是新数据缺乏标记信息. 解决这一问题的有效途径之一是半监督学习(semi-supervised learning)方法, 即同时使用有标记(labeled)和无标记(unlabeled)样本训练模型. 目前半监督学习所使用的主要研究方法有: 生成式模型, 主要涉及对条件密度的估计; 判别式模型, 包括基于低密度分离的半监督学习方法和基于图的半监督学习方法.

对于网络大数据概念间关系的利用, Bart等人[15]提出使用非参贝叶斯的方法构建图像和概念间的层次关系, 使得视觉上相似的图像和概念分布在相同或者相近的子树中, 并呈现了一副与人类视觉感知基本一致的树状关系图. 但是, 概念间关系的发掘并不仅仅为了验证计算得到的概念间关系是否与人类的认知一致, 更是为了利用这些关系优化模型训练的各个环节. 因此, 为了应对大数据的类别和概念数量多的问题, 以及因此产生的巨大测试代价, 有研究者提出利用模型间关系构建多类分类器树, 使得模

型的测试环节的效率大大提升. 另一种发掘和使用模型间关系的方式是: 在训练时, 根据学习得到的概念间的关系, 使相近的概念共享特征、属性和模型, 而使相差较大的概念所使用的特征、属性和模型差异更大.

深度学习是最新的研究热点, 其主要思想是通过神经网络模拟人脑的学习过程, 并进一步模拟人脑的多层抽象机制来实现对数据的抽象表达, 将特征学习和分类器学习整合到统一的学习框架中. “深度学习”这一概念在2006年被Hinton首先提出[16], 算法目的是通过模型的输入来训练整个模型的参数, 从而使得模型逐步的形成对于数据相关的“概念”. 很多成熟的理论或技术被用在构建深度学习系统中. Salakhutdinov和Hinton[17]将玻尔兹曼机(Boltzmann machines)应用到逐层构建的深度网络中, 并且使用变分法逐层优化参数. 当深度网络的所有隐层网络都被构建后, 使用有监督的细优化(fine-tuning)完成深度网络的训练. 之后, Vincent等人[18]使用计算更为简单方便的自动编码器替代玻尔兹曼机, 实现对深度网络中隐层网络的构建. 关于深度网络在实际问题中的应用, 直到最近两年才取得了较大的进展, 使用深度网络构建的特征提取和分类算法在图像识别和语音识别中都取得了非常好的成绩. Dean等人[19]详细报道了采用深度网络在大规模图像识别任务中所采用的技术和所取得的突破性的进展. 为了能够更好地应对大规模数据, 深度网络的学习算法一般都采用并行化方式进行, 有些同时采用了并行化的一阶梯度和二阶梯度优化算法, 并在一定程度上同时实现了模块和数据的并行化[19]. 目前, 包括微软、谷歌、IBM等许多知名的IT公司都致力于开发商用化的深度学习系统. 由Andrew Ng教授主导开发的Google Brain就是通过16000个CPU以及GPU加速实现的大规模深度学习系统, 该系统可以训练具有几十亿个参数的神经网络, 并且通过著名的ImageNet数据集进行测试表明, 该系统的表现超过了所有现存的机器学习算法.

3.2 基于多模态特征融合的大数据模式发现技术

基于多模态特征融合的大数据模式发现技术主要研究异构媒体的统一表示、相似性计算和语义关联分析方法, 为项目最终目标提供数据关联基础和关键技术支撑. 目前国内外已经有了一些相关研究, 下面从异构媒体的统一表示、异构媒体相似性计算和异构语义关联分析方法3个方面分别来阐述国内外研究现状.

特征表示是异构媒体关联与管理的重要基础, 其中一个关键问题是研究不同模态数据的特征融合方法, 即如何以统一的方式表示不同模态的内容. 比较广泛的做法是基于子空间的映射技术, 这一类方法以典型相关分析(canonical correlation analysis, CCA)为代表, 学习出映射子空间, 使得在子空间中两组变量的相关性最大. Wu等人[20,21]通过CCA将音频和图像两种不同媒体的数据映射到子空间中, 从而能够度量不同媒体内容的相似性. 然而, 这些方法往往只考虑了数据之间的一一对应关系, 例如通过CCA或者CFA学习出映射子空间, 使得原始的一一对应的异构数据之间的相关性最大. 但它们不能挖掘更加丰富的语义信息, 例如跨网络数据的语义类别信息. 在Rasiwasia等人[22]的研究中, 跨网络数据之间的关联信息通过CCA进行学习, 高层抽象通过逻辑回归将文本或图像表示为具有相同维度的语义概念向量, 其中每一维表示该多媒体数据属于对应类别的概率. 然而该方法独立表示每种媒体类型, 无法充分挖掘特征之间的关联, 且只局限于两种媒体.

如何基于内容度量不同媒体的相似性是异构媒体关联和管理的核心问题. 现有的异构媒体相似性度量方法主要有两类: 第一类是基于图模型的相似性度量方法[23,24]; 第二类是基于学习的相似性度量方法[25,26]. 基于图模型的相似性度量方法大多基于共生性假设: 如果两个跨媒体文档包含同一个媒体对象, 那么这两个跨媒体文档就具有相同的语义信息. Yang等人[23]提出了以跨媒体文档为结点的图结构, 称为跨媒体文档语义图模型(Multimedia Document Semantic Graph, MMDSG). 由于基于图的方法存在大量参数, 参数的设置也是一个难题. 因此Yang等人[25]提出了一种对参数不敏感的基于局部回归和全局校正(Local Regression and Global Alignment, LRGA)的学习算法, 可以学习出拉普拉斯矩阵用于排序. Jia等人[26]提出了一种多媒体文档随机场来挖掘不同媒体对象之间的关联关系. 这类方法对参数不敏感, 但是依赖共生性假设, 如果两个多媒体文档包含同一个媒体对象, 那么就具有相同的语义信息.

3.3 网络群体行为分析与事件态势感知技术

基于行为心理动力学模型的群体行为分析与事件态势感知技术主要研究多模态异构社交网络数据融合、社交网络群体行为的心理动力学模型、社交网络结构的微观演化机理、社交网络影响力多尺度度量模型、基于多元信息搜索和群体行为分析技术，为项目最终目标提供群体行为分析与事件态势感知的理论与技术支撑. 基于Web的社交网络研究具有重大的科学意义和较高的实际应用价值，因而受到了国内外学术界和产业界的广泛关注. 但由于社交网络服务的兴起还不足10年，因此面向社交网络的理论研究尚处于起步阶段. 下面从社交网络的相关研究关键技术点来阐述国内外研究现状.

在社交网络基础理论方面，传统社会网络的社会学研究已有相当长的研究积累. 最典型的理论有二级传播、弱连带优势、强连带优势以及结构洞理论等. Lazarsfeld等人[27]提出了二级传播理论，描述了人类社会中信息从媒介到受众的通常过程；Granovetter[28]提出了弱连带优势理论，认为关系较疏远的人可能拥有差别较大的有用信息；Krackhardt[29]提出了强连带优势理论，认为强连带提供了人们彼此相互信任的基础；社会学家Burt[30]提出了结构洞理论，研究社会关系网络的结构形态.

在社交网络的结构分析方面，主要包括宏观、中观和微观3个层面的结构分析. 宏观结构分析主要关注网络的统计特性. Barabási和Bonabeau[31]提出了无标度网络模型；Watts和Strogatz[2]研究了网络小世界特性产生的机制. 中观结构分析主要以社区结构分析为主. Newman和Girvan[32]提出了基于模块度的社区度量和社区发现方法；Palla等人[33]提出了基于网络渗流的重叠社区度量和社区发现方法；Rosvall和Bergstrom[34]提出了基于网络压缩编码的社区度量和发现方法；Arenas等人[35]提出了基于网络同步动力学的社区结构度量和发现方法. 微观结构分析研究网络中的显著微观结构模式及其对网络功能的影响. Marvela等人[36]分析了微观三角形结构对网络演化的影响；Milo等人[37]研究了网络构件的显著微观结构模式；Kleinberg等人[38]从博弈论的角度阐述了结构洞的作用.

在社交网络关系的理论分析方面，Leskovec等人[39]定义了符号网络，并从结构平衡和社交状态理论的角度揭示了社交网络中正向和负向关系的形成机理. 从机器学习的角度，目前的研究主要集中在社交关系的预测，如Adamic和Adar[40]通过个人主页之间的链接关系推断现实世界中人与人之间的关系；Wang等人[41]通过论文合著关系预测社交网络中的合作；Liben-Nowell和Kleinberg[5]综述了社交网络中的链接关系预测问题.

在社交网络用户行为分析方面，主要包括心理动因分析、个体用户行为分析和群体行为分析，以及个体和群体的偏好分析. 语义空间表征技术包括潜在语义分析(LSA)和自动化语篇分析(Coh-Metrix)等为研究心理动因提供了有效的工具. 在个体用户行为分析方面，Zeng等人[42]提出利用隐马尔可夫模型对个体用户行为建模，Scott[43]使用图结构对个体用户的性质进行度量. 在群体用户的聚集行为方面，Maia等人[44]对YouTube的用户属性进行聚类分析；Backstrom等人[45]研究了虚拟社区中的自然群组行为；Tan等人[12]通过时间状态模型对用户行为进行预测；Tang等人[46]分析了社交网络中的从众现象. 用户偏好建模主要分析用户个体或者群体的对Web数据内容的感兴趣程度，目前主要有显式和隐式两类建模方式.

在社会网络信息传播规律方面，传统信息传播模型大多基于疾病传染模型，研究社交网络的宏观特征，如Moore和Newman[47]的SIR模型、Kuperman和Abramson[48]的SIRS模型. 研究者们也注意到信息传播过程与社会影响力之间的密切联系. 1967年Milgram[49]提出了六度分离理论、Christakis和Fowler[50]提出了三度影响理论、Kempe等人[6]提出了信息传播的通用阈值和级联模型、Zhang等人[51]研究了信息传播中的从众现象、Yang等人[52]提出了基于角色的信息传播模型. 由于疾病传播与信息传播的性质具有明显的差异，这方面的研究工作存在着一定的缺陷，还存在对个体用户的差异性研究不够系统深入的问题.

在社交网络搜索方面，目前主要包括大规模异构社交网络数据的整合和索引、社交网络搜索在线应用等研究. 数据融合过程中采用了文本数据的相似性连接、实体抽取、图数据结构感知的相似性匹配等技术，但对数据的语义信息的利用还不够. 自动推荐近年在Amazon等电子商务网站和Netflix，Hulu等视频服务网站中有广泛的应用. 如何准确地寻找用户感兴趣的商品和服务、如何挖掘可用数据极少的不活

跃用户(冷启动问题)、如何高效处理快速增长的海量数据, 是目前推荐系统研究的3个主要热点问题, 社交网络的兴起给推荐系统带来新的活力和挑战.

在社交网络分析平台与系统方面, 不仅著名科研机构致力于社交网络分析研究, 各大网络公司也纷纷建立自己的社交网络分析工具, 如斯坦福大学的SNAP系统、卡内基梅隆大学的AutoMap系统、Google公司的Pregel系统(表2). 但大部分系统还只是支持网络宏观分析, 如网络结构分析和可视化分析, 而忽视了网络微观分析(如个体用户行为分析、影响力分析等), 此外大部分系统都忽视了内容分析和高效索引的重要性, 因此难以支撑大规模社交网络的信息传播分析需求.

通过国内外发展现状分析, 可以发现社交网络的研究还存在以下问题:

(1) 现有网络结构分析大都从宏观层面展开, 很少关注网络结构的微观变化;

(2) 社交网络的信息传播模型主要基于传染病模型, 没有考虑用户在信息传播中的个人角色、心理动因和不确定性;

(3) 影响力分析通常只考虑网络节点的全局影响力, 忽视了"影响力"的尺度多样性;

表 2 现有社交网络分析平台和系统

Table 2 Existing social network analysis platforms and systems

系统开发单位	系统名称[a)]	项目来源	平台/系统功能[b)]							
			S	C	T	E	V	H	I	P
斯坦福大学	SNAP	美国自然科学基金、微软、雅虎	●	○	○	○	●	○	○	○
卡内基梅隆大学	Pegasus	美国自然科学基金、雅虎、劳伦斯·利弗莫尔实验室	●	○	●	○	●	●	●	○
卡内基梅隆大学	AutoMap	美国自然科学基金、美国陆军研究院	●	●	●	●	○	○	●	●
卡内基梅隆大学	OddBall	美国国家科学基金、美国能源部的国家实验室	●	○	●	●	●	○	○	○
西北大学(美国)	C-IKNOW	美国自然科学基金、美国国家健康以及陆军研究院	●	●	○	○	●	○	○	●
华盛顿大学	Statnet	美国国家科学基金、海军研究局	●	○	○	●	●	○	○	●
印第安纳大学	NWB	美国自然科学基金	●	○	●	○	●	○	○	○
匹兹堡大学	EpiFast	美国国家科学基金网络技术与系统	●	○	●	○	●	●	○	○
加州大学洛杉矶分校	HCS		●	○	●	○	●	●	●	●
纽约州立大学石溪分校	NetworkX		●	○	●	○	●	○	○	○
罗兰大学	CFinder	匈牙利自然科学基金	●	○	●	○	●	○	●	●
Google	Pregel		●	○	●	○	○	●	○	○
IBM	X-RIME		●	○	●	○	●	●	○	○
微软研究院	CoSBiLab Graph		●	○	●	○	●	●	●	○
Gephi.org	Gephi		●	○	●	○	●	○	●	○
Analytic Technologies	UCINET		●	○	●	○	●	○	●	○
ATOS	MyInfo+	伦敦奥运会合作项目	●	●	●	●	●	●	○	○
开源项目	Pajek		●	○	●	○	●	○	○	○
开源项目	GraphLab		●	●	●	○	○	○	○	○
清华大学	SAE	国家高技术研究发展计划、国家自然科学基金及华为支持项目	●	●	●	○	○	●	●	●

a) 系统名称来源于: SNAP, http://snap.stanford.edu/snap/doc.html; Pegasus, http://www.cs.cmu.edu/~pegasus/what%20is%20pegasus.htm; AutoMap, http://www.casos.cs.cmu.edu/projects/automap/hardware.php; OddBall, http://www.pdl.cs.cmu.edu/PDL-FTP/associated/OddBall.pdf; C-IKNOW, http://ciknow.northwestern.edu/; Statnet, http://statnetproject.org/; NWB, http://nwb.slis.indiana.edu; EpiFast, http://www.epifast.com/2004/menu.htm; HCS, http://hcs.ucla.edu.home.htm; NetworkX, http://networkx.lanl.gov/overview.html; CFinder, http://www.cfinder.org; Pregel, http://www.royans.net/arch/pregel-googles-other-data-processing-infrastructure/; X-RIME, http://xrime.sourceforge.net/; CoSBiLab Graph, http://www.cosbi.eu/index.php/research/prototypes/graph; Gephi, http://gephi.org/; UCINET, https://sites.google.com/site/ucinetsoftware/home; MyInfo+, http://atos.net; Pajek, http://vlado.fmf.uni-lj.si/pub/networks/pajek/; GraphLab, http://graphlab.org/projects/index.html; SAE, https://github.com/actnet/saedb. b) S, 结构分析; C, 内容分析; T, 传播分析; E, 事件检测; V, 可视化分析; H, 高效索引查询; I, 影响力分析; P, 用户行为分析

(4) 用户行为分析主要从宏观层面研究群体用户行为, 缺少对个体行为的建模;

(5) 尽管国内外研发了很多社交网络应用系统, 但总体上仍旧缺乏能够对社交网络信息进行科学管理和有效引导的应用系统.

3.4 网络信息聚合与呈现技术

多层次多维度信息聚合与呈现技术主要研究多源异构数据的统一表示, 媒体信息的语义聚合和热点事件信息发现, 主要关注多源动态信息的聚合和管理以及网络热点事件的挖掘与呈现, 目前在这两方面已有一些研究, 相关国内外研究现状和发展趋势具体如下.

多源海量动态信息的异构性、实时性和不确定性特点为信息的模型定义、预处理与集成、存储与索引、查询分析处理带来了很大的困难, 设计有效的方法和策略, 进行多源海量动态信息的聚合和管理已经成为信息领域研究的当务之急[53]. 以不确定性信息的管理和使用为例, 由于物理信息系统中不仅包含数量信息, 而且包含大量的位置和空间信息, 这些信息的异构性和不确定性为多源海量动态信息的聚合与管理带来了很大的困难. 在实际应用中, 如何有效管理物理信息系统产生的信息、消除不确定性信息产生的矛盾, 对行业的发展具有深远影响. 目前, 对信息聚合服务的研究主要集中在聚合标准、信息聚合对互联网发展的影响、聚合技术在各领域的应用等几个方面. 以RSS为代表的信息聚合平台可以集成多家信息来源, 自动浏览和监视这些来源网站的内容, 将最新内容及时传送给用户. 英国剑桥大学联合利华分子信息学中心、伦敦皇家学院化学系以及荷兰Nijmegen大学分析化学实验室研究人员研究了如何将RSS 1.0规范与CM化学标记语言(chemical marked language)结合起来, 用以实现一种基于元数据的提示服务(a metadata based alerting service). 一些企业, 如谷歌、脸谱公司等已经注意到这方面巨大的应用前景, 并进行了相关的分析综合尝试, 如谷歌公司推出的Google+, 针对手机、移动设备用户信息交流的"扎客"平台等. 但是, 这些尝试还比较初级, 只是将各种信息简单分析后堆砌在一起, 缺乏对于各种平台数据之间本质联系的深层次分析.

网络热点事件的掌控需要人们对事件进行全方位的分析, 因此有必要采用适当的方式将事件呈现出来. 现有的方法大多基于数据可视化的策略进行信息呈现. 针对不同目的, 信息呈现的方法也不一样. 如微软亚洲研究院的"人立方"关系搜索系统(http://renlifang.msra.cn), 采用了任务关系图的策略进行结果的呈现, Yu等人[54]则提出了一种异质网络中检索词驱动的语义相似子图发现的策略.

虽然研究者已经进行了大量的相关研究, 但该领域仍有若干关键技术有待进一步探讨, 包括:

(ⅰ) 异构媒体信息的语义关联方面. 内容关联方式单一, 难以适应跨平台网络事件挖掘的需求, 偏重于基于物理链接的挖掘而忽略潜在因素的关联, 导致挖掘的事件不够全面.

(ⅱ) 网络事件挖掘方面. 主要依赖于搜索方式, 利用网络数据相似度的匹配与排序, 而没有充分考虑到事件潜在的语义结构.

(ⅲ) 媒体内容呈现方面. 往往只针对单一平台网络数据进行分析, 缺乏平台间、异质媒体间的协同呈现.

3.5 网络大数据智能服务系统

大数据智能服务系统及应用示范集成前4点的关键技术, 主要研究如下内容.

3.5.1 网络数据采集、预处理与整合技术

网络数据的采集包括媒体内容的采集与媒体用户行为数据的采集. 针对媒体内容的异构性, 媒体内容的采集需要对不同类型的内容分别进行处理, 提取有效信息. 用户行为数据采集主要有两种方式: 一种方式是通过采集记录内部系统日志来实现, 如Facebook的Scribe, Apache的Chukwa和Cloudera的Flume等; 另一种方式是通过ISP来进行采集, 即通过ISP的路由器来收集数据报文, 并加以分类. 海量的报文数据, 以及报文数据的复杂性和实时性都给报文特征的提取带来较大困难, 而分类算法的准确性是学术界研究的焦点问题.

用户识别技术是用户日志挖掘的基础和研究热点. 目前对用户识别技术的研究主要集中在单个网站内的用户识别, 采用的技术包括: 基于IP地址和浏览器信息的用户识别; 基于cookie技术和扩充属性的用户识别; 基于用户IP、用户代理(agent)、用户会话(session)和引用页(refer)的用户识别. 由于用户上网环境的复杂性, 如何将用户在不同时间、不同地点甚至不同媒体上的行为进行关联仍然是用户识别的一大难题.

3.5.2 面向海量数据的数据存储与处理技术

(ⅰ) 面向海量数据的数据存储技术. 传统关系数据库保证了强一致性(ACID)和高可用性(侧重AC), 在高可伸缩性方面存在难以克服的缺陷, 因此无法应付海量数据的高效存储和高并发访问. 针对该问题, NoSQL技术被提出, 并得到迅速发展. NoSQL 数据库的基本思想是通过牺牲强一致性, 使得系统达到高伸缩性和高可用性(侧重AP), 即在一致性和系统可用性之间做出权衡. 根据存储模型的不同, 目前主流的NoSQL可以分为4类: 基于键/值存储的NoSQL数据库(如Redis)、面向列族的NoSQL数据库(如Hbase, Cassandra)、面向文档的NoSQL数据库(如MongoDB)和基于图的NoSQL数据库(如Neo4J).

异构资源检索平台对数字资源的保存及方便用户对数字资源的使用都有着极其重要的作用. 在数字资源越来越丰富、种类越来越多的今天, 更简洁、更实用、功能更强的异构资源统一检索平台的构建无疑有着重要的实用价值. 数据可以分为3种类型: 结构化、半结构化和无结构化数据, 它们在应用中分别主要体现为关系数据、XML数据和全文数据. 对于这3种类型的数据, 当前都有较成熟的索引模型和查询方法, 并且这些模型和方法在大部分数据库产品中占据了主流地位. 但是实际应用中的数据在很多情况下并不单纯是一种类型, 而且3种数据的异构性导致它们的索引模型之间也存在一定程度的异构性, 所以如何处理混合类型的数据还是一个亟需解决的问题.

(ⅱ) 面向海量数据的数据并行处理技术. MapReduce是2004年由谷歌公司提出的一个用来进行并行处理和生成大数据集的编程模型. Hadoop是MapReduce的开源实现, 受到了产业界和学术界共同关注. Hadoop分布式平台采用Shared-nothing结构, 节点之间彼此独立, 具有高容错性, 能够容忍节点的高失败率. 因此, Hadoop能够部署到由中低端计算机组成的大规模机群中, 并且其可伸缩性在业界已经得到有力验证.

MapReduce框架能够较好地处理大规模的数据计算, 但是在实现需要迭代类算法时, 效率比较低. 针对该问题, 也出现了一些支持迭代计算的框架, 它们或者基于MapReduce进行修改, 或者借鉴了其计算思想进行设计. 典型的有由加州大学伯克利分校开发的Spark, 用来解决MapReduce所不擅长的两类计算: 迭代计算和交互式分析. 基本思想是将数据存在内存, 避免重复的加载.

如何有效地分布式地处理各种海量复杂数据也是目前的研究热点, 也出现了一些针对这些具体任务的计算框架, 它们一般以Hadoop平台为基础, 提供了许多任务特定的操作或功能. 例如, 为了支持海量图数据的查询与匹配, 谷歌公司开发了Pregel, 可以在通用分布式服务器上处理PB级别图数据, 与之对应产业界也推出了开源项目GraphLab.

为满足海量数据的流处理需求, Twitter, Yahoo等公司都各自研发了代表性的流处理平台. Yahoo的S4(Simple Scalable Streaming System)是一个分布式、可扩展、分区容错、可插拔的流式系统. Twitter开源的Storm实时流处理平台为分布式实时计算提供了一组通信原语, 可用于流处理、持续计算和分布式远程程序调用.

(ⅲ) 面向海量数据的数据并行挖掘技术. 数据挖掘通常需要遍历训练数据获得相关的统计信息, 用于求解或优化模型参数, 在大规模数据上进行频繁的数据访问需要耗费大量运算时间. 数据挖掘领域长期受益于并行算法和架构的使用, 使得性能逐渐提升. 过去15年来, 效果尤其显著. 试图将这些进步结合起来, 并且提炼, 使得计算能力呈几何级数增长. 即便是图形处理、游戏编程是公认的复杂, 它们也从并行化受益颇多. 研究显示数据挖掘、图遍历、有限状态机是并行化未来的热门方向.

MapReduce框架已经被证明是运行数据挖掘算法性能的重要工具. 国内中国科学院计算技术研究所2008年底开发的基于Hadoop的并行分布式数据挖掘平台, 也已用于中国移动通信企业TB级实际数据的挖掘.

(ⅳ) 面向海量数据的数据查询与分析技术. 近年来, 传统数据仓库技术难以适应海量数据查询与分析的问题, 受到学术界和工业界的密切关注. 一些工作例如EMC Greenplum, HP Vertica等, 基于传统数据库技术, 并结合并行数据库与OLAP分析型数据库各自的优点, 来实现海量数据的实时查询与分析. 但是, 这种方式在可伸缩性与容错性方面存在不足. 另一方面, 针对MapReduce框架中要求程序员自己实现用于完成具体查询和分析的Map和Reduce任务, 负担过重的问题, 一些工作开始关注基于MapReduce的数据仓库研究, 例如Facebook的Hive、雅虎的Pig、谷歌的Sawzall等, 其基本思想都是通过解析器将用户的查询语句解析为一系列MapReduce任务. 与基本的

MapReduce系统相比，高层查询语言更加容易使用，但是存在效率不足的问题. 与此同时，一些工作开始尝试数据库技术与MapReduce框架的结合. 希望结合两者的优点来实现海量数据的高效查询与分析，例如 Cloudera 推出的 Impala 项目不再使用 Hive+MapReduce批处理的思想，而是通过采用与商用并行关系数据库中类似的分布式查询引擎，可以直接从HDFS或者HBase中用SELECT, JOIN和统计函数查询数据，从而大大降低了延迟.

3.5.3 面向海量数据的数据智能分析技术

在社区结构的挖掘通常被描述为图聚类问题. 由于群体行为和兴趣的多样性，重叠社区结构的研究逐渐成为网络用户社区的研究重点. 模糊C均值、非负矩阵分解、派系过滤算法等方法已经被应用于重叠用户社区的聚类分析. 用户兴趣的变化、交互行为的改变等因素会导致用户社区随时间发生演化. 基于普聚类、张量分析、贝叶斯估计的多种社区演变分析方法得到了深入研究.

4 结论和展望

综上所述，面向社交网络的大数据分析呈现出以下几点趋势: (1) 数据网络化是大社交数据分析的基础. 现有工作大都从宏观层面对社交网络结构开展研究工作，很少关注网络结构微观变化对信息传播的动态影响; (2) 理解数据空间和社交空间之间的交互是理解数据的重要手段. 一方面，用户间的社会影响是信息传播的原动力，现有的影响力分析通常考虑网络节点在全局中的影响力，而忽视了“影响力”本身的尺度多样性，例如，同一节点在不同的社区、不同地域中具有不同的影响力; (3) 尽管国内外研究单位开展了相当数量的大数据和社交网络应用系统的研究，但总体而言，仍旧缺乏对大社交数据进行科学管理和有效管理的实用系统.

近些年，社交网络在我国得到了迅猛的发展，积累了大量的用户数据，为深度挖掘和分析海量异构社交网络带来了巨大的机遇. 我国在计算社会学、网络科学、数据挖掘、数据库和机器学习等相关领域的基础研究和技术积累、研究基地和队伍建设基础，都为大社交数据分析和管理的基础理论及其应用研究提供了良好的学科基础. 未来，针对这方面的深入研究，有助于我们在大数据和社交网络时代占领技术制高点，提升网络信息管理与应用水平.

参考文献

1 Erdős P, Rényi A. On random graphs. Publ Math, 1959, 6: 290–297
2 Watts D J, Strogatz S H. Collective dynamics of ‘small-world’ networks. Nature, 1998, 393: 440–442
3 Barabási A L. Emergence of scaling in random networks. Science, 1999, 286: 509–512
4 Leskovec J, Adamic L, Huberman B. The dynamics of viral marketing. ACM Trans Web, 2007, 1: 5
5 Liben-Nowell D, Kleinberg J. The link prediction problem for social networks. J Am Soc Inf Sci Technol, 2007, 58: 1019–1031
6 Kempe D, Kleinberg J, Tardos E. Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’03), 2003. 137–146
7 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 Med J, 2008, 337: a2338
8 Tang J, Sun J, Wang C, et al. Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’09), 2009. 807–816
9 Sun J, Tang J. A survey of models and algorithms for social influence analysis. In: Social Network Data Analytics. Berlin: Springer, 2011. 177–214
10 Tang J, Lou T, Kleinberg J. Inferring social ties across heterogeneous networks. In: Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM’12), 2012. 743–752
11 Weber C A, Current J R, Benton W C. Vendor selection criteria and methods. Eur J Oper Res, 1991, 50: 2–18
12 Tan C, Tang J, Sun J, et al. Social action tracking via noise tolerant time-varying factor graphs. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD’10), 2010. 1049–1058
13 Hopcroft J, Kannan R. Computer Science Theory for the Information Age. Berlin: Springer, 2011
14 Yang J, Yan R, Hauptmann A G. Cross-domain video concept detection using adaptive SVMs. In: Proceedings of the 15th International Conference on Multimedia (ACM MM’07), 2007. 188–197

15 Bart E, Porteous I, Perona P, et al. Unsupervised learning of visual taxonomies. In: Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), 2008
16 Hinton G, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets. Neur Comp, 2006, 18: 1527–1554
17 Salakhutdinov R, Hinton G. An efficient learning procedure for deep Boltzmann machines. Neur Comp, 2012, 24: 1967–2006
18 Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J Mach Learning Res, 2010, 11: 3371–3408
19 Dean J, Corrado G S, Monga R, et al. Large scale distributed deep networks. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS'12), 2012. 1223–1231
20 Wu F, Zhang H, Zhuang Y. Learning semantic correlations for cross-media retrieval. In: Proceedings of 2006 IEEE International Conference on Image Processing (ICIP'06), 2006. 1465–1468
21 Zhang H, Zhuang Y, Wu F. Cross-modal correlation learning for clustering on image-audio dataset. In: Proceedings of the 15th International Conference on Multimedia (ACM MM'07), 2007. 273–276
22 Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th International Conference on Multimedia (ACM MM'10), 2010. 251–260
23 Yang Y, Zhuang Y, Wang W. Heterogeneous multimedia data semantics mining using content and location context. In: Proceedings of the 16th International Conference on Multimedia (ACM MM'08), 2008. 655–658
24 Zhuang Y, Yang Y, Wu F. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Trans Multim, 2008, 10: 221–229
25 Yang Y, Xu D, Nie F, et al. Ranking with local regression and global alignment for cross media retrieval. In: Proceedings of the 17th International Conference on Multimedia (ACM MM'09), 2009. 175–184
26 Jia Y, Salzmann M, Darrell T. Learning cross-modality similarity for multinomial data. In: Proceedings of 2011 IEEE International Conference on Computer Vision (ICCV'11), 2011. 2407–2414
27 Lazarsfeld P F, Berelson B, Gaudet H. The People's Choice: How the Voter Makes up His Mind in a Presidential Campaign. New York: Columbia University Press, 1948
28 Granovetter M. The strength of weak ties. Am J Sociol, 1973, 78: 1360–1380
29 Krackhardt D. The Strength of Strong Ties: The Importance of Philos in Organizations. Boston: Harvard Business School Press, 1992
30 Burt R S. Structural Holes: The Social Structure of Competition. Boston: Harvard University Press, 1992
31 Barabási A L, Bonabeau E. Scale-free networks. Sci Am, 2003, 288: 56–69
32 Newman M E J, Girvan M. Finding and evaluating community structure in networks. Phys Rev E, 2004, 69: 026113
33 Palla G, Derényi I, Farkas I, et al. Uncovering the overlapping community structure of complex networks in nature and society. Nature, 2005, 435: 814–818
34 Rosvall M, Bergstrom C T. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA, 2008, 105: 1118–1123
35 Arenas A, Díaz-Guilera A, Pérez-Vicente C J. Synchronization reveals topological scales in complex networks. Phys Rev Lett, 2006, 96: 114102
36 Marvela S A, Kleinberg J, Kleinberg R D, et al. Continuous-time model of structural balance. Proc Natl Acad Sci USA, 2011, 108: 1771–1776
37 Milo R, Shen-Orr S, Itzkovitz S, et al. Network motifs: Simple building blocks of complex networks. Science, 2002, 298: 824–827
38 Kleinberg J, Suri S, Tardos E, et al. Strategic network formation with structural holes. In: Proceedings of the 9th ACM Conference on Electronic Commerce (EC'08), 2008. 284–293
39 Leskovec J, Huttenlocher D, Kleinberg J. Signed networks in social media. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI'10), 2010. 1361–1370
40 Adamic L A, Adar E. How to search a social network. Soc Networks, 2005, 27: 187–203
41 Wang C, Han J, Jia Y, et al. Mining advisor-advisee relationships from research publication networks. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10), 2010. 203–212
42 Zeng J, Zhang S, Wu C. A framework for WWW user activity analysis based on user interest. Knowl Based Syst, 2008, 21: 905–910
43 Scott J P. Social Network Analysis: A Handbook. 2nd ed. London: Sage Publications Ltd, 2000
44 Maia M, Almeida J, Almeida V. Identifying user profiles in online social networks. In: Proceedings of the 1st International Workshop on Social Network Systems, 2008
45 Backstrom L, Kumar R, Marlow C, et al. Preferential behavior in online groups. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM'08), 2008. 117–128

46 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 (KDD'13), 2013. 347–355
47 Moore C, Newman M E J. Epidemics and percolation in small-world networks. Phys Rev E, 2000, 61: 5678–5682
48 Kuperman M, Abramson G. Small world effect in an epidemiological model. Phys Rev Lett, 2001, 86: 2909–2912
49 Milgram S. The small world problem. Psychol Today, 1967, 2: 60–67
50 Christakis N A, Fowler J H. Connected: The Surprising Power of Our Networks and How They Shape Our Lives. New York: Little, Brown and Company, 2009
51 Zhang J, Tang J, Zhuang H, et al. Role-aware conformity influence modeling and analysis in social networks. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI'14), 2014. 958–965
52 Yang Y, Tang J, Leung C, et al. RAIN: Social role-aware information diffusion. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15), 2015
53 Wu S, Hofman J M, Mason W A, et al. Who says what to whom on Twitter. In: Proceedings of the 20th International Conference on World Wide Web (WWW'11), 2011. 705–714
54 Yu L, Asur S, Huberman B. What trends in Chinese social media. In: Proceedings of the 5th Workshop on Social Network Mining and Analysis (SNA-KDD'11), 2011

Deep analytics and mining for big social data

TANG Jie & CHEN WenGuang

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Online social networking is one of the most prominent features and main reasons of the big data era. In this paper, we first present the immense opportunities brought by big data in: advancing the technical level of the Chinese IT industry, leading development of the new Internet economy, and accelerating interdisciplinary research between sociology and information science. Next we analyze key issues in online social networks and point out new challenges of big social data research on semantic understanding and analysis, multi-modal association and fusion, group behavior analysis and mining, multidimensional analysis and visualization, and system development and integration. We then focus on introducing key international and domestic achievements in big data and online social networks. We conclude with a look at future trends in big social data analytics and mining.

social network, social influence, community discovery, social behavior prediction

doi: 10.1360/N972014-00954