数据挖掘学科发展报告

唐 杰， 梅俏竹

清华大学计算机系

美国密歇根大学

Recent Advances of Data Mining in China

Jie Tang and Qiaozhu Mei

Department of Computer Science, Tsinghua University

School of Information, University of Michigan

Data mining is concerned with the process of discovering novel, non-trivial, and potentially useful knowledge from large-scale data sets. Because of its broad applications to various domains, the field of data mining has attracted tremendous attention from both academia and industry. In recent years, Chinese institutes have made significant contribution to the research frontier of the data mining community. This report summarizes the recent advances of data mining research in China in respect of theoretical foundations, social network mining, and big data mining. We highlight recent publications by Chinese research groups in the top venues of data mining research. Compared to the developments in leading international research groups which emphasize on theoretical and interdisciplinary research, data mining research in China present a focus on practical real world applications. In general, there is a significant trend of data mining research towards social networks and big data, which provides a considerable opportunity for Chinese researchers to establish novel research directions and to make a broader impact.

1．引言

数据挖掘是知识发现过程中的一个关键步骤，一般是指从大量数据中自动发现隐含的的数据关系，并将其转化为计算机可处理的结构化表示。数据挖掘是计算机学科中的一个交叉研究领域，其研究方法与多个其他科学紧密相连，如：统计、机

器学习、专家系统、信息检索、社会网络、自然语言处理和模式识别等等。

近年，随着各行业对大规模数据处理和深度分析需求的快速增长，数据挖掘引起了研究界和工业界的广泛关注。自 1995 年以来，学术界和工业界共同成立了 ACM 的数据挖掘及知识发现专委会，并组织了国际数据挖掘与知识发现大会（ACM SIGKDD Conference on Knowledge Discovery and Data Mining，简称 KDD [1]），后者发展成为数据挖掘领域的顶级国际会议。至今 KDD 大会已经连续举办了 19 届，论文的投稿量和参会人数呈现出逐年增加的趋势。图 1 给出了自 2001 年以来 SIGKDD 每届接收的论文投稿数和最终录用的论文数的对比。近几年，以社会网络和信息网络为中心的大数据分析成为数据挖掘研究的热点。本报告围绕数据挖掘领域近年最主要的几个研究方向（基础理论、社会网络和大数据），以数据挖掘顶级国际会议 KDD 和国际期刊 IEEE TKDE、ACM TKDD 上发表的论文为基础介绍近几年国内学者在数据挖掘领域的主要研究进展，分析和比较国际国内学科发展趋势，并展望未来发展机遇。

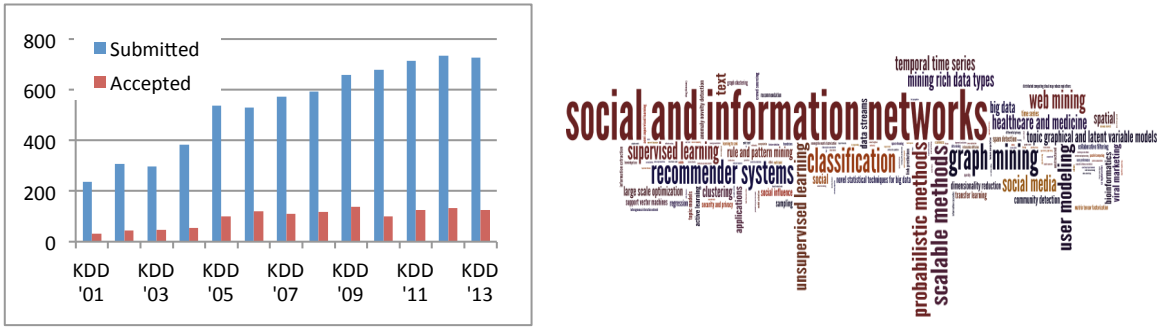

图 1. 数据挖掘国际会议 KDD 历年投稿和论文接收情况(左)；KDD 2013 研究热点(右)

2. 研究现状和主要成果

2.1 数据挖掘基础理论

最早的数据挖掘理论基础主要源于统计，机器学习和数据库系统。经过近 20 年的发展，数据挖掘领域逐渐形成了一套自己的基础理论，主要包括规则和模式挖掘，分类、聚类、话题学习等。近年，随着网络数据的规模和复杂性的快速增长，时间序列和空间数据挖掘、以及基于大规模网络（图）的稀疏学习也得到越来越多的重视。以下我们简要介绍国内学者在数据挖掘基础理论上的最新成果。

在分类学习方面，清华大学的张长水团队研究了多任务的特征学习方法， 提出了名为 rMTFL 的学习方法。该方法首先将多任务和不同特征的关系用矩阵表示，并

基于 Group Lasso 的思想抽取出相关任务的特征空间，并因此找出孤立任务[6]。清华大学的靳晓明等人针对跨域的文本分类，提出跨域的主动学习方法[17]。该方法有效地结合了不同数据源的特征，自动从多数据源中抽取同质特征并区分异构特征，从而有效的选取样本进行主动学习。南京大学的周志华带领的课题组提出分类算法中应使用代价区间（cost interval）而不是精确的代价值，因为实际应用中，用户常常只能判断各类错误的相对严重性而无法给出精确描述。他们提出的 CISVM 算法将 cost interval 应用于 SVM，比使用任何单一代价的标准 SVM 减少了 60%的风险[21]。他们还进一步提出名为 MAHR 的分类算法。该算法可以自动发现分类结果之间的关联关系，从而提高分类精度[13]。在多类标的学习中，由于每个样例可以和多个类标关联，可能的类标集非常多，导致多类标分类和预测常常比较困难。东南大学的张敏灵等人使用贝叶斯网络刻画类标之间的依赖关系，将多类标学习问题分解为一系列的单类标分类问题，从而在多个数据集上超越了现有方法的效果[44]。流数据分类是分类学习中的一个重要分支，集成学习是对流式数据进行分类的常用方法，但线性扫描每个分类器会带来很大的时间开销。中科院的张鹏等人提出了一种新颖的 Ensemble-tree(E-tree)方法，利用类似 R-tree 的高度平衡的结构将流数据分类中集成学习的复杂度由线性降低到次线性[41]。概率图模型是数据挖掘中的重要基础工具，北京大学的宋国杰等人提出基于重叠分解的概率图模型[8]，其基本思路是将原始的概率图分解为若干小的概率图进行求解。其论文给出理论证明，求解出这样的近似分解和对原始概率图模型进行一步正则化处理是等价的。中国科技大学的俞能海等人还将概率图模型应用于个人简介的自动抽取，基本思路是用马尔可夫逻辑网络实现信息抽取并自动生成类似维基百科的页面[20]。

无监督的聚类和话题学习是数据挖掘领域研究的另一个核心问题。清华大学的张长水等人提出了从多重相关、随时间变化的语料库中挖掘文本簇演变的方法。他们通过加入相邻时间片的依赖，将层次化 Dirichlet 过程（HDP）扩展为 evolutionary HDP（EvoHDP）。这种方法可以发现文本簇的产生、消失，以及语料库内部和多语料库之间的演变[42]。浙江大学的蔡登等人研究了非监督学习中特征选择问题[1]。针对传统方法忽略了特征之间的联系，他们提出融合流型学习和一阶正则化方法，选择能使原始数据的簇结构保留得最好的特征，并提出了一个高效的聚类方法 Multi-Cluster Feature Selection。浙江大学的张仲非等人还将半监督的学习方法应用于图片标注，他们提出的半监督的层次化 Dirichlet 过程方法(SSC-HDP)，在图片标注的

实验中比现有的 MoM-HDP 和 Corr-LDA 模型取得更好的效果[28] 。无监督的数据补齐是一个很有挑战的问题，浙江大学的何晓飞和亚利桑那州立大学的叶杰平一起对矩阵补齐问题进行了深入研究，提出解决该问题的一个高效算法 Accelerated Singular Value Thresholding (ASVT)。该方法将原来 SVT 的收敛速度从 $O(1/N)$加速到 $O(1/N^2)$（N 是算法中的迭代次数）[11]。北京大学的张铭等人提出利用话题模型对社交网络中的用户生成内容进行建模的方法。他们通过构建不同的上下文来增强话题模型的效果，避免社交网络中的数据稀疏问题。北京大学的王厚峰和微软的周明等人将话题模型应用于 Twitter 数据以生成面向实体的用户观点摘要 [25]。其基本思路是利用 Affinity Propagation 算法对 Twitter 内容中的 Hashtag 进行聚类，然后再对实体相关的情感进行分类。西安电子科技大学的研究团队也研究了多信息源的半监督学习问题[29]。

从海量数据中挖掘出潜在规则和模式是数据挖掘中的基础问题。清华大学的王建勇研究了不确定性数据上判别模式（Discriminative Pattern）的挖掘问题，提出了 uHARMONY 算法，从数据库中直接找出判别模式，无需进行耗时的特征选择，使用 uHARMONY 的 SVM 相比经典不确定分类算法有 4%~10%的性能提升[5]。哈尔滨工业大学的李建中等人研究非确定图中的频繁子图挖掘问题，引入 ϕ-频繁概率来衡量一个子图的频繁程度。他们提出了一种近似算法，估计并证明了找到解的概率[47]。

2.2 社交网络分析和图挖掘研究

社交网络分析是指利用统计方法、图论等技术对社交网络服务中产生的数据进行定量分析。社交网络分析和图挖掘无疑是近年数据挖掘领域最热的话题，仅今年数据挖掘国际大会 KDD 2013 上面相关的大会报告分会场就有 7 个（共 27 篇论文），其他相关的 Poster 论文还有 30 余篇。从总的趋势来看，数据分析和挖掘的任务变得更加细化。从社交网络分析情况来看，其中三个最热的话题是：网络结构分析、群体行为和影响力建模以及网络信息传播的分析；从图挖掘方面来看，其中最热的研究问题是：图模式挖掘和基于图的学习算法研究；此外也有很多关于社交网络和图挖掘的应用，例如社交推荐、社交搜索等。下面分别从这几个方面总结一下研究进展。

在网络结构分析方面，从宏观的网络聚类系数估计、到中观的网络社区发现、

再到更微观的网络关系挖掘都有不少的研究工作发表。西电的黄健斌等人研究了网络社区发现问题，提出基于网络密度聚类的算法，该算法不仅可以发现任意大小和形状的网络社区，还可以自动检测网络关联节点（Hub）和孤立节点[12]。东北大学的于戈等人深入研究了图聚类算法，该算法也可以应用于网络数据的社区发现问题[7]。清华大学的唐杰等人和 UIUC 大学的韩家炜等人合作对社交网络关系进行深入研究，他们发现通过对网络用户的行为和用户交互进行挖掘，能够自动识别出用户之间的社交网络关系来；他们提出基于时间的概率因子图模型，实现了无监督的自动学习[32]。

网络用户行为，尤其是群体行为是社交网络分析区别于传统信息网络分析的关键因素；社会影响力又是网络用户行为的驱动力。在社会影响力方面，清华大学的唐杰等人提出基于话题的影响力度量模型[33]，模型基于主题的网络关系图，能够定量且细粒度地衡量结点之间的影响。针对大网络数据，他们还基于 Hadoop 设计了模型的并行算法。他们还进一步探讨了如何基于社会影响力对网络用户行为进行预测，提出网络用户行为的容噪预测模型 NTT-FGM。该模型同时对网络结构、用户属性和用户历史行为建模，显著提高了用户行为预测的精度[32]。影响力传播最大化是社会影响力研究中的另一个重要问题，北京大学的宋国杰等人对该问题进行了深入研究，他们基于移动网络的特点，利用账号信息的传播探测网络中的社区，进一步使用动态规划选择一些社区中最有影响力的 K 个节点，并且证明了近似算法的精度界[36]。在用户行为建模方面，中科院和 IBM 中国研究院的项亮和袁泉等人提出了 Session-based Temporal Graph（STG）来刻画用户随时间变化的长期、短期偏好，并基于 STG 模型提出了新颖的时序推荐算法 Injected Preference Fusion，在两个实际数据集上相对经典方法取得了 15%~34%的性能提升[38]。用户行为模型有很多相关的实际应用。例如：清华大学的王建勇等人利用用户兴趣模型来提高 Twitter 数据中的命名实体识别精度[25]。唐杰等人研究了网络用户行为的从众现象，提出名为 Confluence 的概率模型对用户行为进行建模和预测[34]。该模型很好的区分了用户的个体从众性和群体从众性，在多个社交网络数据集上的试验验证了该方法的有效性。

信息传播是社交网络研究中的一个核心问题，传统的信息传播研究主要集中在传播模型的设计和分析上，例如：疾病传染模型 SIR 模型和 SIS 模型以及小世界网络中 SIR 模型等。近年的热点和趋势在于从线社交网络的大规模用户交互数据分析信息在在线网络中的传播机理。清华大学的唐杰等人提出基于网络流差最大化模型

来自动识别网络中控制信息在不同社区间传播的“结构洞”用户(Structural Hole Spanner)[22]，该工作证明了从大规模网络中自动发现结构洞用户的问题是一个NP-Hard 的问题，并提出了具有理论近似度的求解算法，在 Twitter 网络和学术网络上都取得了很好的验证效果。清华大学的崔鹏等人还基于逻辑回归算法对腾讯微博上的信息传播模式进行预测[3]。

社交网络相关的应用研究很多，其中最重要的就是社交推荐，包括信息推荐和好友推荐等。清华大学的唐杰等人研究了社交网络跨领域（跨社区）的推荐问题，提出跨领域话题学习方法(Cross-domain topic learning)[35]。该方法解决了跨领域推荐的三个关键难点：链接稀疏性、知识互补性和话题偏斜性，提高了交叉领域合作者推荐的精度。王建勇等人利用异构网络建模的结果来提高个性化的标签推荐精度，其基本思路是利用有导随机游走模型学习不同类型关系和不同类型节点对标签推荐的重要性[4]。他们还进一步研究了社交网络中基于位置信息的用户群组推荐方法[43]。此外，他们还基于关键词传播的思想设计了网络视频的描述信息补齐和噪音消除算法[40]，该算法结合了文本相似度和时间相似度，在优酷的数据集上取得了很好的效果。李国良等人也研究了基于位置的推荐方法，他们的核心思路是提高数据索引的效率，提出将结构和内容相结合的基于 R-tree 的索引方法 [18]。北京大学的崔斌等人提出名为 LCARS 的推荐模型，LCARS 使用话题模型对社交数据中的位置、内容以及用户兴趣同时进行建模[39]，他们在国内豆瓣网络(Douban.com)的数据上进行了实验验证，得到了更高的推荐精度。南京财经大学的武之昂和北京航空航天大学的吴俊杰研究了推荐系统中的“托攻击”现象，提出基于 MC-Relief 的特征选择方法以及半监督简单贝叶斯的托攻击判别模型[37]。和社交网络相关的其他应用还包括：搜索、情感分类、信息抽取等。举例来说，厦门大学的洪文兴等人通过对信息抽取、用户行为分析建立了厦门市的人才招聘实用系统[10]。 上海交通大学朱燕民等人和惠普实验室合作，基于对长期 GPS 数据的分析建模，实现利用 GPS 数据更新地图数据的功能。他们对比了现有的不同方法，并在上海的 2000 多个出租车数据和芝加哥的多个公交车数据上进行了实验分析[19]。清华大学的唐杰等人研究了社交网络中基于用户层次的情感分类模型，并在 Twitter 的数据上进行了验证[31]。

2.3 大数据挖掘

大数据又称海量数据，指的是所涉及的数据规模巨大，以至于目前已有的软件

工具无法在合理时间内，处理、管理、挖掘这些数据，并将其整理成为帮助企业经营决策更积极目的的信息。[1] 在维克托·迈尔-舍恩伯格及肯尼斯·库克耶编写的《大数据时代》中大数据特指不用随机分析法（抽样调查）这样的捷径，而采用所有数据的方法[24]。大数据的 4V 特点：Volume、Velocity、Variety、Veracity。大数据领域的研究进展主要包括可扩展性、并行性、分布式算法等方面。

中国人民大学的李翠平等人将 GPU 用于 SimRank 算法的加速。SimRank 是用网络结构信息计算节点相似度的经典算法，但算法复杂度较高。通过 GPU 的并行加速，SimRank 可以得到 20 倍的加速比[9]。上海交通大学的余勇团队利用大规模的广告投放数据研究广告投放价格和公司广告预算之间的策略优化问题。他们将该问题模型化为一个带约束的优化问题，通过求解该优化问题可以得到最优的广告投放策略[45]。清华大学的朱军等人提出可扩展的最大化边界的话题模型[46]。该模型不仅可以学习话题的概率分布，还可以同时学习一个预测模型，并通过将两个学习任务结合来提高模型精度。崔鹏等人则针对大范围的相似性搜索提出了基于关系的异构哈希框架（Relation-aware Heterogeneous Hashing），通过对不同数据类型构建 Hamming 空间，做到同时优化同构和异构映射关系。论文利用腾讯微博和 Flickr 的数据集证明了 RaHH 框架较目前的哈希方法的性能有显著提高[27]。北京大学的张岩等人利用机器学习的办法自动从大规模维基百科的数据中抽取概念，用于增强 WordNet 的能力，研发出 WorkiNet。该工具可以及时发现新出现的词，因此取得更高的覆盖率，可以有效的帮助多个文本挖掘任务[15]。

3. 国内外比较与发展趋势

3.1 国际数据挖掘研究趋势

数据挖掘最早的研究可以追溯到数十年前，从 20 世纪 80 年代就有关于知识发现及数据挖掘的研讨会了。早期的数据挖掘研究主要集中在北美，也正是由于这个原因，国际数据挖掘大会 KDD 在 2009 年之前从来没有在美国之外的地方召开。2009 年 KDD 第一次离开北美在巴黎举行。2012 年 KDD 第二次离开美国，来到了北京。这一方面反映了欧洲和中国数据挖掘研究的快速崛起，另一方面也反映数据挖掘更加国际化。今年（2013 年）在芝加哥举行的 KDD 大会吸引了来自全球 50 多个国家

[1] http://en.wikipedia.org/wiki/Big_data

1200 多人参加，打破历届 KDD 大会的参加人数纪录，从而 KDD 成长为 ACM SIG 系列会议中规模最大的会议之一，仅次于 SIGGRAPH。同年的 ACM 数据挖掘专委会换届选举中，华裔学者刘兵教授当选主席。从近年国际数据挖掘的研究趋势来看，社交网络分析和大数据挖掘无疑也是最热也是最重要的两个子领域。

在大数据方面，国际领先的研究机构更加注重底层的基础架构，例如，来自加州伯克利大学的 Canny 和 Zhao 等人提出结合 CPU、GPU 以及全新的算法设计方案，提出名为 BID 的一个大数据处理框架，该框架融合了硬件、软件以及用于支撑大规模数据挖掘的设计模式，将单台 PC 机处理数据的速度提高数十倍[2]。而从上层算法设计来看，国际学者纷纷提出不同的大数据处理方法用以解决传统分布式计算和并行计算中存在的问题。例如加州伯克利大学的 Michael I. Jordan 等人提出了“bag of little bootstraps”的方法以解决子采样的波动性问题[16]，并同时提出了基于 Stein 的方法来解决大规模的矩阵填充问题。康奈尔大学的 Karthik Raman 和 Adith Swaminathan 等人探讨了如何应对大数据上的复杂分析问题，论文将大数据上的复杂分析任务分解为一系列的简单任务。

在另外一个研究热点，即社交网络分析和信息网络分析方面，国际研究的趋势更偏重基础和理论。例如今年KDD的最佳学生论文则是研究如何利用有限内存空间近似估计网络流数据中的网络聚类系数[14]，该论文给出了详细的理论证明和分析。此外，今年的数据挖掘创新奖颁给了康奈尔大学的Jon Kleinberg教授。Jon Kleinberg是美国科学院院士，著名的HITS算法的发明人。他在名为《面向在线社交网络的计算模型》（Computational Perspectives on Social Phenomena in On-Line Networks）的获奖报告中，重点介绍了社交网络中出现的群体行为和用户的网络交互。他着重从社会平衡理论、社会地位理论、以及结构多样性等方面介绍了社交网络的一些本质现象，用他本人的话是“the web knows you better than you know yourself”（网络比您更懂您自己）。

此外，在信息传播和网络内容分析相结合方面，社会网络分析和异构信息网络分析分别占据了很重要的位置。斯坦福大学的Jure Leskovec等人在信息传播方面做了大量工作，近年来，几乎每年的国际数据挖掘大会中，Jure Leskovec等人的文章都在引用次数最高的文章范围之中。在异构信息网络挖掘方面，伊利诺伊大学香槟分校的韩家炜教授无疑是这方面的权威，他在异构网络中的对象排序、对象关系挖掘方面做了大量工作，引领了这方面的研究。密歇根大学的梅俏竹等人在文本挖掘与网

络挖掘的结合上做了很多前沿的工作。他们提出的DivRank算法能在信息网络中自动选取重要且多样化的节点，从而能够广泛应用于社交网络中节点排序和文本摘要等问题[25]。

3.2 国内研究特色与差距

国内数据挖掘的研究和国际上领先的研究机构（卡耐基梅隆大学、康奈尔大学、斯坦福大学、伊利诺伊香槟分校、密歇根大学、微软、Google 等）结合非常紧密。不少在 KDD 上发表的论文都是国内国外合作研究的成果。从总的趋势来看，国内和国外的研究方向仍有一些不同：在有些方面存在一定的差距，同时在另外一些方面也有鲜明的特色。

总的来说，国外的研究更偏重于数据挖掘的理论基础和交叉学科的研究。国内的研究更偏重于实际的应用，以最新的技术解决现实的问题。国内学者在解决问题方面的研究上处于国际前沿水平，发表了大量有学术影响的论文；但在定义全新的科学问题方面还缺乏开创性的成果。同时在基础研究方面，国内学者和国际领先的研究团队相比还有一定的差距。这个差距正在快速缩小。例如清华大学在今年的 KDD 会议就有 8 篇科研论文被录用，这在国际上所有的研究机构里也是十分突出的成绩。KDDCUP 是每年数据挖掘算法竞赛的舞台，近年已经多次看到来自国内的团队（如：上海交通大学、中科院等）登上领奖台。

4. 需求与展望

随着数据挖掘领域的研究不断深入及其愈发广泛的应用，数据挖掘关注的焦点也有了新的变化。总的趋势是，数据挖掘研究和应用更加“社会化”和“大数据化”。数据挖掘的理论和应用在相当一段时间继续保持稳定发展，但有着朝向大规模的社交数据分析和时间序列数据分析方向发展的趋势。在用户层面，移动计算设备的普及与大数据革命带来的机遇使得搜索引擎对用户所处的上下文环境具有了前所未有的深刻认识，但对于如何将认识上的深入转化为用户信息获取过程的便利仍然缺乏成功经验。近年来，以用户个性化、用户交互等为代表的研究论文的数量大幅增加，是这方面的趋势体现。除此之外，社交网络服务的兴起对互联网数据环境和用户群体均将形成关键性的影响，如何更好的面对相对封闭的社交网络数据环境和被社交

关系组织起来的用户群体，也是数据挖掘面临的机遇与挑战。

具体来说，通过对数据挖掘国际会议 KDD 上近年发表的论文分析来看，一方面，数据挖掘领域中偏重理论的研究有所增加，尤其体现在与其他学科的基础研究有了更多密切的合作，譬如通过统计学的方法来区分数据现象下隐藏的因果关系（causality）和关联关系（correlation），通过设计心理学的实验去剖析用户的动机以进行更好的建模，以及通过引入社会学的基本理论来解释从数据中观察到的现象等。KDD 2013 的最佳研究论文和最佳学生论文都是着重关注网络数据挖掘的理论分析。可以预见，会有更多的研究深入探讨数据挖掘的本质问题，不断完善其理论基础。另一方面，数据挖掘的应用中关注“大数据”的趋势越来越明显。在大规模数据下，如何保证现有数据挖掘算法的时间和空间复杂度的应用成为研究热点。人们对数据挖掘技术，除了要求有好的效果之外，往往还要求其能够被高效地实现，以适用于更大规模、更高速演化的网络数据。此外，另一个研究热点将是网络安全、隐私以及软件可靠性和可扩展性。

5. 结束语

本报告围绕数据挖掘的基础理论、社交网络分析和大数据挖掘三个主要方面介绍近三年来国内在数据挖掘领域的主要研究进展，分析国际学科发展趋势，国内的研究特色，以及国内外的差距。总的来说，国内学者在分类模型、无监督学习以及社交网络分析方面的研究上取得了一系列国际水平的成果，但同时也欠缺一些重要的、基础性的研究问题（例如：信息传播模型的理论分析、大数据基础架构等），因而有重大影响的成果还不多，在这方面国内同仁还需要进一步加强研究工作。

此外，从学科发展的角度来说，数据挖掘正处于最佳发展时期。人类社会在经历一场数据革命：放眼四方，皆可见诸多规模庞大的数据集，并且这些数据集还在以惊人速度不断增长。如何更好地利用这些大数据显得愈发重要，对企业、组织如是，对科学、工程、医药乃至社会亦如是。数据挖掘幸运地成为了这场数据革命的中心，也站在了大数据信息时代的潮头浪尖，承载着在未来的科研与应用中发展大数据分析、知识挖掘和数据科学的重任。对数据挖掘领域来说，这是一个机遇与挑战并存的时期。

参考文献

[1] D. Cai, C. Zhang and X. He. Unsupervised Feature Selection for Multi-Cluster Data. *Proc. KDD 2010.* pp. 333-342.

[2] J. Canny and H. Zhao. Big Data Analytics with Small Footprint: Squaring the Cloud. *Proc. KDD 2013.* pp. 95-103.

[3] P. Cui, S. Jin, L. Yu, F. Wang, W. Zhu and S. Yang. Cascading Outbreak Prediction in Networks: A Data-Driven Approach. *Proc. KDD 2013.* pp. 901-909.

[4] W. Feng and J. Wang. Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems. *Proc. KDD 2012.* pp. 1276-1284.

[5] C. Gao and J. Wang. Direct Mining of Discriminative Patterns for Classifying Uncertain Data. *Proc. KDD 2010.* pp.861-870.

[6] P. Gong, J. Ye and C. Zhang. Robust Multi-task Feature Learning. *Proc. KDD 2012.* pp. 895-903.

[7] Y. Gu, C. Gao, G. Cong and G. Yu. Effective and Efficient Clustering Methods for Correlated Probabilistic Graphs. *IEEE Trans. Knowledge and Data Engineering,* 25(7): 2013.

[8] L. Han, G. Song, G. Cong and K. Xie. Overlapping Decomposition for Causal Graphical Modeling. *Proc. KDD 2012.* pp. 114-122.

[9] G. He, H. Feng, C. Li and H. Chen. Parallel SimRank Computation on Large Graphs with Iterative Aggregation. *Proc. KDD 2010.* pp. 543-552.

[10] W. Hong, L. Li, T. Li and W. Pan. iHR: An Online Recruiting System for Xiamen Talent Service Center. *Proc. KDD 2013.* pp. 1177-1185.

[11] Y. Hu, D. Zhang, J. Liu, J. Ye and X. He. Accelerated Singular Value Thresholding for Matrix Completion. *Proc. KDD 2012.* pp. 298-306.

[12] J. Huang, H. Sun, Q. Song, H. Deng and J. Han. Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network. *IEEE Trans. Knowledge and Data Engineering,* 25(8): 1876-1889, 2012.

[13] S. Huang, Y. Yu, Z. Zhou. Multi-Label Hypothesis Reuse. *Proc. KDD 2012.* pp. 525-533.

[14] M. Jha, C. Seshadhri and A. Pinar. A Space Efficient Streaming Algorithm for Triangle Counting Using the Birthday Paradox. *Proc. KDD 2013.* pp.589-597.

[15] S. Jiang, L. Bing, B. Sun, Y. Zhang and W. Lam. Ontology Enhancement and

Concept Granularity Learning: Keeping Yourself Current and Adaptive. *Proc. KDD 2011*.. pp. 1244-1252.

[16] A. Kleiner, A. Talwalkar, S. Agarwal, I. Stoica and M. Jordan. A General Bootstrap Performance Diagnostic. *Proc. KDD 2013*. pp. 419-427.

[17] L. Li, X. Jin, S. Pan, J. Sun. Multi-domain Active Learning for Text Classification. *Proc. KDD 2012*. pp. 1086-1094.

[18] G. Li, Y. Yang, T. Wang and J. Feng. Location-Aware Publish/Subscribe. *Proc. KDD 2013*. pp. 802-810.

[19] X. Liu, J. Biagioni, J. Eriksson, Y. Wang, G. Forman and Y. Zhu. Mining Large-Scale, Sparse GPS Traces for Map Inference: Comparison of Approaches. *Proc. KDD 2012*. pp. 669-677.

[20] X. Liu, Z. Nie, N. Yu and J. Wen. BioSnowball: Automated Population of Wikis. *Proc. KDD 2010*. pp. 969-978.

[21] X. Liu and Z. Zhou. Learning with Cost Intervals. *Proc. KDD 2010*. pp. 403-412.

[22] T. Lou and J. Tang. Mining Structural Hole Spanners Through Information Diffusion in Social Networks. *Proc. WWW 2013*. pp. 837-848.

[23] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. *ACM Trans. KDD*, 7(2): no. 5, 2013.

[24] V. Mayer-Schonberger and K. Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt. 1 edition

[25] Q. Mei, J. Guo, and D. Radev. DivRank: The Interplay of Prestige and Diversity in Information Networks. Proc. KDD 2010. pp. 1009-1018.

[26] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li and H. Wang. Entity-Centric Topic-Oriented Opinion Summarization in Twitter. *Proc. KDD 2012*. pp. 379-387.

[27] M. Ou, P. Cui, F. Wang, J. Wang, W. Zhu and S. Yang. Comparing Apples to Oranges: A Scalable Solution with Heterogeneous Hashing. *Proc. KDD 2013*. pp. 230-238.

[28] Z. Qi, M. Yang, Zhongfei Zhang and Zhengyou Zhang. Mining Partially Annotated Images. *Proc. KDD 2011*. pp. 1199-1207.

[29] F. Shang, L. Jiao and F. Wang. Semi-Supervised Learning with Mixed Knowledge Information. *Proc. KDD 2012*. pp. 732-740.

[30] W. Shen, J. Wang, P. Luo and M. Wang. Linking Named Entities in Tweets with Knowledge Base via User Interest Modeling. *Proc. KDD 2013*. pp. 68-76.

[31] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou and P. Li. User-Level Sentiment Analysis Incorporating Social Networks. *Proc. KDD 2011*. pp. 1397-1405.

[32] C. Tan, J. Tang, J. Sun, Q. Lin and F. Wang. Social Action Tracking via Noise Tolerant Time-Varying Factor Graphs. *Proc. KDD 2010*. pp. 1049-1058.

[33] J. Tang, J. Sun, C. Wang and Z. Yang. Social Influence Analysis in Large-scale Networks. *Proc. KDD 2009*. pp. 807-816.

[34] J. Tang, S. Wu and J. Sun. Confluence: Conformity Influence in Large Social Networks. *Proc. KDD 2013*. pp. 347-355.

[35] J. Tang, S. Wu, J. Sun and H. Su. Cross-domain Collaboration Recommendation. *Proc. KDD 2012*. pp. 1285-1293.

[36] Y. Wang, G. Cong, G. Song and K. Xie. Community-Based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks. *Proc. KDD 2010.*. pp. 1039-1048.

[37] Z. Wu, J. Wu, J. Cao and D. Tao. HySAD: A Semi-Supervised Hybrid Shilling Attack Detector for Trustworthy Product Recommendation. *Proc. KDD 2012*. pp. 985-994.

[38] L. Xiang, Q. Yuan, S. Zhao, L. Chen, X. Zhang, Q. Yang and J. Sun. Temporal Recommendation on Graphs Via Long- and Short-Term Preference Fusion. *Proc. KDD 2010*. pp. 723-732.

[39] H. Yin, Y. Sun, B. Cui, Z. Hu and L. Chen. LCARS: A Location-Content-Aware Recommender System. *Proc. KDD 2013*. pp. 221-230.

[40] J. Zhang, X. Fan, J. Wang and L. Zhou. Keyword-Propagation-Based Information Enriching and Noise Removal for Web News Videos. *Proc. KDD 2012*. pp. 561-569.

[41] P. Zhang, J. Li, P. Wang, B. Gao, X.Zhu and L. Guo. Enabling Fast Prediction for Ensemble Models on Data Streams. *Proc. KDD 2011*. pp. 177-185.

[42] J. Zhang, Y. Song, C. Zhang and S. Liu. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-Varying Corpora. *Proc. KDD 2010*. pp. 1079-1088.

[43] W. Zhang, J. Wang and W. Feng. Combining Latent Factor Model with Location Features for Event-Based Group Recommendation. *Proc. KDD 2013*. pp. 910-918.

[44] M. Zhang and K. Zhang. Multi-Label Learning by Exploiting Label Dependency. *Proc. KDD 2010*. pp. 999-1008.

[45] W. Zhang, Y. Zhang, B. Gao, Y. Yu, X. Yuan and T. Liu. Joint Optimization of Bid

and Budget Allocation in Sponsored Search. *Proc. KDD 2012.* pp. 1177-1185.
[46] J. Zhu, X. Zheng, L. Zhou and B. Zhang. Scalable Inference in Max-Margin Topic Models. *Proc. KDD 2013.* pp 964-972.
[47] Z. Zou, H. Gao and J. Li. Discovering Frequent Subgraphs Over Uncertain Graph Databases Under Probabilistic Semantics. *Proc. KDD 2010.* pp. 633-643.

撰稿人

唐杰，2006 年在清华大学计算机系获得博士学位，曾在康纳尔大学、伊利诺伊香槟分校、香港科技大学进行学术访问。现任清华大学计算机系副教授、博士生导师。研究兴趣包括：社会网络分析和数据挖掘。荣获首届国家自然科学基金优秀青年基金，获 2012 中国计算机学会青年科学家奖、2011 年北京市科技新星。研发了研究者社会网络 ArnetMiner 系统，吸引全球 220 个国家和地区 432 万独立 IP 的访问。

梅俏竹，2009 年在伊利诺伊大学(UIUC)获得博士学位，现任美国密歇根大学信息学院和计算机系助理教授。研究兴趣包括：数据挖掘，信息检索，社会媒体，医学信息等，在相关领域发表论文数十篇。曾获 2006 与 2007 年 SIGKDD 最佳学生论文奖、2010 年 SIGKDD 最佳博士论文奖(certificate of recognition)，并于 2011 年获美国国家自然科学基金会的杰出学术发展奖 (NSF CAREER Award)。

其他作者及致谢：感谢清华大学学生王凝枰和胡晓帮助整理资料。