大数据时代的简约计算

张家琳，孙晓明
中国科学院计算技术研究所，北京 100190

摘要
大数据存储和分析的能力是未来创新型国家的核心战略能力。当前关于大数据的理论研究在共性问题提炼、方法论框架和实时数据算法理论上仍存在一些不足，从大数据“海量、实时、多样”三大特征出发，聚焦网络大数据这一对象，以数据复杂性的度量和约简作为主线，具体从网络链路预测及推荐、动态演化网络上的算法研究、网络小世界模型与信息传播3个问题出发，研究大数据在时间、空间和关联关系上的简约计算。

关键词
时间复杂性；空间复杂性；关系复杂性；数据复杂性

中图分类号：TP3-0 文献标识码：A doi: 10.11959/j.issn.2096-0271.2016037

On the measurements of algorithms in big data era

ZHANG Jialin, SUN Xiaoming
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract
The ability to store and analyze big data is a crucial capability of a powerful country in the new century. The current research of big data contains weakness on common scientific questions, general methodology, and theoretical analysis of real-time algorithm. It started from three key features about big data: volume, variety, and velocity, and the measurement and simplification of time complexity, space complexity, and relationship complexity for big data were focused.

Key words
time complexity, space complexity, relationship complexity, data complexity

1 引言

随着网络、通信、感知等技术的迅猛发展，人类正进入大数据时代：根据国外相关机构预测，全世界数据总量以每两年翻一番的速度增长。近年来大数据的飙升主要来源于互联网服务，并且对大到国计民生小到衣食住行都产生了革命性的影响。因此在互联网上可访问到的人、机、物三元世界产生的网络大数据是大家关注的焦点。

网络大数据具有如下3个特点。

● 海量：网络空间中数据的体量不断扩大，IDC（International Data Corporation，国际数据公司）的研究报告称，2012年网络大数据总量为2.7 ZB，预计到2020年，总量将达到40 ZB。

● 实时：网络大数据通常以流的形式动态、快速地产生，具有很强的时效性，甚至呈现脉冲式的突发涌现，并且这些数据需要快速处理，实时响应。

● 多样：描述同一主题的数据往往来源多样，关联关系复杂，而且包含结构化、半结构化和非结构化等多种数据类型。

网络大数据在经济、社会、政治、科学等多方面都有不可估量的价值。美国政府认为大数据是“未来的新石油”，并把大数据研究上升为国家意志，这必然会在各个领域产生深远的影响。

（1）网络大数据的研究对捍卫国家网络空间的数字主权、维护国家安全和社会稳定有重要作用

信息化时代，国家层面的竞争力将部分体现为一国拥有网络大数据的规模、活性以及对数据的解释与运用的能力。国家在网络空间的数字主权也将是继海、陆、空、天四大空间之后的另一个大国博弈的空间。备受瞩目的“棱镜门”，深刻暴露出一些大国在有计划、有步骤地采集各国的数字“DNA”。2012年3月，美国国家科学基金会提出要“形成一个包括数学、统计基础和计算机算法的独特学科”——大数据科学。该计划还强调，大数据技术事关美国的国家安全，影响科学研究的步伐，还将引发教育和学习的变革。这意味着网络大数据的主权已上升为国家意志，直接影响国家和社会的稳定，事关国家的战略安全。

（2）网络大数据是国民经济核心产业信息化升级的重要推动力量

“人、机、物”三元世界的融合产生了大规模的数据，如何感知、测量、利用这些网络大数据成为国民经济中许多行业面临的共同难题。通过对网络大数据共性问题的分析和研究，使企业能够掌握网络大数据的处理技术或者能够承受网络大数据处理的成本与代价，进而使整个行业迈入数字化与信息化的新阶段。从这个意义上来看，对网络大数据基础共性问题的解决将是新一代信息技术融合应用的新焦点，是信息产业持续高速增长的新引擎，也是行业用户提升竞争力的新动力。

（3）网络大数据技术上的突破将催生出战略性新兴产业

网络大数据技术的突破意味着人们能够理清数据交互连接产生的复杂性，掌握数据冗余与缺失双重特征引起的不确定性，驾驭数据的高速增长与交叉互连引起的涌现性，进而能够根据实际需求从网络数据中挖掘出其蕴含的信息、知识甚至是智慧，最终达到充分利用网络数据价值的目的。网络数据已成为联系各个环节的关键纽带，通过对网络数据纽带的分析与掌握，可以降低行业成本、促进行业效率、提升行业生产力。在网络大数据技术的驱动下，行业模式的革新将

可能催生出数据材料、数据制造、数据能源、数据制药、数据金融等一系列战略性的新兴产业。

(4)大数据正在引起学术界对科学研究思维与方法的一场革命

传统科学研究的范式是从现象中分析提炼理论假设，再利用实验验证相应的理论。大数据的出现催生了一种新的科研模式，即面对大数据，科研人员只需从数据中直接查找、分析或挖掘所需的信息和知识，这些知识表现为概率形态的关联或因果关系，这种关系可能复杂到无法为人类直观掌握，但是可以很好地解释现实、预测未来。图灵奖得主Gray J在他的最后一次演讲中描绘了数据密集型科学研究的“第四范式”，把数据密集型科学从计算科学中单独区分开来。Gray认为，要解决面临的某些最棘手的全球性挑战，“第四范式”可能是唯一系统性的方法。

大数据研究方兴未艾，成果累累，每年仅在Nature及其子刊、Science和PNAS上发表的大数据分析相关论文就有近百篇。其中，网络大数据又扮演中心的角色。从计算机科学的角度看，目前的研究主要有3方面有待进一步加强。首先，目前还缺乏专门针对海量实时流式数据的算法理论、算法设计与评估框架。其次，对于特定数据对象的研究较多，对于共性问题的提炼和分析较少，还缺乏可察觉的方法论的主线。最后，在静态数据或离线数据上的算法测试类研究较多，在真实系统中的大规模实验较少，还缺乏可信赖的效果评估。因此，数据科学，甚至说“第四范式”，都还只是一个模糊的雏形。

本文尝试从数据复杂度的角度进行突破，针对网络大数据所具备的“海量、实时、多样”三大特征，依托国家自然科学基金重点项目“大数据结构与关系的度量与简约计算”，围绕大数据时间、空间、关联复杂性的度量和约简展开，希望探索出符合当前实时海量流式数据处理的，新的算法复杂性理论基本思想和算法设计的基本框架，寻找从时间、空间和特征关联3方面约简数据和处理数据的算法，从而对数据科学基础理论和基本方法论的形成产生贡献。

具体来说，本文将集中关注与网络大数据有关的算法理论和应用问题，围绕重点项目“大数据结构与关系的度量与简约计算”实施一年多来在网络链路预测与推荐、网络小世界模型及信息传播、动态演化网络的相关算法3方面取得的一些进展进行汇报，展示对数据复杂度的认识和理解。

2 相关工作

2.1 网络大数据计算

传统的CPU密集型的计算，数据量不大，算法复杂度往往只要求是多项式级即可，理论研究的焦点也在于区分多项式级和非多项式级的算法。而网络大数据计算动辄面临TB乃至PB级的数据规模，计算从CPU密集型转化为数据密集型。算法设计的关键是保证时间为线性甚至亚线性。另一方面，数据传输（无论从外存读取还是网络上传输）的时间开销远大于CPU处理时间，这使得CPU不再成为计算的瓶颈。因此，计算方法的重点变成了努力降低算法涉及的数据的移动开销。主要思路有3类：分散化、局部化和增量化。

(1)分散化

大机群分布式计算是高效大数据处理的首选，因为单个计算节点的工作负载可以大幅度降低，特别是当数据分散存储的时候，通过分布式计算可以减少数据的跨节点流动，降低数据移动开销。Google（谷歌）公布的MapReduce编程模型在工业界

乃至学术界产生了极大的影响，以至于“谈大数据必谈MapReduce”[1]。

（2）局部化

网络局部性算法最早指的是在网络分布式计算中，每个计算节点的输出仅仅与常数跳范围内的邻居节点有关，与整个网络的规模无关[2]。在网络大数据背景下，网络规模巨大且动态演化，对整个网络结构的存储、快照、访问都需要耗费高昂的成本，此时局部性算法不再强调分布式，而是关注网络以数据流的形式输入，如何实时处理以及如何只访问网络局部的数据就能够获得计算结果。局部算法在时间复杂度上具有明显的优势（亚线性甚至常数时间），在复杂网络的计算中越来越受到关注。

（3）增量化

在动态网络中，每个时刻的网络数据都可以看作在前一时刻数据基础上作了一定的偏移（称为增量）。如果观察间隔较短，那么相对于整个网络规模，增量一般不大。如果基于增量更新网络的特定性质，在理想情况下，更新算法的时间复杂度不依赖于整个网络的规模，仅仅与增量有关，这类算法称为增量式算法。Desikan P等人[3]针对动态网络的Pagerank更新，把网络中的点分类，使得需要重新计算的点数很少，该方法后来被Bahmani B等人[4]推广到Monte Carlo的Pagerank算法。

2.2 网络大数据特征刻画和结构挖掘

复杂网络的特性主要由一些统计值来刻画，如度分布、最短路径长度等，这些宏观特征是由各个节点的动力学行为及其节点之间相互作用产生的集中表现。1998年，Watts D J等人[5]分析了网络中的高聚集性和短特征路径长度等特性，研究了网络“小世界”特性产生的机制。对于静态网络，通常采用拓扑距离刻画网络的最短路径长度，而对于动态变化的时序网络，一般采用时序路径长度进行刻画[6]。Pan R K 等人[7]提出对时序网络中的时序路径进行确切的定义并给出了相应的计算算法。

Newman M E J[8]的研究成果，使得复杂网络中的社区发现成为近几年复杂网络领域的一个研究热点，并形成了复杂网络中一个重要的研究方向。Fortunato S[9]在Physics Reports上给出了社区发现的综述。2004年，Newman M E J[10]提出了基于模块度优化的快速算法。随后，研究者在Newman M E J等人的工作基础上，提出了多种类型的基于模块度优化的算法。

2.3 基于网络的缺失预测和趋势预测

网络中的链路预测是指如何通过已知的网络结构信息来预测网络中尚未产生连边的两个节点之间产生连接的可能性。由于实际应用中通常存在严重的数据缺失问题，链路预测可以通过推断补齐这些缺失连边，从而更加准确地对网络进行分析，链路预测已成为准确分析社会网络和生物网络的有力辅助工具[11]。另外，社会媒体中的推荐问题，譬如Facebook上的朋友推荐和新浪微博中的关注对象推荐，本质上也是链路预测问题[12]。

推荐系统通常包括3个组成要素：用户、对象和推荐方法，其中推荐方法是整个推荐系统的核心。笔者主要考虑基于网络的推荐系统。在简化的情况下，推荐系统可视为二部分图上的链路预测问题。在大数据环境下，推荐系统规模很大，用户和商品数目动辄百千万计，两个用户之间选择的重叠非常少，使得绝大部分基于关联分析的算法（譬如协同过滤）的计算效果都不好。事实上，网络方法很早就应用于推荐系统。例如，Aggarwal C C等人[13]研究

了基于图(网络结构)的协同推荐算法,结果表明基于图的协同过滤方法在计算速度、推荐精度、可扩展性、学习时间等方面均优于传统的协同推荐算法。Huang Z等人[14]用二层图模型刻画客户—产品推荐系统,讨论了二部分图的小世界效应和集聚性质对不同推荐算法的影响。

3 网络链路预测及推荐

3.1 “结构微扰法”链路预测方法

链路预测是网络科学中一个重要的基础问题[15]。精准的预测结果既可以指导生物学的实验,还可以进行社交网络的好友预测。好的预测算法本身还给出了很多网络演化可能机制的暗示。遗憾的是,人们并不知道一个算法是否“足够精确”。针对一个完全随机的网络,“什么都预测不到”可能已经是最好的结果了,但针对一个非常规则的网络,精心设计的方法可能能够100%进行预测。知道了一个网络的链路在多大程度上“能够被预测出来”,能够使得人们去判断算法是否已经接近甚至达到预测的上界,是否还有提升的空间。

事实上,“可被预测的程度”本身也可以看作网络的一种重要性质。为了衡量网络可被预测的难易程度,Lü L等人[16]提出了如下假设:网络越具有某些规律性,越容易被预测。进一步地,如果随机从网络中抽取出一小部分链路,网络的特征向量空间受到的影响很小,就说明网络是具有规律性的。Lü L等人使用类似于量子力学中对哈密顿量做一阶微扰的方法,假定减少或增加少量链接所产生的微扰,只对特征值有影响,而对特征向量没有影响,这样可以观察微扰后通过这种办法重构的邻接矩阵和真实邻接矩阵的差异。Lü L等人提出了一种度量这个差异的参数—结构一致性(structural consistence),被认为可以直接用来刻画网络的“可被预测的程度”[16]。

大量的模拟网络和真实网络实验都支持了上述结论:结构一致性越强的网络越容易被准确预测丢失的链路。Lü L等人利用结构一致性,提出了一种新的名为“结构微扰法”(structural perturbation method)的链路预测方法。这个方法在预测丢失的链路以及甄别网络中添加的噪音边两方面都明显超过了当前主流的方法,包括知名的层次结构法和随机分块法。

3.2 场景自适应的跨领域推荐

数据稀疏是推荐系统面临的一大挑战。跨领域推荐通过融合多个领域的数据来克服数据稀疏问题。现有的跨领域推荐方法主要有两类:第一类基于同质性假设,即假设同一个对象在不同的领域共享同一个表达,这类方法适用于在每个领域都稀疏的对象,但不能刻画领域对对象的影响;第二类基于异质性假设,即假设每个领域有一个领域独有的变换矩阵,每个对象在不同场景中的表达由该对象的全局表达和领域变换矩阵相作用得到,这类模型适用于在部分领域稀疏而在其他领域不稀疏的对象,但对于在每个领域都稀疏的对象效果很差。针对上述问题,Shen H W等人[17]提出了一种场景自适应的跨领域推荐方法(context-adaptive matrix factorization, AdaMF),对象的表达建模为其全局表达和场景相关表达的一个混合分布,采用混合系数来自适应地调节全局表达和场景相关表达的作用。在MovieLens-Netflix数据集上的实验表明,AdaMF在稀疏—稀疏、稀疏—稠密、稠密—稠密等各个场景下都一致性地优于

现有的两类代表性方法。

3.3 基于用户行为的购物推荐

如何对用户下一次的购物数据进行预测是市场分析里的重要问题。传统的方法有两种：一种是基于商品顺序的推荐，这种方式捕获了用户购物的顺序行为，但是忽略了购物推荐的个性化因素，并且缺乏用户对商品整体兴趣的描述；另一种是协同过滤，这种方式忽略了用户交易的特征，将用户所有购买的商品混在一起建模。为了解决以上问题，Lan Y Y等人[18]提出了层次化表达模型（hierarchical representation model）来完成用户的购物推荐。参考文献[18]中假设用户的表达和商品的表达均在同一个连续的空间中，商品的表达可以通过操作符合成交易的表达，用来代表用户购物的顺序行为，用户的表达代表用户的整体兴趣。在模型的第二层使用操作符将两个表达合并在一起作为用户当前的兴趣表达来预测用户下一步购买的商品。在和多个baseline进行比较的实验中，Lan Y Y等人的模型在f-measure、hit-ratio以及NDCG指标上均取得了较好的性能。

4 动态演化网络算法研究

4.1 动态演化网络排序算法

排序作为最基本而经典的算法问题，在大数据时代依然是众多关键应用的基石，如搜索、推荐系统等。笔者研究了访问受限的动态数据模型下的排序和查找问题[19]。借鉴Anagnostopoulos等人提出的动态数据的模型，采用Kendall tau距离作为衡量算法性能的指标。笔者研究了Top$k$ selection问题：在每个时刻t，找出Top$k$的元素并将其排序。之前Anagnostopoulos等人的工作只研究了两个极端情况k=1和k=n。笔者的主要贡献是确定了该问题的“相变点”——k^*，即当k=$o(k^*)$时，该问题可以以1−o(1)的概率无差错地解决。同时笔者证明了当k超过k^*时，对于任何算法，所求得的顺序与真实顺序的Kendall tau距离都至少是k^2/n，而且笔者的算法表明这个界是紧的。笔者还研究了比Top$k$ selection弱的一个问题：Top$k$ set问题。在这一问题中笔者只需要确认Top$k$的元素，而不需要确定它们的顺序，证明了对任意的k，Top$k$ set问题都可以以1−o(1)的概率无差错地求解。

4.2 基于动态距离的网络社区发现算法

社区挖掘是大规模网络分析和挖掘的基础，它在社交网络、生物网络、脑网络等诸多方面都有重要的应用。但如何有效地挖掘大规模网络中存在的社区结构仍然面临着巨大的挑战。针对这个基础理论研究问题，Shao J等人[20]提出了一个新的社团挖掘算法：Attractor算法。该算法的基本思想是将网络看作一个动力学系统，每个节点与周围节点进行交互，提出3种直观的交互模式，通过模拟网络中节点间的距离变化动态地发现社区结构。由于社区检测是基于网络内在的连接模式，因此该算法能找出网络中不同大小的固有社团。同时由于算法的时间复杂度低，因此可以处理大规模网络。大量人工数据集和真实数据集实验都表明Attractor算法相比传统算法更有优势。这一工作为大规模网络中的社区挖掘问题提供了新的思路和方法。

4.3 并行秘书问题在线算法

秘书问题是20世纪60年代提出的经

典在线问题，笔者研究了这个问题的一个一般化变种，并在并行模式下考虑了这个经典的在线优化问题[21]。假设雇主计划从n个完全随机到来的候选人中选择J个人。雇主对于不同的候选人有着不同的评价，想要录取的这些人尽可能是前k好的。这里数据是以流式的方式到来的，每面试完一个候选人，面试官才知道当前候选人的价值，并且要立即决定是否录取这个人，不可反悔。笔者在研究中提出了一个基于观察—选择的确定性算法。这个算法具有高效、易实现的特点，并且从线性规划出发，利用互补松弛定理，可以证明该算法的最优性。笔者的算法同样可以用于解决当各队列的名额是预先指定的情况，从而解决了EC2012上Feldman等人的文章中的一个未解问题。针对两个典型的例子，给出了算法的近似比。

5 网络小世界模型与信息传播

5.1 基于博弈论的小世界模型

小世界模型是复杂网络模型中的一个重要模型。它刻画了各种复杂网络中经常出现的平均距离很短而聚合度较高的现象。2002年Kleinberg J提出了适于通行的小世界网络的概率模型，指出当模型中的随机长边幂率分布系数r等于基准格子网络的维度时，小世界网络才是可通行的。之后的实证研究印证了现实的社交网络的幂率系数r确实接近于网络的有效维度。

Chen W等人[22]从博弈论的角度出发，将网络中的每个节点看作一个网络博弈的玩家，其长边幂率分布系数r是其策略，r值偏大表示该节点侧重于连接其附近的节点，随着r值减小，其连接格子上较远距离节点的概率增加。Chen W等人在这一网络博弈中独创性地引入了一个新的效用函数，使得每个节点的效用是其随机长边的平均格子距离与随机长边有反向边的平均概率的乘积。前者表明，节点想连接远处的节点以得到不同的信息，而后者表明节点倾向于连边的互惠性（reciprocity）以使联系更加稳定。Chen W等人在理论上论证了DRB（distance-reciprocity balanced，距离—互惠平衡）博弈仅有两个纳什均衡，而适于通行的小世界网络是唯一一个稳定的均衡，任何团体都无法通过共谋偏离这个均衡以使得团体的成员获利，而且即使绝大多数节点都随机扰动，节点也能很快回到适于通行的小世界模型状态。他们还通过模拟实验进一步验证了即使节点不了解其他节点的连接偏好，也同样会收敛到适于通行的小世界网络。Chen W等人还通过人人网和美国LiveJournal两个实际网络进行了验证，实验发现DRB博弈仍能很快收敛，而收敛后节点的连接偏好与实测结果的相关度相当高，其平均值也接近网络的有效维度。

5.2 影响力最大化问题

影响力模型和最大化研究大多数基于独立级联模型（independent cascade）的影响力最大化问题，主要考虑单个个体传播或纯竞争性多个体传播，传播过程是一次性的，并且传播结果用期望值作为度量标准。在此基础上，从几个不同的角度对问题进行了推广。

笔者首次提出了基于概率保证的影响力最大化问题[23]，典型的应用是：话题或事件希望能以一定的概率保证覆盖超过一定比例的节点，以此来争夺社交网站上的热点事件或者十大话题等。笔者考察当对同一事件或物品的信息传播反复多次出现时，其影响概率逐渐累积之后，会对节

点决策产生的影响，并基于此提出了基于概率累积的影响力最大化问题[24]。Lu W等人[25]还首次提出了一个比较独立级联模型（comparative independent cascade model, Com-IC model），将双个体在竞争或互补情形下的传播方式统一表述在一个模型下。文中研究了模型的性质，并着重研究了在互补情形下的影响力最大化问题。基于此改进了基于反向可达集合的高效算法，并提出了夹心近似策略，当影响力函数本身不具备子模性（submodularity）时仍能给出一定的近似比。

5.3 基于资源分配的影响力节点发现算法

通过考虑邻居节点的资源以及传播率对目标节点的影响，Shang M S等人[26]提出了一种改进的迭代资源算法来识别影响力节点。该方法认为目标节点的重要性程度受邻居感染情况以及传播率的影响，邻居的影响力资源为基本的中心性，如：度、k核、接近中心性、特征向量中心性等。通过在4个真实网络中的SIR模型结果比较，该方法和原有的方法相比在没有增加参数以及复杂度的情况下，提高了精确度。特别地，在Erdos-Renyi网络里，kendall系数提高了23%左右，在Protein网络里提高了24%左右，效果比较明显。该改进的迭代资源算法考虑了网络结构以及传播属性，可以更好地识别网络中的重要性节点，结合网络结构和传播动力学机制对识别核心节点具有重要的启示作用。

6 结束语

本文聚焦网络大数据这一当前热点领域，从网络链路预测及推荐、动态演化网络算法研究以及网络小世界模型与信息传播3个方面，展示如何从数据复杂度的角度对大数据的算法设计进行突破。希望能通过提出新的算法复杂性理论的基本思想和算法设计的基本框架，对数据科学基础理论和基本方法论的形成产生贡献。

致谢

在本文的撰写过程中，得到了周涛教授、陈卫研究员、陈端兵教授、沈华伟研究员、邵俊明教授等的大力支持和帮助，部分素材来源于他们的研究工作，在此一并表示真诚的感谢！

参考文献：

[1] DEAN J, GHEMAWAT S. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[2] SUOMELA J. Survey of local algorithms[J]. ACM Computing Surveys (CSUR), 2013, 45(2): 94-111.

[3] DESIKAN P, PATHAK N, SRIVASTAVA J, et al. Incremental page rank computation on evolving graphs[C]//The 14th International Conference on World Wide Web, May 10-14, 2005, Chiba, Japan. New York: ACM Press, 2005: 1094-1095.

[4] BAHMANI B, CHOWDHURY A, GOEL A. Fast incremental and personalized PageRank[J]. Proceedings of the VLDB Endowment, 2010, 4(3): 173-184.

[5] WATTS D J, STROGATZ S H. Collective dynamics of “small-world” networks[J]. Nature, 1998, 393(6684): 440-442.

[6] HOLME P, SARAMÄKI J. Temporal networks[J]. Physics Reports, 2011, 519(3): 97-125.

[7] PAN R K, SARAMÄKI J. Path lengths, correlations, and centrality in temporal networks[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2011, 84(1 Pt 2):1577-1589.

[8] NEWMAN M E J. Modularity and community structure in networks[J]. Proceedings of the National Academy of Sciences, 2006, 103(1): 8577-8582.

[9] FORTUNATO S. Community detection in graphs[J]. Physics Reports, 2010, 486(3-5): 75-174.

[10] NEWMAN M E J. Fast algorithm for detecting community structure in networks[J]. Physics Review, 2004, E 69, 066133.

[11] SCHAFER L, GRAHAM J W. Missing data: our view of the state of the art[J]. Psychological Methods, 2002, 7(2): 147-177.

[12] ZHANG Q M, LÜ L, WANG W Q, et al. Potential theory for directed networks[J]. Plos One, 2013, 8(2): e55437.

[13] AGGARWAL C C, WOLF J L, WU K L, et al. Horting hatches an egg: a new graph-theoretic approach to collaborative filtering[C]//The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA, USA. New York: ACM Press, 1999: 201-212.

[14] HUANG Z, ZENG D, CHEN H. Analyzing consumer-product graphs: empirical findings and applications inrecommender systems[J]. Management science, 2007, 53(7):1146-1164.

[15] LÜ L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A Statistical Mechanics & Its Applications, 2011, 390(6):1150-1170.

[16] LÜ L, PAN L, ZHOU T, et al. Toward link predictability of complex networks[J]. Proceedings of the National Academy of Sciences of the United States of America, 2015, 112(8): 2325-2330.

[17] MAN T, SHEN H W, HUANG J M, et al. Context-adaptive matrix factorization for multi-context recommendation[C]// The 24th ACM International Conference on Information and Knowledge Management(CIKM), October 19-23, 2015, Melbourne, Australia. New York: ACM Press, 2015: 901-910.

[18] WANG P F, GUO J F, LAN Y Y, et al. Learning hierarchical representation model for next basket recommendation[C]//The 38th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015), August 9-13, 2015, Santiago, Chile. New York: ACM Press, 2015: 403-412.

[19] SUN X M, ZHANG J, ZHANG J L. Solving multi-choice secretary problem in parallel: an optimal observation-selection protocol[C]//The 25th International Symposium on Algorithms and Computation (ISAAC 2014), December 15-17, 2014, Jeonju, Korea. [S.l.]: Springer International Publishing, 2014: 661-673.

[20] SHAO J, HAN Z, YANG Q, et al. Community detection based on distance dynamics[C]// The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015, Sydney, Australia. New York: ACM Press, 2015:1075-1084.

[21] HUANG Q, LIU X W, SUN X M, et al. How to select the top-k elements from evolving data? [C]// The 26th International Symposium on Algorithms and Computation (ISAAC 2015), August 3-8, 2015, Macao, China. New York: ACM Press, 2015: 60-70.

[22] YANG Z, CHEN W. A game theoretic model for the formation of navigable small-world networks[C]//The 4th International World Wide Web Conference (WWW'2015), May 18-22, 2015, Florence, Italy. [S.l.:s.n.], 2015.

[23] ZHANG P, CHEN W, SUN X M, et al. Minimizing seed set selection with probabilistic coverage guarantee in a social network[C]//The 20th ACM SIGKDD

International Conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York, USA. New York: ACM Press, 2014: 1306-1315.

[24] SHAN X H, CHEN W, LI Q, et al. Cumulative activation in social networks[J]. 2016, arXiv:1605.04635.

[25] LU W, CHEN W, LAKSHMANAN L V S. From competition to complementarity: comparative influence diffusion and maximization[C]//The 42nd International Conference on Very Large Data Bases (VLDB' 2016), September 5-9, 2016, New Delhi, India. New York: ACM Press, 2015: 60-71.

[26] ZHONG L F, LIU J G, SHANG M S. Iterative resource allocation based on propagation feature of node for identifying the influential nodes[J]. Physics Letters A, 2015, 379(38): 2272-2276.

作者简介

张家琳(1983-),中国科学院计算技术研究所副研究员,主要研究方向为在线算法、近似算法、社交网络、算法博弈论等。

孙晓明(1978-),中国科学院计算技术研究所研究员,主要研究方向为算法与计算复杂性、量子计算等。

收稿日期: 2016-06-20

基金项目: 国家自然科学基金资助项目(No.61222202, No.61433014, No.61502449);中组部万人计划青年拔尖人才项目

Foundation Items: The National Natural Science Foundation of China (No. 61222202, No.61433014, No.61502449), The China National Program for Support of Top-notch Young Professionals