人工智能之数据挖掘

Research Report of Data Mining

2020 年第 9 期

清华大学人工智能研究院

北京智源人工智能研究院

清华—中国工程院知识智能联合研究中心

2020 年 12 月

摘要

数据挖掘（Data Mining）旨在从大规模、不完全、有噪声、模糊随机的数据集中自动抽取隐含的、以前未知的、具有潜在应用价值的模式或规则等有用知识的复杂过程，是一类深层次的数据分析方法，也是知识发现的关键步骤。

本报告围绕数据挖掘的概念内涵、关键技术、人才研究、应用场景、发展趋势等方面展开深入研究，主要内容包括：

一、数据挖掘基本概念、发展历程、研究进展、问题与挑战。详细介绍了数据挖掘的基本概念、发展历程、技术研究关键词图谱、研究进展，以及研究过程中面临的问题与挑战。

二、数据挖掘技术研究现状分析。分别对数据挖掘十大经典算法、统计数据分析方法、科技情报挖掘技术、社交网络与图数据挖掘技术、自然语言数据挖掘技术、多媒体数据挖掘技术、大规模数据挖掘技术、数据隐私保护和安全等方面进行了详细介绍和深入分析，并解读了 SIGKDD 会议收录的代表性论文。

三、数据挖掘领域人才现状分析。基于 AMiner 平台提供的论文和学者大数据，从学者分布、学术水平、国际合作、学者流动等维度，对国内外相关研究学者和机构进行了对比分析，总结中国科研学者队伍建设过程中的弱势环节和问题，并提出对策建议。

四、数据挖掘典型应用场景分析。分别介绍了数据挖掘技术在零售业、旅游业、物流业、医学界、金融业、电信业等不同行业的应用场景，并如何助力这些行业的发展。

最后分析了数据挖掘相关技术研究发展趋势和创新热点，以及中国的专利数据和国家自然科学基金支持情况，并展望了数据挖掘未来发展趋势。

目录

图表目录

1　概述篇

1 概述篇

1.1 数据挖掘基本概念

数据挖掘（Data Mining）的广义观点：从数据库中抽取隐含的、以前未知的、具有潜在应用价值的模式或规则等有用知识的复杂过程，是一类深层次的数据分析方法。数据挖掘旨在从数据中挖掘知识，是一种跨学科的计算机科学分支，使用人工智能、机器学习、统计学和数据库等交叉学科领域方法在大规模、不完全、有噪声、模糊随机的数据集中自动搜索隐藏于其中的有着特殊关系性的数据和信息，并将其转化为计算机可处理的结构化表示，是知识发现的一个关键步骤（如图 1 所示）[1]。

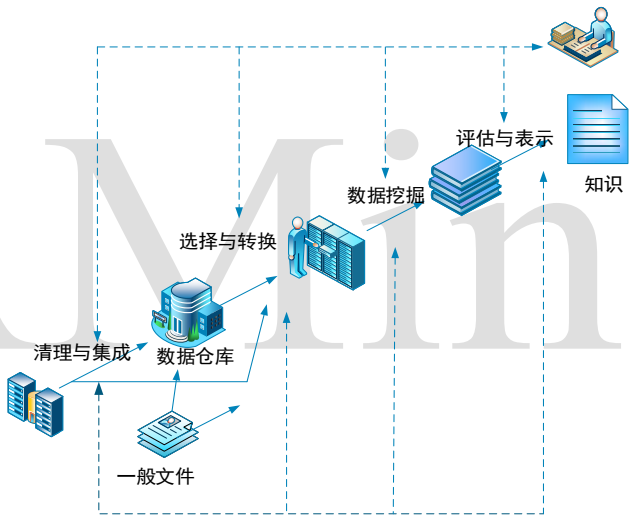


图 1 数据挖掘是知识发现的核心过程

知识发现是从各种媒体表示信息中，根据不同的需求获得知识的过程，向使用者屏蔽原始数据的繁琐细节，直接将发现的知识向使用者报告。知识发现是指从数据中发现有用知识的整个过程，而数据挖掘是指知识发现过程中的特定步骤。数据挖掘是特定算法的应用，用于从数据中提取模式。图 1 展示了知识发现的主要步骤。

（1） 数据清理：消除噪声和删除不一致数据。

（2） 数据集成：将多种数据源组合在一起。

（3） 数据选择：从数据库中提取与分析任务相关的数据。

（4） 数据变换：通过汇总或聚集操作把数据变换、统一成适合挖掘的形式。

（5） 数据挖掘：使用智能方法提取数据模式。

（6） 模式评估：根据某种度量，识别代表知识的模式。

（7） 知识表示：使用可视化与知识表示技术，向用户提供挖掘的知识。

数据挖掘的对象可以是任何类型的数据源，包括数据库数据、数据仓库、事物数据，以及文本、多媒体数据、空间数据、时序数据、web 数据、数据流、图或网络数据等。其中，数据库数据是一种结构化数据，比如关系数据库、图数据库中的数据。数据仓库（Data Warehouse）是一个从多个数据源收集的信息存储库，存放在一致的模式下，并且通常驻留在单个站点上，是决策支持系统和联机分析应用数据源的结构化数据环境。

为了满足用户从多角度多层次进行数据查询和分析的需要，数据仓库通常使用数据立方体（Data Cube）的多维数据结构建模。数据立方体模型中，每个维度对应模式中的一个或一组属性，比如城市（江苏、上海、浙江）、商品（电子产品、日用品、书籍）、季度（一季度、二季度、三季度）、月份（4 月、5 月、6 月）等；每个单元存放某种聚集度量值（比如计数、加和等）。图 2 显示了销售数据的数据立方体模型，通过钻取、上卷、切片、切块、旋转等联机分析处理（Online Analytical Processing, OLAP）操作，允许用户在不同汇总级别观察数据。

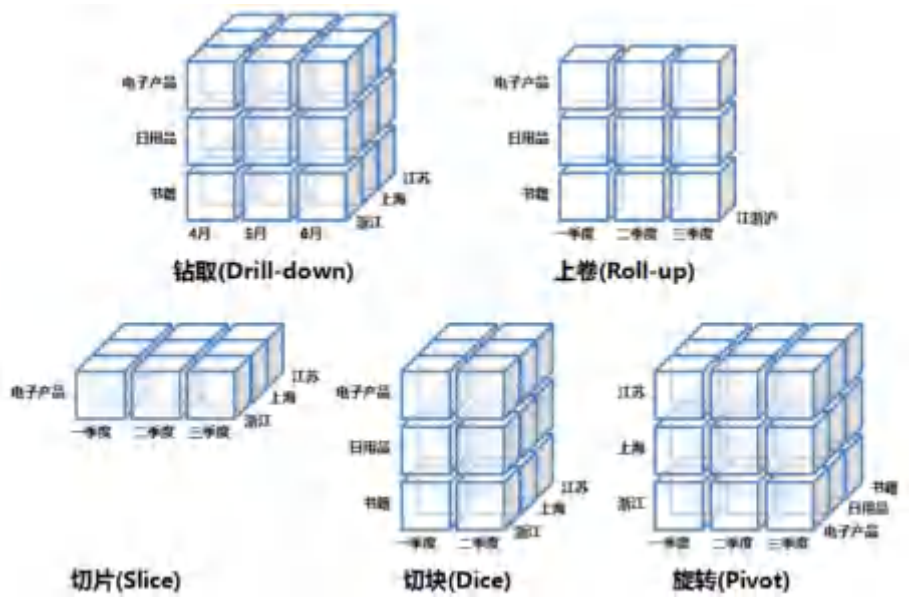


图 2 数据立方体模型示例

事物数据通常表示人类一次活动记录，比如一次购物、一个航班订票等。一个事物数据包含一个唯一的事物标识号，以及一个组成事物的项列表（购买的商品）。从图 2 可以看出，事物数据包含嵌套关系，难以放在关系数据库中，通常存放在表 1 所示的表格式的文件中。

表 1 事物数据库的片段[1]

trans_ID	商品 ID 的列表
T100	I1, I3, I8, I16
T200	I2, I8
…	…

近年随着互联网、物联网、通信网络以及社交网络快速发展，人类进入了大数据时代。为了充分挖掘和利用网络上海量的数据，大数据逐渐发展成为一个交叉研究学科，与数据挖掘紧密相连。一方面大数据包含数据挖掘的各个阶段，即数据收集、预处理、特征选择、模式挖掘、知识表示等；另一方面大数据的基础架构又为数据挖掘提供数据处理的硬件设施；最后大数据的迅速发展也使得数据挖掘对象变得更为复杂，不仅包括人类社会与物理世界的复杂联系，还愈加明显地呈现出高度动态化。要从大量无序数据中获取真正价值，数据挖

掘算法必须满足对真实数据和实时数据的处理能力，这使得很多传统算法不再适用。

大数据具有 4V 特性，对 4V 特性的解释有多种，其中美国国家标准技术研究院（National Institute of Standards and Technology）的解释，即规模庞大（Volume）、种类繁多（Variety）、增长速度快（Velocity）和变化多样（Variability）。IBM 给出了类似的解释，但 Variability 变成了真实性（Veracity），后来将数据价值（Value）引入了进来，成为了大数据的 5V 特性。目前比较公认的定义是麦肯锡全球研究机构（McKinsey Global Institute）也给出的，综合了“现有技术无法处理”和“数据特征定义”，即规模庞大（Volume）、种类繁多（Variety）、数据时效高（Velocity）和价值密度低（Value）。

大数据的快速发展极大地促进了数据挖掘领域的发展。数据挖掘涉及的常见的任务有：

（1） 数据表征

是对目标类数据的一般特征或指定特征的总结。对应于用户指定类的数据通常通过数据库查询收集。例如，要研究上一年销售额增长 10%的软件产品的特征，可以通过执行 SQL 查询来收集与此类产品相关的数据。

（2） 异常检测

数据库可能包含不符合数据一般行为或模型的数据对象，这些数据对象即为异常值。大多数数据挖掘方法将异常值视为噪声或异常但是在诸如欺诈检测等应用中，罕见事件可能比常见的更有价值。异常值数据的分析通常被称为异常值挖掘。

（3） 关联规则学习

搜索变量之间的关系。例如，一个超市可能会收集顾客购买习惯的资料。运用关联规则学习，超市可以确定哪些产品经常一起买，并利用这些信息促进销售，这种学习也被称为市场购物篮分析。

（4） 聚类

发现数据的类别与结构。聚类算法基于最大化类内相似性和最小化类间相似性的原则，将对象进行聚类或分组。也就是说，形成对象集群，使得集群内的对象彼此之间具有较高的相似性，但与其他集群中的对象非常不相似。每个集群都可以被视为一类对象，从中可以派生出规则。

（5） 分类

分类是查找描述和区分数据类别或概念的模型（或函数）的过程，目的是为了能够使用模型来预测未知对象的类别。例如，一个电子邮件程序可能试图将一个电子邮件分类为“合法的”或“垃圾邮件”。

（6） 回归

试图找到能够以最小误差对该数据建模的函数。回归分析是最常用于数字预测的统计方法，还可以根据现有数据预测趋势。

（7） 演化分析

描述并建模对象行为随时间变化的规则或趋势。这种分析具有时间序列数据分析、序列或周期性模式匹配以及基于相似性的数据分析的特征。

1.2 数据挖掘发展历程

随着数据体量的快速增加，人们希望有一种方法可以帮助处理这些纷繁复杂的数据，从中发现有价值的信息或知识为决策服务，数据挖掘在此背景下应运而生，下文将介绍数据挖掘的起源和发展历程[2]。

20 世纪 60 年代，资料搜集阶段。在这个阶段受到资料存储能力的限制，特别是当时还处在磁盘存储的阶段，因此主要解决的是数据搜集的问题，而且更多是针对静态数据的搜集与展现，所解决的商业问题也是基于历史结果的统计资料。

20 世纪 70 年代，资料存储阶段。随着数据库管理系统趋于成熟，存储和查询百万兆字节甚至千万亿位成为可能。而且，数据仓库允许用户从面向事物处

理的思维方式向更注重数据分析的方式进行转变。然而，从这些多维模型的数据仓库中提取复杂深度信息的能力是非常有限的。

20 世纪 80 年代，资料分析访问阶段。关系型数据库与结构性查询语言的出现，使得动态的数据查询与展现成为可能，人们可以用资料来解决一些更为聚焦的商业问题。在这个阶段，数据挖掘走进了历史舞台。GregoryI.Piatetsky-Shapiro（也是 KDnuggets 的创始人）等人于 1989 年 8 月在美国底特律的国际人工智能联合会议（IJCAI）上召开了一个专题讨论会（workshop），首次提出了知识发现（Knowledge Discovery in Database，KDD）这一概念。KDD 涉及数据库、机器学习、统计学、模式识别、数据可视化、高性能计算、知识获取、神经网络、信息检索等众多学科和技术，再后来的 30 年间 KDD 逐渐形成了一个独立、蓬勃发展的交叉研究领域。也正是在这个时期，出现了些成熟的算法，能够“学习”数据间关系，相关领域的专家能够从中推测出各种数据关系的实际意义。

20 世纪 90 年代，数据仓库决策与支持阶段。OLAP 与数据仓库技术的突飞猛进使得多层次的数据回溯与动态处理成为现实，人们可以用数据来获取知识，对经营进行决策，零售公司和金融团体使用数据挖掘分析数据和观察趋势以扩大客源，预测利率的波动，股票价格以及顾客需求。1995 年，在加拿大蒙特利尔正式召开了第一届“知识发现和数据挖掘”国际学术会议 KDD。1995 年在美国计算机 ACM 年会上，开始把数据挖掘视为知识发现的一个基本步骤。随后成立了 ACM 专委会 SIGKDD 以及对应的国际数据挖掘与知识发现大会（ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 简称 SIGKDD），到目前为止 SIGKDD 已是国际数据挖掘领域的顶级会议。

21 世纪至今，真正的数据挖掘的时代。2012 年，美国奥巴马政府在白宫网站上发布了《大数据研究和发展倡议》，旨在提升利用大量复杂数据获取知识和洞见的能力，六大联邦政府机构达成一致，宣布将为此投入 2 亿美元以上经费，支持大力发展对数字化数据的接入、组织和挖掘的工具和技术，并进一步扩展，形成了包括联邦政府 12 个部门和机构的多项研究计划。这一倡议掀起了全球范围内政府推动大数据分析和研究的热潮，大大提升了数据挖掘的研究热

度。大数据时代，信息化的发展非常快，数据也在不断更新，相应的资料研究也越来越难。我们需要对大规模资料进行处理，从中提取出有价值的信息。随着信息技术的发展，数据挖掘已经越来越成熟，成为一门交叉学科。一般来说，数据挖掘结合了数据库、人工智能、模式识别、神经网络、机器学习、统计、高性能计算、数据可视化、空间数据分析和信息检索等很多方面的知识。

1.3 数据挖掘知识图谱

本节分析了近年来数据挖掘领域的高水平学术论文，挖掘出了包括社交网络、大数据、情报分析、聚类分析、文本挖掘、用户行为、推荐系统、离群检测、专家系统等近年来全球活跃的学术研究相关关键词。此外，结合知识图谱技术，本报告将以上研究领域关键词表示为三级图谱结构，具体分析和处理的方法如下：

（1） 使用自然语言处理技术，提取每篇论文文献的关键词，据此，结合学科领域知识图谱，将文章分配到相应领域；

（2） 依据学科领域对论文文献进行聚类，并统计论文数量作为领域的研究热度；

（3） 领域专家按照领域层级对学科领域划分等级，设计了三级图谱结构，最后根据概念热度定义当前研究热点。

图 3 是数据挖掘二级知识图谱的可视化表示，三级详细数据可以参见本报告附录，或到 https://www.aminer.cn/data 中直接下载原始数据。

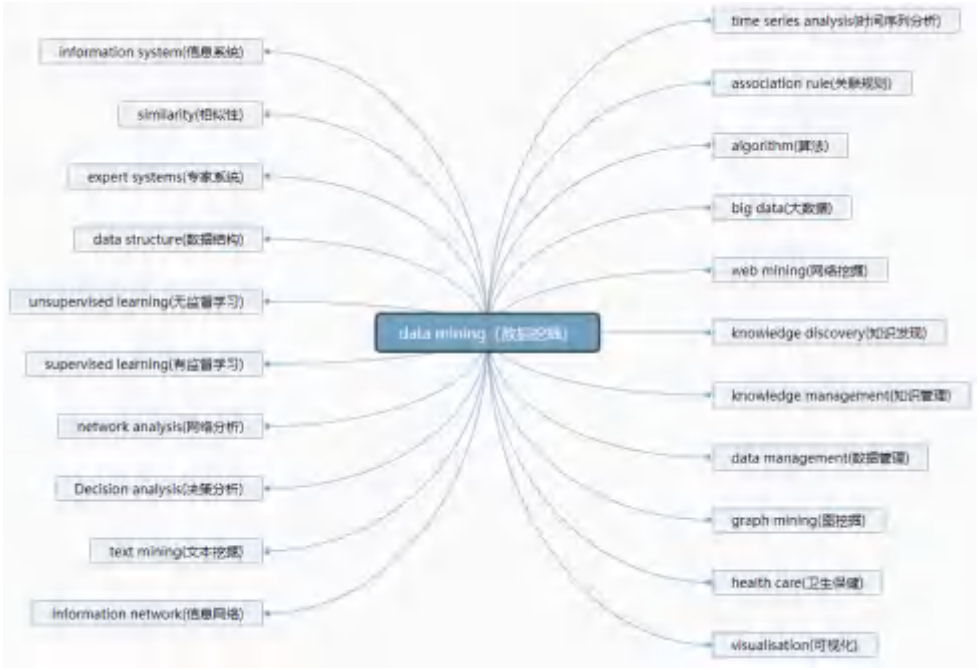


图 3 Data Mining 知识图谱

1.4 数据挖掘研究进展

随着大数据时代的来临，各大互联网企业每天都在产生数以亿计的数据。海量数据隐含着非常有价值的信息，传统数据分析的方法已经不再适用。同时，大量未经处理的数据可能会被人们所忽视。数据挖掘就是想自动地从大规模的数据中挖掘出有意义的知识或者模式。这里，我们将数据挖掘领域近期的主要发展归为两大类：复杂数据挖掘与分布式数据挖掘。

复杂数据包括序列数据、图数据等。在序列数据挖掘中，基于注意力（Attention）机制的 Transformer 模型表现出了巨大的潜力，在机器翻译等任务上取得了非常好的效果。随后，BERT 模型使用双向 Transformer 通过预训练方式在各种自然语言处理的任务上都达到了当时最好的效果。在图数据挖掘研究中，网络表示学习仍然是近年来非常热门的话题。从 DeepWalk 算法开始，基于随机游走的算法在无监督的表示学习任务中表现良好。NetMF 算法将几种基于随机游走的算法统一写成了矩阵分解的形式，给网络表示学习算法提供了理论基础。图卷积神经网络是另一种处理图数据的有效方法，借鉴了图谱论中的图卷积并使用图的拉普拉斯矩阵，在半监督的节点分类任务和图分类任务中都表现出很好的效果。除此之外，异构网络的表示与挖掘也逐渐被大家所关注。

分布式数据挖掘已成为数据挖掘领域非常有前途的方向。随着数据挖掘计算成本的增加和数据隐私保护的问题，分布式数据挖掘开始备受关注。分布式数据挖掘遵循“全局分布、局部集中”的挖掘原则，利用分布式计算方式对分布式资源进行挖掘，通过整合局部知识来获得全局知识，以此降低计算成本并增强数据保密性。由于分布式数据挖掘采用了不同的计算方式，传统的数据挖掘技术很难直接应用于分布式数据挖掘。目前，数据安全与数据隐私开始被大家所关注。2018 年 5 月，通用数据保护条例（GDPR）在欧盟正式生效，这也使得基于隐私保护的分布式数据挖掘方法逐渐被研究者所重视。

数据挖掘技术已经被广泛地应用于各类实际问题，包括金融数据分析、推荐系统等。数据挖掘相关研究需要结合实际问题，注重与机器学习、统计学科等的交叉，从大数据中挖掘出有价值的信息。

1.5 数据挖掘问题与挑战

2005 年，在数据挖掘领域的顶级国际会议 ICDM（International Conference on Data Mining）上，与会专家针对数据挖掘的研究现状整理了十大问题与挑战（如表 2 所示），具有很高的前瞻性和指导意义。本章节通过调研相关资料，对这十大问题的研究现状进行深入分析和总结。

表 2 数据挖掘领域十大问题与挑战

序号	问题与挑战
1	数据挖掘的统一理论框架的构建（Developing a Unifying Theory of Data Mining）
2	高维数据和高速数据流的挖掘（Scaling Up for High Dimensional Data/High Speed Streams）
3	序列和时序数据的挖掘（Mining Sequence Data and Time Series Data）
4	复杂数据中复杂知识的挖掘（Mining Complex Knowledge from Complex Data）
5	网络环境中的数据挖掘（Data Mining in a Network Setting）
6	分布式数据和多代理数据的挖掘（Distributed Data Mining and Mining Multi-agent Data）
7	生物和环境数据的挖掘（Data Mining for Biological and Environmental Problems）

8	数据挖掘过程中的相关问题处理（Data-Mining-Process Related Problems）
9	数据挖掘中数据安全、数据所涉及到的隐私和数据完整性的维护（Security, Privacy and Data Integrity）
10	非静态、非平衡及成本敏感数据的挖掘（Dealing with Non-static, Unbalanced and Cost-sensitive Data）

1.5.1 数据挖掘的统一理论框架的构建

“数据挖掘的统一理论框架的构建”问题的提出，主要是考虑到当时数据挖掘技术一般针对特定领域（比如医学、生物、金融等），或者解决特定问题，没有一个统一通用的理论框架来处理所有领域问题。随着大数据时代的到来，数据挖掘技术可以从海量数据中提取隐藏的、潜在的和有用的知识，并应用到各个行业领域，取得了良好的社会效益和经济效益。但是不同领域具有不同的数据特征（文本、图像、音视频等）和任务需求（分类、聚类、关联规则挖掘等），使用数据挖掘技术时，需要针对不同特点设计不同的挖掘算法，无形中增加了设计成本。因此，相关学者希望构建统一的数据挖掘理论框架，以此实现低成本处理不同行业领域的任务需求。

1.5.2 高维数据和高速数据流的挖掘

（1） 高维数据挖掘

高维数据挖掘和传统的数据挖掘最主要的区别在于数据对象的高维度，也是数据挖掘的难点。信息技术的进步降低了数据采集的难度，导致数据库规模变大、数据复杂性变高。比如金融交易数据、网络文档、用户评论、多媒体数据等的维度通常可以达到成百上千维，甚至更高。由于高维数据存在的普遍性，对高维数据挖掘的研究有着非常重要的意义。但“维灾”的影响也使得高维数据挖掘变得异常困难，必须采用一些特殊的手段进行处理。随着数据维数的升高，高维索引结构的性能迅速下降。在低维空间中，经常采用欧式距离作为数据之间的相似性度量，但相似性的概念在高维空间中通常不复存在，这给高维数据挖掘带来了严峻考验。一方面引起基于索引结构的数据挖掘算法的性能下降，另一方面很多基于全空间距离函数的挖掘方法也会失效。解决的方法包括：

- 将数据从高维降到低维，使用低维数据处理办法；

- 对算法效率下降问题可以通过设计更为有效的索引结构、采用增量算法及并行算法等来提高算法的性能；
- 对失效的问题通过重新定义使其获得合法性。

（2） 高速数据流挖掘

高速数据流挖掘指在高速到达的数据流上发现并提取隐含在其中的有用信息和知识的过程。数据流是一种动态的数据序列，具有实时到达、到达次序无法控制、规模宏大、连续不确定、随时间变化等特点。这些特点使得数据流中的元素难以存入内存或硬盘，导致无法检索其中的某一元素，要么处理所有元素，要么丢弃所有元素。

由于数据流自身的特性，数据流挖掘比传统的数据挖掘要复杂，已经成为一个新的研究方向，需要面临很多挑战。

- 数据流中元素不断产生和内存的存储空间有限存在一定矛盾，通常进行实时处理。因此数据流算法需要考虑如何充分利用内存空间，提高单次处理数据的规模。
- 数据流中元素是实时连续到达的，为了保证在内存中数据元素被覆盖前处理完，需要考虑算法效率来降低处理时间。
- 数据流中元素到达时间无法控制，在挖掘算法设计时需要考虑数据元素只能单遍扫描的特性。聚类、频繁项挖掘、分类等算法也常用到数据流挖掘任务。图 4 展示了一种数据流挖掘流程。针对输入的数据流中元素特征，采用聚类算法获取若干类别，进而挖掘数据隐藏的信息模式，如果模式合理则进行知识表示，从而实现网络数据流的动态挖掘。

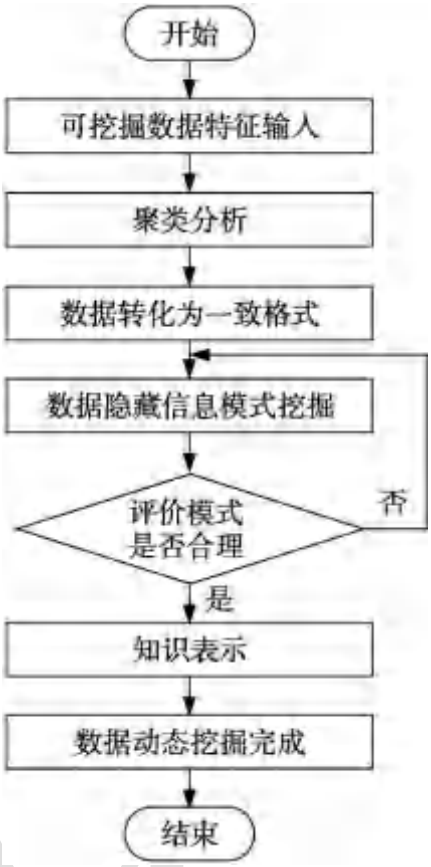


图 4 数据流挖掘流程图[1]

1.5.3 序列和时序数据的挖掘

序列是事件的有序列表。根据事件的特征，序列数据可以分成三类:时间序列数据、符号序列数据和生物学序列数据。时间序列数据由相等时间间隔记录的数值数据的长序列组成，如股票交易数据等。符号序列数据由事件或标称数据的长序列组成，通常不是相等的时间间隔观测，如顾客购买序列、Web 点击流等。生物学序列数据通常很长，具有间隙，携带重要的、复杂的、隐藏的语义，如 DNA 序列、蛋白质序列等。

由于序列数据没有明显特征，使得序列分类是一项具有挑战性的任务。针对序列数据的分类方法主要包括：

（1） 基于特征向量的分类，但高维特征向量也难以捕获序列特征性质；

（2） 基于序列距离的分类，使用距离函数度量序列之间的相似性；

（3） 基于统计模型的分类，比如马尔科夫链和隐马尔科夫模型。

时间序列数据是一种非离散化的数据，用于符号序列的特征选择技术不适用于该类数据，因为离散化可能导致信息损失。将 DTW（dynamic time warping）和 KNN（k nearest neighbors）结合提供了一种处理时间序列分类任务的解决思路。该方法的主要思想是：首先根据序列点之间的距离(欧氏距离)，获得一个序列距离矩阵；然后通过 KNN 进行聚类，以此得到不同类别的时间序列数据。

1.5.4 复杂数据中复杂知识的挖掘

随着信息技术的发展，互联网上出现大量结构各异的复杂数据类型，包括序列数据、图数据、多媒体数据等，如图 5 所示。如何从海量的复杂结构数据中发现隐含其中有价值的知识是数据挖掘领域的研究重点，但是由于数据的海量性、动态性、噪音、缺失和稀疏性等特点，给复杂数据挖掘技术研究带来很大挑战。序列数据挖掘在前述内容已有介绍，这里不再赘述。由于本报告其他章节有图与网络数据、多媒体数据、文本数据、Web 数据等挖掘内容的相关介绍，这里仅对空间数据的定义、相关技术等内容进行阐述。

空间数据挖掘是指对空间数据库中蕴涵的知识、空间关系或其他有意义的但非显式存在的模式等信息的提取。空间数据库中存储了大量与空间有关的数据，通常按复杂的多维空间索引结构组织数据。访问时，经常需要空间推理、地理计算和空间知识表示技术。常用方法包括：基于泛化的知识发现技术、空间聚类方法、空间关联规则分析、空间分类、空间趋势分析和基于证据理论的方法等[4]。

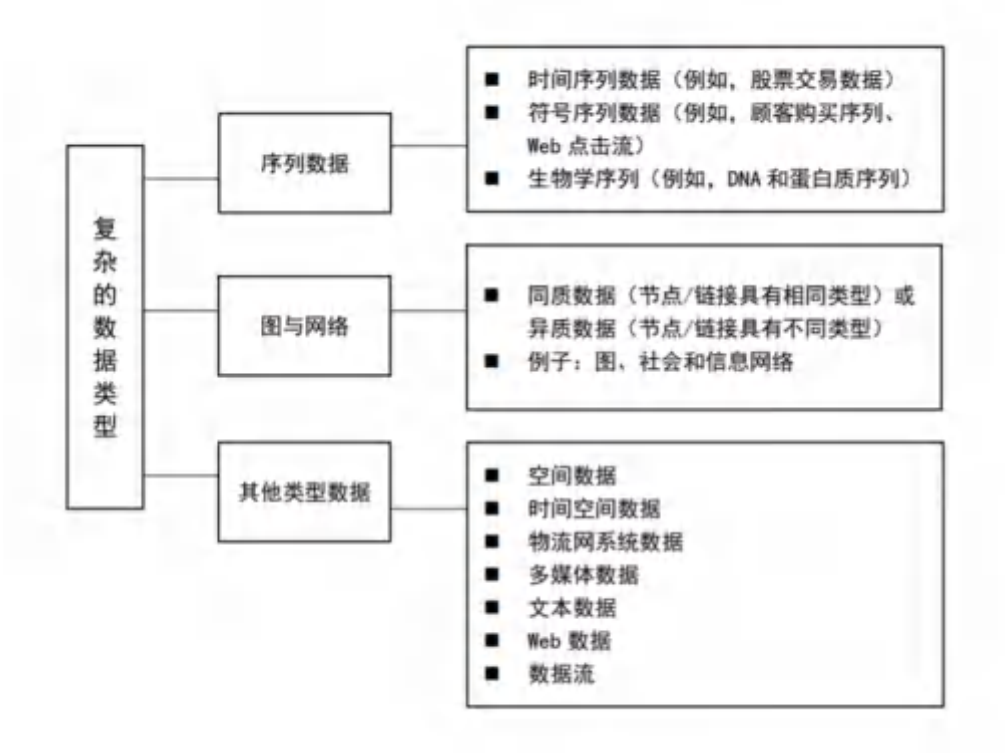


图 5 挖掘的复杂数据类型

1.5.5 网络环境中的数据挖掘

网络数据挖掘即从网络的内容、结构、日志中发现有用信息的过程，比如，Web 页面、网页之间的链接、用户浏览日志等（见表 3）。网络内容挖掘主要指针对 Web 页面数据（包括文本、图像、音频、视频等），挖掘隐含的、有价值的、可理解的知识和模式。由于 Web 页面信息内容丰富，通常具有时空相关性，使得挖掘对象比较复杂，在知识表示和解释方面比较困难，通常采用基于内容检索法进行数据查询和特征提取。网络结构挖掘主要针对网页内部的链接结构和网页之间的链接结构进行挖掘，通过算法发现它们之间的有用信息，比如通过衡量一个站点或网页的入链数量和出链数量，可以反映它们的角色，以及流行程度和重要性。结构挖掘通常需要分析 Web 网络全局数据，因此在个性化搜索引擎领域得到广泛引用。网络日志挖掘的对象主要是 Web 使用记录数据，包括服务器日志、浏览器日志、注册信息、用户查询等。由于这些使用记录数据由用户产生，而用户群体规模庞大，使得记录数据量非常巨大，而且数据类型也比较丰富。通过挖掘日志记录数据，一方面可以发现用户访问 Web 的模式；另一方面可以获取用户的喜好偏向、满意度等，对站点的推荐服务提供参考。

表 3 网络数据挖掘的分类[5]

内容挖掘	Web 页面内容挖掘	从文档内容或描述中抽取知识
	搜索结果挖掘	发现文档中隐含的知识，理解文档的内容和结构
结构挖掘	Web 结构分析	通过页面之间的连接，分析页面的权重，为调整网站结构提供信息
日志挖掘	一般访问模式跟踪	Web 日志挖掘；使用数据挖掘技术理解一般的访问模式和趋势，以明确更好的结构和资源分类
	个性化访问模式跟踪	用户自己定制信息，每次分析一个用户的访问模式，网站则通过学习用户的访问模式，自动进行重组

1.5.6 分布式数据和多代理数据的挖掘

传统的数据挖掘算法假设数据是集中收集的、驻留在内存中的、静态的。然而，由于资源共享的快速增长，在有限的资源下管理和处理大量的数据是具有挑战性的。例如，大量的数据被快速地产生并存储在多个位置。将它们集中在一个地方变得越来越昂贵。此外，传统的数据挖掘算法普遍存在内存限制、处理能力低、硬盘容量不足等问题和挑战。为了解决上述问题，分布式数据挖掘（distributed data mining, DDM）技术应运而生，在许多应用中逐渐成为一种有价值的替代方案[6]。

DDM 也被定义为分布式环境下的数据挖掘，主要利用分布式计算技术来挖掘分布式存储的数据资源，通过整合局部结果（模式）来获取最终结果（全局模式）。图 6 展示了一种分布式数据挖掘框架。与集中式挖掘相比，DDM 因为只需要传输局部模式而非局部资源数据，因此具有网络通信开销较少的优点。此外，DDM 在分布式环境下共享局部模式而非全部数据，可以更好地保护数据隐私性。

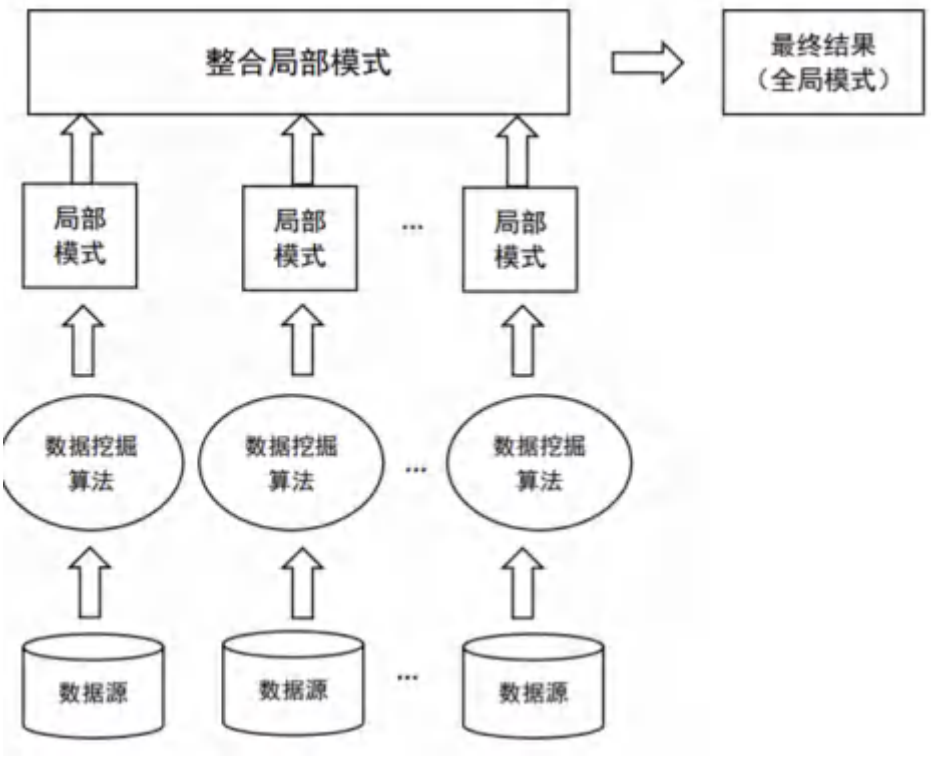


图 6 分布式数据挖掘框架[6]

为了解决 DDM 存在分布式数据环境下的挖掘限制和多算法协助等瓶颈问题，研究学者提出了基于多代理（Multi-Agent）的分布式数据挖掘技术。主要设计理念是利用 Agent 的自治性实现局部挖掘以保护数据私有性；利用 Agent 的主动性减少用户参与以提高挖掘自动化水平；利用 Agent 的协作性实现多算法协同挖掘等[8]。MATEO 等人提出一个智能分布式架构和基于 Agent 的数据挖掘模型来实现自适应机制，以便实施数据挖掘算法和多 Agent 间的高效交互。如图 7 所示，数据挖掘模型基于 Multi-Agent 系统实现，包括 3 项功能：聚类、分类和关联规则挖掘，来实施知识发现和系统需求的采集[9]。基于多代理的分布式数据挖掘技术中，多代理间的通信和协作是影响挖掘效率的重要因素，大部分研究主要针对协作机制的设计，在代理间的通信开销方面考虑较少。

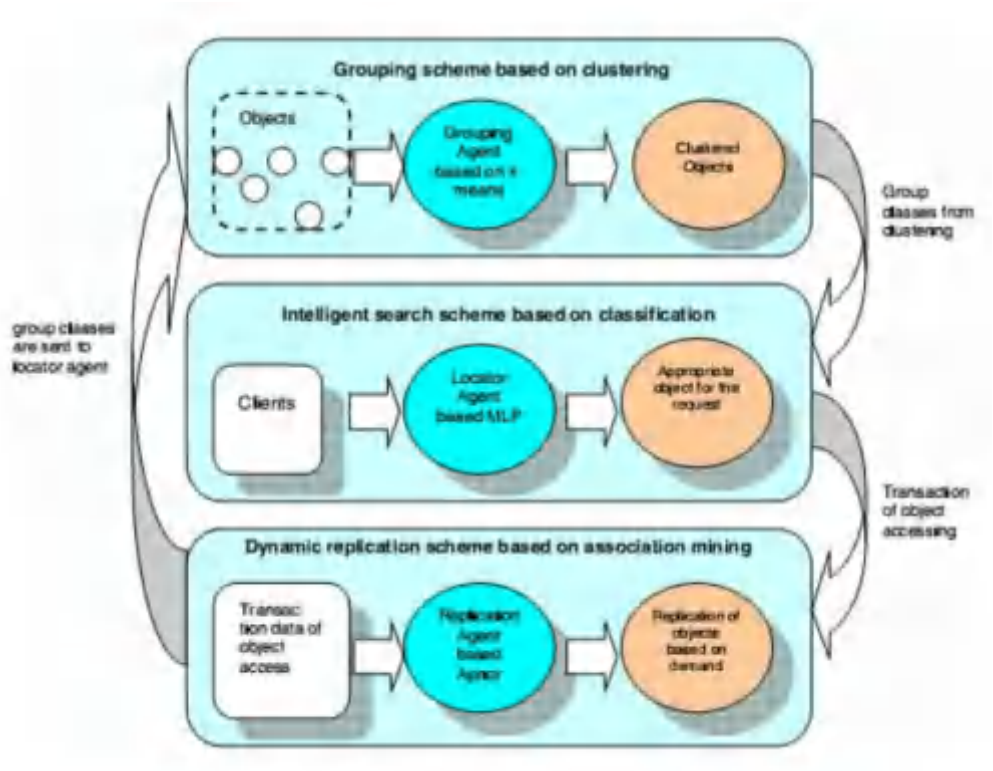


图 7 面向基于 Multi-Agent 间通信和协作的智能分布式框架的数据挖掘模型

1.5.7 生物和环境数据的挖掘

随着生物数据获取技术的发展，生物研究领域积累了大量的不同模态的数据，比如：遗传多组学数据，通过不同成像技术获取的影像数据等。多模态生物数据的积累使得我们从不同尺度全面地研究生物信息成为可能，但是由于生物数据自身的一些特点（比如：多模态，高维度，小样本等），直接应用传统的数据挖掘算法对其进行分析的效果并不令人满意。比如，随着特征维度的增加，数据挖掘模型具有泛化能力的难度会以指数级增加，进而造成数据挖掘模型过拟合的问题。开发有效的数据挖掘算法来帮助从海量的生物多模态数据中挖掘出有意义的信息，并将其转换为生物学知识是当前生物学研究领域的热点和难点问题[10]。

在互联网信息技术快速发展的今天，GIS、遥感、无线网络、web、物联网等技术在环境管理广泛应用，尤其采用物联网技术实现对环境质量、污染源、放射源的实时监管，导致环境数据出现爆发式增长，比如：环保标准和法规数据、环境质量数据、环境评价与环境预则数据、环境事件与事故处理数据等。

因此如何利用数据挖掘技术发现环境数据背后隐藏的价值，提高环境管理决策水平是亟待研究的课题。环境数据挖掘技术主要包括：发现驱动的数据挖掘、交互式数据挖掘、机器学习类数据挖掘、统计分析类数据挖掘等。然而随着信息技术的快速发展，环境数据挖掘技术的发展也变得更加快速，研究工作需要集中在如何研制新的算法或改善现有算法的运行效率和可伸缩性，以及如何提高挖掘过程的自动化以减少人工干预等方面[11]。

1.5.8 数据挖掘过程中的相关问题处理

大数据时代给数据挖掘技术研究带来很多机遇，但由于数据的规模大、来源种类多样、价值密度低、增长速度快、准确性低等特征（详见图 8），也带来很大挑战。为了保证数据挖掘算法捕获正确的数据，从不同的来源解析相同的数据，将数据从一种形式转换为另一种形式进行分析，对数据进行建模，对正确的结果进行可视化理解等，下面总结了数据挖掘过程中每个阶段需要面对的问题挑战[13]。

（1） 数据采集和入库：不同类型的数据存储在一个地方、相同数据在不同的数据源中命名形式不一样、不同数据源中的数据性质不同等会给数据采集和存储带来很大挑战。

（2） 数据清洗：定义和确定错误类型，搜索和识别错误，更正错误，记录错误以及修改数据输入程序以减少结构化、非结构化和半结构化数据会带来数据清理方面的挑战。当数据没有存在噪音、不完整、不一致等问题时，数据挖掘和分析过程会提供正确的信息。

（3） 数据分析与挖掘：数据分析与挖掘是大数据的核心挑战。如果在采集、存储、清理、集成、转换等流程中出现任何问题，会导致容易挖掘到无用的数据。如果想在大数据中获取有价值的信息，有必要研究适用于所有类型数据的挖掘技术。

（4） 数据集成与融合：在大数据挖掘中，不同类型的数据模式集成和融合是最大的挑战。例如，集成的数据模式通常是在不同数据源中获取，而不同数据源中的同一对象通常具有不同的名称表示形式。如何正确将指向同一对象

的不同数据融合到一起，是数据集成和融合技术的巨大挑战。

（5） 数据查询和索引：大数据时代的特征要求采用分布式存储和处理技术，并且实时处理和存储数据。为了减少在分布式环境中对数据的响应时间，需要优化数据查询和索引技术。

图 8 大数据特征[12]

1.5.9 数据挖掘中数据安全、数据所涉及到的隐私和数据完整性的维护

大数据在改变着人类的社会生产和生活方式的同时，也在深刻地改变着人们的思想和观念。大数据时代隐私安全问题的出现使得整个社会的价值观念受到前所未有的挑战，隐私安全问题不仅会造成财产损失，而且会严重危害人的身心健康，以此产生的安全焦虑会严重影响到人们的正常生活和工作，也不利于大数据产业的长期健康发展。在这样的背景下，如何加强对大数据的管理与引导，培养人们的正确的隐私观念需要引起我们的重视。

大数据时代隐私是在信息技术高度发展情况下出现的，是隐私在新阶段的新发展。从个人隐私安全层面来看，大数据将网络大众带入到开放透明的裸奔时代，主要体现在：

（1） 个人提供的数据化形式的个人基本信息，如身份证号码、手机号码、健康信息、社保卡号等；

（2） 个人财产与账户信息，如信用卡、网上银行、手机银行、网上交易账号和密码等；

（3） 互联网社交联系方式，如网络通讯录、电子邮箱地址等；

（4） 个人网络活动及踪迹，如位置信息、购物记录、社交照片等；

（5） 个人在公共场所活动而被数字化记录的影像信息，如视频监控、手机拍摄等[14]。

数据完整性（Data Integrity）是指数据的准确性和可靠性，用于描述存储的所有数据值均处于客观真实的状态。数据挖掘所使用的数据常常是为其他用途收集的，原始数据中出现的问题会对下一阶段的分析过程产生重大的影响，因此在数据清洗阶段需要检验数据完整性。数据对象遗漏一个或多个属性值的情况在数据挖掘任务中屡见不鲜，例如有的人拒绝透露年龄和体重，这时信息收集不全的现象变得十分常见。数据完整性检验通常包括 5 个基本原则：可溯源、清晰、同步、原始或真实复制、准确。针对数据不完整问题，比如遗漏值，有许多应对的策略，包括删除数据对象或属性、估计遗漏值、在分析时忽略遗漏值、使用默认值、使用属性平均值、使用同类样本平均值、预测最可能的值等。

1.5.10 非静态、非平衡及成本敏感数据的挖掘

网络动态数据（Network Dynamic Data, NDD）指的是网络环境中随时间变化的、无规则的、无限增长的动态数据。这类数据往往是非结构化、半结构化或复杂结构化的动态数据，新产生的数据往往会使得先前挖掘出来的知识“过期”或“失效”，因而难以精确地、有效地获取和处理网络环境中新产生的大规模动态数据。传统数据挖掘技术主要针对静态数据集、数据仓库，在挖掘网

络动态数据时，需要面对不断变化的网络环境及各种动态变化的实时数据之外，还需要对整个数据挖掘过程、数据集与关联规则集的更新过程等进行实时分析和处理。这使得采用传统数据挖掘理论技术来分析网络动态数据变得比较困难，对其问题定义、数据采集、数据预处理、数据清理/集成、数据选择/变换/归约、数据挖掘、模式评估、解释和应用等各个环节产生了较大的影响[15]。

随着用户需求在深度、广度层面的不断深化与拓展，基于海量、高维、动态的网络数据处理需求变得越来越迫切，不仅仅要处理随机样本数据，而且要在网络环境中实时处理过去、现在与未来的全体数据。网络动态数据挖掘（Network Dynamic Data Mining, NDDM）研究是一个具有挑战性、前沿性的新兴研究领域，在电子商务、金融管理、交通工程、气象监测、遥感分析、冶金监测、电力监测等行业及竞争情报、计算机仿真、战略管理、网络监控等研究领域有着广泛的应用需求及前景。

在不平衡数据中，人们将拥有较多实例的那一类称为多数类（有时也称为负样本），将拥有相对较少实例的那一类称为少数类（有时也称为正样本）。如图 9 所示，三角形表示少数类，圆形表示多数类。在实际应用中存在着大量的不平衡数据，比如银行坏账的数目、医疗领域患癌症的数目等。少数类检测和基于不平衡数据的学习已经成为数据挖掘领域的难题被关注。在高度不平衡的数据中进行学习时，传统的特征选择方法所选择的特征更加偏向于多数类而忽略少数类，导致分类器很容易倾向于多数类。因此，为了训练出更加适合不平衡数据的分类模型，对不平衡数据进行处理显得格外重要。比如，如何处理噪声点来减少其对分类器性能的影响，如何分析不平衡数据挖掘方法性能评价曲线并且建立一套标准的能够综合度量分类器性能的评估方法等[16]。

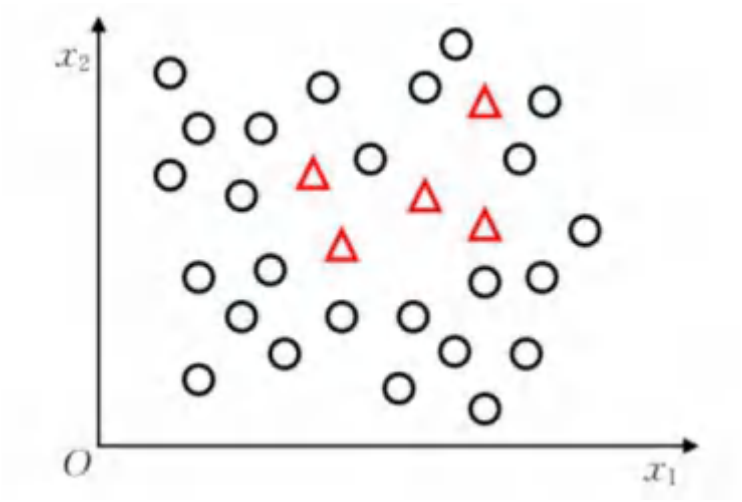


图 9 不平衡数据分布图

信息时代的到来，数据作为信息的载体其重要性也愈加突出，对数据的分析研究可及时发现问题，为决策的可行性评估提供客观依据。但现有的大多算法或过程没有考虑挖掘过程中的“代价”因素。现有分类算法（如决策树和神经网络等）大多假定每个样本的误分类具有同样的代价而致力于提高分类器的泛化精度，分类的结果偏向于大类别样本，即大类别样本的分类精度高于小类别样本。但对很多现实的应用，如医疗诊断、信用卡欺诈检测、网络入侵、故障识别等，不同类别的样本数相差较大，不同样本的误分类代价通常不相等，仅凭全局精度评价分类器的性能优劣是不够的。在极端情况下，不考虑样本的不同误分类代价而建立的模型没有任何意义。例如在医疗诊断中，把“病人”误诊为“健康人”的代价和把“健康人”误诊为“病人”的代价是不同的，前者使“病人”失去治疗的机会，以病情恶化或生命为代价，后者以再次诊断或药物的副作用为代价，显然前者的误分类代价要大于后者。进一步，若样本集包含少量“病人”和大量“健康人”样本，分类器通过把所有样本划分为“健康人”即可得到的分类精度，但这样的诊断不能识别出“病人”，没有实际意义。此时需要引入代价敏感数据挖掘技术（Cost Sensitive Data Mining, CSDM），即考虑“病人”和“健康人”的不同误分类代价，以挖掘结果的期望代价而不是误差率最小为目标。上面的例子针对的是误分类代价，代价还有很多其它类型，如样本获取的代价、考虑分类器稳健性的拒识代价、考虑噪音数据对挖掘算法影响的噪音代价[17]。

2　技术篇

2 技术篇

在当今的大数据时代，数据在社会中扮演着重要的角色。然而数据通常并不能直接被人们利用，如何从大量的看似杂乱无章的数据中揭示出其中隐含的内在规律，发掘出有用的知识以指导人们进行科学的推断与决策，是需要对这些纷繁复杂的数据进行分析的。数据挖掘从一个新的视角将数据库技术、统计学、机器学习、信息检索技术、数据可视化和模式识别与人工智能等领域有机结合起来，它组合了各个领域的优点，因而能从数据中挖掘到运用其他传统方法不能发现的有用知识。这里从数据挖掘的应用领域这一视角来阐述数据挖掘的方法。

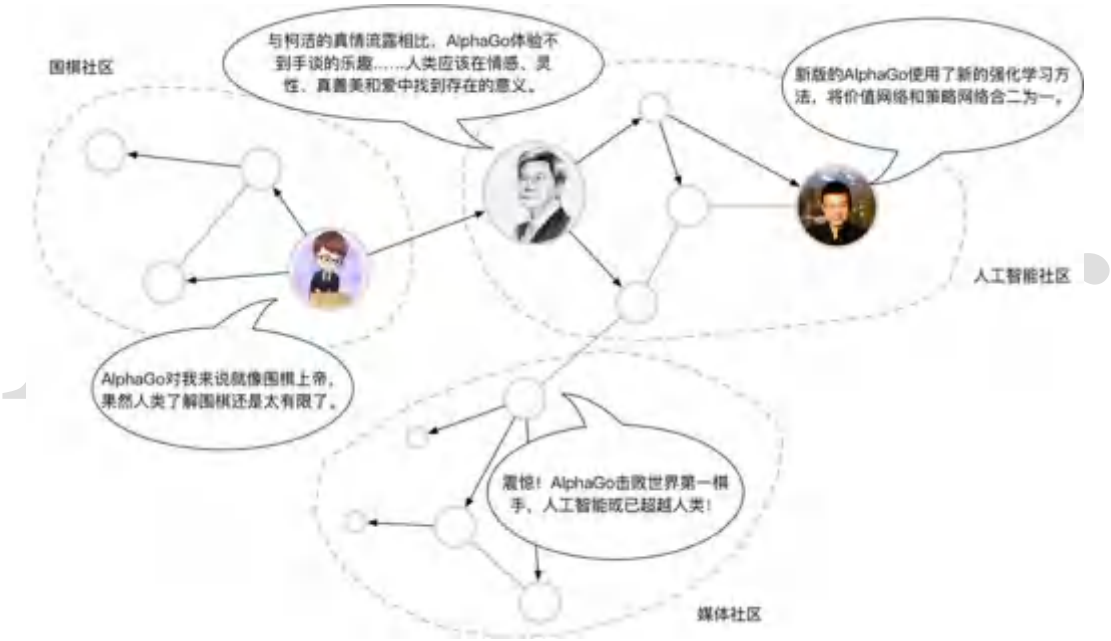


图 10 柯洁乌镇大战 AlphaGo 憾负的微博热议

其实大数据对我们来说并不陌生，举一个例子，新浪微博就是拥有海量数据的资讯平台，截止到 2017 年 12 月，微博已拥有接近 4 亿活跃用户，内容存量超千亿，“大 V”的一举一动和社会热点话题都会引起大量的评论与转发，掀起一股“数据风暴”。如图 10 所示，2017 年 5 月 27 日，世界围棋第一人柯洁九段对战 AlphaGo 的三番棋落下帷幕，柯洁以 0:3 惨败，他赛前的豪言壮志与赛后绝望的泪水令网友动容。“人工智能”瞬间成为微博上热议的话题，引起了千万级别的评论与转发。到底是哪些人对人工智能感兴趣呢？为了形象直观地了解关注者群体的年龄、性别比例、职业等，我们需要对数据进行数据描

述性分析，平均数、中位数、分位数、方差等统计指标可以帮助我们粗略了解数据分布，峰度、偏度等则描述了更细致的特征。在关注程度层面，很多人仅仅是转发，有的用户则是有感而发，年龄、职业等因素是否会影响对该话题的关注程度呢？回归分析、方差分析等方法则可以帮助我们解决这个疑惑。

2.1 数据挖掘十大经典算法

国际权威的学术组织 the IEEE International Conference on Data Mining（ICDM）在 2006 年 12 月评选出了数据挖掘领域的十大经典算法，如图 11 所示。这些算法包括监督和无监督的，可以解决分类和聚类问题，适用于数值型和标称型数据，具有实现简单、泛化错误率低等优点，被广泛应用于多种任务中。本章节通过资料调研[18]，详细介绍了十大经典算法的实现原理和步骤、案例分析、优缺点等内容。

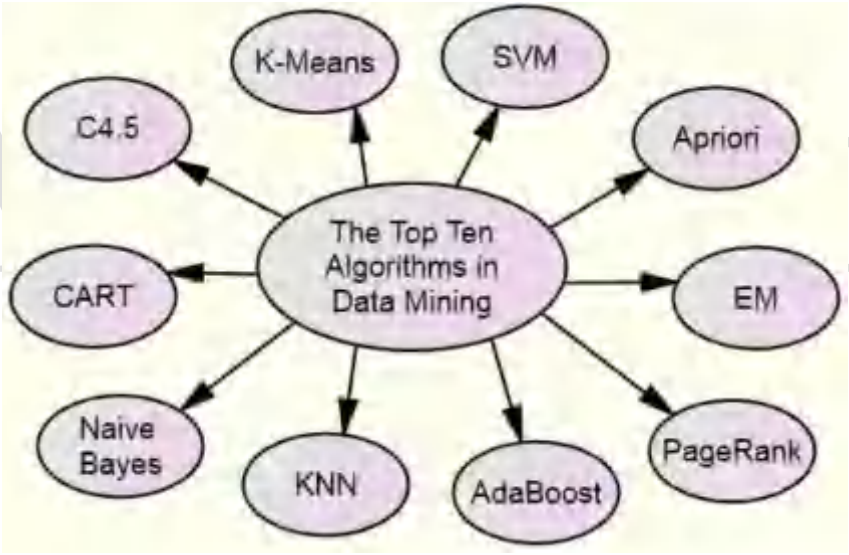


图 11 数据挖掘十大经典算法

2.1.1 C4.5

C4.5 算法是由 Ross Quinlan 在 ID3 算法的基础上，开发的用于产生决策树的算法，通常用于统计分类。C4.5 算法与 ID3 算法一样使用了信息熵的概念，并和 ID3 一样通过学习数据来建立决策树。C4.5 算法对 ID3 算法的改进之处在于：

（1） C4.5 算法用信息增益率来选择属性，克服了 ID3 算法用信息增益选

择属性时偏向选择取值多的属性的不足；

（2） 在树构造过程中进行剪枝；

（3） 能够完成对连续属性的离散化处理；

（4） 能够对不完整数据进行处理。

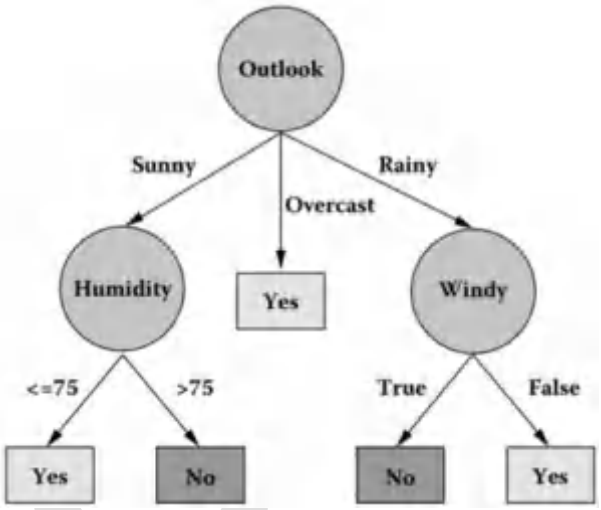


图 12 C4.5 算法生成的决策树[19]

给定一个数据集，其中的每一个元组都能用一组属性值来描述，每一个元组属于一个互斥的类别中的某一类。C4.5 的目标是通过学习，找到一个从属性值到类别的映射关系，并且这个映射能用于对新的类别未知的实体进行分类。这里使用打篮球的案例来研究 C4.5 算法的执行流程。给定一些训练样本，具有如下 4 个属性：Outlook（天气）、Temperature（温度）、Humidity（湿度）、Windy（是否刮风），来判断今天是否适合出去打篮球。图 12 展示了 C4.5 算法根据训练样本生成的决策树。根节点是 Outlook 属性，划分出了三个组合（Sunny、Overcast、Rainy）。其中，Overcast 划分中的集合是“纯”的，故此子树就停止生长，表示 Outlook 的属性值为 Overcast 时，适合出去打篮球。而 Sunny 和 Rainy 的属性值划分样例集合中包含 Humidity 和 Windy 的不同属性值，因此它们不“纯”，需要继续使用子树来表示子集，直到子集为“纯”的（即子集中的所有实例都属于同一个类别），树才停止生长。根据 Yes，得出只有符合以上属性值要求（比如 Outlook 为 Sunny，Humidity<=75；或者 Outlook 为 Rainy，Windy 为 False），才能适合出去打篮球。

把哪个属性作为根节点，是 C4.5 算法研究的重点，它采用信息增益率来选择属性。信息增益率使用“分裂信息”值将信息增益规范化，选择具有最大增益率的属性作为分裂属性。信息增益率的计算公式如下所示：

$$GainRatio(A)=\frac{Gain(A)}{SplitInfo(A)} \quad (1)$$

$$SplitInfo_A(D)=-\sum_{j=1}^{v}\frac{|D_j|}{|D|}*\log_2\left(\frac{|D_j|}{D}\right) \quad (2)$$

$$Gain(A)=Info(D)-Info_A(D) \quad (3)$$

$$Info(D)=-\sum_{i=1}^{m}p_i\log_2(p_i) \quad (4)$$

$$Info_A(D)=\sum_{j=1}^{v}\frac{|D_j|}{|D|}*Info(D_j) \quad (5)$$

公式（1）用于计算属性 A 的信息增益率。公式（2）通过将训练数据集 D 划分成对应于属性 A 测试的 v 个划分产生的信息。信息增益定义为原来的信息需求（即仅基于类比例）与新需求（即对 A 划分之后得到的）之间的差，使用公式（3）进行计算。Info（D）称为熵，使用公式（4）来计算 D 中的元组分类所需求的期望信息。假定按照属性 A 划分 D 中的元组，且属性 A 将 D 划分成 v 个不同的类。在该划分之后，为了得到准确的分类还需要的信息由公式（5）进行度量。

C4.5 算法的优点是：产生的分类规则易于理解，准确率较高。缺点是：在构造树的过程中，需要对数据集进行多次的顺序扫描和排序，因而降低了算法的执行效率[20]。

2.1.2 K-Means

K-Means 算法的思想是对于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个簇，让簇内的点尽量紧密地连在一起，而让簇间的距离尽量的大。如果用数据表达式表示，假设簇划分为（$C_1,C_2,...C_k$），则目标是最小化平方误差 E：

$$E=\sum_{i=1}^{k}\sum_{x\epsilon C_i}||x-\mu_i||_2^2 \quad (1)$$

其中 μ_i 是簇 C_i 的均值向量，有时也称为质心，表达式为：

$$\mu_i = \frac{1}{|C_i|}\sum_{x \epsilon C_i} x \quad (2)$$

求上式的最小值采用的是启发式的迭代方法，图 13 可以形象描述。图（a）表达了初始的数据集，假设 k=2。在图（b）中，我们随机选择了两个 k 类所对应的类别质心，即图中的红色质心和蓝色质心，然后分别求样本中所有点到这两个质心的距离，并标记每个样本的类别为和该样本距离最小的质心的类别，如图 c 所示，经过计算样本和红色质心和蓝色质心的距离，我们得到了所有样本点的第一轮迭代后的类别。此时我们对当前标记为红色和蓝色的点分别求其新的质心，如图（d）所示，新的红色质心和蓝色质心的位置已经发生了变动。图（e）和图（f）重复了图（c）和图（d）的过程，即将所有点的类别标记为距离最近的质心的类别并求新的质心。最终我们得到的两个类别如图（f）。

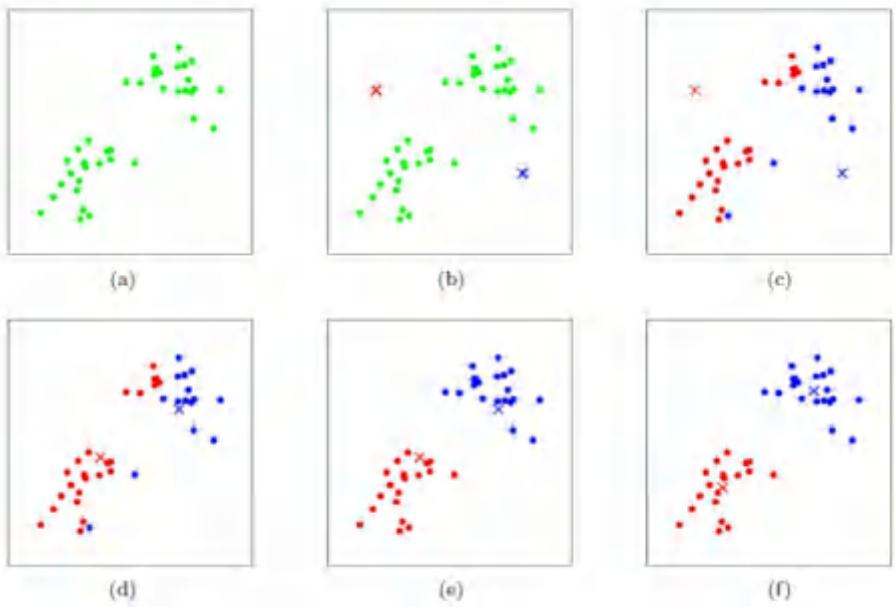


图 13 K-Means 算法效果图[21]

K-Means 的主要优点有：原理比较简单，实现也很容易，收敛速度快；算法的可解释度比较强。缺点有：K 值的选取不好把握；采用迭代方法，得到的结果只是局部最优；对噪音和异常点比较敏感。

2.1.3 SVM（Support Vector Machine）

支持向量机（Support Vector Machine）是一种监督式学习的方法，它广

泛的应用于统计分类以及回归分析中。支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。该算法的基本思想是：找到集合边缘上的若干数据（称为支持向量），用这些点找出一个平面（称为决策面），使得支持向量到该平面的距离最大。

支持向量机之所以坚持寻找最大边缘超平面，是因为它具有最好的泛化能力。它不仅使训练数据具有最佳的分类性，而且为测试数据的正确分类留下了很大的空间。为确保实际找到最大边距超平面，支持向量机会试图最大化关于向量 w 和 b 的目标函数[22]：

$$L_P=\frac{1}{2}||\vec{w}||-\sum_{i=1}^{t}\alpha_i y_i\left(\vec{w}\cdot\vec{x_i}+b\right)+\sum_{i=1}^{t}\alpha_i \quad (1)$$

其中 t 是训练样本数，α_i（i=1,…,t）是L_P在α_i的导数为零的非负数。α_i是拉格朗日乘子，L_P是拉格朗日方程。在这个方程中，向量 w 和常数 b 定义了超平面。

在分类问题中，很多时候有多个解，如下图左边所示，在理想的线性可分的情况下其决策平面会有多个。而 SVM 的基本模型是在特征空间上找到最佳的分离超平面使得训练集上正负样本间隔最大，SVM 算法计算出来的分界会保留对类别最大的间距，即有足够的余量，如图 14 右边所示[23]。

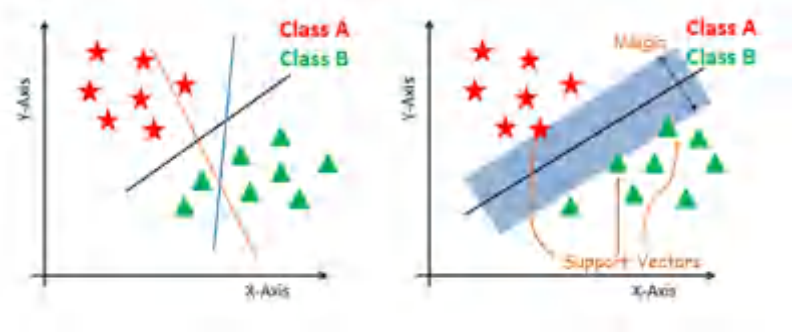


图 14 SVM 的决策平面

在解决线性不可分问题时，它可以通过引入核函数，巧妙地解决了在高维空间中的内积运算，从而很好地解决了非线性分类问题。如下图所示，通过核函数的引入，将线性不可分的数据映射到一个高维的特征空间内，使得数据在

特征空间内是可分的。如图 15 所示：

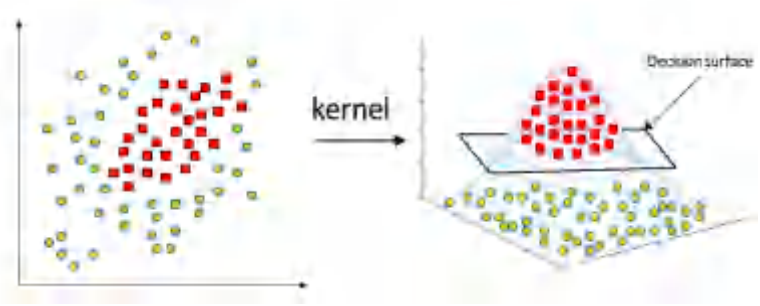


图 15 SVM 的核函数

支持向量机的优点在于：有严格的数学理论支持，可解释性强，不依靠统计方法，从而简化了通常的分类和回归问题；能找出对任务至关重要的关键样本；采用核技巧之后，可以处理非线性分类/回归任务。缺点包括：训练时间长；当支持向量的数量较大时，预测计算复杂度较高。因此支持向量机目前只适合小批量样本的任务，无法适应百万甚至上亿样本的任务。

2.1.4 Apriori

Apriori 算法是一种挖掘关联规则的频繁项集算法，其核心思想是通过候选集生成和情节的向下封闭检测两个阶段来挖掘频繁项集。Apriori 算法已被广泛的应用到商业、网络安全等各个领域。通过该算法我们可以对数据集做关联分析——在大规模的数据中寻找有趣关系的任务，这些关系可以有两种形式：频繁项集、关联规则。频繁项集指经常一块出现的物品集合；关联规则暗示两种物品之间可能存在很强的关系，用表 4 作为示例：

表 4 超市购物清单样例

交易编号	商品
0	豆奶、莴苣
1	莴苣、尿布、葡萄酒、甜菜
2	豆奶、尿布、葡萄酒、橙汁
3	莴苣、豆奶、尿布、葡萄酒
4	莴苣、豆奶、尿布、橙汁

超市中的葡萄酒、尿布、豆奶经常出现在一起，从上面的数据集中也可以找到尿布->葡萄酒的关联规则，这意味着有人买了尿布，那很有可能他也会购买葡萄酒。定义和表示频繁项集和关联规则还需要引入支持度和可信度（置信

度）的概念。**支持度**是指数据集中包含该项集的数据所占数据集的比例，度量一个集合在原始数据中出现的频率。上图中，豆奶的支持度为 4/5，（豆奶、尿布）为 3/5。支持度是针对项集来说的，因此可以定义一个最小支持度，只保留最小支持度的项集；**置信度**是针对一条关联规则来定义的，a->b 的置信度=支持度{a|b}/支持度{a}，a|b 表示 ab 的并集。如{尿布}->{葡萄酒}关联规则的置信度为：支持度{尿布，葡萄酒}/支持度{尿布}，其中{尿布，葡萄酒}的支持度为 3/5，{尿布}的支持度为 4/5，所以“尿布->葡萄酒”的可行度为 3/4=0.75，这意味着尿布的记录中，我们的规则有 75%都适用[24]。

关联分析有两个目标：发现频繁项集，发现关联规则。实际上，Apriori 就是通过排除法来选择频繁项集和关联规则：

（1） 如果某个项集是频繁的，那么它的所有子集也是频繁的；

（2） 如果某个项集是非频繁的，那么它的所有超集也是非频繁的；

（3） 基于此，Apriori 算法从单元素项集开始，通过组合满足最小支持度的项集来形成更大的集合。

首先，找出所有的频繁项集，这些项集出现的频繁性至少和预定义的最小支持度一样。然后由频繁项集产生强关联规则，这些规则必须满足最小支持度和最小可信度。然后使用第一步找到的频繁项集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右边只有一项，一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。为了生成所有频繁项集，使用了递归的方法。

Apriori 算法的优点包括：算法简单明了，没有复杂的理论推导，也易于实现。缺点包括：对数据库的扫描次数过多，算法会产生大量的中间项集，算法的适应面比较窄等。

2.1.5 EM（Expectation Maximization）

EM（Expectation-Maximum）算法也称期望最大化算法，EM 算法受到缺失思想影响，最初是为了解决数据缺失情况下的参数估计问题，其算法基础和收敛有效性等问题在 Dempster、Laird 和 Rubin 三人于 1977 年所做的文章

“Maximum likelihood from incomplete data via the EM algorithm”中给出了详细的阐述。EM 算法是最常见的隐变量估计方法，在机器学习中有极为广泛的用途，例如常被用来学习高斯混合模型的参数、隐式马尔科夫算法、LDA 主题模型的变分推断等。

EM 算法是一种迭代优化策略，由于它的计算方法中每一次迭代都分两步，其中一个为期望步（E 步），另一个为极大步（M 步），所以算法被称为 EM 算法。其基本思想是：根据已经给出的观测数据，估计出模型参数的值；再依据上一步估计出的参数值估计缺失数据的值；然后根据估计出的缺失数据加上之前已经观测到的数据重新再对参数值进行估计，反复迭代，直至最后收敛，迭代结束[25]。

假设有 100 个男生和 100 个女生的身高，但是不知道这 200 个数据中哪个是男生的身高，哪个是女生的身高，即抽取得到的每个样本都不知道是从哪个分布中抽取的。这个时候，对于每个样本，就有两个未知量需要估计[26]：

（1） 这个身高数据是来自于男生数据集合还是来自于女生？

（2） 男生、女生身高数据集的正态分布的参数分别是多少？

图 16 EM 算法要解决的问题

那么，对于具体的身高问题使用 EM 算法求解步骤如图 17 所示。

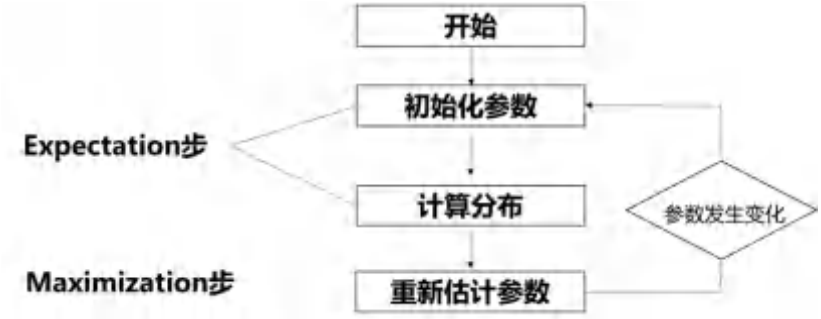


图 17 身高问题 EM 算法求解步骤

（1） 初始化参数：先初始化男生身高的正态分布的参数：如均值=1.65，方差=0.15；

（2） 计算每一个人更可能属于男生分布或者女生分布；

（3） 通过分为男生的 n 个人来重新估计男生身高分布的参数（最大似然估计），女生分布也按照相同的方式估计出来，更新分布；

（4） 这时候两个分布的概率也变了，然后重复步骤（1）至（3），直到参数不发生变化为止。

EM 算法的流程如下[26]：

输入：观察到的数据 $x=(x_1,x_2,\cdots x_n)$ 联合分布 $(x,z;\theta)$，条件分布 $p(z|x,\theta)$，最大迭代次数J。

算法步骤：

（1） 随机初始化模型参数θ的初值 θ_0。

（2） J=1,2,…,J开始 EM 算法迭代：

E 步：计算联合分布的条件概率期望：

$$Q_I(z_i)=p(z_i|x_i,\theta_j) \quad (1)$$

$$l(\theta,\theta_j)=\sum_{i=1}^{n}\sum_{z_i}Q_i(z_i)\log\frac{p(x_i,z_i;\theta)}{Q_I(z_i)} \quad (2)$$

M 步：极大化 $l(\theta,\theta_j)$，得到 θ_{i+1}：

$$\theta_{j+1}=\arg\max l(\theta,\theta_j) \quad (3)$$

如果 θ_{i+1}已经收敛，则算法结束。否则继续进行 E 步和 M 步进行迭代。

输出：模型参数 θ。

传统 EM 算法对初始值敏感，聚类结果随不同的初始值而波动较大。总的来说，EM 算法收敛的优劣很大程度上取决于其初始参数。针对传统 EM 算法对初始

值敏感的问题，许多研究者在 EM 算法初始化方面做了许多研究，如果感兴趣可以自己查阅相关资料。

2.1.6 PageRank

PageRank 算法是 Google 排名运算法则（排名公式）的一个非常重要的组成部分，其用于衡量一个网站好坏的标准。PageRank 根据网站的外部链接和内部链接的数量和质量衡量网站的价值。PageRank 背后的概念是：每个到页面的链接都是对该页面的一次投票，被链接的越多，就意味着被其他网站投票越多。这个就是所谓的“链接流行度”。PageRank 这个概念引自学术中一篇论文的被引述的频度——即被别人引述的次数越多，一般判断这篇论文的权威性就越高。

PageRank 通过网络浩瀚的超链接关系来确定一个页面的等级。Google 把从 A 页面到 B 页面的链接解释为 A 页面给 B 页面投票，根据投票来源（甚至来源的来源，即链接到 A 页面的页面）和投票目标的等级来决定新的等级。简单的说，一个高等级的页面可以使其他低等级页面的等级提升。

假设一个由 4 个页面组成的小团体：A，B，C 和 D。如果所有页面都链向 A，那么 A 的 PR 值将是 B，C 及 D 的 PR 值总和。

$$PR(A)=PR(B)+PR(C)+PR(D) \quad (1)$$

继续假设 B 也有链接到 C，并且 D 也有链接到包括 A 的 3 个页面。一个页面不能投票 2 次。所以 B 给每个页面半票。以同样的逻辑，D 投出的票只有三分之一算到了 A 的 PageRank 上。

$$PR(A)=\frac{PR(B)}{2}+\frac{PR(C)}{1}+\frac{PR(D)}{3} \quad (2)$$

换句话说，根据链出总数平分一个页面的 PR 值。

$$PR(A)=\frac{PR(B)}{L(B)}+\frac{PR(C)}{L(C)}+\frac{PR(D)}{L(D)} \quad (3)$$

最后，所有这些被换算为一个百分比再乘上一个系数。由于“没有向外链接的页面”传递出去的 PageRank 会是 0，所以，Google 通过数学系统给了每个页面一个最小值：

$$PR(A)=\left(\frac{PR(B)}{L(B)}+\frac{PR(C)}{L(C)}+\frac{PR(D)}{L(D)}+\cdots\right)d+\frac{1-d}{N} \quad (4)$$

在 Sergey Brin 和 Lawrence Page1998 年原文中给每一个页面设定的最小值是 1-d，而不是这里的（1-d）/N，所以一个页面的 PageRank 是由其他页面的 PageRank 计算得到。Google 不断地重复计算每个页面的 PageRank，如果给每个页面一个随机 PageRank 值（非 0），那么经过不断地重复计算，这些页面的 PR 值会趋向于稳定，也就是收敛的状态。这就是搜索引擎使用 PageRank 算法计算网页排名的流程。

2.1.7 AdaBoost

Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器（弱分类器），然后把这些弱分类器集合起来，构成一个更强的最终分类器（强分类器）。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。

该算法其实是一个简单的弱分类算法的提升过程，这个过程通过不断地训练，可以提高对数据的分类能力[27]。

（1） 先通过对 N 个训练样本的学习得到第一个弱分类器；

（2） 将分错的样本和其他的新数据一起构成一个新的 N 个的训练样本，整个训练过程如此迭代地进行下去；

（3） 将第 1、2 步都分错了的样本加上其他的新样本构成另一个新的 N 个的训练样本，通过对这个样本的学习得到第三个弱分类器；

（4） 最终经过提升的强分类器。即某个数据被分为哪一类要由各分类器权值决定。

图 18 描述了 AdaBoost 的执行，包含一个弱分类器和 boosting 组件。弱分类器在一维的数据中尝试去寻找最理想的阈值来将数据分离为两类。boosting 组件迭代调用分类器，经过每一步分类，它改变了错误分类示例的权重。因此，

创建了一个级联的弱分类器，它的行为就像一个强分类器。

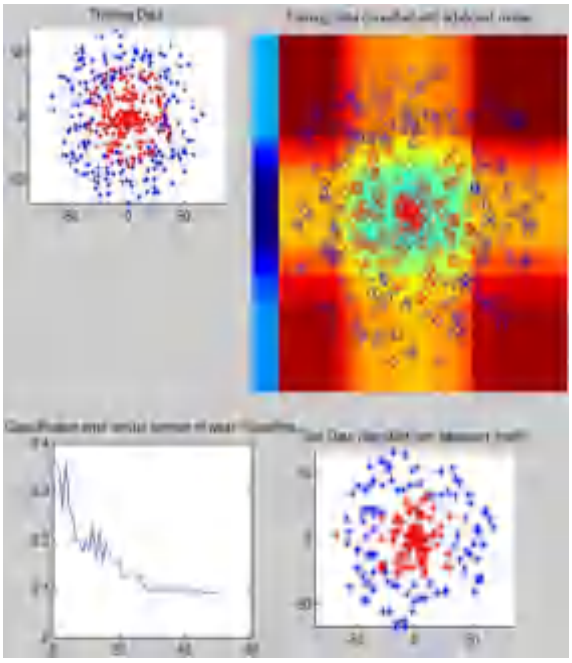


图 18 AdaBoost 结果

目前，对 Adaboost 算法的研究以及应用大多集中于分类问题，同时近年也出现了一些在回归问题上的应用。Adaboost 系列算法主要解决了：两类问题、多类单标签问题、多类多标签问题、大类单标签问题和回归问题。它用全部的训练样本进行学习。

Aadboost 算法系列具有较高的检测速率，且不易出现过拟合现象。但是该算法在实现过程中为取得更高的检测精度则需要较大的训练样本集，执行效果依赖于弱分类器的选择，搜索时间随之增加，故训练过程使得所用时间非常大，也因此限制了该算法的广泛应用。

2.1.8 KNN（K-Nearest Neighbor）

最邻近规则分类算法（K-Nearest Neighbor, KNN）是一个理论上比较成熟的方法。KNN 的核心思想是：如果一个样本在特征空间中的 K 个最相邻的样本中的大多数属于某一个类别，则该样本也属于这个类别，并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。KNN 方法在做类别决策时，只与极少量的相邻样本有关。由于 KNN 方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法

来确定所属类别的，因此对于类域的交叉或重迭较多的待分样本集来说，KNN 方法较其他方法更为适合。

简单来说，KNN 算法可以描述为以下几个步骤[28]：

（1） 计算测试数据与各个训练数据之间的距离；

（2） 按照距离的递增关系进行排序；

（3） 选取距离最小的 K 个点；

（4） 确定前 K 个点所在类别的出现频率；

（5） 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。

如果一个样本在特征空间中的 K 个最邻近的样本中的大多数属于某一个类别，则该样本也划分为这个类别。KNN 算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。通过找出一个样本的 K 个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。图 19 是 KNN 算法中，K 等于不同值时的算法分类结果。

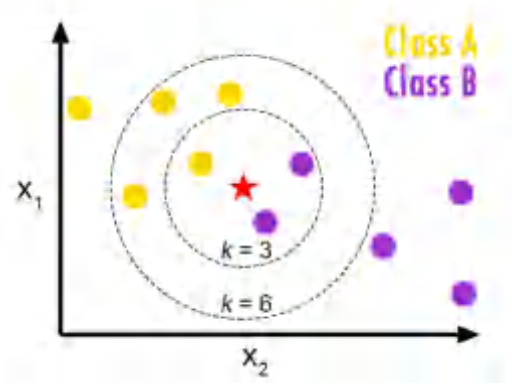


图 19 KNN 算法简单示例

同理，图 20 中我们要确定绿点属于哪个颜色（红色或者蓝色），要做的就是选出距离目标点距离最近的 K 个点，看这 K 个点的大多数颜色是什么颜色。当 K 取 3 的时候，我们可以看出距离最近的三个，分别是红色、红色、蓝色，因此得到目标点为红色[29]。

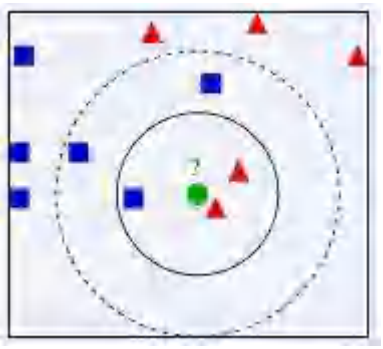


图 20 KNN 算法分类示例

KNN 算法的优点是简单、容易理解，通过 K 值的增大可具备噪音数据的鲁棒性。缺点是需要大量的空间去储存已知的实例，算法复杂度高。

2.1.9 Naive Bayes

朴素贝叶斯法（Naive Bayes）是基于贝叶斯定理与特征条件独立假设的分类方法，和决策树模型是使用最为广泛的两种分类模型。相比决策树模型，朴素贝叶斯模型所需估计的参数很少，对缺失数据不太敏感，算法也比较简单，具有稳定的分类效率。但是在实际应用中，属性之间相互独立的假设往往是不成立的，给朴素贝叶斯模型的分类准确度带来一定影响。

其数学原理很容易理解：如果你看到一个人总是做好事，则会推断那个人多半会是一个好人。这就是说，当你不能准确判断时候，可以依靠事物特定本质相关的事件出现的多少（概率）作为判断依据，贝叶斯定理：

$$P(A|B)=P(B|A)*P(A)/P(B) \quad (1)$$

该公式表示在 B 发生的条件下 A 发生的条件概率，等于 A 事件发生条件下 B 事件发生的条件概率乘以 A 事件的概率，再除以 B 事件发生的概率。公式中，P(A)叫做先验概率，P(A|B)叫做后验概率。

朴素贝叶斯的基本方法：在统计资料的基础上，依据条件概率公式，计算当前特征的样本属于某个分类的概率，选择最大的概率分类[30]。

对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。其计算流程表述如下：

（1） x={a1, a2,..., am}为待分类项，每个 ai 为 x 的一个特征属性；

（2） 有类别集合 C={y1, y2, ..., yn}；

（3） 计算 P(y1|x), P(y2|x), ..., P(yn|x)；

（4） 如果 P(yk|x)=max{P(y1|x)}，则$x \in y_k$。

举个例子，下面的训练数据集反映了学生出去玩（Play）和天气（Weather）之间的关系。

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

图 21 Naïve Bayes 算法分类示例

将数据集转换为频率表，然后创建概率表，比如 P(sunny) =0.36，用朴素贝叶斯计算后验概率，后验概率大的为预测分类。如果天气是 sunny 就出去玩，这样说是否正确？可以根据后验概率来确定以上说法是否正确，即 P(Yes|Sunny)=P(Sunny|Yes)*P(Yes)/P(Sunny)=0.60>0.5，所以天气好就可以出去玩。

朴素贝叶斯算法的优点包括：算法的逻辑性十分简单，当数据呈现不同的特点时，朴素贝叶斯的分类性能不会有太大的差异，算法的鲁棒性比较好。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。缺点包括：属性独立性的条件同时也是朴素贝叶斯分类器的不足之处。数据集属性的独立性在很多情况下是很难满足的，因为数据集的属性之间往往都存在着相互关联，如果在分类过程中出现这种问题，会导致分类的效果大大降低[31]。

2.1.10 CART（Classification and Regression Trees）

分类与回归树（Classification and Regression Trees, CART）是判定树的一个实现方式，可用于分类与回归。由 ID3，C4.5 演化而来，是许多基于树的 bagging、boosting 模型的基础。

CART 是在给定输入随机变量 x 条件下输出随机变量 y 的条件概率分布，与 ID3 和 C4.5 的判定树所不同的是，ID3 和 C4.5 生成的判定树可以是多叉的，每个节点下的叉数由该节点特征的取值种类而定，比如特征年龄分为（青年，中年，老年），那么该节点下可分为 3 叉。而 CART 为假设判定树为二叉树，内部结点特征取值为“是”和“否”，左分支取值为“是”，右分支取值为“否”。这样的判定树等价于递归地二分每一个特征，将输入空间划分为有限个单元，并在这些单元上预测概率分布，也就是在输入给定的条件下输出条件概率分布。CART 算法的优点是可以对复杂和非线性的数据建模，缺点是结果不易理解。

CART 算法由以下两步组成[32]：

（1） 树的生成：基于训练数据集生成决策树，生成的决策树要尽量大；

（2） 树的剪枝：用验证数据集对已生成的树进行剪枝并选择最优子树，这时损失函数最小作为剪枝的标准。

决策树的生成就是通过递归地构建二叉决策树的过程，对回归树用平方误差最小化准则，对分类树用基尼指数最小化准则，进行特征选择，生成二叉树[33]。

CART 回归的流程包括：

（1） 遍历每个特征，对于特征f_i，遍历每个取值 s，用切分点 s 将数据集分为两份，计算切分后的误差；

（2） 求出误差最小的特征及其对应的切分点，此特征即被选中作为分裂结点，切分点形成左右分支；

（3） 递归地重复以上步骤。

CART 算法相比 C4.5 算法的分类方法，采用了简化的二叉树模型，同时特征选择采用了近似的基尼系数来简化计算。当然 CART 树最大的好处是还可以做回归模型，这个 C4.5 没有。下表给出了 ID3、C4.5 和 CART 的一个比较总结。

表 5 ID3、C4.5 和 CART 的比较总结

算法	支持模型	树结构	特征选择	连续值处理	缺失值处理	剪枝
ID3	分类	多叉树	信息增益	不支持	不支持	不支持
C4.5	分类	多叉树	信息增益比	支持	支持	支持
CART	分类、回归	二叉树	基尼系数、均方差	支持	支持	支持

CART 算法的缺点在于：

（1） 在做特征选择的时候都是选择最优的一个特征来做分类决策，但是大多数，分类决策不应该是由某一个特征决定的，而是应该由一组特征决定的。

（2） 如果样本发生一点点的改动，就会导致树结构的剧烈改变。

2.2 统计数据分析

在统计数据分析中，最简单而直接的方式是对数据进行宏观层面的数据描述性分析，例如均值、方差等。

2.2.1 基本统计分析方法

在大数据分析中，我们获取到数据后，第一时间所想的往往是从一个相对宏观的角度来观察一下这些数据长什么样子，也就是分析一下它们的特征。比如对于微博上的名人，我们可以通过近三个月来发布的消息数量来描述他们的活跃度，或者通过平均每条消息被转发的数量来评价它们在粉丝群体中的受欢迎程度。这些能够概括数据位置特性，分散性关联性等数字特征，以及能够反映数据整体分布特征的分析方法，我们称之为数据描述性分析。

表 6 两个比较受欢迎的微博名人在 2018 年 3 月到 2018 年 5 月间的一部分微博数据

微博名人 W			微博名人 Z		
发布时间	点赞数量	转发数量	发布时间	点赞数量	转发数量
18.3.18	3056	187	18.5.8	3398	1175
18.4.29	1169	511	18.3.6	4849	253
18.3.3	2743	177	18.3.18	4246	211

18.4.29	1616	215	18.4.28	4342	113
18.2.22	2391	92	18.3.14	3464	206
18.3.19	930	119	18.5.2	1819	1067
18.4.8	968	331	18.5.1	2300	1056
18.5.2	1011	51	18.4.8	2955	120
18.5.10	1386	36	18.4.17	3023	104
18.4.18	936	38	18.3.14	2560	229

在表示数据的统计特征方面，均值是最常用的指标之一。我们假设一组数据共有n个一维数据，分别是x_1，x_2，…，x_n，则均值$\bar{x}$可以表示为：

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

均值可以用来反映数据的平均水平，比如在表 5 的两个名人的 20 条微博数据中，W 的微博平均点赞数量是 1620.6，Z 的平均点赞数量是 3295.6。也就是说 Z 的微博受欢迎程度看起来更高一些。需要注意的是，影响这一数值的包含很多因素，比如Z发表的微博往往带有图片，同时Z本身有着更多的粉丝数量。因此单从均值来看只能得到一些最直观的信息，若想得到更为细节性的结论仍需对数据进一步的考察。

与均值有着类似表达能力的还有中位数。中位数顾名思义，是数据中按照大小顺序排序后，处于中间位置的数。中位数M可以表示成：

$$M=\begin{cases}x_{\frac{n+1}{2}}，\ n\text{ 为奇数}\\ \frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right)，\ n\text{ 为偶数}\end{cases}$$

相较均值，中位数有着更好的抗扰性。均值虽然能够反映数据的平均表现，但是如果数据本身的差距较大，均值会极大地受到影响。比如在 99 个年收入 10 万的人中加入 1 个年收入 1000 万的可以把平均年收入提高到 19.9 万，但这一均值实际上并没有很好地反映出这个人群的收入特征。而中位数对于这种问题并没有那么敏感。

如果我们把中位数的概念推广可以得到p分位数M_p，也就是排在序列长度p（0≤p≤1）位置的数。M_p可以表示成：

$$M_p=(1+\lfloor(n+1)p\rfloor-(n+1)p)x_{\lfloor(n+1)p\rfloor}+\left((n+1)p-\lfloor(n+1)p\rfloor\right)x_{\lfloor(n+1)p\rfloor+1}$$

其中四分位数（p=0.25和p=0.75）最为常用。箱型图即是四分位数的一种常见应用。两个微博名人的微博点赞数据的箱型图如图 22 所示。

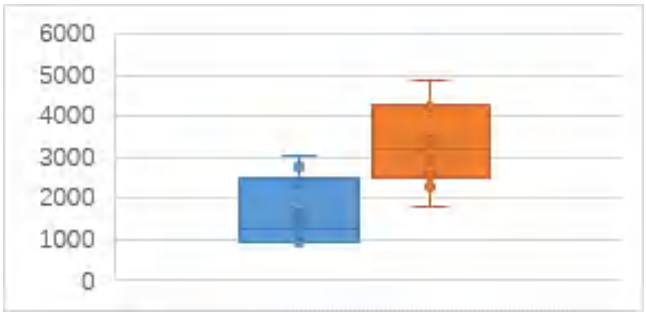


图 22 两个微博名人的微博点赞数据的箱型图

回到上述的年收入例子，对于两组数据：一组包含 99 个年收入 10 万的人和 1 个年收入 1000 万的人；另一组包含 60 个年收入 10 万的人和 40 个年收入 34.75 万的人，我们发现这两组数据无论是均值还是中位数都完全一样。然而第二组数据虽然数据间的差异也很大，但比第一组还是要“均匀”了很多。方差s^2就是用来反映这种数据分散性程度的最常用的一种指标。其算术平方根被称为标准差s。这两种指标的值越大，数据的分散性程度越高。

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$s=\sqrt{s^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

两组数据前一组的方差是 9801，而后一组则是 148.5，可见前一组的数据比后一组的数据要分散得多。反映到年收入的问题上，第一组的 100 个人整体

上的贫富差距更加悬殊。另一种忽略了数据内部差异而仅关注数据上下界的指标是极差R，它被定义为最大值x_{max}与最小值x_{min}之差。

$$R=x_{max}-x_{min}$$

上述的一些指标更多地从单个数值的大小上来反映数据的特征。然而从上述的年收入的例子中我们可以看到，这些指标即使在一些差别很大的数据上仍有可能呈现出相同结果，如果想要更加准确地把握住数据的整体情况，我们不能忽视其分布特征。

对于有限的数据，我们可以通过频率分布直方图来观察数据的分布。如图 23 所示，首先将数据取值范围划分成若干区间（一般取等间隔，间隔大小称为组距），统计数据落入每一区间的频率。当数据量足够多的时候，如果我们把组数不断加大，让组距小到趋近于 0，把纵坐标的频率除以组距，我们可以得到概率密度函数（Probability density function）f_X。其在某一区间上的积分就是数据落在这一范围内的概率。

$$P(a\leqslant X\leqslant b)=\int_a^b f_X(x)\,dx$$

进一步我们有累积分布函数（Cumulative distribution function）F_X。

$$F_X(x)=\int_{-\infty}^{x} f_X(u)\,du$$

从而有：

$$P(a\leqslant X\leqslant b)=F_X(b)-F_X(a)$$

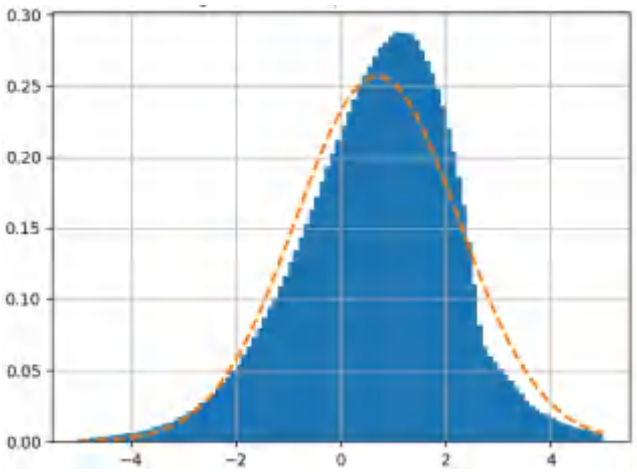


图 23 组数较大组距较小的频率分布直方图

这里，我们其实是在尝试拟合获得的样本数据背后的分布。这是什么意思呢？比如拿微博数据来举例，我们拿到的可能只是 10 条微博数据，然而一个人在过去三个月可能有成百上千条微博数据，我们称之为总体；这 10 条获取到的数据，我们称为样本。我们想通过样本来了解总体的信息，就可以通过样本的数值信息去估计总体的信息。

这里涉及到较多的概率论与数理统计的内容。通过一些推导可以发现，样本均值$\bar{x}$是总体数学期望 μ 的一种无偏估计，而总体方差 σ^2 的一种无偏估计是：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

需要注意的是，这和我们之前提到的数据的方差有些许区别。透过概率密度函数和累积分布函数，我们可以更加直观地看到数据的分布。有时数据的分布会偏向于某种特定分布，其中正态分布（高斯分布）是其中最为常见的一种。数据的分布往往不是对称的。偏度g_1是用来衡量数据对称性的一种指标。

$$g_1=\frac{n}{(n-1)(n-2)s^3}\sum_{i=1}^{n}(x_i-\bar{x})^3$$

在偏度近于零时，数据分布相对比较对称；偏度大于零时，比均值更小的数据更多一些，反之则是比均值更大的数据较多。

另一方面，如果我们以正态分布作为标准，不同分布的数据在均值附近的集中程度也不同：有的分布可能会显得“平坦”一些，有更多的数据分布在两侧；有的分布则看起来比较“尖锐”，数据更多地集中在均值附近。峰度g_2是用来度量分布形状的指标。峰度为正意味着有更多的数据分布在两侧极端，峰度为负则意味着数据较多地集中在均值附近。

$$g_2=\frac{n(n+1)}{(n-1)(n-2)(n-3)s^4}\sum_{i=1}^{n}(x_i-\bar{x})^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

QQ 图（Quantile-Quantile Plot）可以用来帮助判断数据的分布是否服从某一指定分布，同时还可以获得数据的偏度和峰度信息。对于数据分布的检验问题以及其他相关内容涉及较多概率论与数理统计的基础知识，有兴趣的读者可以深入学习。

上述的各种度量指标都是针对一维数据的。然而最开始给出的微博数据，实际上包含了三个维度：发布时间、点赞数量和转发数量。在现实中，我们做数据分析处理时，往往面对的是多维数据，不光有数值型数据（比如年收入），也有具有固定几个属性值的数据（比如性别和职业）；可能有二值型数据（比如是否是会员），也可能有日期这样特殊格式的数据；数值型数据不光有连续值的（比如身高、体重），也有离散值的（比如年龄）。这里我们采用统一的方法来表示多维数据。设一组数据中有n个k维数据x_1，x_2，…，x_n。其中每一个数据$x_i(1\leqslant i\leqslant n)$可以用一个k维向量（$x_{i1}$，$x_{i2}$，…，$x_{ik}$）^{T}表示。其中二维数据的两个维度常常用（X,Y）^{T}表示其总体，每一个数据可以用（x_i,y_i）^{T}来表示。类似一维数据，多维数据也有均值向量$\bar{x}$，以及对应于方差的协方差矩阵（Covariance Matrix）S。

$$\bar{\mathbf{x}}=\left(\frac{1}{n}\sum_{i=1}^{n}x_{i1},\ \frac{1}{n}\sum_{i=1}^{n}x_{i2},\cdots,\frac{1}{n}\sum_{i=1}^{n}x_{ik}\right)^{T}$$

$$\mathbf{S}=[s_{ij}]_{k\times k},\quad s_{ij}=\frac{1}{n-1}\sum_{u=1}^{n}\left(x_{ui}-\bar{x}_{i}\right)\left(x_{uj}-\bar{x}_{j}\right)$$

其中根据协方差矩阵可以计算 Pearson 相关矩阵R。

$$\mathbf{R}=[r_{ij}]_{k\times k},\ r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$$

协方差矩阵和 Pearson 相关矩阵都是对称矩阵，而且 Pearson 相关矩阵对角线为 1。当维度为 2 时，s_{12}被称作X, Y之间的协方差，r_{12}被称作 Pearson 相关系数。这里的相关系数和相关矩阵，反映的是两变量之间的线性相关性。直观地，如果两变量同步上升同步下降则表现出正向线性相关性；如果一个上升另一个下降，两变量朝相反方向变化，则表现出负向线性相关性。与均值向量类似，多维数据中也有中位数向量。在相关性方面，除了上述 Pearson 相关矩阵，还有采用秩的 Spearman 相关矩阵，对应二维情况有 Spearman 相关系数。

对于多维总体，我们也可以将一维情况下的分布函数以及概率密度函数扩展到多维情况下，感兴趣的读者可以阅读相关书籍了解。

2.2.2 回归分析方法

相比于上节所讲的单个变量的统计分析或两个变量相关分析，在实际生产生活中存在着更多的变量。而在多变量的数据分析过程中，我们有时候会对这些变量之间的作用关系感兴趣。比如房价问题，在一个时间区间内，一个房子的价格会受到其空间大小、卧室数量、卫生间数量、所处层数等数值变量的影响，还有朝向、地理位置等其它变量的影响。那么，我们直观上会认为，越大的房间越贵，拥有更多卧室的房间会更贵一些。那么这些因素是如何综合影响

房价的呢？我们可以简单地建立这样的模型：房价Y是由空间大小X_1，卧室数量X_2，卫生间数量X_3等 k 个变量决定的。也就是：

$$Y=f\left(X_1, X_2, \cdots, X_k\right)+\varepsilon$$

其中ε是由于我们没有考虑进去的各种因素所产生的误差。特别地，当f是关于$X_1, X_2, \cdots, X_k$的线性函数时，我们有：

$$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k+\varepsilon$$

称作线性回归模型。其中$\beta_0, \beta_1, \cdots, \beta_k$是未知参数，也称为回归参数或回归系数。我们假设 ε 是数学期望为 0 的随机误差。

对于线性回归模型，当我们有了 n（n 一般较大，n≥k）个房子的数据时，我们便可以使用这些数据去估计未知参数。设第 i 个房子的数据为$（x_{i1}, x_{i2}, \cdots, x_{ik}）^T$，对应房价$y_i$（实际上Y也可以是多维变量）。采用向量的方式表示就是：

$$Y=X\beta+\varepsilon$$

我们采用最小二乘法来估计回归参数。即最小化：

$$S(\beta)=\varepsilon^T\varepsilon=\sum_{i=1}^{n}（y_i-\sum_{j=1}^{k}\beta_j x_{ij}）^2$$

分别对$\beta_0, \beta_1, \cdots, \beta_k$求偏导并令其得零。得到正规方程：

$$X^T X\beta=X^T Y$$

由于$X^T X$是实对称正定矩阵，我们可以得到β的最小二乘估计为：

$$\widehat{\beta}=（X^T X）^{-1} X^T Y$$

由于$S(\beta)$的 Hessian 矩阵正定且上述最小二乘估计为唯一解，因此$S(\widehat{\beta})$为$S(\beta)$的极小值点。代回得到：

$$\widehat{Y}=X\widehat{\beta}+\varepsilon$$

为回归方程。此时，我们就完成了对于未知参数$\beta_0,\beta_1,\cdots,\beta_k$的估计，接下来我们便能使用这个回归方程去根据新房子的大小等信息估计或者预测它的房价了。

我们的分析到这一步，其实就已经完成了大部分。在应用中，我们用上述的过程建立模型，利用已有数据拟合模型并应用在未知的数据上。最小二乘法保证了，当我们的模型采用上述的线性模型时，上述的解就是最优解。然而，即使是最优解也会存在误差。毕竟，现实中这些变量之间的关系总是复杂的，不一定像我们假设的那样是简单的线性关系。那么，在我们研究的问题上，线性模型究竟合适不合适呢？接下来，我们分析一下这个线性模型对于训练数据究竟能够拟合到什么程度。

我们在得到回归方程后，将估计参数时使用的观测值，代回方程，将观测值Y与拟合值$\widehat{Y}$作差得到残差向量：

$$\widehat{\varepsilon}=Y-\widehat{Y}=Y-X(X^TX)^{-1}X^TY=(I-H)Y$$

进而我们可以定义残差平方和SSE（Sum-of-Square Error）来衡量误差大小。

$$SSE=\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2=\widehat{\varepsilon}^T\widehat{\varepsilon}$$

SSE 越小，拟合得到的模型对于拟合时使用的数据估计越准确。进而定义复相关系数R：

$$R^2=1-\frac{SSE}{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}$$

来衡量模型的线性显著性。R越接近 1，就代表线性模型对于该问题越合适。如果R很小，那么我们就需要去考虑是不是少考虑了某些重要因素，或者说问题本身就无法用线性模型来解决。

上述过程在大数据分析中，观测样本数 n 往往较大，求解回归方程时采用迭代法求解会更快更高效一些。而对于定性变量与因变量之间的关系，我们可以使用方差分析。对统计感兴趣的读者可以阅读相关书籍中线性回归关系的显著性检验、假设检验一般方法以及方差分析深入了解。

2.2.3 关联分析

关联分析方法可以发现隐藏在大型数据集中有意义的联系。这种联系可以用关联规则来表示。在使用关联规则时，需要考虑两个问题：一是从大数据集中发现模式可能效率很低；二是所发现的某些关联可能是毫无意义的。支持度这一度量可以删除那些毫无意义的关联规则，置信度可以度量规则的可能性大小。关联分析的算法主要有：Apriori 算法、DHP 算法、DIC 算法和 FP-增长算法等。其中最常用的是 Apriori 算法，下面对它做一介绍。

关联模式中最著名的是 Apriori 算法，它由 R. Agrawal 等人首先提出来[34]，其算法思想是：首先找出频繁性至少和预定义的最小支持度一样的所有频集，然后由频集产生强关联规则。最典型的例子就是沃尔玛尿布和啤酒事件，在此例中，商家利用统计这两种商品在一次购买中共同出现的频数，将出现频数多的搭配转化为关联规则。

Apriori 算法的实现是通过对数据库的多次扫描来发现所有的频繁项目集。在每一次扫描中只考虑具有同一长度的所有项目集，在进行第一次扫描中，Apriori 算法计算数据库中所有单个项目的支持度，生成所有长度为 1 的频繁项目集；在后续的每一次扫描中，首先以第 1 次扫描所生成的所有项目集为基础产生新的候选项目集，然后扫描数据库，计算这些候选项目集的支持度，删除

其支持度低于用户给定的最小支持度的项目集；最后生成频繁项目集。重复以上过程直到再也发现不了新的频繁项目集为止。

由此可见，若要提高 Apriori 算法的效率，可以减少对数据库的扫描次数或者减少不必要的频繁项目集的生成，对 Apriori 算法的改进主要方法有：基于划分的方法；基于抽样的方法；增量更新方法；概念层次的方法和基于散列和压缩技术的方法[35]。

2.2.4 聚类分析

聚类分析是将数据划分成具有意义的组进行多元统计分析，是一种定量方法。讨论的对象是大量的样本，要求能够合理地按照各自的特性进行分类，没有任何模式可供参考或依循，即是在没有先验知识的情况下进行的。

聚类分析的基本思想是认为研究的样本或变量之间存在着程度不同的相似性（亲疏关系）。根据一批样本的多个观测指标，指出一些能够度量样本或变量之间相似程度的统计量，以这些统计量作为分类的依据，把一些相似程度较大的样本聚合为一类，直到把所有的样本都聚合完毕，形成一个由小到大的分类系统。选择用哪种聚类算法由数据类型、聚类目的和应用决定。主要的聚类方法有以下几种。

（1） 划分聚类

给定一个有 N 条记录的数据集，以及要生成簇的数目 K。划分方法首先给出一个初始的分组方法，然后通过反复迭代的方式改变分组，使得每一次改进之后的分组方案都比前一次好。该方法常用的算法有：K-Means 算法、K-medoids 算法和 CLARANS 算法等。K-Means 算法已在 2.1.2 小节进行详细介绍，这里不再阐述。

- K-medoids 算法

K-medoids（中心点）聚类法与 K-means 算法类似，区别在于中心点的选取，K-means 中选取的中心点为当前类簇中所有点的重心，而 K-medoids 法选取的中心点为当前类簇中存在的一点，准则函数是当前类簇中所有其他点到该中心点的距离之和最小，如图 24 所示，这就在一定程度上削弱了异常值的影响，但缺

点是计算较为复杂，耗费的计算时间比 K-means 多。

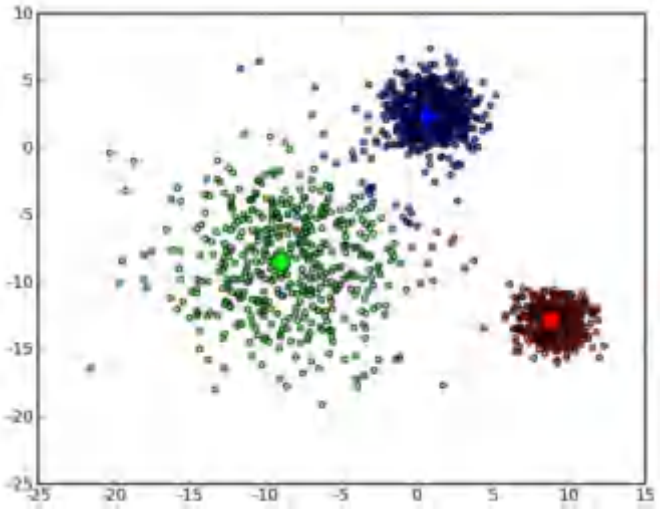


图 24 K-medoids 算法样例

具体的算法流程如下：

a) 在总体 n 个样本点中任意选取 k 个点作为 medoids；

b) 按照与 medoids 最近的原则，将剩余 n-k 个点分配到当前最佳的 medoids 代表的类中；

c) 对于第 i 个类中除对应 medoids 点外的所有其他点，按顺序计算当其为新的 medoids 时，准则函数的值，遍历所有可能，选取准则函数最小时对应的点作为新的 medoids；

d) 重复 2-3 的过程，直到所有的 medoids 点不再发生变化或已达到设定的最大迭代次数；

e) 产出最终确定的 k 个类。

- CLARANS 算法

CLARANS（Clustering LARge Applications，大型应用中的聚类方法）是分割方法中基于随机搜索的大型应用聚类算法。在分割方法中最早提出的一些算法大多对小数据集合非常有效，但对大的数据集合没有良好的可伸缩性，例如一种典型的 k-medoids 算法 PAM（partitioning around medoids）。

CLARA 能处理比 PAM 大的数据集合，其有效性取决于样本的大小，但当某个采样得到的中心点不属于最佳的中心点时，CLARA 不能得到最佳聚类结果。CLARANS 是在 CLARA 算法的基础上提出来的，与 CLARA 不同，CLARANS 没有在任一给定的时间局限于任一样本，而是在搜索的每一步都带一定随机性地选取一个样本。CLARANS 的时间复杂度大约是 O（n^2），其中 n 是对象的数目。此方法的优点是一方面改进了 CLARA 的聚类质量，另一方面拓展了数据处理量的伸缩范围，具有较好的聚类效果。但它的计算效率较低，且对数据输入顺序敏感，只能聚类凸状或球型边界。

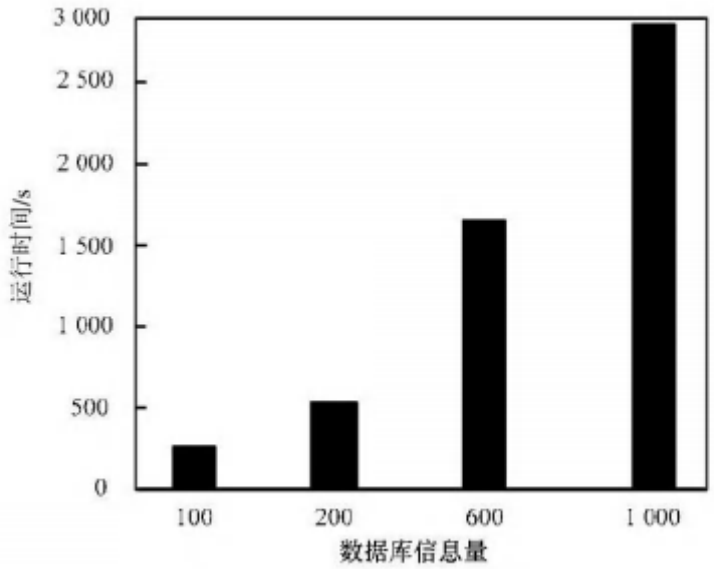


图 25 不确定性目标的 CLARANS 聚类算法对于不同大小数据库的运行时间比较[36]

具体的算法流程如下[37]：

a) 输入参数 numlocal 和 maxneighbor；

b) 从 n 个目标中随机地选取 k 个目标构成质心集合，并令它们作为 current；

c) 令 j 等于 1；

d) 从 2 中剩下的 n－k 个目标集中随机选取一个目标，并用之替换质心集合中随机的某一个质心可得到一个新的质心集合，计算两个质心集合的代价差；

e) 如果新的质心集合代价较小则将其赋给 current，重置 j=1，否则 j+=1；

f）直到 j 大于等于 maxneighbor，则 current 为此时的最小代价质心集合；

g）重复以上步骤 numlocal 次，取其中代价最小的质心集合为最终质心集合；

h）按照最终质心集合进行划分并输出。

（2） 层次聚类

层次方法是对给定的数据对象集合进行层次分解，层次方法可以分为凝聚和分裂。该方法在合并、分裂的时候要检测大量的记录和簇，因而伸缩性比较差。比较常见的方法有四种：BIRCH、CURE、ROCK 和 Chameleon。

- BIRCH 算法

利用层次结构的平衡迭代归约和聚类（Balanced Iterative Reducing and Clustering using Hierarchies，BIRCH） 是为大量数值数据聚类设计的，它将层次聚类（在初始微聚类阶段）与诸如迭代地划分这样的其他聚类算法（在其后的宏聚类阶段）集成在一起。它克服了凝聚聚类方法所面临的两个困难：a）可伸缩性；b）不能撤销先前步骤所做的工作。

BIRCH 使用聚类特征来概括一个簇，使用聚类特征树（CF-树）来表示聚类的层次结构，BIRCH 方法对新对象增量或动态聚类也非常有效。

给定有限的主存，BIRCH 一个重要的考虑是最小化 I/O 时间。BIRCH 采用了一种多阶段聚类技术：数据集的单遍扫描产生一个基本的好聚类，而一或多遍的额外扫描可以进一步地改进聚类质量。它主要包括两个阶段：

a）BIRCH 扫描数据库，建立一棵存放于内存的初始 CF-树，它可以被看做数据的多层压缩，试图保留数据的内在聚类结构；

b）BIRCH 采用某个（选定的）聚类算法对 CF-树的叶结点进行聚类，把稀疏的簇当做离群点删除，而把稠密的簇合并为更大的簇[38]。

图 26 BIRCH 流程图[39]

- CURE 算法

CURE（Clustering Using Representative）采用了一种新型的层次聚类算法，这种算法介于“单链”和“组平均”之间，他克服了这两种层次聚类算法的不足之处，可以处理大型数据、离群点和具有非球形大小和非均匀大小的簇的数据。

CURE 算法的基本流程可以表示为图 27，首先从原始数据集中随机抽取一部分样本点作为子集，再对该子集进行划分，在这些划分后的集合上运行 CURE 聚类算法得到每个集合的簇，并删除其中的离群点，然后对这些簇进一步进行 CURE 层次聚类，并删除其中的离群点，最后对磁盘中剩余的数据集样本点进行划分。

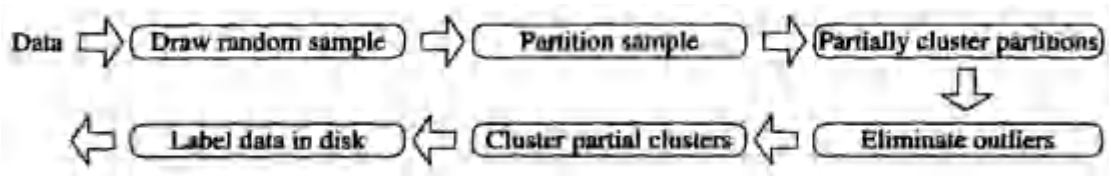


图 27 CURE 算法的基本流程[40]

绝大多数聚类算法或者擅长处理球形和相似大小的聚类，或者在存在孤立点时变得比较脆弱。CURE 采用了一种新颖的层次聚类算法，该算法选择基于质

心和基于代表对象方法之间的中间策略。它不同于单个质心或对象来代表一个类，而是选择数据空间中固定数目的具有代表性的点。一个类的代表点通过如下方式产生：首先选择类中分散的对象，然后根据一个特定的分数或收缩因子“收缩”或移动它们。在算法的每一步，有最近距离的代表点对（每个点来自于一个不同的类）的两个类被合并。

每个类有多于一个的代表点使得 CURE 可以适应非球形的几何形状。类的收缩或凝聚可以有助于控制孤立点的影响。因此，CURE 对孤立点的处理更加健壮，而且能够识别非球形和大小变化比较大的类。针对大型数据库，CURE 采用随机取样和划分两种方法组合：一个随机样本首先被划分，每个划分被部分聚类。

CURE 算法的步骤如下：

a）从源数据对象中抽取一个随机样本 S；

b）将样本 S 分割为一组划分；

c）对每个划分局部的聚类；

d）通过随机取样剔除孤立点，如果一个类增长太慢，就去掉它；

e）对局部的类进行聚类，落在每个新形成的类中的代表点根据用户定义的一个收缩因子收缩或向类中心移动，这些点代表和捕捉到了类的形状；

f）用相应的类标签来标记数据。

- ROCK 算法

ROCK（RObust Clustering using linKs）聚类算法是一种鲁棒的用于分类属性的聚类算法，该算法属于凝聚型的层次聚类算法。之所以鲁棒是因为在确认两对象（样本点/簇）之间的关系时考虑了他们共同的邻居（相似样本点）的数量，在算法中被叫做链接（Link）的概念。而一些聚类算法只关注对象之间的相似度[42]。

ROCK 算法适用于类别型数据，比如，关键字、比尔值、和枚举值等，其核心思想是利用链接作为相似性的度量，而不是仅仅依赖于距离。但相似度阈值需要预先指定，阈值对聚类质量影响很大，在对数据集没有充分了解的前提下

很难给出合理的阈值。另外，在 ROCK 算法中，相似度函数仅被用于最初邻居的判断上，只考虑相似与否，而未考虑相似程度，使算法对相似度阈值过于敏感。

ROCK 算法的具体流程如下[43]：

a）输入聚类个数 k 值和相似度阈值 θ；

b）计算点与点之间的相似度，生成相似度矩阵；

c）计算邻居矩阵 A；

d）计算链接矩阵 L= AxA；

e）计算 Goodness Measure，将相似性最高的两个对象合并；

f）回到第 3 步进行迭代更新，直到形成 k 个聚类，或者聚类的数量不再发生变换。

- Chameleon 算法

Chameleon（变色龙）是一种层次聚类算法，它采用动态建模来确定一对簇之间的相似度。在 Chameleon 中，簇的相似度依据如下两点评估：a）簇中对象的连接情况；b）簇的邻近性。也就是说，如果两个簇的互连性都很高并且它们之间又靠得很近就将其合并。这样，Chameleon 就不用依赖于一个静态的、用户提供的模型，能够自动地适应被合并簇的内部特征。这一合并过程有利于发现自然、同构的簇，并且只要定义了相似度函数就可应用于所有类型的数据。图 28 解释了 Chameleon 如何运作。

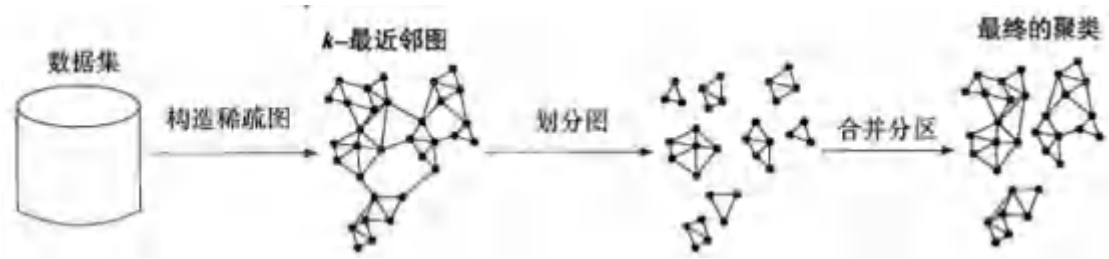


图 28 Chameleon 运作过程示意图

Chameleon 采用 k-最近邻图的方法来构建一个稀疏图，其中，图的每个顶点代表一个数据对象，如果一个对象是另一个对象的 k 个最相似的对象之一，

那么这两个顶点（对象）之间就存在一条边，这些边加权后反映对象间的相似度。Chameleon 使用一种图划分算法，把 k-最近邻图划分成大量相对较小的子簇，使得边割最小。也就是说，簇 C 被划分成子簇 C_i 和 C_j，使得把 C 二分成 C_i 和 C_j 而被切断的边的权重之和最小，它评估簇 C_i 和 C_j 之间的绝对互连性。然后，Chameleon 使用一种凝聚层次聚类算法，其基于子簇的相似度反复地合并子簇。为了确定最相似的子簇对，它既考虑每个簇的互连性，又考虑簇的邻近性（closeness）。更确切地说，Chameleon 根据两个簇 C_i 和 C_j 的相对互连度 RI（C_i，C_j）和相对接近度 RC（C_i，C_j）来决定它们的相似度[38]。

（3） 基于密度的方法

基于密度的方法与其他方法的一个本质区别是：它不是基于距离作为相似性度量的，而是基于密度的。这样就能克服基于距离的算法只能发现类球状聚类的缺点。最具代表性的是 DBSCAN 算法、OPTICS 算法和 DENCLUE 算法。

- DBSCAN 算法

DBSCAN（Density-Based Spatial Clustering of Applications with Noise）是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。

DBSCAN 中的几个定义：a） E 邻域：给定对象半径为 E 内的区域称为该对象的 E 邻域；b）核心对象：如果给定对象 E 邻域内的样本点数大于等于 MinPts，则称该对象为核心对象；c）直接密度可达：对于样本集合 D，如果样本点 q 在 p 的 E 邻域内，并且 p 为核心对象，那么对象 q 从对象 p 直接密度可达；d）密度可达：对于样本集合 D，给定一串样本点 p_1，$p_2 \cdots p_n$，$p=p_1$，$q=p_n$，假如对象 p_i 从 p_{i-1} 直接密度可达，那么对象 q 从对象 p 密度可达；e）密度相连：存在样本集合 D 中的一点 o，如果对象 o 到对象 p 和对象 q 都是密度可达的，那么 p 和 q 密度相联。

DBSCAN 算法的流程描述如下[44]：

输入：包含 n 个对象的数据库，半径 e，最少数目 MinPts；

输出：所有生成的簇，达到密度要求。

方法：

a) Repeat；

b) 从数据库中抽出一个未处理的点；

c) IF 抽出的点是核心点 THEN 找出所有从该点密度可达的对象，形成一个簇；

d) ELSE 抽出的点是边缘点（非核心对象），跳出本次循环，寻找下一个点；

e) UNTIL 所有的点都被处理。

- OPTICS 算法

为了克服在聚类分析中使用一组全局参数的缺点，提出了 OPTICS（Ordering points to identify the clustering structure）聚类分析方法。OPTICS 并不显式地产生数据集聚类，而是输出簇排序（cluster ordering），这个排序是所有分析对象的线性表，并且代表了数据的基于密度的聚类结构。较稠密簇中的对象在簇排序中相互靠近，这个排序等价于从广泛的参数设置中得到的基于密度的聚类。这样，OPTICS 不需要用户提供特定密度阈值，簇排序可以用来提取基本的聚类信息（例如，簇中心或任意形状的簇），导出内在的聚类结构，也可以提供聚类的可视化[38]。

OPTICS 的两个重要概念：a）核心距离：对象 p 的核心距离是指 p 成为核心对象的最小 E′，如果 p 不是核心对象，那么 p 的核心距离没有任何意义；b）可达距离：对象 q 到对象 p 的可达距离是指 p 的核心距离和 p 与 q 之间欧几里得距离之间的较大值，如果 p 不是核心对象，p 和 q 之间的可达距离没有意义。

OPTICS 算法描述如下[45]：

输入：样本集 D，邻域半径 E，给定点在 E 领域内成为核心对象的最小领域点数 MinPts；

输出：具有可达距离信息的样本点输出排序。

方法：

a) 创建两个队列，有序队列和结果队列（有序队列用来存储核心对象及其核心对象的直接可达对象，并按可达距离升序排列；结果队列用来存储样本点的输出次序）；

b) 如果所有样本集 D 中的点都处理完毕，则算法结束。否则，选择一个未处理（即不在结果队列中）且为核心对象的样本点，找到其所有直接密度可达样本点，如果该样本点不存在于结果队列中，则将其放入有序队列中，并按可达距离排序；

c) 如果有序队列为空，则跳至 2，否则，从有序队列中取出第一个样本点（即可达距离最小的样本点）进行拓展，并将取出的样本点保存至结果队列中，如果它不存在结果队列当中的话：

✓ 判断该拓展点是否是核心对象，如果不是，回到 3，否则找到该拓展点所有的直接密度可达点；

✓ 判断该直接密度可达样本点是否已经存在结果队列，是则不处理，否则下一步；

✓ 如果有序队列中已经存在该直接密度可达点，如果此时新的可达距离小于旧的可达距离，则用新可达距离取代旧可达距离，有序队列重新排序；

✓ 如果有序队列中不存在该直接密度可达样本点，则插入该点，并对有序队列重新排序；

d) 算法结束，输出结果队列中的有序样本点。

- DENCLUE 算法

DENCLUE（DENsity-baused CLUstEring，基于密度的聚类）是一种基于一组密度分布函数的聚类算法。该算法的主要思想是：一个样本的影响可以用一个数学函数形式化建模，该函数称为影响函数（influence function），描述数据点对其邻域的影响；数据空间的整体密度可以用所有数据点的影响函数的和来建模；簇可以通过识别密度吸引点数学确定，其密度吸引点是全局密度函数

的局部最大值。

该算法对有巨大噪声的数据集有良好的聚类特性，允许对高维数据集中的任意形状的聚类给出简洁的数学描述，比现存的算法快的多（如 DBSCAN）。不过该算法需要大量的参数，且参数对结果的影响巨大。

DENCLUE 算法的步骤如下[46]：

a) 对数据点占据的空间推导密度函数；

b) 通过沿密度增长最大的方向（即梯度方向）移动，识别密度函数的局部最大点（这是局部吸引点），将每个点关联到一个密度吸引点；

c) 定义与特定的密度吸引点相关联的点构成的簇；

d) 丢弃与非平凡密度吸引点相关联的簇（密度吸引点 x′称为非平凡密度吸引点，如果 f*（x′）＜η （其中 f 是密度函数，η 是指定的阀值）；

e) 若两个密度吸引点之间存在密度大于或等于 η 的路径，则合并所代表的簇。对所有的密度吸引点重复此过程，直到不再改变时算法中止。

（4） 基于网格的方法

这种方法首先将数据空间划分成有限个单元的网格结构，所有的处理都是以单个的单元为对象。这么处理的一个明显优点就是处理速度很快，通常这是与目标数据集中记录的个数无关的，它只与把数据空间划分的单元数量有关。代表算法有：STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法。

- STING 算法

STING（STatistical INformation Grid,统计信息网格）是一种基于网格的多分辨率的聚类技术，它将输入对象的空间区域划分成矩形单元。

图 29 显示了 STING 聚类的一个层次结构。高层单元的统计参数可以很容易地从低层单元的参数计算得到。这些参数包括：属性无关的参数 count（计数）;属性相关的参数 mean（均值）、stdev（标准差）、min（最小值）、max（最大值），以及该单元中属性值遵循的 distribution（分布）类型，如

normal（正态的）、uniform（均匀的）、exponential（指数的）或 none（如果分布未知）。这里，属性是一个选作分析的度量，如住宅对象 price。当数据被加载到数据库时，最底层单元的参数 count、mean、stdev、min 和 max 直接由数据计算。如果分布的类型事先知道，则 disribution 的值可以由用户指定，也可以通过假设检验（如 x^2 检验）来获得。较高层单元的分布类型可以基于其对应的低层单元多数的分布类型，用一个阈值过滤过程的合取来计算。如果低层单元的分布彼此不同，阈值检验失败，则高层单元的分布类型被置为 none。

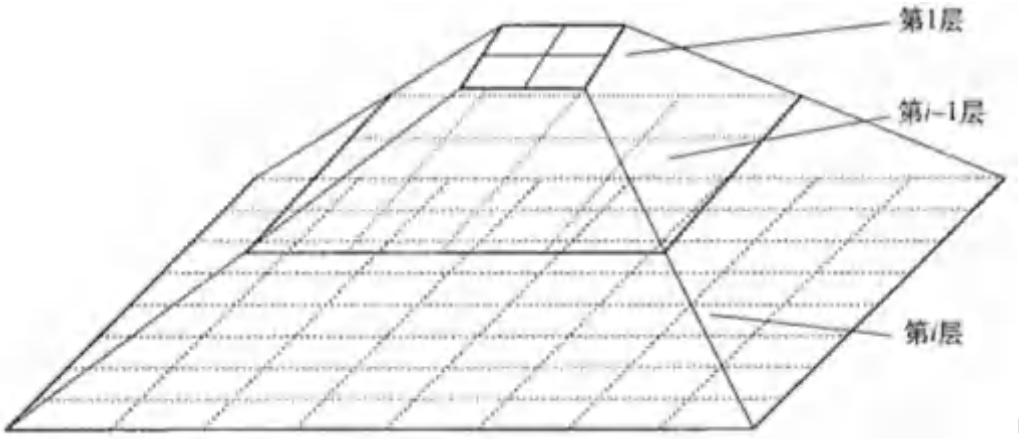


图 29 STING 聚类层次结构

STING 有几个优点：a）基于网格的计算是独立于查询的，因为存储在每个单元中的统计信息提供了单元中数据汇总信息，不依赖于查询；b）网格结构有利于并行处理和增量更新；c）该方法的主要优点是效率高。STING 扫描数据库一次来计算单元的统计信息，因此产生聚类的时间复杂度是 O（n），其中 n 是对象数。在层次结构建立后，查询处理时间是 O（g），其中 g 是最底层网格单元的数目，通常远远小于 n。

由于 STING 采用了一种多分辨率的方法来进行聚类分析，因此 STING 的聚类质量取决于网格结构的最底层的粒度。如果最底层的粒度很细，则处理的代价会显著增加；然而，如果网格结构最底层的粒度太粗，则会降低聚类分析的质量。此外，STING 在构建一个父亲单元时没有考虑子女单元和其相邻单元之间的联系。因此，结果簇的形状是 isothetic,即所有的簇边界不是水平的，就是竖直的，没有斜的分界线。尽管该技术有较快的处理速度，但可能降低簇的质量和精确性[38]。

- CLIQUE 算法

CLIQUE 算法是基于网格的空间聚类算法，但它同时也非常好地结合了基于密度的聚类算法，因此既能够发现任意形状的簇，又可以像基于网格的算法一样处理较大的多维数据。

CLIQUE 算法把每个维划分成不重叠的社区，从而把数据对象的整个嵌入空间划分成单元，它使用一个密度阈值来识别稠密单位，一个单元是稠密的，如果映射到它的对象超过密度阈值。该算法需要两个参数：一个是网格的步长，第二个是密度的阈值。网格步长确定了空间的划分，而密度阈值用来定义密集网格。

CLIQUE 算法的聚类思想：a）首先扫描所有网格。当发现第一个密集网格时，便以该网格开始扩展，扩展原则是若一个网格与已知密集区域内的网格邻接并且其其自身也是密集的，则将该网格加入到该密集区域中，直到不再有这样的网格被发现为止。b）算法再继续扫描网格并重复上述过程，直到所有网格被遍历。以自动地发现最高维的子空间，高密度聚类存在于这些子空间中，并且对元组的输入顺序不敏感，无需假设任何规范的数据分布，它随输入数据的大小线性地扩展，当数据的维数增加时具有良好的可伸缩性。

CLIQUE 算法的聚类过程[41]：

a）把数据空间划分若不重叠的矩形单元，并计算每个网格的密度，根据给定的阀值,识别稠密网格和非稠密网格，且置所有网格初始状态为“未处理标记”；

b）遍历所有网格,判断当前网格是否有“未处理标记”，若没有，则处理下一个网格，否则进行如下步骤 3-7 处理，直到所有网格处理完成，转步骤 8；

c）改变网格标记为“已处理”若是非密集网格，则转步骤 2；

d）若是密集网格,则将其赋予新的簇标记，创建一个队列，将该密集网格置入队列；

e）判断队列是否为空，若空，则转处理下一个网格，转步骤 2，否则进行

如下处理：

- ✓ 取出队头的网格元素，检查其所有邻接的有“未处理标记”的网格；
- ✓ 更改网格标记为“已处理”；
- ✓ 若邻接网格为密集网格，则将其赋予当前簇标记，并将其加入队列；
- ✓ 转步骤 5；

f）密度连通区域检查结束，标记相同的密集网格组成密度连通区域，即目标簇；

g）修改簇标记，进行下一个簇的查找，转步骤 2；

h）遍历整个数据集，将数据元素标记为所在网格簇标记值。

- WAVE-CLUSTER 算法

WAVE-CLUSTER 方法在聚类分析中引入了小波变换的原理，主要应用于信号处理领域，主要思想是把多维数据看作一个多维信号来处理。它首先将数据空间划分成网格结构，然后通过小波变换将数据空间变换成频域空间，在频域空间通过与一个核函数作卷积后,数据的自然聚类属性就显现出来。WAVE-CLUSTER 方法是一个多分辨率的算法，高分辨率可以获得细节的信息，低分辨率可以获得轮廓信息。方法的时间复杂度是 D（n），其中 n 是数据库中对象的个数。

（5） 基于模型的方法

基于模型的方法假设数据集是由一系列的概率分布所决定的，给每一个聚类假定一个模型，然后寻找数据对给定模型的最佳拟合。这样的一个模型可能是数据点在空间中的密度分布函数，它由一系列的概念分布决定，也可以是通过基于标准的统计来自动求出聚类的数目。典型的算法包括：EM 算法、COBWEB 算法、SOM 算法。EM 算法已在 2.1.5 小节进行详细介绍，这里不再阐述。

- COBWEB 算法

基于模型的聚类算法中的阶层式方法是一含有渐增式学习以建立概念阶层架构的非监督式机器学习方法。其概念阶层架构为一树状架构，根节点为概念

架构的最高层观念，因此根节点包含了所有实例的信息。COBWEB 算法就是属于这种阶层式概念聚类算法，该算法是由机器学习研究者在 20 世纪 80 年代提出的，用于在对象-属性数据集处理方面。它产生聚类树状图，这个树被命名为分类树，树的各个节点都是实例属性等信息的描述。该树用概率描述来刻画整个聚类，采用了一个启发式估算度量——分类效用来指导树的构建。

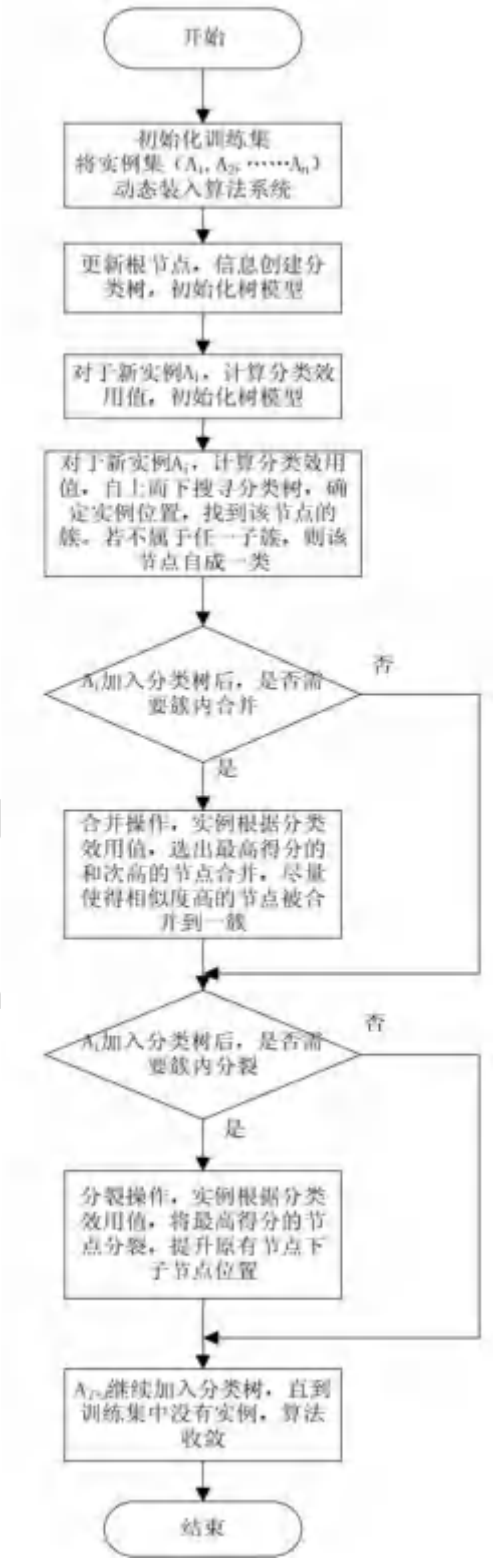


图 30 COBWEB 算法逻辑流程图

COBWEB 算法步骤如下[47]：

输入：实例集（A1,A2.....An）；

输出：分类树及形成的概念层次。

方法：

a）初始化训练集，将实例集（$A_1, A_2 \ldots\ldots A_n$）动态装入算法系统中；

b）更新根节点信息，创建分类树，初始化树模型；

c）对于新实例 A_i 计算分类效用值（CU），自上而下搜寻分类树，确定实例位置，找到该节点的簇。若不属于任一子簇，则该节点自成一类；

d）A_i 加入分类树后，是否需要簇内进行合并或分裂操作，若合并执行 5，分裂跳至 6；

e）合并操作，实例根据分类效用值，选出最高得分的和次高的节点进行合并，尽量使得相似性高的节点被合并到一个簇中；

f）分裂操作，实例根据分类效用值，将最高得分的节点分裂，提升原有节点下子节点位置；

g）A_{i+1} 继续加入分类树，直到训练集中没有实例，算法收敛。

- SOM 算法

SOM（Self Organized Maps）神经网络是由芬兰神经网络专家 Kohonen 教授提出的，该算法假设在输入对象中存在一些拓扑结构或顺序，可以实现从输入空间（n 维）到输出平面（2 维）的降维映射，其映射具有拓扑特征保持性质，与实际的大脑处理有很强的理论联系。自组织映射神经网络，即 Self Organizing Maps（SOM），可以对数据进行无监督学习聚类。它的思想很简单，本质上是一种只有输入层—隐藏层的神经网络。隐藏层中的一个节点代表一个需要聚成的类。训练时采用“竞争学习”的方式，每个输入的样例在隐藏层中找到一个和它最匹配的节点，称为它的激活节点，也叫“winning neuron”。紧接着用随机梯度下降法更新激活节点的参数。同时，和激活节点临近的点也根据它们距离激活节点的远近而适当地更新参数。

SOM 的一个特点是，隐藏层的节点是有拓扑关系的。这个拓扑关系需要我们确定，如果想要一维的模型，那么隐藏节点依次连成一条线；如果想要二维

的拓扑关系，那么就形成一个平面，如图 31 所示。

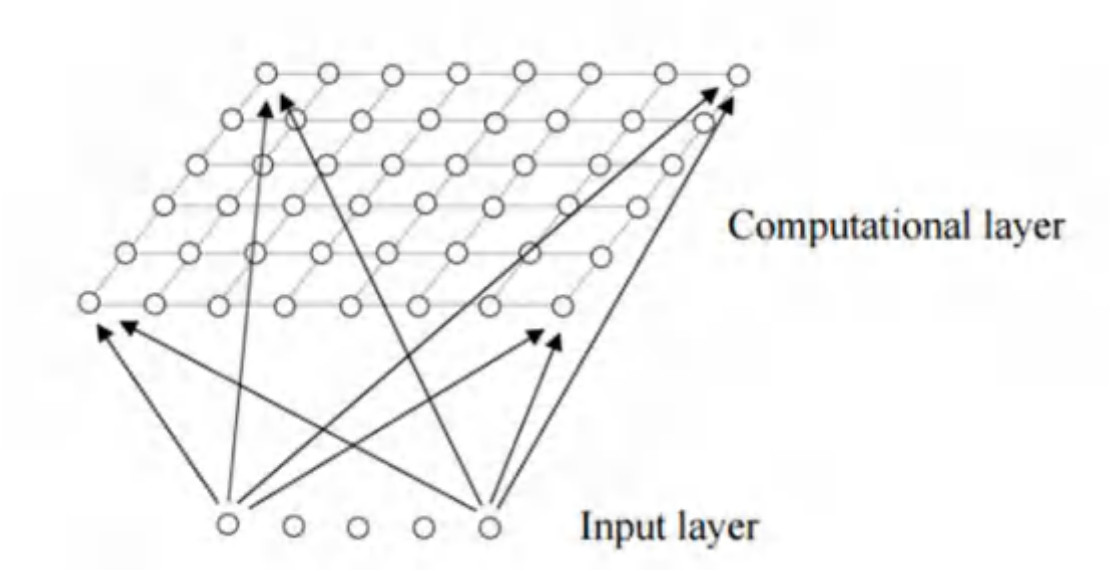


图 31 Kohonen Network

SOM 网络包含输入层和输出层。输入层对应一个高维的输入向量，输出层由一系列组织在 2 维网格上的有序节点构成，输入节点与输出节点通过权重向量连接。学习过程中，找到与之距离最短的输出层单元，即获胜单元，对其更新。同时，将邻近区域的权值更新，使输出节点保持输入向量的拓扑特征。

既然隐藏层是有拓扑关系的，所以我们也可以说，SOM 可以把任意维度的输入离散化到一维或者二维（更高维度的不常见）的离散空间上。Computation layer 里面的节点与 Input layer 的节点是全连接的。

SOM 算法的流程如下[48]：

a) 网络初始化，对输出层每个节点权重赋初值；

b) 将输入样本中随机选取输入向量，找到与输入向量距离最小的权重向量；

c) 定义获胜单元，在获胜单元的邻近区域调整权重使其向输入向量靠拢；

d) 提供新样本、进行训练；

e) 收缩邻域半径、减小学习率、重复，直到小于允许值，输出聚类结果。

SOM 算法优点：

将相邻关系强加在簇质心上，有利于聚类结果的解释；具有降维功能；可

视化；自组织；无监督学习；拓扑结构保持；概率分布保持。

SOM 算法缺点：

用户必选选择参数、邻域函数、网格类型和质心个数；一个 SOM 簇通常并不对应单个自然簇、可能有自然簇的合并和分裂；缺乏具体的目标函数；SOM 不保证收敛。

2.3 科技情报挖掘技术

2.3.1 知识溯源

知识溯源，即追本溯源，探寻知识的根本、源头。溯源问题在近年来得到了大量的研究与关注，产生了如基于传染源中心性的溯源算法、基于置信传播的溯源算法、基于蒙特卡洛（Monte-Carlo）模拟的溯源算法、基于最小描述长度的溯源算法等。其中，基于置信传播的溯源算法主要通过因子图（factor graph）上的消息传递来做概率图模型上的概率推断，从而计算未观察节点上的边际概率分布。基于最小描述长度的溯源算法认为真正的源头应该使得描述源头和传播结果的长度最短。相关的技术方法可以用于大规模传染病的疫病控制、社交网络中的谣言溯源、广告营销中的最大化营销效果等任务中。比如，在大规模的病毒爆发与传染过程中，检测到最有可能的感染源头将会对感染的控制、感染机制研究、传播影响最大化问题提供帮助与见解。

在社交网络的影响传播中，独立级联（IC）模型和线性阈值模型（LT）是两种经典的影响模型。这两个模型中处于激活状态的点感染邻居的方式是不同的。独立级联模型中处于激活状态的节点通过某一成功概率试图激活邻居节点，如果失败，该影响被抛弃，即只有一次感染邻居的机会。然而在线性阈值模型中，当一个已经激活的节点去试图激活邻居节点而没有成功时，其对邻居节点的影响力被累积而不是被舍弃，这个贡献直到节点被激活或传播过程结束为止，这是一个影响累积的过程。其中，集合中每个点随机分配一个阈值（或者按一定分布分配），累计影响超过阈值时节点被感染。初始激活节点集合中每个节点任意分配一个影响力值，即对邻居节点有多大影响力[49]。

2.3.2 趋势分析

技术是产业结构升级、经济发展的根本动力，技术创新已成为决定企业和国家竞争力的重要因素。在全球化的背景下，技术竞争越来越激烈，提前把握技术未来发展趋势，对于国家抢占战略制高点具有重要意义。技术趋势分析的常用定量方法包括基于回归分析的方法、基于引文网络的方法、基于时序主题的方法，一般以历史数据为分析基础，根据曲线拟合进行趋势外推或里程节点变化进行趋势研究。常用技术趋势分析方法的优缺点对比如表 7 所示[50]。

表 7 常用技术趋势分析方法的优缺点对比

方法	优点	缺点
基于回归分析的方法	简单易行	未考虑技术特性，只是浅显的趋势（增长/递减）描述
基于社会网络分析的方法	能较好反映技术之间的发展关联关系	脱离文本语义信息、路径的复杂性或局限性、节点过多或过度选择节点等
基于时序主题的方法	清晰的描述各阶段技术主题	技术描述过于笼统，无法深入理解技术发展趋势，缺少技术之间的发展关联分析

统计回归分析方法，通过历史数据的拟合来判断变量间存在的定量关系，线性回归是一种简单的回归分析模型，其一般计算表达式如公式（1）。其中，Y为因变量用于表示数量，t为自变量用于表示时间，β0为截距，β1为斜率，εi 为时间 i 时的偏差值。β1 的值越大，Y 变化越快，拟合曲线示例如下图所示。该方法认为技术发展受到某种自然属性的限制，当技术突破限制后会快速发展，之后当技术发展达到一定程度后又逐渐趋缓，直至达到饱和状态又遇到新的瓶颈，这时技术的增速再次放缓，直到出现渐进式创新或颠覆性创新。

$$Y_i=\beta_0+\beta_1 t_i+\varepsilon_i,\ i=1,2,\cdots,n \quad (1)$$

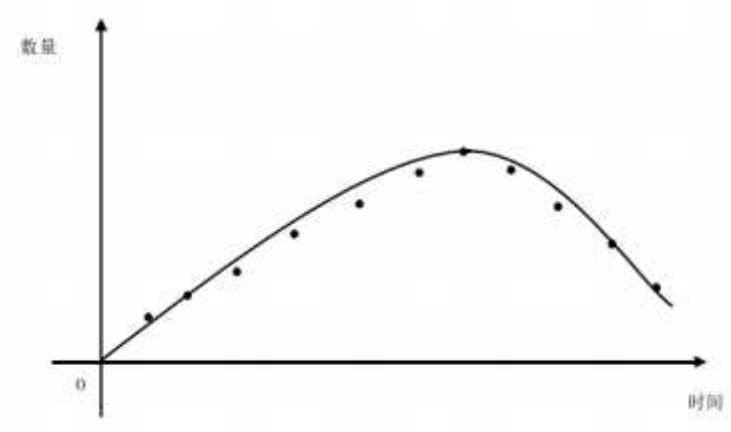


图 32 基于回归分析的趋势拟合曲线示例

基于社会网络分析的方法以技术演进路径作为研究视角，一般基于专利或文献数据，利用它们的引用关系，描绘技术领域的发展历史和演化路径。从庞杂的专利或文献整体引用网络中识别出技术发展变化和趋势存在较大困难，因此，有学者考虑从局部网或核心网中分析技术发展趋势。基于引文关系的主路径分析主要是通过发现整体网络中连接度最大的文献路径来描绘技术领域的发展态势。很多学者逐渐将主路径分析方法应用于技术的趋势分析上。基于专利/文献引用的技术演进轨迹的一般分析流程包括数据获取、数据清洗、技术阶段划分、主路径提取、技术演进轨迹分析等五个步骤，如图 33 所示。

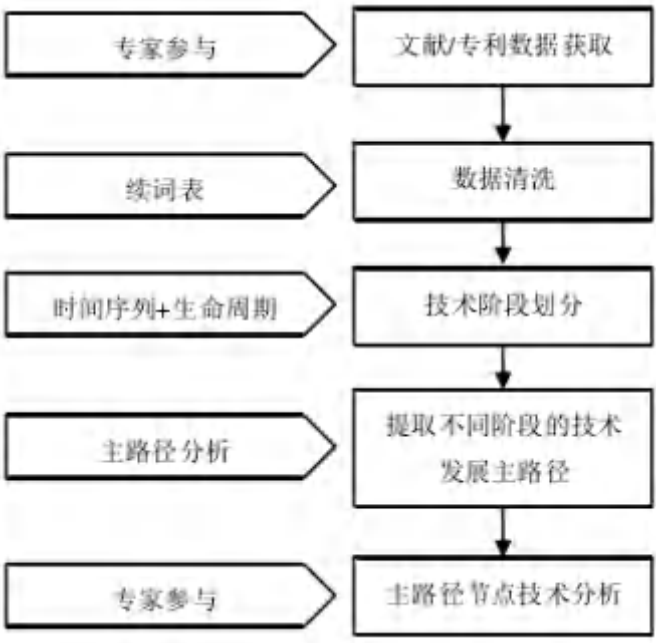


图 33 基于引用关系的技术演变路径分析流程

基于时序主题的方法将技术视作一个动态发展的环境，通过不同阶段技术主题的变化，从而识别技术的发展趋势。基于时序主题的常用方法包括专利地图、技术路线图等。廖君华等[51]基于博客、微博、新闻、网友评论等网络数据，利用 LDA 和时间标签，识别了与钓鱼岛相关各主题随时间变化情况，对于舆情监控提供了一种分析方法。Venugopalan 等[52]基于主题分类分别绘制每年对应的专利地图，揭示技术变化的隐藏模式。基于时序主题分析的方法仅凝练出各阶段的技术发展主题，缺少细化技术的解读和技术之间关系的分析。

2.3.3 前沿预测

前沿技术预测，是指在广泛收集技术相关信息的基础上，采取人机紧密结合方法，通过技术发展现状分析、动向跟踪和趋势拟合，研判未来某个时间点技术的应用情况，以及对经济社会发展影响程度，为管理部门开展科技战略决策提供参考和借鉴[53]。

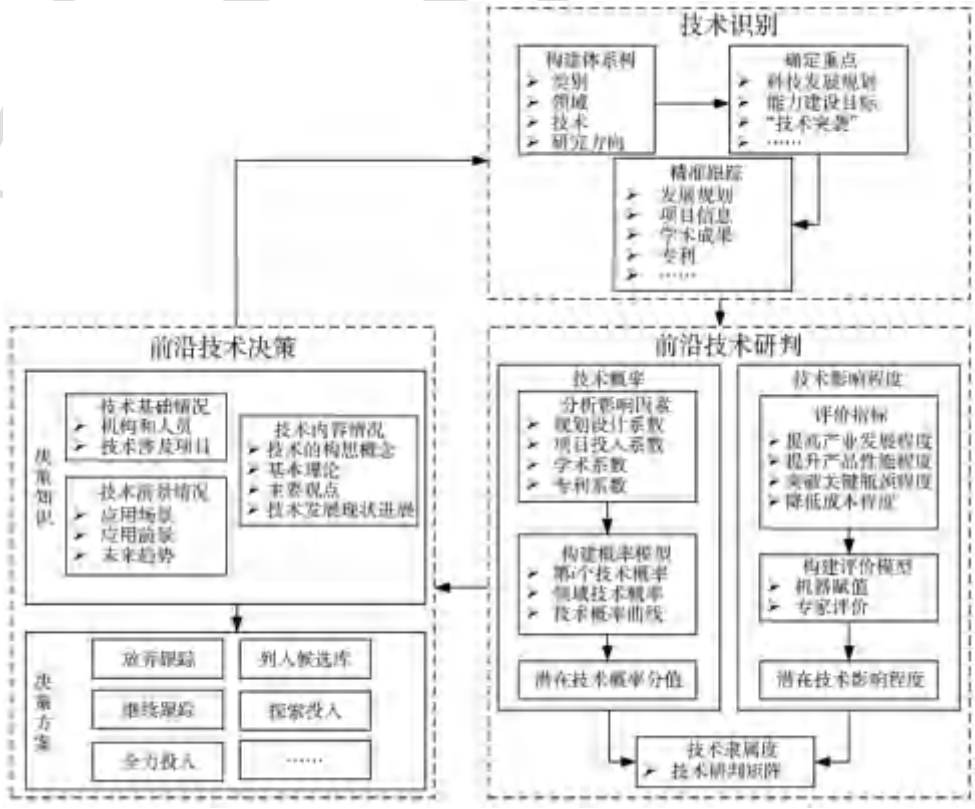


图 34 基于 IRD 的前沿技术预测总体思路

前沿技术预测以发现和研判前沿技术为目标，按照紧贴需求、突出重点、形成体系、人机结合的原则，借鉴直觉推断法、趋势分析法、机器学习、模拟

仿真等技术预测方法，设计由“识（Identity）—研（Research）—决（Decision）”（IRD）构成的前沿技术预测总体思路，判断未来某个时间点技术的状态和属性，实现技术的精准跟踪、前沿技术的科学研判和技术的全息展示。主要包括技术识别、前沿技术研判和前沿技术决策等 3 个主要步骤，如图 34 所示[54]。

（1）技术识别，主要是构建覆盖全范围、全方位、全要素的技术体系，识别拟预测的技术主要领域和技术方向，为前沿技术研判提供标准化的跟踪样本和参照。

（2）前沿技术研判。动态跟踪识别后的重点技术领域或技术，从某项技术概率和技术影响两个维度出发，综合分析判断该项技术隶属于前沿技术程度。

（3）前沿技术决策，推送技术研判结果给科技管理部门，并动态展示该项技术的发展趋势、重点、机构和应用前景等。管理部门根据国家科技发展战略，提出科学可行的技术决策方案。

2.3.4 命名排歧

实际环境中，人名实体存在歧义性，主要体现在两个方面：第一，名称变体，同一个人名人物实体具有多个实体指称项，如笔名、别名等；第二，同一人名实体指称项可指向多个真实世界的人物实体，即同名现象。在信息工程、知识库构建、信息融合等任务中，常常需要预先确定一个人名实体指称项所指向的真实世界中具体的人物，即通常的人名消歧。

按任务目标划分，人名消歧包括基于聚类的人名消歧（知识库没有给定）和基于实体链接的人名消歧（知识库给定）。由于知识库没有给定，基于聚类的人名消歧将所有指向同一个目标实体的实体指称项聚在同一个类别下，每一个类别仅仅对应一个目标人物实体。基于实体链接的人名消歧是指将属于知识库的人名指称项映射到知识库中；将不属于知识库的指称项中所有指向同一个人物实体的指称项聚到一类；指向非人物实体的指称项标记为 Other 类。比如，对于每篇科技文献，遍历其中涉及的所有主要学者，一般是文献作者、项目负责人和参与人员等，将其匹配到知识库中已有的学者上，或将其识别为一个新

发现的学者进行入库。Zhang 等人采用一种基于语义的命名实体排歧方法，基本思想是首先利用融合全局和局部信息的表示学习方法将实体投影到低维空间计算文档之间的相似度；然后使用随机采样方式建立伪训练集来预估候选集重名个数；最后采用群智学习策略提高数据结果的准确性，方法框架如图 35 所示[55]。

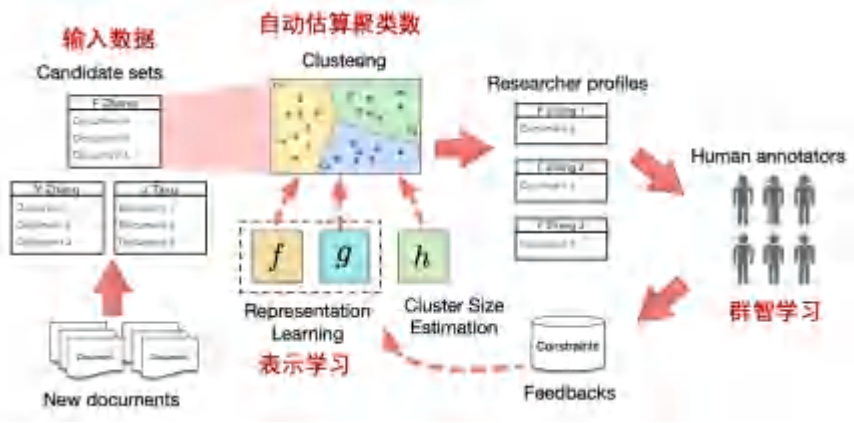


图 35 命名实体消歧架构图

2.3.5 决策支持

决策支持系统的目标是支持管理者的决策过程，在信息技术的支持下实现信息分析。借助于各种数据处理方式，决策支持系统能够辅助管理者进行决策，实现信息增值。信息技术的飞速发展是推动决策支持系统发展的强大动力。综观国内外决策支持系统的研究领域，决策支持系统的发展演变过程有很强的逻辑基础，从时间的先后来看，具有十分明显的分界线（如图 36 所示）[56]。

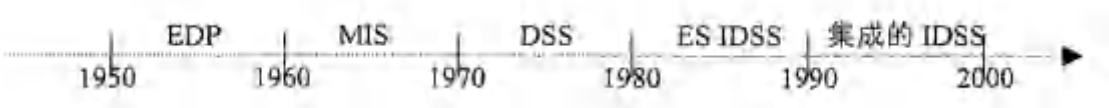


图 36 决策支持系统的发展演变过程

事务处理系统（TPS/EDP）开始于 20 世纪 50 年代中期，一些大公司对于其本身重复性的事务工作开始采用计算机，利用事务处理系统能够提高计算机单项数据的处理能力。管理信息系统（MIS）产生于关系数据库技术与直接访问存储设备结合的背景，从各种来源获取系统中信息处理所需的全面数据，并对其进行检索、筛选和组合以便及时地为管理决策过程提供必要的信息。

决策支持系统（DSS）最早于 20 世纪 70 年代,由美国 M.S.Scott Morton 教授在《智能决策系统》一文中首先提出。它是以管理科学、信息学、控制论和行为科学为基础，以计算机技术、仿真技术和信息技术为手段，面对复杂的决策问题辅助中高层决策者决策活动的，具有一定智能行为的人机交互系统。

智能决策支持系统（IDSS）将人工智能（包括：专家系统、自然语言处理（NLP）和人工神经网络（ANN）等）和决策支持系统结合，融合定性分析和定量分析方法，使得决策支持的效果大大改善。专家系统（ES）是 20 世纪 50 年代人工智能的进一步发展，利用专家的知识在计算机上进行推理，达到专家解决问题的能力。

集成化的智能决策支持系统（IIDSS）是决策支持系统发展的最新趋势，以数据仓库为基础，以联机分析处理和数据挖掘技术为手段的决策支持系统解决方案。它可以解决决策支持系统中信息分析面临问题，可以按照信息分析的需要组织数据，能够建立分析型处理环境的数据存储和组织技术，可以为信息分析提供全面、统一的信息源，进而改善决策的效果与效率。

2.3.6 人才情报

国际竞争的实质是以经济和科技实力为基础的综合国力的较量，而综合国力的较量归根到底是人才的竞争，人才是科技创新最关键的因素。大力培养和吸引科技人才已成为世界各国赢得国际竞争优势的战略性选择。

科技人才评价是科技创新活动的重要内容，是构建创新型科技人才管理制度的重要基础性研究工作。赵伟等人以胜任力模型理论和个体创新行为理论为基础，提出了创新型科技人才个体素质评价冰山模型，并通过建立个体素质与创新行为的映射关系，从创新知识、创新能力、创新技能、创新动力、管理能力和影响力等六个方面构建了创新型科技人才评价的理论模型[57]。陈苏超等人运用模糊层次分析法筛选出对高层次创新型科技人才有影响力的重要指标，包含 4 个二级指标（知识层次、创新能力、社会贡献、综合能力）、18 个三级指标，引入模糊数或模糊语言变量将其量化，利用模糊神经网络模型对高层次创新型科技人才进行评价[58]。盛楠等人将科技人才分为科技创新人才和科技创业人才，并根据人才评价理论和人岗匹配原理分别构建了包含 3 个一级指标（基

本素质、创新能力、创新成果）的科技创新人才评价指标体系和 3 个一级指标（基本素质、创业能力、创业项目）的科技创业人才评价指标体系[59]。李瑞等人针对工程技术类高层次创新型科技人才开展深度访谈，结合胜任力模型理论，设计并检验调查量表，运用因子分析方法构建了包含 6 个一级指标（创新知识、创新技能、影响力、创新能力、创新动力、管理能力）、16 个二级指标和 47 个三级指标的工程技术类高层次创新型科技人才评价指标体系[60]。

基于前述现状分析，可以发现现有的科技人才评价体系多以定性方法为主，一般采用同行评议，但是同行评议方法需要邀请领域专家，一方面费时费力，另一方面细分领域的专家难以甄别，并且评议结果容易参杂专家的主观感受，影响了结果的客观性。此外，同行评议方法能够评审的对象数量有限，不利于大规模的人才输出。因此，有必要研究一种可计量的科技人才评价方法，弥补大规模科技人才评价中的不足，有望为高层次科技人才的引进和培养提供科学决策的理论依据，从而提高科技人才引进的效率并降低风险。

当前，新一轮产业变革正在兴起，以科技创新为核心的新技术革命深入推进，谁掌握了科技创新规律，谁就掌握了未来经济发展引擎。清华大学唐杰教授带头研发的“科技情报大数据挖掘与智能服务平台”（简称 AMiner，https://aminer.org/），为人才情报深入挖掘提供了可能。该平台以学术活动、科研人员、科技文献三大类数据为基础，构建三者之间的关联关系，深入分析挖掘，面向全球科研机构及相关工作人员，提供学者、论文文献等学术信息资源检索，以及如学术评价、专家搜索/推荐、学者地图、学者关系网络分析、研究兴趣发展变化分析、技术发展趋势分析等专业知识服务[61]。

2.3.7 科学计量

科学计量学是对科学活动过程与管理实施量化评估、刻画和预测的科学学分支学科，研究科学学中定量方面的问题。科学计量学用定量方法处理科学活动的投入，如科研人员、科研经费、论文数量、被引数量、信息传播和交流网络的形成等。

科学计量学的发展历程包括以下几个时期：

（1）萌芽时期（19 世纪下半叶到 20 世纪初），代表人物包括：德堪多（Alphose de Candolle）、高尔顿（Francis Galton）、科尔（F.J. Cole）；

（2）奠基时期（20 世纪初到 60 年代末），代表人物包括：洛特卡（A. J. Lotka）、布拉福德（S.C.Bradford）、加菲尔德（Eugene Garfied）；

（3）发展时期（20 世纪 70 年代后），前苏联学者纳利莫夫（V. V. Nalimov）和穆利钦科（A.M.Mulchenko）合著的第一部科学计量学著作《科学计量学：把科学作为情报过程来研究科学的发展》，标志着奠基时期基本完成，从此进入发展时期。

科学计量学研究的数据来源包括：科技图书、科技期刊、科技报告、会议文献、专利文献、标准文献、学位论文、产品资料、技术档案、科技报纸、光盘数据、网络数据。下面图表展示了科学计量学与文献计量学、情报计量学的区别与联系。

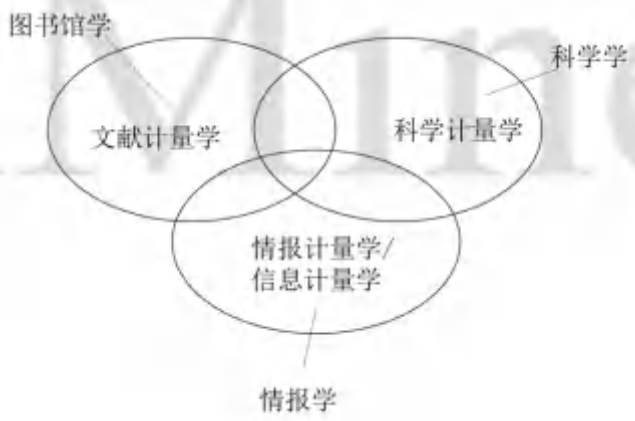


图 37 文献计量学、科学计量学和情报计量学（信息计量学）的联系与区别

表 8 科学计量学与文献计量学、信息计量学的关系

	含义	研究目的	研究对象	研究方法	不同点
文献计量学	1969 年，英国普理查德，把数学和统计学用于图书和其他通讯媒介物的一门科学	（1）探讨科学文献本身规律，提高文献情报管理科学性；（2）通过文献规律的探讨，进而揭示科学技术的规律性	科学文献及其数量	数学、统计学	侧重于文献情报管理和决策
科学	1969 年苏联弗纳里莫夫应	定量探讨科学发展的内在规律	情报本身	数学、统计学	侧重于科学管理和决策

计量学	用定量方法研究科学学的一门学科，或是以科学发展过程的各种定量方法总和				服务、文献指标、人才、经费和设备指标
情报计量学	1929 年，德国昂纳克：采用定量方法来描述和研究情报的现象和规律的一门学科	定量研究情报这一社会现象的动态规律	情报本身	数学、统计学	侧重于定量描述情报过程和规律
网络计量学	应用文献计量学、科学计量学以及信息技术、分析各种信息媒介、信息交流的一种方法	（1）为参与网络分析与定量描述的研究者提供科学交流工具；（2）提供网络研究与发展内容进展和分布的研究参考的来源；（3）测量科学交流活动、发展状况、规律及模式；（4）挖掘科学技术情报资源、促进社会、经济发展	网页中的文献计量学方法、万维网的电子数据库分析、主页、电子引文分析、电子媒介和资源研究、数字图书馆、虚拟图书馆、域名分布	数学、统计学、图论、数据分析软件、计算机科学等	侧重于网络这一媒介物

科技期刊是反映科学技术产出水平的窗口，一个国家科技水平的高低可通过期刊的状况得以反映。科技期刊的评价指标包括：（1）总被引用次数：所评价期刊理念发表的论文在评价当年被其他期刊和该期刊本身引用的总次数，以表明该期刊在科学交流中被使用的程度。（2）影响因子：可测度近年期刊的学术影响力。该项指标用论文平均被引率反映了期刊近期在科学发展和文献交流中所起的作用。公式：影响因子=某刊前两年发表的论文的总被引用次数/该刊前两年发表的论文的总数。科技论文评价指标：论文的类型、论文发表的期刊影响、文献发表的期刊的国际显示度、论文的基金资助情况、论文合著情况、论文的即年被引用情况、论文的合作者数、论文的参考文献数、论文的获奖情况等。

2.4 社交网络与图数据挖掘技术

本节以社交网络为例，介绍数据挖掘算法在实际数据上的具体应用。社交网络的主要结构形式是图，图数据不同于简单的连续性或离散型数据，其结点

之间的关系由于图的拓扑结构而变得复杂，其分析方法也不同与一般的统计和机器学习的数据分析。本节主要介绍一下社会网络中数据分析的基本技术，即社交网络的主要组织形式：图结构的度量算子和一些基本的算法。

2.4.1 图的度量算子

为了解决“社会网络中，谁是中心角色（具有影响力的用户），谁是志趣相投的用户，如何找到这些相似的个体”这类问题，需要量化用户中心性，用户相似度的度量方案。这些度量方案的输入信息通常是表征社会媒体交互信息的图结构。通过定义中心性度量方案，可以识别不同类型的中心结点。

（1） 中心性

中心性定义了网络中一个结点的重要性。这里主要介绍以下几个中心性度量，包括度中心性、特征向量中心性、PageRank 和中间中心性。

- 度中心性

在真实世界的交互中，一般认为具有很多连接关系的人是重要的，度中心性就是利用了这种思想。在无向图中，结点v_i的度中心性$C_d(v_i)$定义为：

$$C_d(v_i)=d_i$$

其中d_i是结点v_i的度。因此对于有更多连接关系的结点，度中心性度量方法认为他们具有更高的中心性。在有向图中，可以利用入度或者出度，也可以将两者结合作为度中心性值。为了使度中心性度量方法可以用来比较不同网络中的中心性值，通过使用度数之和来进行归一化，其中 m 是边的总数：

$$C_d^{sum}(v_i)=\frac{d_i}{\sum_j d_j}=\frac{d_i}{2m}$$

- 特征向量中心性

在度中心性度量中，我们认为具有较多连接的结点更重要。然而更进一步，拥有很多朋友并不能确保这个人就是重要的，拥有更多重要的朋友才能提供更

有力的信息。因此可以用邻居结点的重要性来概括本结点的重要性。设$C_e(v_i)$表示结点v_i的特征向量中心性，则其求解公式如下：

$$C_e(v_i)=\frac{1}{\lambda}\sum_{j=1}^{n}A_{j,i}C_e(v_j)$$

其中$\mathbf{A}$是临接矩阵,λ是某个固定的常数。这是个递归定义的函数，设$C_e=\left(C_e(v_1),C_e(v_2),\cdots,C_e(v_n)\right)^T$是所有结点的中心向量，上式可以重写为：

$$\lambda C_e=A^T C_e$$

其中C_e是邻接矩阵 A 的特征向量，λ是对应的特征值。

- PageRank

PageRank 度量方法被谷歌搜索引擎用来对网页的检索结果进行排序。网页以及它们之间的连接关系是一个巨大的网络，PageRank 为网络图中的所有结点（网页）定义了中心性度量。当用户在谷歌中检索时，和查询相匹配的且具有较高的的网页将最先显示。

$$C_p=\beta\left(I-\alpha A^T D^{-1}\right)\cdot I$$

其中$D=diag\left(d_1^{out},d_1^{out},\cdots d_n^{out}\right)$是一个关于度的对角矩阵，$\alpha,\beta$是两个常数。

- 中间中心性

另一种中心性度量方法是考虑结点在连接其他结点时所表现出的重要性，它是指网络图中某一结点与其他各点之间相间隔的程度，表示一个人在多大程度上是图中其他结点的“中介”，这类结点通常具有沟通桥梁的作用。其中一种方法是计算其他结点通过结点v_i的最短路径的数目，也就是说，我们在度量

结点v_i在连接结点 s 和结点 t 时所表现的重要性。这种度量方法称为中间中心性：

$$C_b(v_i)=\sum_{s\neq t\neq v_i}\frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

其中，σ_{st}是从结点 s 到结点 t 的最短路径数目，$\sigma_{st}(v_i)$是从 s 到 t 经过v_i的最短路径数目。为了对结点进行跨网络的比较，我们将$\sigma_b(v_i)$除以中间中心性的最大值（n-1）（n-2），得到归一化后的中间中心性：

$$C_b^{norm}=\frac{C_b(v_i)}{(n-1)(n-2)}$$

（2） 相似性

本节来讨论网络中两个结点间相似度的方法，在社会媒体中，这些结点可以表示关系网络中的个体或者是一些相关的事物。这些相关联的个体的相似度既可以基于它们所嵌入的网络，即结构相似度，也可以基于他们所产生的内容，即内容相似度来计算。采集网络信息时，结点间的相似度可以通过计算他们的结构等价性来获得。此外，SimRank 相似度也是一种流行的计算结点相似度的方法。

- 结构等价性

为了计算结构等价性，需要考虑两个结点间共有的邻接结点。从网络拓扑结构角度出发，考虑两个结点的共同邻居数，是基于网络半结构信息定义相似度的最简单的方法，共同邻居数越大，说明两个结点越相似。结点的相似度定义如下：

$$\sigma(v_i,v_j)=\left|N(v_i)\cap N(\upsilon_j)\right|$$

对于规模较大的网络，结点可能共享很多邻接结点，所以这个值会迅速增长。一般的，相似度被看做一个有界的值，通常是在[0,1]范围内。对此可以使用多种归一化方法，例如 Jaccard 相似度或者余弦相似度：

$$\sigma_{jaccard}(v_i, v_j)=\frac{|N(v_i)\cap N(v_j)|}{|N(v_i)\cup N(v_j)|}$$

$$\sigma_{cosine}(v_i, v_j)=\frac{|N(v_i)\cap N(v_j)|}{\sqrt{|N(v_i)||N(v_j)|}}$$

- SimRank 相似度和近似 SimRank 相似度

Simrank 算法的思路是“指向相似结点的结点也相似”：

$$S(a,b)=\frac{C}{|I(a)||I(b)|}\sum_{i=1}^{|I(a)||I(b)|}\sum_{j=1}^{}S\left(I_i(a),I_j(b)\right)$$

其中 C 为 0 到 1 之间的一个常数。时间复杂度是 $O\left(In^2d^2\right)$，其中I是迭代次数，n是网络中所有结点数，d是结点的平均度数。其效率太低，若考虑不再直接计算任意两个结点之间的 simrank 值，定义：

$$Partial_{I(a)}^{j}=\sum_{i\in I(a)}S(i,j),\ (\forall\in I(b))$$

那么有：

$$S(a,b)=\frac{C}{|I(a)||I(b)|}\sum_{j\in I(b)}Partial_{I(a)}^{j}$$

那么其算法的时间复杂度是 $O\left(In^2d\right)$，其中C在0~1之间。

为了更进一步加快相似度计算，可以使用基于随机游走的 SimRank 算法，是对结点相似度的一种较高精度的估计算法。

算法描述如下：

a）随机生成 R 条长度为 T 的路径 P，建立倒排索引 Index；

b）对结点 v，根据 Index 获得 v 所在的路径 P’；

c）对于 P’，遍历每条路径可以找到所有与 v 共同出现的结点。通过累加频次计算两个点之间的近邻相似度，得到 O’；

d）获得与结点 v 的相似度最大的 K 个结点的列表 L。

2.4.2 社交网络上的算法

这一小节主要介绍社交网络上的流行算法，主要包括网络上的行为分析算法，社区发现算法。社会网络中的聚类被称为社区发现，许多精心设计的高效算法可以很好地处理上亿用户的大规模网络。

（1） 行为分析算法

个体在社交网络中所表现出不同的行为，作为个体或者更大范围的群体行为的一部分。当讨论个体行为时，我们的关注点集中在一个个体。在个体行为分析方面，最典型的应用是用户行为的传播，比如在一个社交网络中，一个用户转发了一个信息，之后他的朋友看到这条信息就有可能转发或评论这条信息，那么这条消息传播的过程中是否转发的行为，就体现了用户行为的激活与否。

在一个网络中，对于一个结点 v 而言，有两个状态，激活和未激活。结点 v 的激活可能导致其邻居结点 w 的激活。接下来我们介绍两种影响力传播模型，包括线性阈值模型和独立级联模型来模拟影响在社会网络中的传播过程。

- 线性阈值模型

每个结点 v 都有一个随机的阈值 $\theta_v \sim U[0,1]$，表征当 v 的邻居结点被激活时，v 被激活的潜在趋势。结点 v 被它的每个邻居 w 所影响，根据权重 $b_{v,w}$，且有约束条件为：

$$\sum_{w} b_{v,w} \leqslant 1$$

若 v 的已激活邻居结点w_a的权重和大于等于其阈值时，v 将被激活，即：

$$\sum_{w_a} b_{v,w_a} \leqslant \theta_v$$

最开始，给定一个随机的阈值，和一个初始集合A_0,在步骤 t，所有步骤 t-1 的激活的结点仍是激活的，且通过上面的条件激活新的结点，直到不再有新的结点被激活为止。

- 独立级联模型

与线性阈值模型模型步骤基本相同，区别在不再通过结点 v 的所有已激活邻居的权重和达到阈值来激活，而是对每个已激活结点 v,都有单个机会来激活其未激活邻居结点w，其激活概率为$P_{v,w}$，不管是否成功，在之后的步骤汇总，v 都不能再次激活其他结点。

从另一个角度来考虑，在公司推广产品时，如何利用有限的资金和资源找到那些影响力较大的用户来推广产品，通过“口碑效应”和“病毒式营销”的推广方式使得新产品的影响达到最大化也是一个热点问题。影响力最大化的目的是，对于一个参数 K，确定如何选择一个具有 K 个结点的集合A_0，使其具有最大的传播影响力(A)。

- 链接预测问题

社交网络上的行为导致新链接的产生，比如交朋友的行为，它可以链接到用户；比如购买行为，可以链接到实体；如加入行为，可以链接到社区。因此，我们可以将这样的行为形式化为链接预测问题。

设$G(V_{train}, E_{train})$为训练的图，则链接预测生成一个$V_{train} \times V_{train} - E_{train}$中最有可能的边的排序列表。列表中第一条边是算法认为最有可能很快出现在

图中的边。链接预测算法给$V_{train} \times V_{train} - E_{train}$中的每一条边$e(x,y)$赋一个分值$\sigma(x,y)$,通过对各个边的分值进行递减排序，就能得到顶点预测结果的排序列表。$\sigma(x,y)$的预测可以通过不同的方法得到，因而计算两个结点之间相似度的任何方法都可以用于链接预测。

（2） 社区发现算法

社区发现是网络研究中的重要课题，吸引了众多研究者的关注。给定一个表征网络的图数据，社区往往指代不同集合的结点，其中同一社区的结点之间的连通性往往高于不同社区间的结点。例如在社会网络中，一个社区可以表征在一起上学、工作、生活的人们。在这里主要介绍 Girvan-Newman 算法，标签传播算法以及 Louvain 算法。

- Girvan-Newman 算法

直观来看，在社区内部结点之间相互连接的边密度较大，因此，通过边来识别社区是一种较为直观的社区发现算法。Girvan-Newman 算法即在该启示下发展而言，如果去除社区之间连接的边，留下的就是社区。对于社区而言，较先去除的边，中心性较低，而中间中心性则较大。因此，逐步去除中间中心性最大的边，直至结束。

Girvan-Newman 算法的详细步骤为：

a) 计算网络中所有边的中间中心性；

b) 去除中间中心性最高的边；

c) 重新计算去除边后的网络中所有边的中间中心性；

d) 跳至步骤 b，重新计算，直至网络中没有边存在。

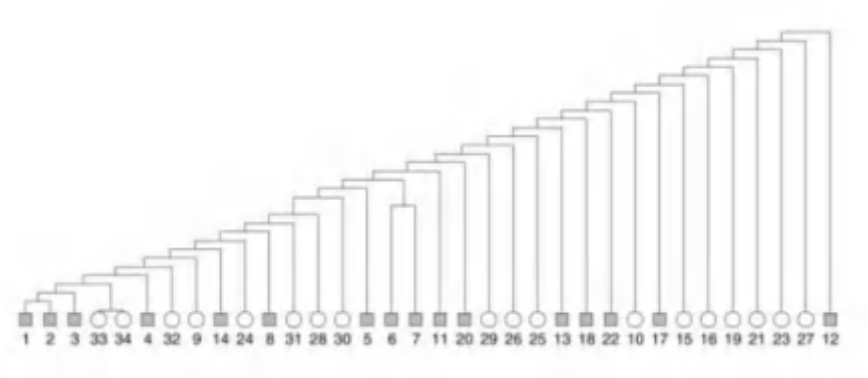


图 38 Girvan-Newman 算法结果

Girvan-Newman 算法所得到的结果实质上是网络中结点的树图，如图 38 所示。该算法给出了如何去除边得到社区结构。但在得到最终的社区数之前，还有一个问题没有得到解决，即如何确定合适的社区数，使社区划分结果最优。他们随后提出了模块度 Q 的概念：

$$Q=\sum_{i}^{K}\left(e_{ii}-a_{i}^{2}\right)^{2}$$

考虑一个将网络分成 K 个社区的分割算法，定义一个 k×k 矩阵 e，其中 e_{ij} 代表连接到社区 i 中的点的边的个数与社区 j 中的边数之比，而 $a_i=\sum_j^K e_{ij}$。

Q 值能够体现网络划分为社区后社区结构的质量，该值越逼近于 1，说明社区结构越明显，该值逼近于 0，则社区结构不明显。对于同一个网络而言，不同算法可能得到的 Q 值不同，Q 值高则代表了该算法较优。

利用 Q 值寻找合适的社区数的思路如下：在上面 Girvan-Newman 算法中每去除一次边，则计算一下所得社区结构的 Q 值，寻找到 Q 值最大时的社区数量。一般而言，计算时不可能会在所有去边过程中都计算 Q 值，往往是寻找某一区间的 Q 值，取得局部最大值即可。在下面的图 39 中可以看到，当社区数量为 4 时，所得 Q 值最大，因此该网络划分为 4 个社区最优。

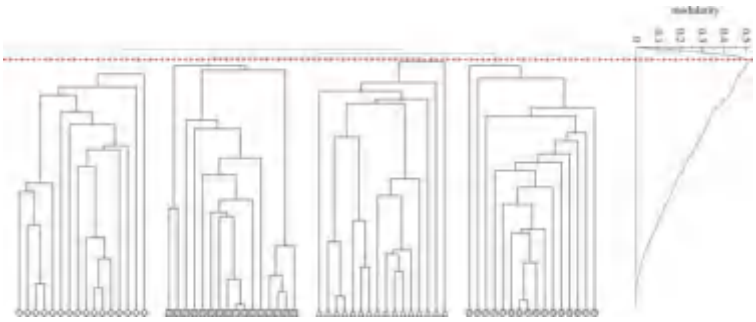


图 39 基于优化 Q 值的算法结果

- 标签传播算法

标签传播算法很简单，其核心思想是相似的结点应该具有相同的标签。

标签传播算法的详细步骤为：

a）初始化所有结点，为每个结点分配唯一的标签；

b）随机选择一个结点，令其及其所属社区的结点标签更换为它的大多数邻居所属的标签，若有几个这样的标签，随机选择一个；

c）若每个结点标签都已和其多数邻居的相同，则停止算法，否则重复步骤 2。

标签传播算法最吸引人的地方，在于它的简单核心思路的简洁，并且很容易实现。一个简单的算法意味着易于理解，并且容易在它之上做很多有针对性的改良。它的另一个极大的优点在于可伸缩性，标签传播算法非常适合用来处理大规模图数据，因为算法的实质是以结点为中心的，所以其实是可以在 Map-Reduce 上实现它。标签传播算法不需要预先给定社区的数量，基于聚类的算法一般都需要预先给定聚类个数，而标签传播不需要任何的先验知识。

- Louvain 算法

Louvain 算法是基于模块度（Modularity）的社区发现算法，该算法在效率和效果上都表现比较好，并且能够发现层次性的社区结构，其优化的目标是最大化整个图属性结构（社区网络）的模块度 Q：

Louvain 算法的详细步骤为：

a）对于每个结点，令其所属社区的结点标签更换为它的其中一个邻居所属的标签，选择方式是选择使其 Q 增益最大，直到所有结点的社区标签都不再改变，计算 Q；

b）若 Q 不再改变，则停止算法，否则重构网络，令同一个社区中的点“凝聚”成一个结点，社区内的边成为新网络的自环，社区之的边成为新网络的边，这时边的权重为两个结点内所有原始结点的边权重之和，重复步骤 b。

迭代这两个步骤直至算法稳定，它的执行流程如图 40 所示：

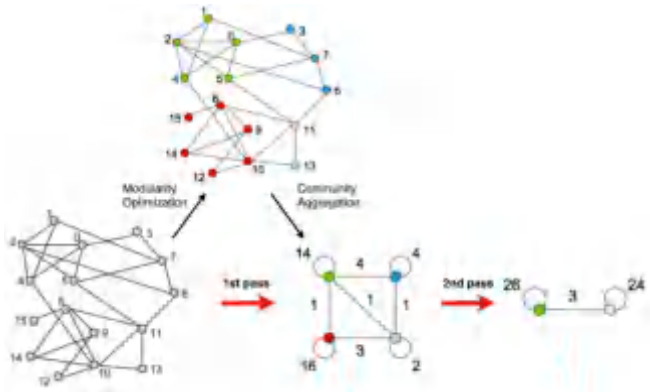


图 40 Louvain 算法步骤

2.5 自然语言数据挖掘技术

自然语言处理是人工智能的一个重要领域，也是数据挖掘的一个重要应用载体，本节以自然语言为例介绍与自然语言相关的数据挖掘算法和应用。语言是人类区别于动物的根本标志，因此自然语言处理体现了人工智能的最高任务与境界。至今，自然语言处理的发展与真正的语义理解仍然差之甚远，但这并不妨碍自然语言数据的研究，如果采取有效的分析方法，我们仍然可以从中获得知识来帮助我们的各项任务。本节将从词、句、话题三个层次简介自然语言中数据分析的基本方法。

2.5.1 词表示分析

文本中的单词是自然语言的基本结构，对于单词的研究，除了简单的词频统计等，词的表示学习饱受关注。词的表示学习，又称为词嵌入（word embedding），指为每个单词找到一个向量表示，理想状况下向量之间的距离和

线性关系可以反映单词之间的语义联系。通过词向量，我们可以通过可视化分析词的关联，也有利于进一步分析。下面介绍三种最为重要的词表示方法：

（1） 词袋模型

词袋模型（Bag of Words）是最简单的词向量表示方法。该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。词袋模型使用一组无序的单词来表达一段文字或一个文档。基于文本的词袋模型的一个简单例子如下：

首先给出两个简单的文本如下：

The more the data, the better the performance of machine learning algorithms.

Which one is more important in machine learning, data or algorithms?

基于上述两个文档中出现的单词，构建如下一个词典：

{“the”：1,“more”：2,“data”：3,“better”：4,“performance”：5,”of”：6,“machine”：7,“learning”：8,”algorithms”：9,“which”：10,“one”：11,“is”：12,“important”：13,“in”：14,“or”：15}

上面的词典中包含 10 个单词,每个单词有唯一的索引,那么每个文本我们可以使用一个 10 维的向量来表示，如下：

[4, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]

[0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]

该向量与原来文本中单词出现的顺序没有关系，而是词典中每个单词在文本中出现的频率。词的这个表示方法很容易想到，但是却又不少缺点，如频率与文本长度有关，导致不同长度文本难以比较；词的重要性不突出，例如“algorithm”可以表明这很可能是计算机科学相关的文本，而“the”“is”这种词却不能。

（2） tf-idf 模型

TF-IDF（Term Frequency-Inverse Document Frequency）是一种用于信息检索常用加权表示法。在一段文本中，词频（Term Frequency，TF）指的是某一个给定的词语在该文件中出现的频率。这个数字是对词数的归一化，以防止它偏向长的文件（同一个词语在长文件里可能会比短文件有更高的词数，而不管该词语重要与否）。对于在某一特定文件里的词语t_i来说，它的重要性可表示为：

$$f_{ij}=\frac{n_{ij}}{\sum_k n_{kj}}$$

上式n_ij是该词t_i在文件d_j中出现的次数，而分母是在文件d_j中所有字词出现的次数之和。

逆向文件频率（Inverse Document Frequency，IDF）是一个词语普遍重要性的度量。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：

$$idf_i=\log（\frac{|D|}{|\{j：t_i\in d_j\}|}）$$

其中$|D|$为语料库中文件的总数，$|\{j：t_i\in d_j\}|$为语料库中包含t_i的总数。之后可以得出词语的 tf-idf 为：

$$tfidf_{ij}=tf_{ij}\cdot idf_i$$

（3） Word2Vec 模型

词袋模型和 tf-idf 都是 20 世纪出现的稀疏向量词表示方法，但是这种方法具有很大的弊端，首先就是当语料库中的词语数量很多时，会受到“维度灾难”现象的影响，导致很多分析不准确；另外这种表示方法忽略了语义关联，也就失去了进一步分析语义的可能性。

为了解决这些缺陷，Hinton 等人提出了词语的分布式表示（Distributed Representation），将单词转换为低维度连续空间的向量。这种分布式表示方法可以将语义信息融合到词向量中，根据两个词向量之间的距离就可以判断两个单词语义相关程度。

Word2Vec 模型就是由谷歌提出的一类高效训练词语分布式表示的模型，其在神经网络语言模型（Neural Network Language Model，NNLM）基础上进行了改进。该模型通过不断地“阅读”文本，并且“拉近”文本中相邻较近的模型对应的词向量。

下面我们具体地介绍最常用的word2vec模型，Skip-Gram模型。Skip-Gram模型分为输入层、投影层和输出层三部分，如图 41 Skip-Gram模型结构所示。我们称单词的“上下文”是该单词在句子中附近小窗口内的其他单词。模型是跟据当前词w_t预测其上下文w_{t-2}，w_{t-1}，w_{t+1}，w_{t+2}出现的概率。

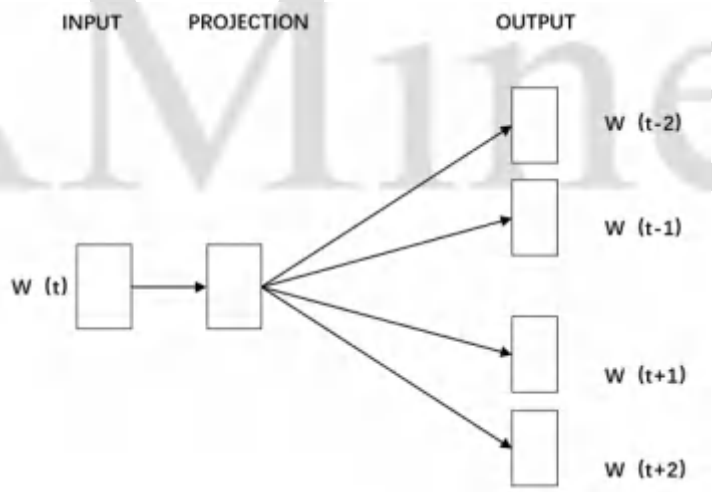


图 41 Skip-Gram模型结构

给定一个长度为T的单词序列$w_1, w_2, \cdots, w_T$，Skip-Gram 模型的目标函数如下：

$$\text{argmax}\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leqslant j\leqslant c, c\neq 0}\log p\left(w_{t+j}\middle|w_t\right)$$

其中，c表示上下文窗口的大小，在有充足的训练数据时，增大窗口c的大小一般可以获得更高的准确率，不过对应的训练时间也会加长。Skip-Gram 模型往往使用 Softmax 函数来计算$p\left(w_{t+j}\middle|w_t\right)$：

$$p\left(w_0\middle|w_I\right)=\frac{\exp\left(v'^{T}_{w_0}v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left(v'^{T}_{w_0}v_{w_I}\right)}$$

上式中v_w和v'_w分别为单词w的输入和输出的词向量表示，W表示整个语料库中所有单词的数目。通过 2.2.2 监督学习方法中神经网络一节介绍的梯度下降的方法优化目标函数，不过由于每次计算条件概率需要 Softmax，枚举每个词典中的每个词，需要的时间代价非常大（与W的大小成正比），在实际中应用非常困难。

为了解决这个问题，可以使用层次化 Softmax（Hierarchical Softmax）来计算$p\left(w_0\middle|w_I\right)$，降低计算复杂度。层次化的 Softmax 利用哈夫曼编码的方式构建了一个包含W个单词的哈夫曼树，并用其叶子节点来作为输出层。

通过使用层次化的 Softmax 可以把$p\left(w_0|w_I\right)$的计算复杂度降低到$O\left(L(w)\right)$，算法的平均复杂度即变成了$O\left(\log(w)\right)$，大幅提升了时间效率。除了层次化的 Softmax 之外，还可以使用负采样（Negative Sampling）的方法对模型做出优化，负采样的目标函数定义如下：

$$\log\sigma\left(v'^{T}_{w_0}v_{w_I}\right)+\sum_{i=1}^{k}E_{w_i\sim P_n(w)}\left[\log\sigma\left(-v'^{T}_{w_0}v_{w_I}\right)\right]$$

用此公式代替原 skip-gram 中目标函数的每一个$\log p\left(w_0|w_I\right)$项，这样优化目标即转变为使用逻辑回归从 k 个噪声分布为$P_n(w)$的负样本中找出最优的目标单词w_0。

通过使用层次化的 Softmax 和负采样两种技术，可以大幅降低 Word2Vec 的时间复杂度，提升词向量训练效率。

2.5.2 语言模型

在实际应用中，可能需要解决这样一类问题：如何计算一个句子的概率？例如在中文输入法中，若打出“nixianzaiganshenme”这一字符串时，其目的是打出“你现在干什么”，但是“你西安在干什么”也是合乎音字转换的句子，从语义上判断,显然 P（你现在干什么|nixianzaiganshenme）应该大于 P（你西安在干什么|nixianzaiganshenme）。

形式化地，对于一个含有 n 个词的句子S=$\{w_1, w_2, \cdots, w_n\}$,语言模型可以计算产生 S 的概率，即：

$$P(S)=P(w_1, w_2, \cdots w_n)=P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, w_2, \cdots, w_{n-1})$$

在估计模型参数时，通过采用极大似然估计，我们知道计算$P(w_i|w_1, w_2, \cdots, w_{i-1})$的方法是对子串$\{w_1, w_2, \cdots, w_{i-1}\}$和$\{w_1, w_2, \cdots, w_i\}$在语料库中统计其出现的次数$C(w_1, w_2, \cdots, w_{i-1})$和$C(w_1, w_2, \cdots, w_i)$，并做除法：

$$P(w_i|w_1, w_2, \cdots, w_{i-1})=\frac{P(w_1, w_2, \cdots, w_{i-1}, w_i)}{P(w_1, w_2, \cdots, w_{i-1})}=\frac{C(w_1, w_2, \cdots, w_{i-1}, w_i)}{C(w_1, w_2, \cdots, w_{i-1})}$$

但是这里有两个重要的问题：（1）对很多子串来说，在语料库中根本没有出现，其条件概率值为 0，因此数据稀疏问题很严重；（2）要计算一个新句子的概率，需要若干项条件概率，而要计算并存储它们有很大的代价，无法实用。

通常，我们会放宽标准，使用马尔科夫假设（Markov Assumption）来计算 S 的近似概率：下一个词的出现仅依赖于它前面的一个或几个词。具体来说，若假设每个词出现的独立于其他词，即依赖于前面的 0 个词，则为 unigram 模型：

$$P(S)=P(w_1)P(w_2)P(w_3)\cdots P(w_n)$$

若假设下一个词的出现依赖它前面的一个词，则为 bigram 模型：

$$P(S)=P(w_1)P(w_2|w_1)P(w_3|w_2)\cdots P(w_n|w_{n-1})$$

假设下一个词的出现依赖它前面的两个词，即 trigram 模型：

$$P(S)=P(w_1)P(w_2|w_1)P(w_3|w_1,w_2)\cdots P(w_n|w_{n-2},w_{n-1})$$

以此类推，下一个词的出现依赖前面 n-1 个词的模型被称为 n-gram 模型。n 越大的话，对下一个词出现的约束信息更多，模型具有更大的辨别能力。n 越小的话，在训练语料库中出现的次数更多，具有更可靠的统计信息。在实际应用中，由于计算资源的限制，往往使用 bigram 或 trigram。n-gram 参数估计的方法也是极大似然估计。

2.5.3 话题模型

分析文本的话题是一种重要的数据分析手段，通过区分微博中不同话题文本量，可以了解社会热点；通过联系话题和常见词，可以加深对词的理解，优化词向量学习；文本话题对于构建用户肖像、优化推荐系统等任务也至关重要。

本世纪前十年，以隐含狄利克雷分布（LDA）为代表的概率生成模型在话题挖掘方法取得了突破性的进展，这类模型也被直接冠以“话题模型”（topic model）的称号。LDA 是一个描述离散数据（这里以文本数据为例来阐述，每个单词可以看作一个随机变量，在所有的单词词典里取值）的生成模型。它假定文本中的每个单词是由一些混合的“话题”产生的，每个“话题”都有一定的权重，即$p(w)=\sum p(w|z)p(z)$，而$p(z)$又由一个狄利克雷（Dirichlet）分布产生。LDA 是概率隐含语义分析（PLSA）的延伸。

提出这些模型的动机都是为了分析文本的潜在结构，也就是“话题”。这些话题能够捕获文本的语义信息，帮助找到文本里意思相同或相近的单词，这使得信息检索时能够找到和查询中的单词语义相关的文档而不只是包含查询的单词的文档。我们定义超参数 α 和 β，生成单词的过程可以使用一个概率图模型表示出来，如表 9 与图 42。

表 9 LDA 模型中的变量和标记

标记	含义
M	语料库中的文档个数 Number of documents in the corpus
K	潜在的话题个数 Number of latent topics
V	词典中单词个数 Number of unique words in vocabulary
α	话题先验分布的参数 Hyper-parameter on the topic （K-vector or scalar if symmetric）
β	单词先验分布的参数 Hyper-parameter on the word （V-vector or scalar if symmetric）
θ	文档生成话题的概率 Notation for p（z|d）, document-topic distribution, size M * K
Φ	话题生成单词的概率 Notation for p（w|z）, word-topic distribution, size K * V
N	文档中的单词个数 Number of words in one document
z	文档中每个单词的话题 Topic assigned for each word in each document
w	文档中的单词 Word in each document

图 42 中，箭头表示生成的前提，超参数 α 和 β 分别可以确定分布 θ=p（z|d）和 Φ=p（w|z），分布 θ 进行采样可以得到话题 z，已知 z 和条件分布 Φ 则可以采样生成单词。只考虑话题的话，我们可以抛弃掉单词之间的顺序，认为每个单词都是通过如上采样过程产生的。

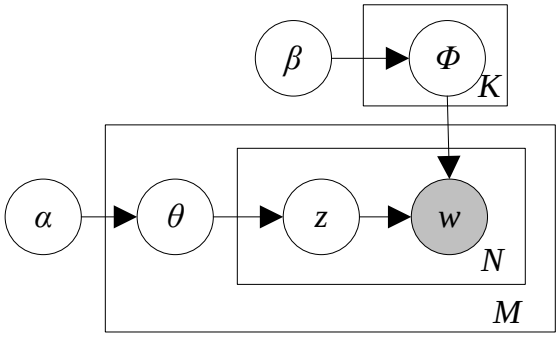


图 42 话题模型的概率图

2.6 多媒体数据挖掘技术

SMAC（社交媒体，移动，分析和云）技术范式的快速发展和采用，已经产生了大量以人为中心的实时、多模式、异构的数据，比如文本、音频、图像、视频等。近年来，多媒体数据挖掘作为一种基于技术的解决方案已经引起了广泛关注。从社交网络应用程序/服务中分析，建模和发现知识的挖掘机会不仅限

于基于文本的大数据，而且还扩展到部分未知的复杂图像，音频和视频结构。多媒体数据挖掘是一个交叉学科领域，涉及图像处理和理解、计算机视觉、数据挖掘和模式识别。表 10 总结了多媒体数据挖掘技术的 SWOT 矩阵[62]。

表 10 多媒体数据挖掘的 SWOT 分析表

优势	**劣势**
•提升业务绩效 •目标用户/客户 •快速准确的分析 •快速准确的分析 •改善战略决策 •客户/用户满意度	•错误结论的风险 •客户隐私受到威胁 •缺乏专业人员 •数据集减少可能导致错误分析
机遇	**威胁**
•检测谣言、欺凌、虚假新闻等 •教育与卫生部门 •欺诈识别	•隐私权 •身份盗用 •费用 •社会认同

多媒体数据挖掘技术由于需要访问大量的个人信息，在数据隐私和安全方面面临着大量挑战，但是在环境安全、教育卫生等领域具有积极促进作用。为了充分发挥多媒体数据挖掘技术的作用，需要针对不同种类的数据使用不同种类的技术。广泛来说，根据数据性质的差异可将数据分为四种类型：文本数据、音频数据、图像数据、视频数据。

2.6.1 文本挖掘

多媒体文本数据挖掘（MTDM），就是从大量的非结构化多媒体文本数据中发现有意义的模式的过程。对多媒体文本数据挖掘最行之有效的途径就是将多媒体文本数据结构化后，再对结构化数据采用数据挖掘方法。多媒体文本数据挖掘的过程如图 43 所示。

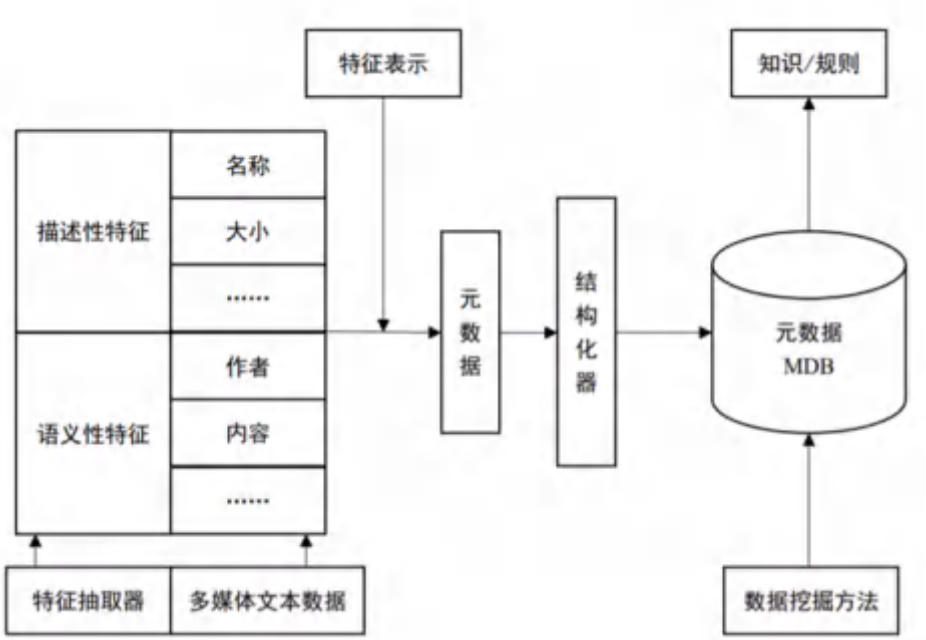


图 43 多媒体文本数据挖掘的过程

文本挖掘算法通常会用到监督、半监督、无监督的机器学习技术。监督机器学习技术包括支持向量机、隐马尔可夫模型、决策树、朴素贝叶斯、k-最近邻算法，半监督机器学习技术的例子包括 bootstrapping、snowball 等，聚类算法是无监督机器学习技术的例子[62]。典型的文本挖掘方法包括自动文摘、文本分类、文本聚类、关联分析等。其中，自动文摘是指从文档中抽取关键信息，用简洁的形式对文档内容进行摘要或解释。这样，用户不需要浏览全文就可以了解文档或文档集合的总体内容。文本分类是指按照预先定义的主题类别为文档集合中的每个文档确定一个类别。这样，用户不但能够方便地浏览文档，而且可以通过限制搜索范围来使文档的查找更为容易。文本聚类与文本分类的不同之处在于，聚类没有预先定义好的主题类别，它的目标是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能地大，而不同簇间的相似度尽可能地小。关联分析是指从文档集合中找出不同词语之间的关系。比如，研究从大量文档中发现一对词语出现的模式的算法，并用于在 Web 上寻找作者和书名的出现模式，有助于从网站上发现新书籍[63]。

2.6.2 音频挖掘

音频挖掘通过分析音频形式的数据（如图 44），以从这些音频信号中获取所需的信息。应用场景包括：在医疗保健中，正确分析婴儿的声音可以告诉他/她的健康状况；在呼叫中心，进行音频/语音分析以提高其服务质量。音频挖掘技术中比较常见的是语音识别和语音合成。

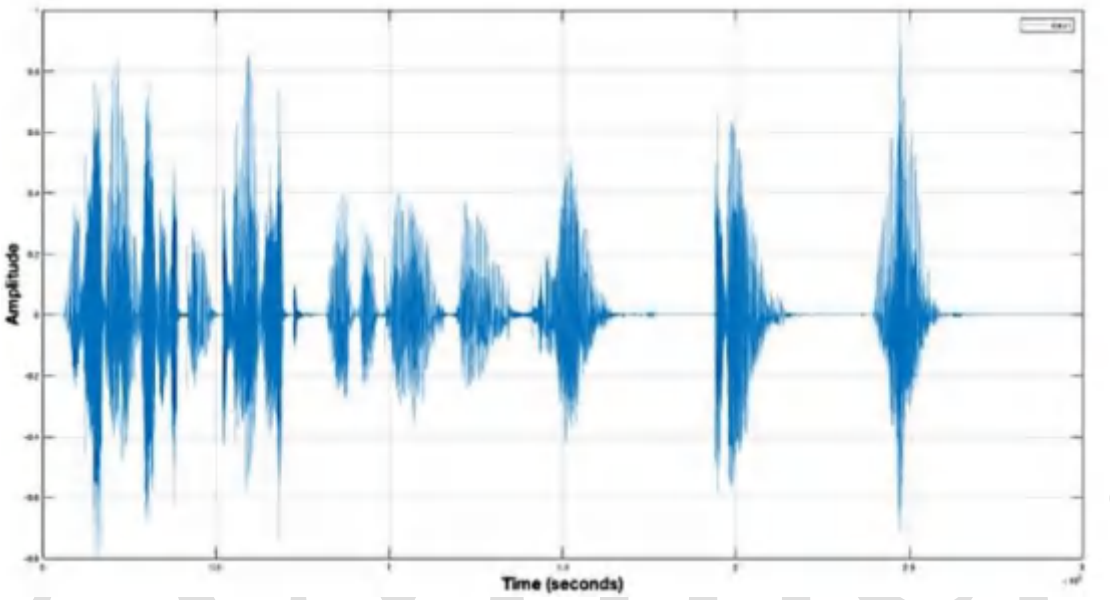


图 44 音频波形图

语音识别技术是让机器通过识别和理解把语音信号转变为相应的文本或命令的技术，主要包括特征提取技术、模式匹配准则及模型训练技术三个方面，所涉及的领域有信号处理、模式识别、概率论和信息论、发声机理和听觉机理、人工智能等。根据识别的对象不同，语音识别任务大体可分为 3 类：即孤立词识别、关键词识别、和连续语音识别。语音识别的应用领域非常广泛，常见的应用系统有语音输入系统、语音控制系统、智能对话查询系统等[64]。

语音合成，又称文语转换（Text to Speech）技术，是通过机械的、电子的方法产生人造语音的技术，涉及声学、语言学、数字信号处理、计算机科学等多个学科。语音合成技术是实现人机语音通信，建立一个有听和讲能力的口语系统所必需的一项关键技术，解决的主要问题就是如何将文字信息转化为可听的声音信息[65]。

2.6.3 图像挖掘

如今，大多数可用的非结构化数据都是图像或视频形式，因此提出有效的图像挖掘技术变得非常重要。图像挖掘是从包含图像的数据集中提取有意义的信息，主要应用包括面部识别、运动分析等。为了适应图像数据的高速产生，需要研究高效的技术和基础架构，以更好地、准确地分析图像形式的数据。图像数据挖掘的一个十分关键的问题是图像数据本身的表示问题，这也是图像处理和模式识别的关键问题。一般而言，可以用颜色、纹理、形状和运动向量等基本特征来表示图像的基本特征。高级概念可以看成是一种特征模式，比如，河流可以认为是具有某种颜色特征的长条形；大片庄稼区可以认为是具有某种颜色分布和纹理特征的大片图像区域。高级概念是人们所关心的，它可能是某种物体的存在、某种现象的发生等。底层的基本特征与高层概念之间必然存在着某种映射关系，这种关系可以用数据挖掘的方法来发现。这样，图像数据挖掘的基本过程可以用图 45 所示的图来表示[66]。

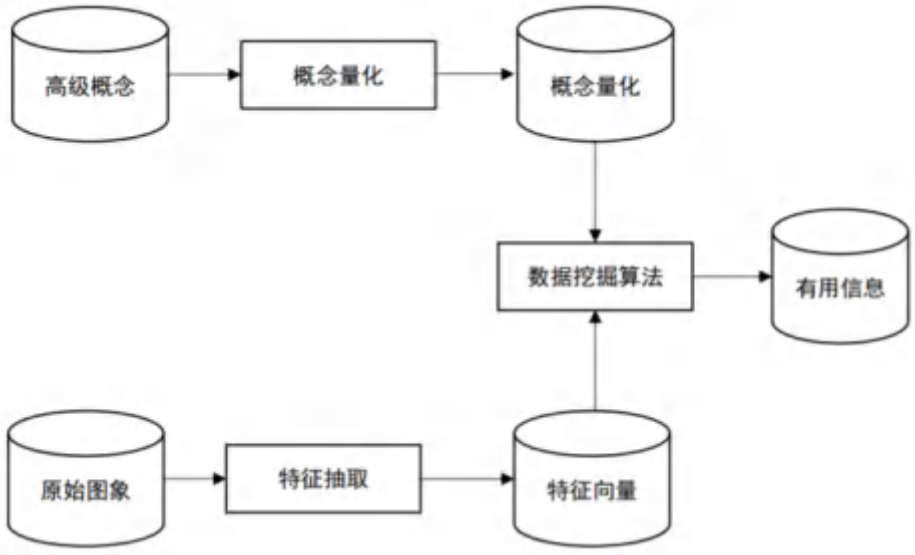


图 45 图像数据挖掘的基本过程

2.6.4 视频挖掘

随着电子技术的快速发展以及 internet 的应用和普及，视频数据急剧增加，在海量的视频数据中寻找感兴趣的目标或信息成为现实生活中人们急需解决的问题。视频信息作为现代社会一种表达信息的主要载体,有着它自己独立的结构。

一般来说，一段视频由一些描述独立故事单元的场景构成；一个场景由一些语义相关的镜头组成；而每个镜头是由一些连续的视频帧构成，它可以由一个或多个关键帧表示。一段视频的典型结构如图 46 所示。

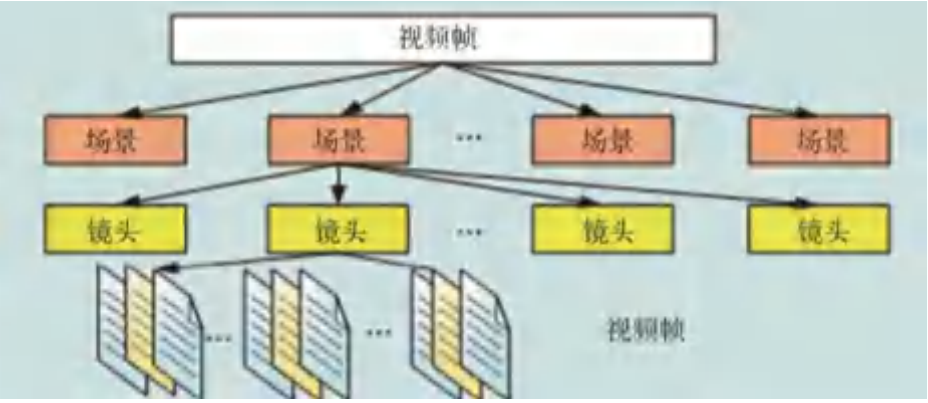


图 46 典型视频结构图

传统的基于文本的检索方式无法克服其人工标注的主观性和工作量异常庞大的缺点，因此研究自动的、智能化程度高的视频检索与挖掘技术成为当前科技工作者们研究的热点。基于内容的视频检索与挖掘（Content-Based Video Retrievaland Ming），就是根据视频的内容和上下文关系等低层特征，对大规模视频数据库中的视频数据进行分析与检索。它是在没有人工参与的情况下，自动提取并描述视频的特征和内容，设计可靠有效的检索算法、系统结构以及友好的人机界面。图 47 表示了基于内容的视频检索与挖掘的处理过程，视频首先被分割成场景，场景接着被分成镜头，由于镜头是由一系列语义上相关和逻辑上独立的帧组成的，且这些帧存在大量的时间冗余，为了表示与分析的方便，通过选择一个镜头中的具有代表性的若干帧作为关键帧来表示一个镜头。对得到的几个关键帧进行特征提取，依提取的特征进行相似比较可进行视频检索与挖掘[67]。

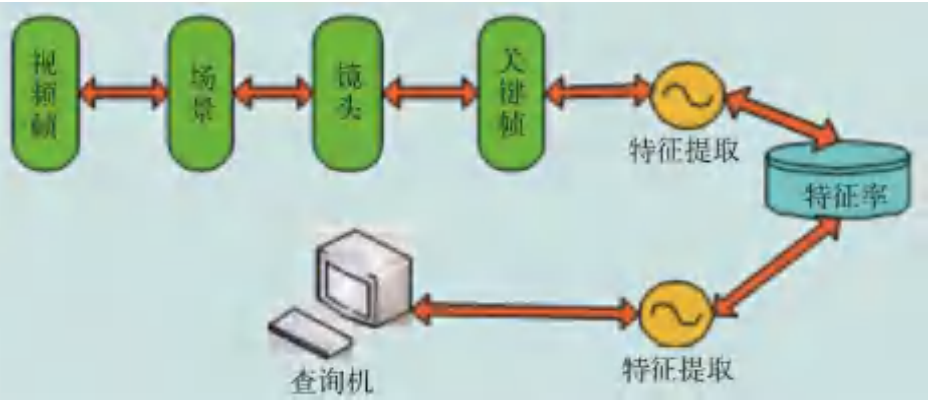


图 47 基于内容的视频检索与挖掘结构图

2.7 大规模数据挖掘技术

“大数据”一词今天对研究人员和实践者都非常重要。它被定义为“需要成本效益高、创新形式的信息处理以增强视野和决策能力的高容量、高速度和高多样性的信息资产”。大数据用于表示数据集的集合，这些数据集太大、太复杂，无法使用传统的数据处理工具和应用程序进行处理。在可接受的时间限制内，用于捕获、管理、处理数据的通用软件工具无法处理数据集的大小。

表 11 大数据的特征

特征	说明
容量 （数据的规模）	它决定了流入的数据量，这些数据可以被存储和生成进一步。视情况而定根据存储的数据量，决定它是否属于“大数据”的范畴
多样性 （数据的不同形式）	指定了不同类型和各种数据源，包括结构化和非结构化数据。例如文件、电子邮件、图像、视频、音频等。
流量 （流数据分析）	它是处理移动中数据的速度及其分析的一个方面，其中的内容流被假定为连续和一致的。除了考虑输入的速度外，它还考虑从中产生有用信息的便利性
可信度 （数据的不确定性）	大数据的这一特性涉及可靠性和数据分析是否准确的首要问题，因此最终会产生可靠和高质量的解决方案
可变性 （数据的上下文含义）	它是指存储的信息不一致。简单地说，它处理与数据相关联的快速变化和可替代的含义
可视化 （呈现数据的方式）	可视化恰好成为大数据的一个重要特征，因为所有被存储的数据都需要以易于阅读和理解的方式进行分类和查看
价值 （使用数据的成本）	它涉及检索数据有用性的实践。人们认为，原始数据本身根本就没有价值。在分析中，“价值”特性处理的是如何将数据转化为知识和信息。

大数据，顾名思义，表示大规模数据集合（如表 11 所示），由于海量性、复杂性、多样性等特征，难以使用传统的数据处理工具和应用程序进行处理，

需要探索新的挖掘技术，从大量复杂的数据源中发现隐藏的信息。随着 Web 技术的发展和数据的大量产生，对大数据的讨论是不可避免的。以下是主要的大数据类型：

（1） 社交网络（人类来源的信息）：这类数据主要来自社交媒体平台，如 Facebook、博客文章和评论、个人文档、Instagram、Flickr 等的图片、YouTube 的视频、电子邮件、网络搜索结果和手机内容短信等。

（2） 传统业务系统（流程中介数据）：这类数据主要来自如客户注册、产品制造、下单和接单等都由这些系统记录和监控产生的业务事件，以及由公共组织生成的数据，如医疗记录和商业交易数据、银行和股票记录、电子商务和信用卡/借记卡等。

（3） 物联网（机器生成数据）：这类数据主要由大量传感器和机器生成，数据结构良好。例如，来自家庭自动化、天气/污染传感器、交通传感器/网络摄像头和安全/监控视频/图像等固定传感器的数据、移动传感器（跟踪）的数据（如移动电话位置）、汽车和卫星图像以及来自计算机系统（如日志和网络日志）的数据等。

2.7.1 大数据平台架构

从技术架构角度，大数据处理平台可划分为 4 个层次：数据采集层、数据存储层、数据处理层和服务封装层，详见图 48。

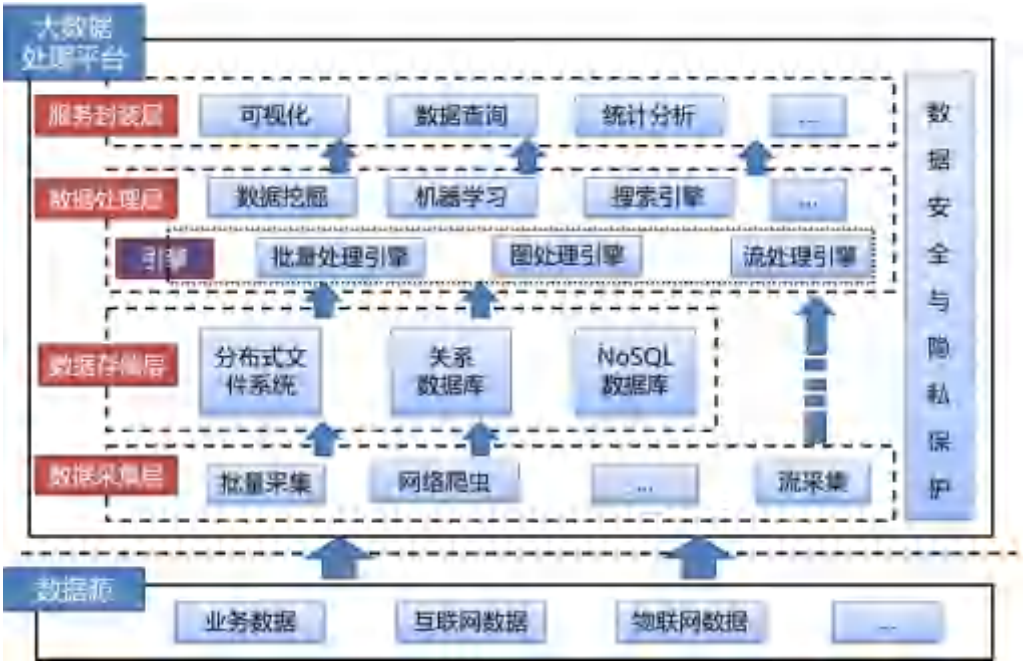


图 48 大数据处理平台技术架构图

（1）数据采集层。主要负责从各种不同的数据源采集数据。常见的数据源包括：业务数据、互联网数据、物联网数据等。对于不同的数据源，通常需要不同的采集方法。对于存储在业务系统中的数据，一般采用批量采集的方法，一次性地导入到大数据存储系统中。对于互联网上的数据，一般通过网络爬虫进行爬取。对于物联网产生的实时数据，一般采用流采集的方式，并存储到大数据存储系统中或是直接发送到流式处理系统进行处理分析。

（2）数据存储层。主要负责大数据的存储和管理工作。大数据处理平台中的原始数据通常存放在分布式文件系统（例如：HDFS）或是云存储系统（例如：Amazon S3）。为了便于对大数据进行访问和处理，大数据处理平台通常会采用一些非关系型（NoSQL）数据库对数据进行组织和管理。针对不同的数据形式和处理要求，可以选用不同类型的非关系型数据库。常见的非关系型数据库有键值（Key-Value）存储数据库（例如：Redis）、列存储数据库（例如：HBase）、文档型数据库（例如：MongoDB）、图形（Graph）数据库（例如：Neo4J）等。

（3）数据处理层。主要负责对大数据的处理和分析工作。针对不同类型的数据，一般需要不同的处理引擎。对于批量的静态数据，一般采用批量处理引擎（例如：MapReduce）。对于动态的流式数据，一般采用流处理引擎（例如：Storm）。对于图数据，一般采用图处理引擎（例如：Giraph）。基于处理引擎

提供的各种基础性的数据计算和处理功能，大数据处理平台中通常会有一些提供复杂数据处理和分析的工具，例如：数据挖掘工具、机器学习工具、搜索引擎等。

（4）服务封装层。主要负责根据不同的用户需要对各种大数据处理和分析功能进行封装并对外提供服务。常见的大数据相关服务包括：数据的可视化、数据查询分析、数据的统计分析等。

除此之外，大数据处理平台一般还包括数据安全和隐式保护模块，这一模块贯穿大数据处理平台的各个层次。

2.7.2 大数据平台实例

基于上述技术架构，我们可以设计并实现一个基于开源系统的大数据处理平台，如图 49。该平台能够支持对批量数据、流式数据和图数据等不同类型大数据的处理和分析。根据具体的应用场景和需求，可以对该开源平台进行裁剪，例如：如果不需要对图数据进行处理，可以裁剪掉相应的模块和子系统（Giraph 和 GraphX）。下面将参照下图的技术架构设计，具体介绍平台各层中的开源系统。

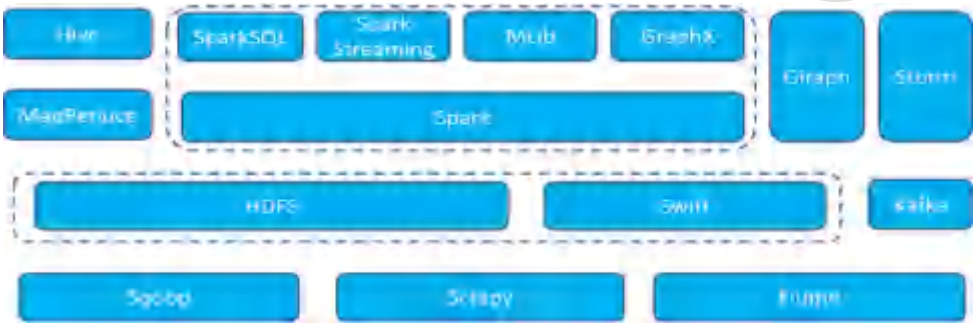


图 49 基于开源系统的大数据处理平台架构

（1） 数据采集系统：

Sqoop 是一个用于在 Hadoop 与关系型数据库间进行数据转移的开源工具，可以将一个关系型数据库（例如：MySQL,Oracle,Postgres 等）中的数据导进到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导出到关系型数据库中。

Scrapy 是一个基于 Python 语言开发的开源 Web 并行爬取框架，它能够快速爬取 Web 站点并从页面中提取自定义的结构化数据。Scrapy 使用 Twisted 异步

网络库来处理网络通讯，用户只需要在 Scrapy 框架的基础上进行模块的定制开发就可以轻松实现一个高效的爬虫应用。

Flume 是一个高可用的、高可靠的、分布式的海量日志采集、聚合和传输的开源系统，主要作用是数据的收集和传输，支持多种不同的输入输出数据源，并提供对数据的简单处理。

（2） 数据存储系统：

HDFS（Hadoop Distributed File System）是参考 Google 的 GFS（Google File System）实现的一个开源分布式文件系统，是 Hadoop 框架中的一个核心模块，具有高容错性、高吞吐量等特点。

Swift（OpenStack Object Storage）是 OpenStack 开源云计算项目的子项目之一，是 OpenStack 云存储服务的重要组件。Swift 能在比较便宜的标准硬件存储基础设施之上提供高可用、分布式、持久性、大文件的对象存储服务。

Kafka 是一种分布式的、基于发布/订阅的消息系统，其功能类似于消息对列。Kafka 可以接收生产者（例如 Webservice、文件、HDFS、HBase 等）的数据，并将其缓存起来，然后发送给消费者（例如 Storm、Spark Streaming、HDFS 等），进而起到缓冲和适配的作用。

（3） 计算引擎：

MapReduce 是开源大数据处理框架 Hadoop 的核心计算引擎，它是 Google MapReduce 的开源实现，主要用于对批量数据的处理。

Storm 是 Twitter 支持开发的一款分布式的、开源的、实时的大数据流式计算系统。Storm 能够快速可靠地处理源源不断的消息，并具有良好的容错机制。

Giraph 是一个采用 BSP 模型（Bulk Synchronous Parallel，整体同步并行计算模型）的开源并行图处理系统，主要参照 Google 的 Pregel 系统并基于 Hadoop 实现。

Spark 是一个专为大规模数据处理而设计的快速、通用的计算引擎，是 UC Berkeley AMP 实验室开源的类 Hadoop MapReduce 的通用并行框架。Spark 扩展

了 MapReduce 计算模型，高效地支持除了批量计算以外的更多计算模式，包括流式计算、图计算等。

（4） 数据分析工具：

Hive 是基于 Hadoop 的一个数据仓库工具。所有 Hive 的数据都存储在 Hadoop 兼容的文件系统（例如，HDFS、Amazon S3）中。Hive 提供了一系列的工具，可以用来进行数据提取、转化、加载（ETL）以及通过类 SQL 查询语言 HiveQL 进行统计分析和查询工作。Hive 将 SQL 语句转换为 MapReduce 任务进行运行。

Spark SQL 是基于 Spark 的一个数据仓库工具，其架构和功能与 Hive 类似，只是把底层的 MapReduce 替换为 Spark。

Spark Streaming 是 Spark 提供的对实时数据进行流式计算的工具，支持对实时数据流的可扩展（scalable）、高吞吐（high-throughput）、容错（fault-tolerant）的流处理。支持从多种数据源获取数据，包括 Kafka、ZeroMQ、Kinesis 以及 TCP sockets。从数据源获取数据之后，可以使用诸如 map、reduce、join 和 window 等高级函数进行复杂算法的处理。

MLib 是一个基于 Spark 的机器学习函数库，它是专门为在分布式集群上并行运行的情况而设计的。MLlib 中包含许多常用的机器学习算法，可以在 Spark 支持的所有编程语言中使用。

GraphX 是一个基于 Spark 的分布式图（Graph）处理工具，提供大量进行图计算和图挖掘的简洁易用的接口，极大地方面了用户对分布式图处理的需求。

2.8 数据隐私保护和安全

2.8.1 数据隐私保护

随着云计算、物联网和社交媒体技术的快速发展，数据量的快速增加，大数据挖掘和分析成为未来知识发现的重要手段，与此同时，数据隐私泄露问题日趋严重。比如，政府和商业组织收集掌握了大量的用户数据，其中包括个人信用信息、健康状况、财务状况和个人偏好，供一些社交网络、银行业务和医

疗保健系统使用，来建模和预测与人类有关的现象，例如犯罪、流行病和社会物理学中的重大挑战。如何保护用户隐私和防止敏感信息泄露成为面临的最大挑战。由于大数据具有规模大、多样性、动态更新速度快等特点，许多传统的隐私保护技术不再适用。

隐私保护数据挖掘（Privacy Protection Data Mining, PPDM）提供了在不泄露隐私信息的情况下使用数据挖掘方法的可能性。大数据下的隐私保护数据挖掘技术主要关注以下两个方面：（1）是如何对原始数据集进行加密和匿名化操作，实现敏感数据的保护；（2）是探究新的数据知识产权保护模式，限制对敏感知识的挖掘。在隐私数据整个生命周期过程中，主要涉及数据收集、数据转换、数据挖掘分析和模式评估四个阶段，包括隐私保护数据属性、各种参与者角色和各种数据化操作，它们之间的关系如图 50 所示[68]。

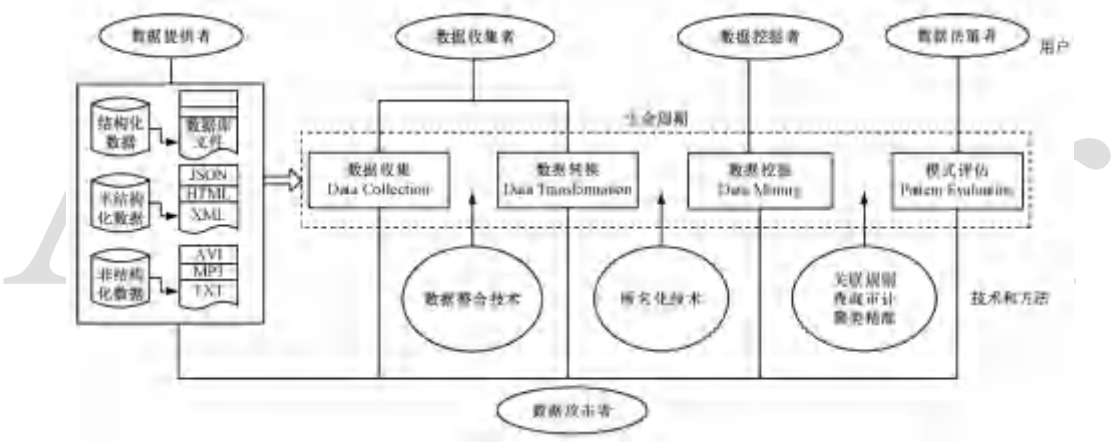


图 50 隐私保护数据挖掘生命周期模型

数据挖掘的隐私保护技术主要包括输入隐私和输出隐私，在输入隐私保护方面，主要涉及收集数据的匿名化。匿名是一组数据对象集中每一单独个体无法被识别的状态。数据匿名性主要针对数据的标记属性、准标记属性或关键属性和敏感属性。在输出隐私方面，为了保护隐私，会干扰或审核结果。

（1） 数据挖掘的输入隐私保护技术

数据挖掘的输入隐私保护的方法主要包括：K－匿名方法、L－多样性方法、T 闭合方法、差分隐私保护方法等。

- K－匿名方法

K－匿名方法（K－anonymity）由 Sweeney 最早提出[69]，它的基本思想是对原始数据进行更加概括、抽象的描述，通过隐匿技术处理后，发布精度较低的数据，要求每个等价类至少包含K条记录，使得同一个准标识符至少有K条记录，使观察者无法通过准标识符连接记录。每条记录与其他K－1条记录的准标志属性不可区分，致使数据挖掘者不能唯一识别隐私信息所属的个体，从而实现用户的隐私保护。K－匿名方法虽然解决了链接攻击问题，但是对数据的敏感属性没有作任何约束，K－匿名性的主要缺点是防止属性泄露，Homayoun[70]通过P敏感匿名解决了此问题。

- L-多样性方法

攻击者可通过敏感数据与个体进行关联分析，实现对个体隐私的背景知识性攻击和匿名性的同质性攻击[71]。L－多样性方法（L-diversity）是对K－匿名方法的扩展，如果对敏感属性至少存在一个有很好代表性的多样化值，则等效类被称为具有L－多样性。如果表的每个等价类均具有L－多样性，则该表被称为L－多样性。等价类E的熵（Entropy）定义如式（1）。

$$\mathrm{Entropy}(E)=-\sum_{s\in S}p(E,s)\log p(E,s) \quad (1)$$

公式（1）中，S是敏感属性的域，而 p(E,s)是E中具有敏感值 s 的记录的一部分。如果对于每个等价类E，Entropy(E)≥log l，则表示具有熵－多样性。如果缺乏多样性，K匿名数据集容易被攻击。

- T 闭合方法。

鉴于匿名的局限性，当由总体数据分布创建的偏斜度攻击偏斜时，满足多样性不会阻止披露。文献[72]提出了T－闭合（T－closeness）方法，认为如果一个类别中的敏感属性的分布与整个表中该属性的分布之间的距离不超过阈值 t，则认为等价类具有 t 紧密性。T－闭合方法是L－多样性的增强概念，适用于发布数据集的敏感属性分布要尽可能贴近整个数据集的敏感属性分布。针对属性值分布不规则、属性值范围很小或者已被分类的数据集，为防止概率性推导，要求任何等价类中敏感属性的分布与整个数据集中相应属性的分布之间

的距离小于阈值 t 。 T －闭合的局限性是要求任何等价类中敏感属性的分布接近整个数据敏感属性的分布。

- 差分隐私保护方法。

差分隐私（differential privacy）的概念最早由 Dwork[73]提出，是密码学中的一种手段，保护的是数据源中一点微小的改动导致的隐私泄露问题。其基本思想通过添加噪声的方法，将一些随机噪声添加到数据库中，确保删除或者添加一个数据集中的记录并不会影响分析的结果，以此来减少基于数据关联性的攻击。差异隐私的目的是提高统计数据库查询的准确性，并减少识别个人记录的可能性。差分隐私提供了有用的数据访问权限，同时使用 Laplace 噪音和指数机制防止了对单个记录的隐私侵害。数据分析师查询数据库，但数据库没有直接回答查询，而是由中间隐私卫士提供响应。隐私卫士分析分析师给出的查询，并从数据库中获取结果，并对结果增加噪音，并以一些嘈杂的响应回复分析师。差异隐私由于可以保证数据隐私，添加的噪音极少，非常接近原始结果，因此正成为一个流行的研究领域。

（2） 数据挖掘的输出隐私保护技术

数据挖掘的输出隐私保护技术主要包括：关联规则、查询审计、分类和聚类等。

- 关联规则的隐私保护

关联规则的隐私保护主要有变换和隐藏两类方法：变换方法主要是修改支持敏感规则的数据，并通过对规则的支持度和置信度小于一定阈值来隐藏规则；隐藏方法不会修改敏感规则的数据，而是隐藏会生成敏感规则的频繁项目集。这两类方法都对非敏感规则的挖掘具有一定的负面影响[74]。

- 数据查询审计技术

在云存储环境中，用户将失去对存储在云服务器上的数据的控制。如果云服务提供商不受信任，则它可能会篡改并丢弃数据，但会向用户声明数据是完整的。数据查询常采用云存储审计技术，即数据所有者或第三方组织对云中的

数据完整性进行审核，从而确保数据不会被云服务提供商篡改和丢弃，并且在审核期间不会泄露用户的隐私[75]。

- 分类结果的隐私保护

分类方法会降低敏感信息的分类准确性，并且通常不会影响其他应用程序的性能。分类结果可以帮助发现数据集中的隐私敏感信息，因此敏感的分类结果信息需要受到保护。决策树分类是建立分类系统的重要数据挖掘方法。在保护隐私的数据挖掘中，挑战是从被扰动的数据中开发出决策树，该决策树提供了一种非常接近原始分布的新颖重构过程[76]。

- 聚类结果的隐私保护

与分类结果的隐私保护类似，保护聚类的隐私敏感结果也是当前研究的重要内容之一。Vaidya[77]等人提出了一种分布式 K-means 聚类方法，该方法专门面向不同站点上存有同一实体集合的不同属性的情况。使用此聚类方法，每个站点可以学习对每个实体进行聚类，但在学习过程中并不会获知其他站点上所存属性的相关信息，从而在信息处理的过程中保障了数据隐私。

随着人工智能和深度学习的兴起，大数据时代数据挖掘与隐私保护之间的技术博弈将成为常态，保护用户隐私将成为人工智能发展的关键。由于多源数据挖掘技术本身的局限性，基于隐私保护的相关研究还处于起步阶段，大数据的种种特性给数据挖掘中的隐私保护提出了不少难题和挑战。人们需要改进数据挖掘的隐私保护方法，并建立新的隐私保护框架和机制。

2.8.2 数据安全

随着人工智能、云计算、移动互联网和物联网等技术的融合发展，以大数据为代表的数据化、数字化是全球信息技术发展趋势之一。大数据 5V 的特性和新的技术架构颠覆了传统的数据管理方式，在数据来源、数据处理使用和数据思维等方面带来革命性的变化，这给大数据安全防护带来了严峻的挑战。大数据的安全不仅是大数据平台的安全，而是以数据为核心，围绕数据全生命周期的安全。数据在全生命周期各阶段流转过程中，在数据采集汇聚、数据存储处理、数据共享使用等方面都面临新的安全挑战。

大数据的安全技术体系是支撑大数据安全管理、安全运行的技术保障。以“密码基础设施、认证基础设施、可信服务管理、密钥管理设施、安全监测预警”五大安全基础设施服务，结合大数据、人工智能和分布式计算存储能力，解决传统安全解决方案中数据离散、单点计算能力不足、信息孤岛和无法联动的问题。大数据的总体安全技术框架如图 51 所示[78]。

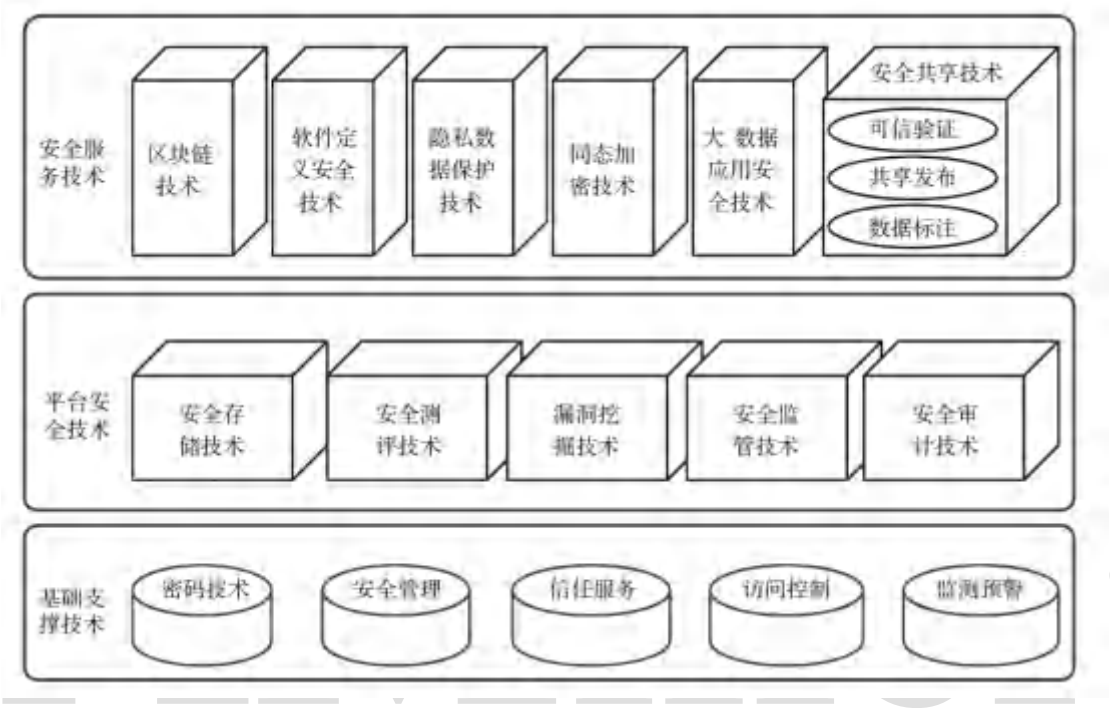


图 51 大数据安全技术框架

大数据环境下数据的安全成为防护的核心，新的安全威胁如数据泄漏、数据隐私保护、数据机密性和完整性保护、恶意内部人员、高级持续性威胁（APT）、数据丢失、数据滥用和恶意使用、数据共享等都面临着新的技术挑战。智能、便捷、高效、透明成为数据安全防护的必然需求。数据是流动的价值，需要在大数据安全中树立数据全生命周期管控理念，执行多维度防护措施。在云端数据和应用安全防护方面，要针对云端应用特点提供不同粒度防护。

2.9 数据挖掘论文主题分析

为了展示数据挖掘领域的研究主题分布情况，旨在为充分了解该领域学术发展进程和学术研究现状提供参考，本章节针对 AMiner 平台上收录的 SIGKDD 在 2010-2020 年期间的 2733 篇论文，采用文档主题生成模型（Latent Dirichlet Allocation, LDA），分析这些论文的研究主题分布情况。

LDA 假定文本中的每个词由一些混合的主题产生的，每个主题都有一定的权重，即$p(w)=\sum_z p(w|z)p(z)$，其中p（z）又是一个Dirichlet 分布产生。LDA 的贝叶斯网络结构如图 52 所示，图中 K 为主题个数，M 是论文总数，N 是某个论文中单词总数，α和β分别是每个主题下词的多项式分布和每个论文下主题的多项分布的 Dirichlet 先验参数。LDA 模型中有一组隐含变量z，参数求解采用吉布斯采样，构建 Markov 链，逼近目标概率分布。获取参数后可以计算论文的主题关键词，$p(w|d)=\sum_z p(w|z)p(z|d)$。

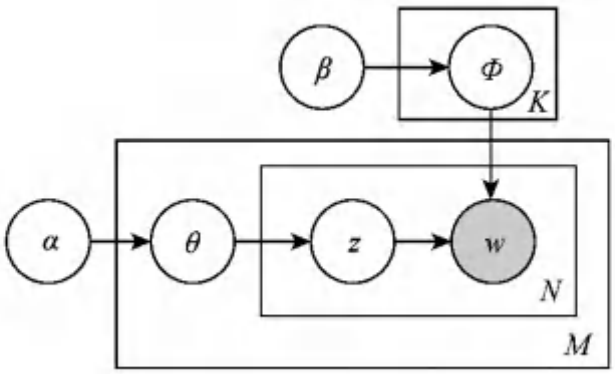


图 52 LDA 结构图

本报告设置主题数 K=5，通过对论文的标题和摘要进行分析，获取了这些论文的研究主题包括：数据挖掘（Data Mining）、社交网络（Social Network）、机器学习（Machine Learning）、推荐系统（Recommender System）、大数据（Big Data）。表 12 展示了每个主题下具有代表性的 3 篇论文及作者信息。

表 12 数据挖掘领域论文主题分布

研究主题	相关论文
数据挖掘	标题：Searching and mining trillions of time series subsequences under dynamic time warping 作者：Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista 出版信息：SIGKDD, 2012
	标题：Active Data Mining 作者：Rakesh Agrawal, Giuseppe Psaila 出版信息：SIGKDD, 2017
	标题：Trajectory pattern mining 作者：Fosca Giannotti, Mirco Nanni, Fabio Pinelli, Dino Pedreschi 出版信息：SIGKDD, 2017

社交网络	标题：Scalable influence maximization for prevalent viral marketing in large-scale social networks 作者：Wei Chen, Chi Wang, Yajun Wang 出版信息：SIGKDD, 2010
	标题：Friendship and mobility: user movement in location-based social networks 作者：Eunjoon Cho, Seth A. Myers, Jure Leskovec 出版信息：SIGKDD, 2011
	标题：COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency 作者：Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei 出版信息：SIGKDD, 2015
机器学习	标题：DeepWalk: online learning of social representations 作者：Bryan Perozzi, Rami Al-Rfou', Steven Skiena 出版信息：SIGKDD, 2014
	标题：XGBoost: A Scalable Tree Boosting System 作者：Tianqi Chen, Carlos Guestrin 出版信息：SIGKDD, 2016
	标题：Structural Deep Network Embedding 作者：DAIXIN WANG, Peng Cui, Wenwu Zhu 出版信息：SIGKDD, 2016
推荐系统	标题：Collaborative topic modeling for recommending scientific articles 作者：Chong Wang, David M. Blei 出版信息：SIGKDD, 2011
	标题：Collaborative Deep Learning for Recommender Systems 作者：Hao Wang, Naiyan Wang, Dit-Yan Yeung 出版信息：SIGKDD, 2015
	标题：Graph Convolutional Neural Networks for Web-Scale Recommender Systems 作者：Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai 出版信息：SIGKDD, 2018
大数据	标题：Understanding urban human activity and mobility patterns using large-scale location-based data from online social media 作者：Samiul Hasan, Xianyuan Zhan, Satish V. Ukkusuri 出版信息：SIGKDD, 2013
	标题：Forecasting Fine-Grained Air Quality Based on Big Data 作者：Yu Zheng, Xiuwen Yi, Ming Li, Ruiyuan Li 出版信息：SIGKDD, 2015
	标题：Petuum: A New Platform for Distributed Machine Learning on Big Data 作者：Eric P. Xing, Qirong Ho, Wei Dai, Jin Kyu Kim 出版信息：SIGKDD, 2015

2.10 数据挖掘经典论文概况

国际知识发现与数据挖掘大会（ACM SIGKDD Conference on Knowledge Discovery and Data Mining，简称 SIGKDD）是数据挖掘领域的顶级国际会议，由 ACM 的数据挖掘及知识发现专委会负责协调筹办。会议内容涵盖数据挖掘的基础理论、算法和实际应用，SIGKDD 主会期间，除了学术研究论文，自 2010 年起还设有面向工业和政府应用的专题研讨会，以及工业应用博览的邀请报告会。

SIGKDD 发展的历史可以追溯到 1989 年，一系列关于知识发现及数据挖掘的研讨会从那时开始组织。自 1995 年以来，SIGKDD 以大会的形式连续举办了 24 届，论文的投稿量和参会人数呈现出逐年累增的趋势。由于 SIGKDD 的学科交叉性和广泛应用性，吸引了来自统计、社会网络分析、机器学习、大数据挖掘、数据库、万维网、生物信息学、多媒体、自然语言处理、人机交互及高性能计算等众多领域的学者。

SIGKDD 每年的大会都会吸引大量的研究界和工业界的投稿。图 53 和图 54 分别给出了 KDD 近几年研究性论文（Research Track）和工业界论文（Industrial Track，最近改为 Applied Data Science Track）的投稿和录用情况。总的来说研究性论文投稿相对比较稳定，录用率也一直在 14%-20%之间。而工业界论文投稿量近年呈现明显快速增长趋势，尤其 2019 和 2020 年达到 700 篇和 756 篇投稿。但是投稿录用率稍有下降，2015 年录用率最高，约为 34%；2020 年的录用率最低，约为 16%。这与近年来深度学习、人工智能、大数据等相关算法在工业界大量应用密不可分。

图 53 2013-2020 KDD 研究性论文投稿与接收情况

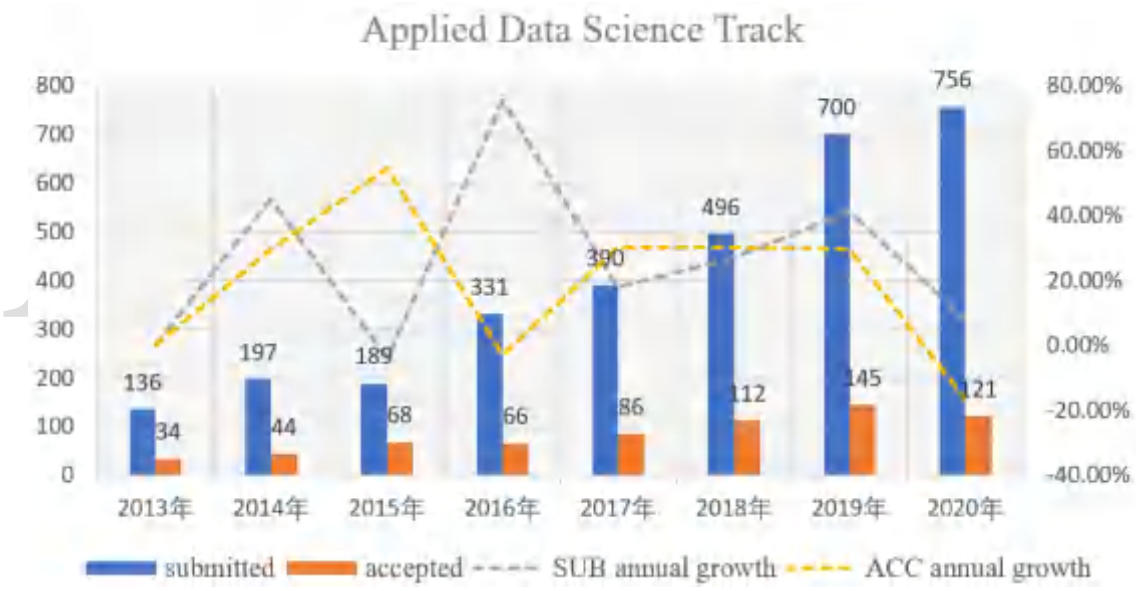


图 54 2013-2020 KDD 工业界论文投稿与接收情况

下面就近几年 SIGKDD 大会分别做一个简单概述，以及对代表性论文进行解读，相信读者能够对整个 SIGKDD 乃至数据挖掘领域有个更深入的了解。

2.10.1 SIGKDD 2013

2013 年 8 月 11-14 日，第 19 届知识发现与数据挖掘大会（SIGKDD 2013）在美国芝加哥召开，此次大会的主题是大数据挖掘。

SIGKDD 2013 大会的主席是前通用汽车高级研究经理 Ramasamy Uthurusamy 博士和伊利诺伊芝加哥大学的 Robert L. Grossman 教授，研究性程序委员会主席由来自德克萨斯奥斯丁大学的 Inderjit S. Dhillon 教授和 Google 的 Yehuda Koren 博士担任，另外有 50 名高级程序委员会委员和 300 名程序委员负责论文评审。吸引了来自全球 50 多个国家 1200 多人参加。

SIGKDD 2013 大会收到来自 50 多个国家的 726 篇论文投稿，每篇文章由至少 3 名审稿人评审，然后由相应领域的高级程序委员协调讨论并推荐，最后由程序主席决定是否接收。最终录用论文 125 篇（录用率约 17.2%），其中大会报告论文 66 篇（仅 9.1%）、张贴报告论文（Poster Presentation）59 篇。

SIGKDD 2013 大会邀请了微软的 Raghu Ramakrishnan、斯坦福的 Andrew Ng、威斯康辛的 Stephen J. Wright 以及 Google 的 Hal Varian 四位在产业界与学术界均产生重要影响的专家进行大会主题报告。

SIGKDD 2013 的具体获奖情况如下：

- 最佳论文

✧ 最佳研究型论文

标题：Simple and Deterministic Matrix Sketching（简易及确定性矩阵草图）

作者：Edo Liberty

论文解读：该论文研究的问题是针对给定的矩阵 A，如何能找到一个更小的压缩矩阵 B 对原矩阵 A 进行估计，这是基于社交网络大数据分析中的一个关键问题。

地址：

https://www.aminer.cn/archive/simple-and-deterministic-matrix-sketching/53e9b500b7602d970402e5e6

标题：Querying Discriminative and Representative Samples for Batch Mode Active Learning（批量模式主动学习的查询判别样本和代表性样本）

作者：Zheng Wang, Jieping Ye

论文解读：本文将 ERM 原则推广到主动学习环境中，开发了一种实用的批量模式主动学习方法，解决了在主动学习中如何能够挑选同时具有代表性又具有区分性的样本。

地址：

https://www.aminer.cn/archive/querying-discriminative-and-representative-samples-for-batch-mode-active-learning/53e9bba7b7602d97047f32a6

标题：A space efficient streaming algorithm for triangle counting using the birthday paradox（一种利用生日悖论的空间高效三角形计数流算法）

作者：Madhav Jha, C. Seshadhri, Ali Pinar

论文解读：该论文设计了一种节省空间的算法，该算法近似于传递性（全局聚类系数）和总三角形计数，只有一次通过作为边缘流给出的图形。模型基于经典的概率结果，即生日悖论，作为单通道流式算法，通过存储极小部分的边缘来维持图的三角形的传递性/数量的实时估计。

地址：

https://www.aminer.cn/archive/a-space-efficient-streaming-algorithm-for-triangle-counting-using-the-birthday-paradox/53e9999eb7602d97021e514e

标题：Recursive Regularization for Large-scale Classification with Hierarchical and Graphical Dependencies（具有层次结构和图形依赖性的大规模分类的递归正则化）

作者：Siddharth Gopal, Yiming Yang

论文解读：分层分类中的两个关键挑战是利用类标签之间的分层依赖关系来提高性能，并同时保持跨大型分层结构的可伸缩性。论文提出了用于大规模

层次分类的正则化框架，可以解决这两个问题。具体来说，将类标签之间的层次依赖关系合并到参数的正则化结构中，从而鼓励层次结构中附近的类共享相似的模型参数。此外，论文还将方法扩展到以下情况：类标签之间的依赖关系以图形而不是层次结构的形式进行编码。为了进行大规模培训，研究者开发了一种并行迭代优化方案，该方案可以处理具有数十万个类和数百万个实例的数据集，并可以学习数 TB 的参数。实验表明，与其他竞争方法相比，该方法得到了持续改进，并在基准数据集上获得了最新的结果。

地址：

https://www.aminer.cn/pub/53e9aac3b7602d97034430d3/recursive-regularization-for-large-scale-classification-with-hierarchical-and-graphical-dependencies

✧ 最佳应用型论文

标题：Amplifying the Voice of Youth in Africa via Text Analytics（通过文本分析放大非洲年轻人的心声）

作者：Prem Melville 及 UNICEF 的多名研究人员

论文解读：U-report 是由联合国儿童基金会乌干达运营的开源短信平台，旨在让社区成员就影响他们的问题发表意见。鉴于数据流的数量和速度很高，对所有消息进行手动检查已不再可持续。文章介绍了 IBM 在 UNICEF 部署的自动消息理解和路由系统，利用数据挖掘方面的最新进展，以充分利用带标签的培训数据，同时结合专家的领域知识，讨论了在实际部署中应用此类技术时的权衡，设计选择和挑战。

地址：

https://www.aminer.cn/pub/53e9b09fb7602d9703b07333/amplifying-the-voice-of-youth-in-africa-via-text-analytics

标题：Query Clustering based on Bid Landscape for Sponsored Search Auction Optimization（基于竞价格局的查询聚类对赞助搜索竞价优化）

作者：微软公司以 Ye Chen 为代表的一众人

论文解读：文章提出了聚类概率分布的形式化方法，并将其应用于查询聚类。首先导出用于聚类高斯密度的 k 均值变体，然后开发了一种用于对高斯混合密度进行聚类的算法，该算法可以对单个高斯进行泛化，并且通常是针对现实世界数据的更现实的参数假设。推导了一种变分 EM 算法，该算法将整个集群内 KL 散度的上限最小化。聚类算法已成功部署到生产中，极大地提高了收入，并提高了现有生产系统的点击率。

地址：

https://www.aminer.cn/pub/53e9af2cb7602d9703964e4f/query-clustering-based-on-bid-landscape-for-sponsored-search-auction-optimization

✧ 最佳博士论文

标题：Mining Heterogeneous Information Networks（挖掘异构信息网络）

作者：Yizhou Sun 博士，其导师为韩家炜（Jiawei Han）教授

论文解读：论文从许多将互连数据视为同构图或网络的现有网络模型中，研究了挖掘异构信息网络的原理和方法，半结构化异构信息网络模型利用了类型化节点和链接的丰富语义。这种半结构化异构网络建模为挖掘互联数据提供了一系列新原则和强大的方法。

地址：

https://www.aminer.cn/archive/mining-heterogeneous-information-networks-a-structural-analysis-approach/53e9ae3cb7602d9703851185

标题：Machine Learning in Health Informatics: Making Better use of Domain Experts（健康信息学中的机器学习：更好地利用领域专家）

作者：Byron Wallace 博士，其导师为 Carla Brodley 教授

论文解读：论文集中关注了临床信息学和具有大量信息的典型案例，发现因为临床信息学内在特性的原因，解决相关问题需要专业领域的知识，而该领域现有的机器学习技术普遍性能表现较弱。该论文目的旨在使用新奇的机器学习与数据挖掘方法，让现实世界中的学习系统更加有效率。

地址：

https://www.aminer.cn/archive/machine-learning-in-health-informatics-making-better-use-of-domain-experts/56d890a4dabfae2eeee4ca4d

✧ SIGKDD 创新和服务大奖

创新贡献奖：Jon Kleinberg（康奈尔大学）

2013 年创新奖（Innovation Award）由康奈尔大学的 Jon Kleinberg 教授获得，以表彰他在社交网络和信息网络分析中的传递行为以及用户个体行为建模方面的工作。Jon Kleinberg 是社会网络分析方向的知名学者，三十余岁时即当选为美国科学院、工程院、艺术与科学院院士，代表算法是 HITS。

SIGKDD 创新奖是知识发现与数据挖掘领域（KDD）的最高荣誉，授予对这一领域做出重大技术贡献的研究人员。根据评审要求，其研究成果必须在数据挖掘理论或商业数据挖掘系统的开发上能够产生深远的影响。这是 SIGKDD 每年最重头的奖项，也是数据挖掘领域最高技术奖项。

服务贡献奖：Gabor Melli（索尼互动娱乐）

2013 年杰出服务奖（Service Award）由 Gabor Melli（索尼互动娱乐）获得。Gabor Melli 博士因其对数据挖掘实践和应用的重大技术贡献，以及对全球 KDD 社区的杰出服务而受到表彰。他长期为 KDD 社区服务，从 1997 年开始，他在 datasetgenerator.com 上发布了合成数据集生成器。在 2005 年，他与伙伴共同创立了数据挖掘案例研究研讨会系列（dataminingcasestudies.com），并成为 SIGKDD 信息总监。从那时起，他参与组织了许多 KDD，ICDM 和 CIKM 国际会议，担任出版主席，研讨会主席，赞助主席，竞赛主席，展览主席和演示主席等职务，同时担任 KDD，CIKM，PAKDD 的评委和 ACM 计算调查。2013 年，

Gabor Melli 博士启动了一个半自动创建广泛的 KDD 概念网络知识库（www.gabormelli.com/RKB/）的项目。

SIGKDD 杰出服务奖主要授予在知识发现及数据挖掘领域作出重大服务贡献的个人或团队，包括主办会议、主持学术团体等服务性工作，并在数据挖掘教学及财务性事务等方面的工作。该奖主要奖励对知识发现及数据挖掘领域作出重大服务贡献的个人或团队，考察的因素主要包括主持学术团体、主办会议等服务性工作，教育学生、研究者和实践者，资助研发活动，为传播技术信息提供专业志愿服务，并通过知识挖掘应用为社会做出贡献，改善全球性医疗、教育、灾难 / 危机管理及环境等议题。

✧ KDD CUP 竞赛

第一名：Algorithm 队

第二名：Dmitry & Leustagos & BS Man 队

KDD CUP 2013 由微软的学术搜索系统提供数据集，共开设两个专题，任务分别是识别作者与论文的对应关系，和作者的名字消歧，分别吸引了 561 和 241 支队伍参赛。最终两个专题的冠军皆被国立台湾大学的林智仁教授、林守德教授和林軒田教授带领的团队斩获。

● 代表性论文

KDD 2013 大会上，最具代表性的两个 Session 是大数据和社交网络分析，在大数据分析框架的 Session 中，来自伯克利大学的论文提出结合 CPU、GPU 以及全新的算法设计来提高大数据挖掘能力。来自康奈尔大学的论文则探讨了如何应对大数据上的复杂分析问题，论文试图将大数据上的复杂分析任务分解为一系列的简单任务。来自韩国 POSTECH 大学的文章则提出一个并行图计算引擎：TurboGraph。以下是这三篇论文的介绍：

标题：Big Data Analytics with Small Footprint:Squaring the Cloud

作者：John Canny, Huasha Zhao

论文解读：论文提出名为 BID 的大数据处理引擎，并基于该引擎开发了 BIDMat 工具包用于矩阵计算和 BIDMath 用于机器学习，论文以聚类和分类算法为例，证明了这种新设计可以将单台 PC 机处理数据的速度提高数十倍。

地址：

https://static.aminer.org/pdf/20160902/loc-conf/KDD/KDD-2013-1103.pdf

标题：Beyond Myopic Inference in Big Data Pipelines

作者：Karthik Raman, Adith Swaminathan, Johannes Gehrke, Thorsten Joachims

论文解读：论文提出一个概率管道模型（Probabilistic Pipeline Model），该模型能够将复杂的概率推断问题分解为多个简单的概率计算过程，有效的提高了概率模型的推断计算过程，论文分别分析了 Top-K 推断、Fixed Beam 推断以及 Adaptive 推断的实际效果。

地址:

https://static.aminer.org/upload/pdf/1365/1623/470/552e6c2545ce5a3f931832a0.pdf

标题：TurboGraph:A Fast Parallel Graph Engine Handling Billion-scale Graphs in a Single PC

作者：Wook-Shin Han, Sangyeon Lee, Kyungyeol Park, Jeong-Hoon Lee, Min-Soo Kim1, Jinha Kim, Hwanjo Yu

论文解读：相比传统的图计算引擎（如：GraphChi），TurboGraph 有效的提高了输入输出的并发性。论文还提出一个叫做 pin-and-slide 的并行执行模式，从引擎底层实现了大规模图的并发分析。

地址:

https://static.aminer.org/upload/pdf/365/1693/1669/552e7c5145ce5a3f931832ac.pdf

在另一个大数据分析算法的 Session 中，四篇着重探讨了如何对现有算法进行改进从而应对大规模数据带来的挑战。一个思路是如何在有限内存中实现传统挖掘算法在大数据上的计算和学习；另外还有两篇文章则对传统的矩阵分解方法进行扩展，使其能够处理大规模异构网络。

在社交网络分析方面，三个最热的话题是：网络用户行为建模、用户影响力以及网络信息传播。在用户行为建模和分析方面，既有单独探讨用户分享行为可信度验证的问题，也有论文开始探讨如何将不同社交网络网站的用户集成起来进行统一的用户行为建模；还有探讨网络用户群体智能的产生过程。在用户影响力方面，几篇有意思的文章分别探讨了网络用户行为的隐含影响力学习问题，社会影响力中的从众现象，数据中的影响力学习，还有一些工作开始将社会影响力应用到其他数据挖掘问题中，如利用社会影响力提高聚类精度和利用影响力分析方法做话题发现。在信息传播方面，今年比较有意思的一篇文章是讨论信息传播对网络结构演化的影响，另外也有文章讨论网络用户在传播信息时的可信度问题。

2.10.2 SIGKDD 2014

2014 年 8 月 24-27 日，第 20 届国际知识发现与数据挖掘大会（SIGKDD 2014）在美国纽约市召开。本届大会的主题为“Data Science for Social Good”（数据科学推动社会进步），旨在呼吁和推动数据科学家投身和致力于解决实际社会问题。

本届大会的主席由 Facebook 公司的数据科学家 Sofus Macskassy 博士和 Dstillery 公司的首席科学家 Claudia Perlich 博士共同担任，研究性程序委员会主席由斯坦福大学 Jure Leskovec 教授和加州大学洛杉矶分校的 Wei Wang 教授共同担任，工业界程序委员会委员会主席由芝加哥大学 Rayid Ghani 教授（曾任奥巴马政府首席科技顾问）和 Social Alpha 创始人兼首席执行官 Prem Melville 博士共同担任。吸引了 2320 人注册参加。

SIGKDD 2014 共收到 1036 篇研究性论文和 197 篇工业和政府应用性论文投稿，双双高于 SIGKDD 2013 相应投稿数量的 40%以上。每篇投稿文章由至少 3 名审稿人评审及 1 名相应领域的高级程序委员协调讨论并推荐，最终由程序委员会主席决定是否录取。在 46 名高级程序委员和 340 名评审人的共同努力下，本次大会最终录取 151 篇研究性论文（录用率约 14.6%）和 44 篇工业和政府应用性论文（录用率约 22%）。

中国大陆学者作为第一作者在本届大会共发表 13 篇相关研究论文，作者来自清华大学、南京大学、浙江大学、上海交通大学、中国人民大学、吉林大学等科研院校，共有 80 余位大陆学者参加了 SIGKDD 2014。

SIGKDD 2014 邀请了艾伦人工智能研究所首席执行官 Oren Etzioni 博士，微软雷蒙德研究院院长 Eric Horvitz 博士，伊坎基因组织学和多尺度生物学研究所所长 Eric Schadt 博士，哈佛大学经济系 Sendhil Mullainathan 教授以及彭博资讯公司首席执行官 Dan Doctoroff 先生进行大会主题报告。

SIGKDD 2014 除了经典的最佳论文、最佳学生论文等奖项外，还设置了与会议主题有关的论文奖项，具体的获奖情况如下：

- 最佳论文

✧ 最佳研究型论文

标题：Reducing the Sampling Complexity of Topic Models（减少主题模型的采样复杂度）

作者：Aaron Q.Li, Amr Ahmed, Sujith Ravi, Alexander J. Smola

论文解读：主题模型的推论通常涉及将潜在变量与观察结果相关联的采样步骤，随着数据量的增加，生成模型失去了稀疏性。文章提出了一种算法，该算法与文档中实际实例化的主题数 kd 成线性比例。建立大型文档集合和结构化层次模型 kdllk，产生了数量级的加速，可以有效地近似密集，缓慢变化的分布。

地址:

https://www.aminer.cn/pub/5550453745ce0a409eb549bd/reducing-the-sampling-complexity-of-topic-models

标题：An Efficient Algorithm For Weak Hierarchical Lasso（弱层次套索的一种有效算法）

作者：Yashu Liu, Jie Wang, Jieping Ye

论文解读：文章通过使用通用迭代收缩和阈值（GIST）优化框架直接解决非凸弱层次 Lasso，这已证明对于解决非凸稀疏公式是有效的。文章还开发了一种用于计算具有线性时间复杂度的子问题的算法，扩展了该技术，以执行基于优化的成对交互作用的二元分类问题的分层测试。仿真研究表明，当主要效果和交互作用之间存在层次结构时，非凸层次测试框架的性能优于凸松弛。

地址：

https://www.aminer.cn/pub/5550453745ce0a409eb549c3/an-efficient-algorithm-for-weak-hierarchical-lasso

标题：Targeting Direct Cash Transfers to the Extremely Poor（定位极度贫困人口直接发放现金）

作者：Enigma, Give Directly 的研究者们

论文解读：他们的论文主要针对社会问题，通过分析肯尼亚贫困村庄的卫星遥感数据来鉴别极度贫穷家庭，以此为根据为他们提供无条件的人道主义关怀和金钱资助。

地址：

https://www.aminer.cn/archive/targeting-direct-cash-transfers-to-the-extremely-poor/5550453645ce0a409eb549aa

✧ 最佳应用型论文

标题：Style in the Long Tail: Discovering Unique Interests with Latent Variable Models in Large Scale Social E-commerce（长尾设计：在大型社交电子商务中通过潜在的变量模型发现独特的兴趣）

作者：Diane J. Hu, Rob Hall, Josh Attenberg

论文解读：该论文描述了在 Etsy 站点上部署两个基于样式的新推荐系统的方法和实验。其使用了 Latent Dirichlet Allocation（LDA）来发现 Etsy 上的趋势类别和样式，然后用它们来描述用户的“兴趣”配置文件。还探索了散列方法，以便在 map-reduce 框架上执行快速最近邻搜索，以便有效地获取建议。这些技术已经成功实施，大大改善了许多关键业务指标。

地址：

https://www.aminer.cn/archive/style-in-the-long-tail-discovering-unique-interests-with-latent-variable-models-in-large-scale-social-e-commerce/5550453645ce0a409eb5491d

✧ 最佳博士论文

标题：Reconstruction and Applications of Collective Storylines from Web Photo Collections（将来自网页照片集合中的集合故事线重构、应用）

作者：Gunhee Kim 博士, 其导师为 Eric Xing 教授

论文解读：该论文的目标是通过联合推断图像集的时间趋势和重叠内容，来创建集体故事情节，还利用重建的照片故事情节，来探索新颖的计算机视觉和数据挖掘应用程序。提出了分支故事情节图的重建算法。

地址:

https://www.aminer.cn/pub/56d8b7d4dabfae2eee16b8dc/reconstruction-and-applications-of-collective-storylines-from-web-photo-collections

✧ 时间检测奖（Test of Time，也就是十年最佳论文）

SIGKDD 大会从 2014 开始设立 Test of Time 最佳论文奖，旨在表彰过去 20 年 KDD 大会上发表得有重大影响力的优秀论文，该奖项最初两年各颁发给三篇论文，之后将每年颁发给一篇论文。

以下三篇论文在 2014 年获此殊荣：

标题：A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise （KDD 1996）（一种基于密度的具有噪声的大型空间数据库集群发现算法）

作者：Martin Ester, Hans-Peter Kriegel, Joerg Sander, Xiaowei Xu

论文解读：针对聚类算法在大型空间数据库应用上的局限，文章提出了一种新的聚类算法 DBSCAN，它依赖于基于密度的聚类概念，该概念旨在显示任意形状的聚类。 DBSCAN 仅需要一个输入参数，并支持用户为其确定合适的值。实验结果表明 DBSCAN 在发现任意形状的集群方面比众所周知的算法 CLARANS 更加有效，并且 DBSCAN 在性能方面比 CLARANS 高 100 倍。

地址：

https://www.aminer.cn/pub/53e9bd7bb7602d9704a1d173/a-density-based-algorithm-for-discovering-clusters-in-large-spatial-databases-with

标题：Integrating Classification and Association Rule Mining （KDD 1998）（集成分类和关联规则挖掘）

作者：刘兵, Wynne Hsu, LinkedIn 公司的 Yiming Ma

论文解读：文章建议将关联规则挖掘和分类规则挖掘这两种挖掘技术整合在一起，通过专注于挖掘关联规则的一个特殊子集（称为类关联规则（CAR））来完成集成。还给出了一种有效的算法，用于基于发现的 CAR 集合构建分类器。实验结果表明，以这种方式构建的分类器通常比最新的分类系统 C4.5 生成的分类器更准确。另外，这种集成有助于解决当前分类系统中存在的许多问题。

地址:

https://www.aminer.cn/pub/53e9aefcb7602d9703930008/integrating-classification-and-association-rule-mining

标题：Maximizing the Spread of Influence through a Social Network（KDD 2003）（通过社交网络最大化影响力扩散度）

作者：David Kempe, Jon Kleinberg, Eva Tardos

论文解读：使用基于子模块函数的分析框架，文章证明了自然贪婪策略所获得的解决方案对于几种类型的模型而言，可证明在最优值的 63%之内；文章提出了一种针对社交网络中此类影响力问题的算法的性能保证进行推理的通用方法，并在大型协作网络上提供了计算实验，近似算法明显优于性能基于对社交网络领域的度中心性和距离中心性的深入研究的节点选择启发式算法。

地址:

https://www.aminer.cn/pub/53e9b6d1b7602d970425d749/maximizing-the-spread-of-influence-through-a-social-network

✧ SIGKDD 创新和服务大奖

创新贡献奖：Prof. Pedro Domingos （University of Washington）

SIGKDD 2014 创新奖由华盛顿大学的 Pedro Domingos 教授获得，以表彰他在流式数据分析、马尔可夫逻辑网等方面的基础性工作。在此之前，Pedro Domingos 教授曾获 SIGKDD、AAAI、UAI 和 NIPS 等国际会议的最佳论文奖。

服务贡献奖：Dr. Ted Senator （Leidos）

Senator 教授在为数据挖掘社区服务方面有着悠久的历史。他的服务影响了一些主要会议的方向，这些会议帮助定义了 KDD 的研究与应用之间的区别，并让人们认识到了那些使用 KDD 解决实际业务、政府和社会价值问题的人所面临的独特挑战和成就。他曾担任 KDD2003 主席，并于 2010、2011 和 2013 年担任工业政府跟踪计划联合主席。他曾于 1996 年和 1997 年担任该计划联合主席和人工智能创新应用（IAAI）主席。自 1993 年以来，他一直是计划委员会成员。自 2003 年以来，他一直是 AAAI 的财务主管，还是 KDD2014 的财务主管。

✧ KDD CUP

KDD CUP 2014 由在线慈善机构 Donors Choos 提供数据集，它要求参赛队伍识别具有潜力的优秀学生项目申请，以此为根据提供经济和其它方面的支持。此次比赛吸引了 652 名人员组成的 472 支队伍参与，比赛冠军最终由“STRAYA”团队获得。

- 代表性论文

2014 年 KDD 上社交网络、机器学习和数据挖掘等领域吸引了最多的投稿，以社交网络为例，大会主会共设有 6 个专题分会场报告讨论其最新进展。为响应今年 SIGKDD 大会主题“Data Science for Social Good”的号召，越来越多的数据科学研究人员致力于使用数学挖掘技术帮助解决社会性问题，例如，此次大会设有 3 个面向医疗领域的专题分会场以及多个面向环境、教育、交通、劳动力等方面的分会场。与此同时，大会还吸引了联合国官方媒体全球脉搏的关注和报道。联合国官方媒体全球脉搏（United Nations Global Pulse）参加了当年的 SIGKDD 大会，并从 200 余篇文章中选中 5 篇具有代表性的数据科学如何解决社会性问题的文章进行逐一报道。该报道认为这 5 篇文章在全球可持续发展及人道主义关怀大背景下具有现实世界的可应用性。以下是 5 篇论文简介：

标题：Inferring User Demographics and Social Strategies in Mobile Social Networks

作者：Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, Nitesh V. Chawla

地址：

https://www.aminer.cn/archive/inferring-user-demographics-and-social-strategies-in-mobile-social-networks/5550453645ce0a409eb54966

论文解读：该文章发现了人类社交策略随着其人生不同阶段社交交往需求的变化过程，例如年轻人更趋向于不断扩展他们的朋友圈，并且积极与异性保持联系；随着年龄增长，当人至中年或老年，则更多趋于保持小数量的但是更亲密的同性朋友圈。基于此发现，该团队根据人类手机的通话和短信使用模式准确地预测了使用者的性别和年龄信息。

标题：EARS（Earthquake Alert and Report System）: a Real Time Decision Support System for Earthquake Crisis Management

作者：Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, Maurizio Tesconi

地址：

https://www.aminer.cn/archive/ears-earthquake-alert-and-report-system-a-real-time-decision-support-system-for-earthquake-crisis-management/5550453545ce0a409eb548e1

论文解读：该团队设计了一个基于实时社交媒体数据的地震预警和损害评估系统，该系统可以实时地报告地震中人口和基础设施层面的信息，并提供决策支持。

标题：Prediction of Human Emergence Behavior and their Mobility following Large-scale Disaster

作者：Xuan Song, Quanshi Zhang, Yoshihide Sekimoto, Ryosuke Shibasaki

地址：

https://www.aminer.cn/pub/5550453645ce0a409eb54982/prediction-of-human-emergency-behavior-and-their-mobility-following-large-scale-disaster

论文解读：准确预测人类的紧急行为及其流动性将成为规划有效的人道主义救援，灾难管理和长期社会重建的关键问题。文章建立了一个大型的人类流动性数据库和几个不同的数据集，以捕获和分析人类紧急行为及其流动性。实证分析发现，大规模灾害后的人类行动有时与正常时期的行动方式相关，并且受到其社会关系、灾害强度、破坏程度、政府的强烈影响。基于此，文章开发了一种人类行为模型，可以准确地预测大规模灾难后的人类应急行为及其流动性。

标题：Targeting Direct Cash Transfers to the Extremely Poor

作者：Brian Abelson, Kush R. Varshney, Joy Sun

地址：

https://www.aminer.cn/archive/targeting-direct-cash-transfers-to-the-extremely-poor/55504 53645ce0a409eb549aa

论文解读：文章主要针对社会贫困问题，通过分析肯尼亚贫困村庄的卫星遥感数据来鉴别极度贫穷家庭，以此为根据为他们提供无条件的人道主义关怀和金钱资助。

标题：Inferring Gas Consumption and Pollution Emission of Vehicles throughout a City

作者：Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, Yong Yu

地址：

https://www.aminer.cn/archive/inferring-gas-consumption-and-pollution-emission-of-vehicles-throughout-a-city/5550453545ce0a409eb548e2

论文解读：该论文基于北京32000辆出租车的GPS轨迹数据分析实时路况，预测不同时间和地区内该市的汽油消耗和污染排放情况。

KDD 2014 程序委员会主席 Jure Leskovec 教授和 Wei Wang 教授以及 SIGKDD 创始人之一 Gregory Piatetskyr 的报告显示，社交和信息网络分析是今年 SIGKDD 大会上最热的话题，其流行趋势已经在 SIGKDD 大会上延续了近 10 年。

2014 年有六个直接相关的大会报告专场，包括研究报告专场：网络信息传播、社交网络分析、网络社团发现、社交与信息网络，工业和政府报告专场：社交媒体以及工业和政府应邀报告专场：社交和新兴业务，相关报告论文 25 篇，图挖掘也有 3 个大会报告专场。总体来说，网络信息传播以及社团发现分别有一个独立的报告专场，成为本次大会社交网络研究中最流行的话题。在网络信息传播方面，网络结构和信息传播的相互影响成为一大热点，另外社会影响力

建模仍然占据着一席之地。在社团发现方面，研究人员从基于节点属性、网络用户相关性、热核的社团发现和动态社团演化等方向上探讨了最新进展。而其他社交网络分析专场中，网络用户建模、微观网络结构及动态网络演化分析则都有最新研究成果。

此外，随着大数据浪潮的推进，大数据挖掘吸引着数据科学家和与会者的高度关注，其中 4 个专题分会场报告大数据技术发展的最新动向。

表 13 专题分会场报告主题

Session 1	第一个为基于统计技术的大数据挖掘专场，该专场聚焦于并行吉布斯采样，非参贝叶斯模型，在线中国餐馆过程，以及概率知识融合等；
Session 2	第二个为可扩展大数据分析方法专场，主要探讨如何将数据挖掘算法应用在大规模数据上；
Session 3	第三个为可扩展大规模图算法专场，该专场报告了最新的图节点 Betweenness 属性，PageRank 算法，图采样以及图切分算法在大规模网络或图结构中如何扩展并进行低时间复杂度计算；
Session 4	第四个为大规模优化和学习技术专场，其主要重心为讨论如何在大规模及流式数据上进行随机优化，数据摘要以及在线机器学习等技术和模型的发展。

2.10.3 SIGKDD 2015

2015 年 8 月 10-13 日，第 21 届国际知识发现与数据挖掘大会（SIGKDD 2015）在澳大利亚悉尼市召开。本届的主题是数据科学和大数据。

本届大会主席由澳大利亚悉尼科技大学工程与信息技术学院操龙兵教授和悉尼科技大学信息技术学院 Chengqi Zhang 教授共同担任，研究性程序委员会主席由康奈尔大学 Thorsten Joachims 教授和莫纳什大学的 Geoff Webb 教授共同担任，工业界程序委员会委员会主席由波音科技高级研究员 Dragos D. Margineantu 博士和 Pexip 公司董事会成员 Graham Williams 共同担任。

SIGKDD 2015 大会收到了 1008 篇稿件，其中研究论文 819 篇，录用 160 篇（录用率约 19.5%）；政府和工业届应用论文 189 篇，录用 68 篇（录用率约 35.9%）。中国大陆学者作为第一作者在本届大会共发表 14 篇相关研究论文，作者来自清华大学、北京大学、中国科技大学、上海交通大学、西南交通大学、中国科学院、以及电子科技大学等科研院校，共有 90 余位大陆学者参加了本届大会。

本届 SIGKDD 大会邀请了来自微软的杰出科学家 Ronny Kohavi 博士，来自悉尼大学的 Hugh Durrant-Whyte 教授，Coursera 的总裁和联合创始人 Daphne Koller 博士，斯坦福大学商学院的 Susan Athey 教授进行大会的主题报告。

SIGKDD 2015 的具体获奖情况如下：

- 最佳论文

✧ 最佳研究型论文

标题：Efficient Algorithms for Public-Private Social Networks（公共—私人的社交网络高效算法）

作者：Flavio Chierichetti, Alessandro Epasto, Ravi Kumar, Silvio Lattanzi, Vahab Mirrokni

论文解读：该论文介绍图的公私模型。在此模型中有一个公共图，并且公共图中的每个节点都有一个关联的私有图。论文研究利用草图化和采样这两种大规模图计算中有效的范式来加速公开和私密相结合的社交网络中的计算问题。

地址：

https://www.aminer.cn/pub/5736973c6e3b12023e62b598/efficient-algorithms-for-public-private-social-networks

标题：Edge-Weighted Personalized PageRank: Breaking a Decade-Old Performance Barrier（边缘加权个性化网页排序：打破十年来的性能障碍）

作者：Wenlei Xie, David Bindel, Alan Demers, Johannes Gehrke

论文解读：自十多年前个性化 PageRank 问世以来，基于边缘权重的个性化一直是一个悬而未决的问题。该论文在当边缘权重被个性化时，描述了第一种用于在一般图形上计算 PageRank 的快速算法。基于模型约简的方法比现有方法的性能高出近五个数量级。与以前的工作相比，这种巨大的性能提升能够以交互速度解决边缘权重个性化的学习排名问题，而这是以前无法实现的。

地址：

https://www.aminer.cn/archive/edge-weighted-personalized-pagerank-breaking-a-decade-old-performance-barrier/5736973c6e3b12023e62b916

✧ 最佳博士论文

标题：Mining Latent Entity Structures from Massive Unstructured and Interconnected Data（从大量非结构化和互连的数据中挖掘潜在的实体结构）

作者：Chi Wang 博士，其导师为韩家炜（Jiawei Han）教授。

论文解读：大数据时代背景下，信息量之庞大，覆盖人类日常生活的方方面面，但有价值且相连的信息通常隐藏在大量无结构和组织的数据背后，该论文目的是提供一种挖掘框架，整合、解决一系列的数据挖掘关联问题。

地址：

https://www.aminer.cn/archive/mining-latent-entity-structures-from-massive-unstructured-and-interconnected-data/555048b845ce0a409eb70bc0

标题：Modeling Large Scale Social Network in Context（从语境中模拟大型社交网络）

作者：Qirong Ho 博士，导师为 Eric Xing 教授

论文解读：该论文提出的方法将统计模型的灵活性与数据挖掘和社会网络社区的关键思想和经验观察相结合，并得到基于原始分布式系统研究的集群计算软件库的支持，最终形成了一个新的混合成员三角形 motif 模型，该模型可以轻松扩展到仅在几个集群机器上就有超过 1 亿个节点的大型网络，并且可以很容易地扩展到使用其他技术来适应网络上下文。

地址：

http://reports-archive.adm.cs.cmu.edu/anon/ml2014/CMU-ML-14-100.pdf

标题：Computing Distrust in Social Media（社交媒体的不信任度计算）

作者：Jiliang Tang 博士，其导师 Huan Liu 教授

论文解读：该论文的主要目的是去解决发现一些创新性的研究及新奇的方法。一种带有新奇框架的预测不信任被提出，让不可见不信任变得可见，开发有原则的方法去应用不信任在各大社交软件。因为不信任是消极链接的一种特殊类型，论文论证了消极链接的不信任算法和特性概括。

地址：

https://repository.asu.edu/items/28550

✧ 时间检测奖（Test of Time Award）

标题：Mining High-Speed Data Streams （KDD 2000）（挖掘高速数据流）

作者：Pedro Domingos, Geoff Hulten

论文解读：许多组织拥有的庞大数据库每天以无数记录的速度无限增长，该论文提出了 VFDT 这一实时数据流挖掘系统，用于快速处理不断增长的数据记录,使用霍夫丁定界来保证其输出在渐近性上与传统学习者的输出几乎相同。文章将 VFDT 应用于挖掘整个华盛顿大学主校区的连续 Web 访问数据流。

地址：

https://www.aminer.cn/archive/mining-high-speed-data-streams/53e9b055b7602d9703ab8cf1

标题：Optimizing Search Engines Using Click-Through Data （KDD 2002）（使用可点击数据来最优化搜索引擎）

作者：Thorsten Joachims

论文解读：该论文率先提出了通过挖掘利用用户点击数据从而提高搜索引擎结果准确率的方法，该方法激发了一系列后继工作的发展并成为目前学术界与工业界搜索引擎研发的公认基础方法。

地址：

https://www.aminer.cn/archive/optimizing-search-engines-using-clickthrough-data/53e9b042b7602d9703aa0d88

标题：Mining and Summarizing Customer Reviews （KDD 2004）（挖掘与汇总消费者点评）

作者：Minqing Hu, 刘兵教授

论文解读：随着电子商务变得越来越流行，产品收到的客户评论数量迅速增长。该文章旨在挖掘并总结产品的所有客户评论，主要内容包括：（1）挖掘客户评论过的产品功能；（2）在每次评论中确定观点句子，并确定每个观点句子是肯定的还是否定的。该方法能够帮助潜在客户和产品生产者更好理解产品。

地址：

https://www.aminer.cn/archive/mining-and-summarizing-customer-reviews/53e9af46b7602d97039827ec

✧ SIGKDD 创新和服务大奖

创新贡献奖：Prof. Hans Peter Kriegel （University of Munich）

创新奖（Innovation Award）由慕尼黑大学的 Hans Peter Kriegel 教授获得，以表彰他在数据挖掘，尤其是聚类分析、异常值检测以及高维数据分析等工作中做出的杰出贡献。在此之前，Hans Peter Kriegel 教授曾获 2014 年 KDD Test of Time 论文奖，并当选 ACM fellow。

服务贡献奖：Prof. Jian Pei （Simon Fraser University）

服务贡献奖（Service Award）授予西蒙佛莱逊大学的 Jian Pei 教授，用以表彰他为推动数据挖掘领域技术发展所作出的杰出成就以及为领域学术进步所作出的杰出服务，包括组织 SIGKDD、ICDM、SDM 以及 CIKM 等相关会议、担任 TKDE 主编以及其它一些主要刊物的副主编等。

✧ KDD CUP 竞赛

第一名：the Intercontinental Ensemble

KDD CUP 2015 由清华大学的唐杰和以色列海法大学的 Ron Bekkerman 一起担任共同主席，由中国大规模在线教育网站学堂在线（XuetangX.com）承办，竞赛的任务是要求参赛队伍预测在线课程中学生退课的情况。此次比赛吸引了来自 26 个国家的 1263 名参赛者组成的 821 支队伍参与。比赛冠军最终由多国军团“the Intercontinental Ensemble”团队获得。

- 代表性论文

在研究热点方面，KDD 2015 最热的论文包括：社交网络挖掘、Web 挖掘以及流数据挖掘、城市计算、图采样算法以及一些新应用场景下的推荐算法等，这些话题都吸引了众多投稿。以图采样算法为例，大会主会共设有 5 个专题分会场报告讨论其最新进展。而一些传统研究如推荐算法，则更多地向一些新的具体应用转换，例如地点推荐和路径推荐等。此外，随着近年知识图谱的迅猛发展，实体挖掘、实体关系映射等传统问题又被重新重视起来。

社会网络及图挖掘近年来是 KDD 领域非常火热的研究方向，当年获得最佳论文奖及最佳学生论文奖的两篇论文均出自该主题。除此以外，来自清华大学知识工程实验室的论文，研究了社交网络中隐藏元素（如用户社会角色，社区等）及可见元素（如用户，用户间的关联关系，以及用户行为等）的关联关系，可应用于提高用户行为预测和社区发现等精确度。清华大学和美国圣母大学合作的耦合网络上的链接预测问题是有别于传统链接预测的新型链接预测问题。

标题：CoupledLP: Link Prediction in Coupled Networks

作者：Yuxiao Dong, Jing Zhang, Jie Tang, Nitesh V. Chawla, Bai Wang

论文解读：该论文提出了一个统一的框架 CoupledLP，利用原子传播规则在目标网络中自动构建隐式链接，以解决目标网络不完整的问题。提出了一个耦合因子图模型，以合并从网络耦合部分提取的元路径。在疾病基因（DG）和移动社交网络这两个数据集上，CoupledLP 框架都优于几种替代方法。所提出的耦合链接预测问题和相应的框架说明了生物学和社交网络中的科学和商业应用。

地址:

https://www.aminer.cn/pub/5ef1b98b9e795e111756d4e6/coupledlp-link-prediction-in-coupled-networks

现今图数据规模往往十分庞大，为了能够高效地处理图数据，可以采用采样的方式，近似计算大规模图的目标算法。如来自德州卡菲基恩大学的工作，描述了如何利用图采样，计算图中 3-profile（即任意三个结点的各种组成结构，如三角形）的出现次数。此外，来自卡耐基梅隆大学的工作，研究了如何利用采样计算图中某个结点与其它结点的最长距离。来自哈佛大学的工作，研究了利用采样算法识别大规模网络中的稠密子图。再比如来自韩国科学技术院的工作，设计了一个动态采样算法，用于统计网络中的局部三角形个数。清华大学和亚利桑那州立大学合作研究的基于图上的随机路径采样方法来快速计算大规模图中两个结点之间的相似度。该方法在一个大约 50 万个结点，500 万条边的图上比已有方法快大约 300 倍，并且能够成功在 10 亿条边的图上计算任意一个结点的 Top-5 相似结点。

Web 挖掘（Web Mining）专题分会。互联网是当前学术和工业界共同关注的热点话题，在互联网环境下的数据挖掘问题同样也是 KDD 会议中的一大关注热点。在当年 KDD 中专门设立了一个有关 Web Mining 的主题研讨会。以下是该专题部分论文简介：

标 题：ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering

作者：Xiang Ren, Ahmed El-Kishky, Chi Wang, Fangbo Tao, Clare R Voss, Heng Ji, Jiawei Han.

论文解读：实体识别是一个重要但具有挑战性的研究问题。该论文研究了具有远程监督的实体识别（ER），并提出了一种新的基于关系短语的 ER 框架，称为 ClusType，该框架运行数据驱动的短语挖掘以生成实体提及候选对象和关系短语，根据在语料库上计算的共现关系短语的类型签名和其表面名称的类型

指示符，预测每个实体提及的类型。与最佳比较方法相比，ClusType 的 F1 得分平均提高了 37%。

地址:

https://www.aminer.cn/pub/5736973b6e3b12023e62b1e3/clustype-effective-entity-recognition-and-typing-by-relation-phrase-based-clustering

标题：TimeMachine: Timeline Generation for Knowledge-Base Entities

作者：Tim Althoff, Xin Luna Dong, Kevin Murphy, Safa Alai, Van Dang, Wei Zhang

论文解读：介绍了基于数据的时间轴自动建立方法。这一方法将问题形式化为数据压缩的问题，提出了高效的 TIMECRUNCH 算法寻找动态图中的时序模式。这一方法有着很多现实的应用场景。

地址:

https://www.aminer.cn/archive/timemachine-timeline-generation-for-knowledge-base-entities/ 5736973b6e3b12023e62b0b7

标题：Entity Matching across Heterogeneous Sources

作者：Yang Yang, Yizhou Sun, Jie Tang, Bo Ma, Juanzi Li

论文解读：论文介绍了跨领域的实体匹配问题。这一问题的主要挑战在于由于表达方式不同，不同领域对实体的介绍性文档在内容上的重叠度很低。文章提出了交叉采样方法用于从不同数据源抽取话题，并利用话题层面对实体进行关联度匹配。

地址：

https://www.aminer.cn/archive/entity-matching-across-heterogeneous-sources/5736973b6e3b12023e62b38a

标题：COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency

作者：Yutao Zhang, Jie Tang, Zhilin Yang, Jian Pei, Philip S. Yu

论文解读：文章提出利用结合局部与全局一致性的方法，将用户配对相似度，网络相似度以及跨网路匹配的传递性进行统一建模。文章提出的方法已被成功应用于 AMiner 学术网络分析与挖掘在线系统中。

地址：

https://www.aminer.cn/archive/cosnet-connecting-heterogeneous-social-networks-with-local-and-global-consistency/5736973b6e3b12023e62b502

流数据挖掘（Stream Data Mining）专题分会。随着大数据爆炸式增长，流数据类型以及采样方法的研究持续受到关注。以下是专题相关的部分论文：

标题：Stream Sampling for Frequency Cap Statistics

作者：Edith Cohen

地址：

https://www.aminer.cn/archive/stream-sampling-for-frequency-cap-statistics/5736973b6e 3b12023e62b384

论文解读：论文是基于 Google 研究院的流数据的统计量计算方法，例如某一属性的唯一值数量以及属性数值之和这两种统计量。他们的主要贡献是提出一个针对非聚集流数据类型的通用采样架构。其核心思想是定义随机打分函数，用于有效地近似估计统计量的值。

标题：Efficient Online Evaluation of Big Data Stream Classifiers

作者：Albert Bifet, Gianmarco de Francisci Morales, Jesse Read, Geoff Holmes, Bernhard Pfahringer

论文解读：文章指出流数据上分类器的评估存在几个关键问题，例如已有的交叉验证和显著性检测等验证方法，直接用在流数据环境下会带来错误估计，流数据的不平衡特征导致样本偏向于一个类别，从而使 F1 和 Accuracy 等检测指标也不再适用等。他们的主要贡献是提出适应流数据的方法来分别解决上述问题。

地址：

https://www.aminer.cn/pub/5736973b6e3b12023e62b516/efficient-online-evaluation-of-big-data-stream-classifiers

标题：A PCA-Based Change Detection Framework for Multidimensional Data Streams

作者：Abdulhakim Qahtan, Basma Alharbi, Suojin Wang, Xiangliang Zhang

论文解读：文章提出一个基于主成分分析方法的检测多维流数据变化的框架。其主要思想是将数据映射到低维空间，使其能够促进密度估计和分数变化的计算。

地址：

https://www.aminer.cn/archive/a-pca-based-change-detection-framework-for-multidimensional-data-streams-change-detection-in-multidimensional-data-streams/5736973c6e3b12023e62ba04

大数据（Big Data）专题分会。大数据研究仍然在升温。以下是该专题部分论文：

标题：Large-Scale Distributed Bayesian Matrix Factorization using Stochastic Gradient MCMC

作者：Sungjin Ahn, Anoop Korattikara, Nathan Liu, Suju Rajan, Max Welling

论文解读：主题是用随机梯度蒙特卡罗法求解大规模分布式贝叶斯矩阵分解。他们的算法基于分布式 Gradient Langevin Dynamics，不但能够达到标准蒙特卡罗（如 Gibbs 采样）的精准度，而且效率更高，媲美随机梯度下降法。此外，他们对算法进行了并行化，使其能够适应大规模数据的贝叶斯矩阵分解。

地址：

https://www.aminer.cn/archive/large-scale-distributed-bayesian-matrix-factorization-using-stochastic-gradient-mcmc/5736973c6e3b12023e62baf7

标题：Scaling Up Stochastic Dual Coordinate Ascent

作者：Kenneth Tran, Saghar Hosseini, Lin Xiao, Thomas Finley, Mikhail Bilenko

地址：

https://www.aminer.cn/pub/5736973b6e3b12023e62b35d/scaling-up-stochastic-dual-coordinate-ascent

论文解读：SDCA 是通过最小化凸优化方法来求解大规模监督学习问题的经典方法。该研究提出了两种针对该问题的对偶问题进行提速的方法并通过两组数据进行了有效验证。

标题：Network Lasso: Clustering and Optimization in Large-Scale Graphs

作者：David Hallac, Jure Leskovec, Stephen Boyd

论文解读：文章提出了一种通用的 network lasso 方法对图数据上的聚类和优化问题进行建模和快速求解。主要介绍了将组套索推广到允许同时进行聚类和图优化的网络套索，开发了一种基于乘数交替方向法（ADMM）的算法，还研究了这种方法的非凸扩展，可用于解决大型优化问题。

地址：

https://www.aminer.cn/pub/5736973c6e3b12023e62b9e3/network-lasso-clustering-and-optimization-in-large-graphs

标题：Petuum: A new Platform for Distributed Machine Learning on Big Data

作者：Eric P. Xing, Qirong Ho, Wei Dai, Jin Kyu Kim, Jinliang Wei, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimanu Kumar, Yaoliang Yu

论文解读：文章研究并实现了一个大数据环境下的通用分布式机器学习平台。其核心思想是同时考虑了容错性，动态结构和不一致的收敛性，使其不同于传统的面向操作的程序。实验结果展示了在 100 多台机器上运行的结果。第五篇论文是来自威廉与玛丽学院和微软研究院的对大规模分布式深度神经网络模型进行快速求解的研究工作。

地址：

https://www.aminer.cn/archive/petuum-a-new-platform-for-distributed-machine-learning-on-big-data/573696486e3b12023e5555e1

推荐系统（Recommender Systems）专题分会。推荐算法与系统的热度依然不减。以下是专题部分论文简介：

标题：A Collective Bayesian Poisson Factorization Model for Cold-start Local Event Recommendation

作者：Wei Zhang, Jianyong Wang

论文解读：文章提出了一个集体贝叶斯泊松分解（CBPF）模型，该模型综合考虑了用户反馈，社交关系、以及事件的组织者、地点和文本内容等因素来求解事件推荐的冷启动问题。

地址：

https://www.aminer.cn/archive/a-collective-bayesian-poisson-factorization-model-for-cold-start-local-event-recommendation/5736973b6e3b12023e62b427

标 题： Inferring Networks of Substitutable and Complementary Products

作者： Julian McAuley, Rahul Pandey, Jure Leskovec

论文解读： 该研究同时考虑了产品的替代和补充方面的因素，将产品推荐问题形式化为链接预测问题，并采用主题模型对该问题进行建模和求解，同时考虑了商品的细粒度分类类别。在 Amazon 的真实数据上进行了实验验证，实验结果极大地优于其它传统方法。

地址：

https://www.aminer.cn/archive/inferring-networks-of-substitutable-and-complementary-products/5736973b6e3b12023e62b11d

标 题： Regularity and Conformity: Location Prediction Using Heterogeneous Mobility Data

作者： Yingzi Wang, Nicholas Jing Yuan, Defu Lian, Linli Xu, Xing Xie, Enhong Chen, Yong Rui

论文解读： 该研究同时考虑了用户行为的规律性和从众性特征，并使用了三种行为数据来建模，使用时间相关的矩阵分解来学习从众特性，使用 Sparse group Lasso 模型来学习规律特性。

地址：

https://www.aminer.cn/archive/regularity-and-conformity-location-prediction-using-heterogeneous-mobility-data/5736973b6e3b12023e62b330

标题： SCRAM: A Sharing Considered Route Assignment Mechanism for Fair Taxi Route Recommendations

作者： Shiyou Qian, Jian Cao, Frédéric Le Mouël, Issam Sahel, Minglu Li

论文解读：该工作一方面为竞争的出租车司机提供了公平的推荐平台，另一方面不牺牲他们的运作效率。该工作使用大量的出租车历史路线数据来测试模型的性能，结果表明该模型在这两个方面都优于之前的方法。

地址：

https://www.aminer.cn/archive/scram-a-sharing-considered-route-assignment-mechanism-for-fair-taxi-route-recommendations/5736973c6e3b12023e62baa2

2.10.4 SIGKDD 2016

2016 年 8 月 13-17 日，第 22 届国际知识发现与数据挖掘大会（SIGKDD 2016）在美国旧金山市召开。大会主题包括：图挖掘和社交网络、深度学习、推荐系统、时间序列和流数据挖掘、应用数据科学等。

SIGKDD 2016 大会主席由 IBM 公司的 Balaji Krishnapuram 博士和 BOSCH 公司的 Mohak Shah 博士担任，研究性程序委员会主席（PC Co-Chairs）由卡耐基梅隆大学（现 Amazon 公司）的 Alex Smola 教授和 IBM 公司的 Charu Aggarwal 博士共同担任；今年开始还将原来的工业和政府 Track 改名为应用数据科学 Track（Applied Data Science Track），本届主席由 IBM 公司的 Rastogi Rajeev 博士和中国百度的 Dou Shen 博士共同担任。此次大会吸引了来自 88 个国家的 2726 人注册参加，创下历史记录。

SIGKDD 2016 大会收到 784 篇研究性论文和 331 篇应用数据科学论文投稿，其中研究性论文的数量相比 2015 年的 1008 篇和 2014 年的 1036 篇有所减少，而应用数据科学论文则相比往年有了很大的增加。本次大会最终录取 142 篇研究性论文（录用率约 18.1%）和 66 篇应用数据科学论文（录用率约 20%）。

SIGKDD 2016 大会邀请了微软新英格兰研究院和纽约研究院院长 Jennifer Chayes 博士，加州大学伯克利分校电子工程与计算机科学系 Joe Hellerstein 教授，全球风险投资公司恩颐投资（NEA）合伙人 Greg Papadopoulos 博士，和牛津大学教授、Google DeepMind 首席科学家 Nando de Freitas 博士，进行了

四场大会主题报告。2015 年 ACM 图灵奖得主、斯坦福大学 Whitfield Diffie 教授受 SIGKDD 邀请在本届大会作了图灵特邀报告。

SIGKDD 2016 具体获奖情况为：

- 最佳论文

✧ 最佳研究型论文

标题：FRAUDAR: Bounding Graph Fraud in the Face of Camouflage（欺诈者：面部伪装中的边界图片欺诈）

作者：Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, Christos Faloutsos

论文解读：文章提出了一种具有抗伪装性的算法 FRAUDAR，为网络欺诈者的有效性提供了上线，并且在真实数据中是有效的。FRAUDAR 在检测伪装和非伪装欺诈方面的准确性均优于顶级竞争对手。追随者跟随图的边缘为 14.7 亿，FRAUDAR 成功地应用到 Twitter 帐户的检测中。

地址：

https://www.aminer.cn/archive/fraudar-bounding-graph-fraud-in-the-face-of-camouflage/57aa28de0a3ac518da9896dc

标题：Ranking Causal Anomalies via Temporal and Dynamical Analysis on Vanishing Correlations（基于消除相关性并通过临时、动态分析排序来排序异常现象的因果关系）

作者：Wei Cheng, Kai Zhang, Haifeng Chen, Guofei Jiang, Zhengzhang Chen, Wei Wang

论文解读：该论文提出了一个基于网络扩散的框架来识别重要的因果异常并对其进行排序。该方法能有效地模拟故障在整个不变网络中的传播，并能同时对结构模式和时变断裂不变性模式进行联合推理。因此，它可以定位真正导致相关性消失的高置信度异常，并可以弥补系统中的非结构化测量噪声。在综

合数据集、银行信息系统数据集和燃煤电厂网络物理系统数据集上的大量实验证明了我们的方法的有效性。

地址：

https://www.aminer.cn/archive/ranking-causal-anomalies-via-temporal-and-dynamical-analysis-on-vanishing-correlations/57aa28de0a3ac518da9896ca

标题：TRIÈST: Counting Local and Global Triangles in Fully-dynamic Streams with Fixed Memory Size（在固定内存大小的全动态流中计算本地三角形和全局三角形）

作者：orenzo De Stefani 等人

论文解读：该论文提出了一种单遍历流式算法，该算法用于对动态网络中全局和局部三角形数量的无偏、低方差、高质量的估计，实验结果显示，该算法在准确性上优于现有的最先进的方法，而且更新时间更短。

地址：

https://www.aminer.cn/archive/tri-st-counting-local-and-global-triangles-in-fully-dynamic-streams-with-fixed-memory-size/599ee0589c05d3074f0123ac

标题：Predicting Matchups and Preferences in Context（预测语境下的匹配与偏好）

作者：Shuo Chen, Thorsten Joachims

论文解读：该论文提出一个综合的概率框架，用于预测两两匹配的结果（两人体育赛事）和偏好配对（产品偏好），与现有模型不同，该模型不仅学习在更具表现力的潜在向量空间中的项目表示，而且还对上下文如何修改匹配和偏好结果进行建模。该模型能够处理任何对称的游戏/比较问题。

地址：

https://www.aminer.cn/archive/predicting-matchups-and-preferences-in-context/57aa28de0a3ac518da9896cb

✧ 最佳应用数据科学论文

标题：Ranking Relevance in Yahoo Search（排序雅虎搜索中的关联性）

作者：Dawei Yin, Yuening Hu, Jiliang Tang, Tim Daly Jr., Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, Jean-Marc Langlois, Yi Chang

论文解读：现代搜索引擎中的相关性已经远远超出了文本匹配，现在面临着巨大的挑战。查询和 URL 之间的语义鸿沟是提高基础相关性的主要障碍。这篇论文总结了雅虎搜索 20 多年的探索和努力，介绍了排序函数、语义特征、查询重写三大关键技术，以及时间、位置相关的搜索。

地址：

https://www.aminer.cn/archive/ranking-relevance-in-yahoo-search/57aa28de0a3ac518da98974b

标题：Matrix Computations and Optimization in Apache Spark（Apache Spark 中的矩阵计算与最优化）

作者：Reza Bosagh Zadeh, Xiangrui Meng, Alexander Ulanov, Burak Yavuz, Li Pu, Shivaram Venkataraman, Evan Sparks, Aaron Staple, Matei Zaharia

论文解读：该论文中主要阐述了一种集群程序框架运算 Apache Spark 中可用的矩阵计算。论文中的主要贡献已经合并入 Apache Spark，在 Spark 安装时属默认设置可用。

地址：

https://www.aminer.cn/archive/matrix-computations-and-optimization-in-apache-spark/56d845aedabfae2eeeac4830

标题：Contextual Intent Tracking for Personal Assistants（用于个人助理的上下文意图追踪）

作者：Yu Sun 等人

论文解读：这篇文章研究了如何为了解决新的推荐范式意图跟踪问题，文章提出了卡尔曼滤波器正则化 PARAFAC2（KP2）临近预报模型，该模型紧凑地表示了上下文和意图的结构以及协同运动。KP2 模型利用了用户之间的协作能力，并为每个用户学习了个性化的动态系统，该系统可以有效地预测用户的意图。KP2 模型可为个人助理软件提供准确的用户意图分析。

地址：

https://www.aminer.cn/archive/contextual-intent-tracking-for-personal-assistants/57aa28de0a3ac518da98974c

标题：Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta（Firebird（数据库）：预测亚特兰大火灾风险并预先进行火险安全检查）

作者：Michael Madaio, Shang-Tse Chen, Oliver L. Haimso, Wenwen Zhang, Xiang Cheng, Matthew Hinds-Aldrich, Duen Horng Chau, Bistra Dilkina

论文解读：该论文开发了 Firebird 框架，使用机器学习、地理编码和信息可视化，以帮助亚特兰大市政消防部门识别和优先安排商业财产火灾检查。Firebird（数据库）计算了该市 5000 多座建筑的火灾风险得分，实际正确率高达 71%。根据 AFRD 的检验标准，已经确定了 6096 处潜在的商业地产进行检验。此外，通过交互式地图，Firebird 整合和可视化火灾事件、财产信息和风险评分，帮助 AFRD 对火灾检查做出知情的决定，并被美国国家消防协会（NFPA）列为使用数据通知消防检查的最佳实践。

地址：

https://www.aminer.cn/archive/firebird-predicting-fire-risk-and-prioritizing-fire-inspections-in-atlanta/573695fe6e3b12023e511d3c

✧ 最佳博士论文

标题：Dissertation: Exploring and Making Sense of Large Graphs（毕业论文：探索理解大规模图形）

作者：Danai Koutra 博士，其导师为 Christos Faloutsos 教授

论文解读：图形代表的信息从网页之间的链接到大脑神经元之间的连接，通常跨越数十亿个节点。在海量的数据中，我们如何找到它最重要的结构？我们如何能检测到关键事件，如计算机系统攻击，或人类大脑中的疾病形成？目前我们已经将我们的方法应用于大量数据，包括 66 亿条边的网络图，18 亿条边的 Twitter 图，以及 9000 万条边的大脑图。

地址：

http://reports-archive.adm.cs.cmu.edu/anon/2015/CMU-CS-15-126.pdf

标题：Mining Disparate Sources for Question Answering（挖掘不同来源以供解答）

作者：Huan Sun 博士，其导师为 Xifeng Yan 教授

论文解读：论文中完善的方法论和框架为一系列研究铺垫了道路，如通过使用不同和互补的数据源，进行各种领域的问答，如健康医疗，商务智能。同时结合人类智能与机器智能，通过问答交互解决一系列问题并做出决策。

地址：

http://web.cse.ohio-state.edu/~sun.397/ResearchProj.html

标题：Scalable Multivariate Time Series Analysis（可扩展的多元时间序列分析）

作者：Taha Bahadori 博士，其导师为 Yan Liu 教授

论文解读：时间序列数据已经在很多应用领域无处不在，如气象科学、社交媒体和健康医疗。利用各种领域大规模收集到的时间序列数据的分析创造出了多种新的挑战和机遇。基于这个论点，我们对大规模时间序列数据进行分析的关键挑战进行了研究，提出了不同的且具扩展性的解决方案。

地址:

http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll3/id/593708

✧ 时间检测奖（Test of Time Award）

标题：Graphs over time: densification laws,shrinking diameters and possible explanations（KDD 2005）（时间推移图：致密化法律、缩小直径和可能的解释）

作者：Jure Leskovec 博士, Jon Kleinberg, Christos Faloutsos

论文解读：这篇论文研究了真实网络如何随时间进行演变，发现随着时间增长，真实网络通常变得越来越稠密，即边数关于点数非线性增长，并且两点之间的平均距离逐渐缩小，而非传统认为的缓慢增长。

地址：

https://www.aminer.cn/archive/graphs-over-time-densification-laws-shrinking-diameters-and-possible-explanations/53e9a515b7602d9702e37973

✧ SIGKDD 创新和服务大奖

创新贡献奖：Philip S.Yu 教授（伊利诺伊大学芝加哥分校）

2016 年创新奖（Innovation Award）颁发给伊利诺伊大学芝加哥分校 Philip S. Yu 教授，以表彰他在大数据挖掘、融合和匿名性上富有影响力的研究工作。此前，Philip S. Yu 教授还获得过国际数据挖掘大会（ICDM）颁发的研究贡献奖。

服务贡献奖：Wei Wang 教授（加州大学洛杉矶分校）

服务奖（Service Award）用于表彰为数据挖掘领域的学术交流和团体发展作出突出贡献的学者，今年的获奖人为加州大学洛杉矶分校 Wei Wang 教授，以表彰她为数据挖掘学术界所作的专业贡献。

✧ KDD CUP 竞赛

第一名：burebistas 队（Adform 公司 Vlad Sandulescu, Bitdevelop 公司 Mihai Chiru）

第二名：T310B 队（清华大学 Yujie Qian, Yinpeng Dong, Ye Ma, Hailong Jin, Juanzi Li）

第三名：browniepointsreturns 队（LatentView Analytics 公司 Mohan Manivannan, Nachiappan Palaniappan）

KDD CUP 2016 竞赛由微软公司组织，竞赛题目为“谁的论文被录用最多？致力于研究机构的影响力度量”。组织者预先指定了 8 个计算机科学不同领域的顶级会议，让参赛队伍预测今年这些会议上各个研究机构发表论文情况的排名。与往届的 KDD CUP 不同，今年的比赛是个名副其实的预测问题，在比赛结束之前组织者和参赛者都不知道会议的真实论文录取情况，因此在比赛过程中没有数据可以用于检验预测的准确性，参赛者需要自己设计并评价预测算法，这使得比赛更加有趣和富有挑战。在综合三个阶段的比赛排名之后，最终丹麦 Adform 公司 Vlad Sandulescu 等人的队伍获得冠军，清华大学 Yujie Qian 等人的队伍获得亚军，印度 LatentView 公司 Mohan Manivannan 等人的队伍获得季军。获奖队伍在 KDD CUP 研讨会上分享和讨论了比赛中使用的模型和算法。

- 代表性论文

下面介绍 KDD 2016 会议的几个主题，包括图挖掘和社交网络、深度学习、推荐系统、时间序列和流数据挖掘、应用数据科学等。

深度学习专题分会。本次 SIGKDD 专门设置了一个深度学习和表示学习（Deep Learning and Embedding）的专题讨论当下炙手可热的深度学习在数据挖掘—尤其是图数据挖掘中的应用。以下是深度学习专题分会部分论文：

标题：node2vec: Scalable Feature Learning for Networks

作者： Aditya Grover, Jure Leskovec

论文解读：文章介绍了一种新的网络数据表示学习的算法。通过重新定义节点的上下文（Context），node2vec 可以同时学习出节点之间距离的信息和节点的局部网络结构的信息。

地址：

https://www.aminer.cn/archive/node-vec-scalable-feature-learning-for-networks/57aa28de 0a3ac518da9896d5

标题：Smart Reply: Automated Response Suggestion for Email

作者：Anjuli KannanF, Karol KurachF, Sujith RaviF, Tobias KaufmannF, Andrew Tomkins, Balint Miklos, Greg Corrado, László Lukács, Marina Ganea, Peter Young, Vivek Ramavajjala

论文解读：研究者通过深度学习实现了一个端到端的自动生成简短邮件回复的模型，这个模型已经在 Gmail 上进行使用，现在已帮助约 10%的邮件进行自动回复。还有一个有趣的工作是来自清华大学和加拿大西蒙弗雷泽大学团队的“Asymmetric Transitivity Preserving Graph Embedding”，他们提出了一个保留网络非传递性的表示学习算法，可以更好地刻画图的结构信息。

地址：

https://www.aminer.cn/archive/smart-reply-automated-response-suggestion-for-email/57aa28dd0a3ac518da9896a6

推荐系统分会。推荐系统一直是数据挖掘领域的一个重要的话题。在 SIGKDD 大会中的推荐系统（Recommender Systems）专题中，让人看到了令人耳目一新的文章：

标题：The Limits of Popularity-Based Recommendations, and the Role of Social Ties

作者：Marco Bressan, Stefano Leucci, Alessandro Panconesi, Prabhakar Raghavan, Erisa Terolli

论文解读：这篇来自罗马大学和谷歌的文章通过经济学的角度尝试回答这样一个问题：推荐系统会多大程度上扭曲市场（改变原来的供求关系），而用户之间的社会联系在其中又扮演了什么样的角色？

地址：

https://www.aminer.cn/pub/57aa28dd0a3ac518da9896aa/the-limits-of-popularity-based-recommendations-and-the-role-of-social-ties

时间序列和流数据挖掘分会。本次 SIGKDD 大会中为期 2 天的专题和研讨会报告中，专门设置了一个时间序列挖掘和学习研讨会（SIGKDD Workshop on Mining and Learning from Time Series），这也迎合了时间序列和流数据挖掘近几年热门的趋势。

在研究性论文中，有一篇论文名为“Recurrent Marked Temporal Point Processes： Embedding Event History to Vector”：

标题：Recurrent Marked Temporal Point Processes: Embedding Event History to Vector

作者：Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel, Le Song

论文解读：尝试使用递归神经网络对时间序列数据进行建模，同时刻画时间和事件的信息。而来自日本熊本大学的工作“Regime Shifts in Streams: Real-time Forecasting of Co-evolving Time Sequences”则讨论了如何对若干个同时进化（Co-evolving）的时间序列进行建模和预测，并同时保证算法的高效和高可扩展性。

地址：

https://www.aminer.cn/archive/recurrent-marked-temporal-point-processes-embedding-event-history-to-vector/57aa28de0a3ac518da9896e7

应用数据科学。和研究性论文不同，应用数据科学的论文更加强调数据挖掘的技术、实践和应用，也更加强调对实际数据的理解和对实际问题的解决。在录取的应用数据科学论文中，无论是老牌的科技公司如微软和雅虎，还是业界新锐如 Pinterest 都和大家分享了他们在实践中和产品中与大数据打交道的心得和技术。

2.10.5 SIGKDD 2017

2017 年 8 月 13-17 日，第 23 届国际知识发现与数据挖掘大会（SIGKDD 2017）在加拿大哈利法克斯召开。

SIGKDD 2017 大会主席由渥太华大学的终身教授、达尔豪斯大学大数据分析研究所所长 Stan Matwin 教授，和 LinkedIn 科学家 Shipeng Yu 博士共同担任，副主席由 IBM Watson Watson Health 小组的首席科学家和高级经理 Faisal Farooq 博士担任；研究性程序委员会主席由 Google 的 Ravi Kumar 博士，与美国东北大学的 Tina Eliassi-Rad 副教授共同担任；工业界程序委员会主席由 Google 的工程总监和研究科学家 Roberto J. Bayardo 博士，与亚马逊研究科学家、加州大学圣地亚哥分校计算机科学与工程系的 Charles Elkan 教授共同担任。KDD 2018 有来自 51 个国家 1656 名注册参会人员。

SIGKDD 2017 研究领域的审核总论文数为 748 篇，收录 130 篇，包括 64 篇 oral，66 篇 poster，录用率分别占 8.6%及 8.8%。应用数据科学领域共审核 390 篇论文，收录 86 篇，包括 36 篇 oral，50 篇 poster，录用率分别占 9.2%和 12.6%。

SIGKDD 2017 大会邀请了微软研究院杰出科学家、哈佛大学的 Cynthia Dwork 教授，加州大学伯克利分校电子工程与计算机科学系的 Bin Yu 教授，多伦多大学计算机科学系的 Renée J. Miller 教授进行大会主题报告。

SIGKDD 2017 具体获奖情况为：

- 最佳论文

✧ 最佳研究型论文

标题：Accelerating Innovation Through Analogy Mining（通过类比挖掘加速创新）

作者：Tom Hope, Joel Chan, Aniket Kittur, Dafna Shahaf

论文解读：这篇论文探讨了学习大型概念资源库更简单的结构表征的可行性和价值，特别是“问题模式”，它规定了产品的目的，以及实现该目的的机制。论文中的方法结合众包和 CNN，提取产品描述中的目的和机制向量表示。

地址：

https://www.aminer.cn/archive/accelerating-innovation-through-analogy-mining/5992a1185ba2006b76482dcf

标题：Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data（多变量时间序列数据的 Toeplitz 逆协方差聚类）

作者：David Hallac, Sagar Vare, Stephen Boyd, Jure Leskovec

论文解读：文章利用期望最大化（EM）算法的一个变化，通过交替最小化来解决 Toeplitz Inverse Covariance-based Clustering（TICC）问题。通过动态规划和交替方向乘子方法，分别推导出了可扩展高效求解这两个子问题的闭型解。通过一系列综合实验将 TICC 与几个最先进的基线进行比较来验证该方法，然后在一个汽车传感器数据集上演示如何在真实场景中使用 TICC 来学习可解释的集群。

地址：

https://www.aminer.cn/archive/toeplitz-inverse-covariance-based-clustering-of-multivariate-time-series-data/599c798a601a182cd264948c

✧ 最佳应用数据科学论文

标题：HinDroid: An Intelligent Android Malware Detection System Based on Structured Heterogeneous Information Network（HinDroid：基于结构化的异构信息网络的智能 Android 恶意软件检测系统）

作者：Shifu Hou, Yanfang Ye, Yangqiu Song, Melih Abdulhayoglu

论文解读：针对安卓手机及安卓软件的爆炸式增长而带来的安卓恶意软件的增加问题，该论文使用基于元路径的方法来描述应用程序和 API 的语义相关性。并使用每一个元路径来制定一个在安卓应用程序上的相似性度量，且使用多核学习聚合不同的相似性。然后利用学习算法对每个元路径进行自动加权，进行预测。这是第一个使用结构化 HIN 进行安卓恶意软件检测的工作。实验表明本文开发的系统 HinDroid 优于其他的安卓恶意软件检测技术

地址：

https://www.aminer.cn/archive/hindroid-an-intelligent-android-malware-detection-system-based-on-structured-heterogeneous-information-network/59a02e2db161e8ad1a7b6db4

标 题 ： DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution（DeepSD：基于单个图像的超级解析生成高解析度的气候变换预测）

作者：Thomas Vandal, Evan Kodra, Sangram Ganguly, Andrew Michaelis, Ramakrishna Nemani, Auroop Ganguly

论文解读：论文提出了 DeepSD，一个广义堆叠超级分辨率卷积神经网络（SRCNN）框架用于气候变量的统计降级。DeepSD 为 SRCNN 增加了多尺度的输入通道，以最大限度地提高统计缩减的可预测性。该论文提供了美国大陆日降水量从 1 度（~100km）降至 1/8 度（~12.5km）的偏差校正空间分解和三种自动统计降水量方法的比较。此外，还讨论了一个使用美国国家航空航天局地球交换（NEX）平台的框架，用于在多个发射场景下缩小 20 多个 ESM 模型。

地址：

https://www.aminer.cn/archive/deepsd-generating-high-resolution-climate-change-projections-through-single-image-super-resolution/599c796a601a182cd263b2d1

✧ 最佳博士论文

标 题： Local Modeling of Attributed Graphs: Algorithms and Applications（属性图的局部建模：算法与应用）

作者： Bryan Perozzi 博士, 其导师为 Steven Skiena 教授

论文解读： 本文重点研究了属性图的可扩展算法和模型。数据可以被看作是离散的或者连续的，论文对两种情况都做了检验。具体来说，利用近期深度学习的研究成果展示线上学习算法，制作出丰富的图嵌入。这种新方法编码的社会关系多重尺度对于网络中的标签分类和回归任务是有用的。

地址：

http://perozzi.net/publications/16_thesis.pdf

标题： User Behavior Modeling with Large-Scale Graph Analysis（用户行为：基于大规模图形分析的建模）

作者： Alex Beutel 博士, 导师为 Alex Smola 教授和 Christos Faloutsos 教授

论文解读： 论文通过建模图理解用户行为。用户间的交互不只是在线上社交网络中，同时也直接影响着他们所处的世界，包括支持政治家、看电影、买衣服、搜索餐厅和看病。这些交互通常伴随具有深度信息的属性，例如交互时间、变化率或评价。

地址：

http://alexbeutel.com/papers/CMU-CS-16-105.pdf

标题： Mining Large Multi-Aspect Data: Algorithms and Applications（挖掘多方面数据：算法与应用）

作者：Evangelos E. Papalexakis 博士，其导师为 Christos Faloutsos 教授

论文解读：这篇论文中，提到了开发出具备快速、可扩展性与可解释性的算法（着重强调张量分析），并用于解决多种类、多方面的数据问题。

地址：

http://h5ip.cn/EDpe

✧ 时间检验奖（Test of Time Award）

标题：Training Linear SVMs in Linear Time，KDD 2006

作者：Thorsten Joachims 教授

论文解读：论文提出了第一种线性 SVM 的通用训练算法，这种缩放对于高维和稀疏数据非常有吸引力。该算法利用简单的切割平面算法来训练线性 SVM，该 SVM 被证明是收敛的。与以前的方法相比，该算法具有几个优点。第一，它非常简单易于实施；其次，它比现有的大分类问题分解方法快几个数量级；第三，该算法具有与训练误差直接相关的有意义的停止标准，这避免了将解决优化问题的时间浪费在不必要的更高精度上；最后，该算法可以处理大数据集的序数回归问题，开辟了几个新的研究领域。

地址：

https://www.aminer.cn/archive/training-linear-svms-in-linear-time/53e9a5afb7602d9702edbd06

✧ SIGKDD 创新和服务大奖

创新贡献奖：裴健博士（西蒙弗雷泽大学）

创新奖（Innovation Award）由裴健博士（西蒙弗雷泽大学，SFU）获得。他因对数据挖掘和应用基础的开创性贡献而受到认可，特别是在模式挖掘和空间数据挖掘方面。他是几种模式增长方法的主要发明者，包括 FP-growth 和 PrefixSpan，它们已被业界广泛使用，并被数据挖掘教科书和开源软件工具包

所采用。作为数据挖掘中被引用最多的作者之一，他的多产出版物被引用了数万次之多。

服务贡献奖：杨强博士（香港科技大学）

杰出服务奖（Service Award）由杨强博士（香港科技大学，HKUST）获得。为表彰杨强博士在服务和推广数据挖掘和人工智能领域的杰出的贡献。他曾担任 ACM KDD 2010 的 PC 联合主席，北京 ACM KDD 2012 大会主席和 2015 年 IJCAI PC 主席。以及 ACM IUI 2009，ACM RecSys 2013 和 IEEE 等会议的联合主席。他还担任过数据挖掘和人工智能的许多委员会主席，包括 2017 年 ACM SIGKDD 时间试卷奖委员会、2017 年 IJCAI 奖委员会和 2017 年 IEEE AI Ten-to-Watch 委员会。他是 ACM 智能系统与技术交易（ACM TIST）的创始主编，ACM TIST 已成为近期 ACM 中引用最多的期刊之一。他还创办了 IEEE 大数据交易期刊，并担任主编，是香港科技大学大数据研究所的创始董事。杨博士是中国数据挖掘、机器学习和 SIGKDD 的强力支持者。

✧ KDD CUP 竞赛

旅行时间预测（Travel Time Prediction）组：

第一名：Convolution 队（微软的 Ke Hu，北京航空航天大学的 Huan Chen，微软的 Pan Huang，美团公司的 Peng Yan）

第二名：好想有个队友队（浙江大学的 Huang Yide）

体积预测（Volume Prediction）组：

第一名：Convolution 队（微软的 Ke Hu，北京航空航天大学的 Huan Chen，微软的 Pan Huang，美团公司的 Peng Yan）

第二名：Black-Swan 队（京东的 Yitian Chen，华东师范大学的 Jie Zhou，南京理工大学的 Jie Lin，腾讯的 Hao Lin，京东的 Yang Guo）

KDD CUP 2017 由阿里巴巴集团云计算部门阿里巴巴云的团队组织，今年的竞赛题为“高速公路收费公路交通流量预测”，旨在通过数据驱动的先发制人措施，为交通管理部门提供支持，并为交通瓶颈的整体和现实解决方案铺平道

路。该竞赛旨在利用先进的数据技术解决所有行业的现实。在交通运输领域，京东开发了先进的能力，提供实时交通预测和中国旅游路线的建议，京东期待与世界各地的人才合作，以应对交通拥堵挑战。KDD CUP 2018 共分为旅行时间预测（Travel Time Prediction）与体积预测（Volume Prediction）两个子竞赛，共有来自世界各地的 3547 支队伍参加了比赛。Convolution 队（微软的 Ke Hu，北京航空航天大学的 Huan Chen，微软的 Pan Huang，美团公司的 Peng Yan）包揽两个冠军。

- 代表性论文

2017 年的 SIGKDD 大会的研究热点包括：机器学习、推荐系统和社交网络等，下文将介绍这些研究热点领域的代表性文章。

图 55 SIGKDD2017 论文研究热点的词云图

机器学习（Machine Learning）专题的代表性文章：

标题：Industrial Machine Learning

作者：Josh Bloom

论文解读：随着机器学习自动化和洞察力与复杂的工业数据系统相结合，效率、改善和挽救生命的真正革命将发生。利用从航空到运输的真实世界用例的系统视图，该论文对比了消费者和工业机器学习的需求和方法，并将重点放

在三个关键领域：将基于物理的模型与数据驱动的模型相结合、差异隐私和安全 ML（包括边缘到云策略）以及模型预测的可解释性。

地址:

https://www.aminer.cn/pub/59ae3c262bbe271c4c71f482/industrial-machine-learning

标题：TensorFlow Estimators: Managing Simplicity vs. Flexibility in High-Level Machine Learning Frameworks

作者：Heng-Tze Cheng, Zakaria Haque, Lichan Hong, Mustafa Ispir, Clemens Mewald, Illia Polosukhin, Georgios Roumpos, D. Sculley, Jamie Smith, David Soergel, Yuan Tang, Philipp Tucker, Martin Wicke, Cassandra Xia, Jianwei Xie

论文解读：该论文提出了一个用于指定、训练、评估和部署机器学习模型的框架，目的是为从业者简化尖端机器学习，以便将此类技术投入生产。该框架允许用户编写代码来定义模型，还提供了一个统一的估计器接口，使得编写独立于模型实现的下游基础设施成为可能。通过在不同的抽象层次上提供 api，使通用模型体系结构开箱即用，同时提供一个实用程序库来加速模型体系结构的实验，从而平衡对灵活性和简单性的竞争需求。

地址:

https://www.aminer.cn/pub/59a02d5eb161e8ad1a7b6d50/tensorflow-estimators-managing-simplicity-vs-flexibility-in-high-level-machine-learning-frameworks

社交网络（Social Network）专题的代表性文章：

标题：The Co-Evolution Model for Social Network Evolving and Opinion Migration

作者：Yupeng Gu, Yizhou Sun, Jianxi Gao

论文解读：社会网络演化和节点属性迁移通常被视为两个正交问题，分别进行研究。该论文提出了一个通过对这两个现象进行建模来闭合回路的协同进化模型，包含当节点属性已知时的网络生成模型和当社会网络结构已知时的属性迁移模型。仿真结果表明，该模型可以模拟广泛的现象，允许通过一系列因素。对国会立法法案联合赞助预测的应用证明了模型的有效性，预测结果优于一些最先进的基线。

地址：

https://www.aminer.cn/pub/59ae3c262bbe271c4c71f4f4/the-co-evolution-model-for-social-network-evolving-and-opinion-migration

标题：On Finding Socially Tenuous Groups for Online Social Networks

作者：Chih-Ya Shen, Liang-Hao Huang, De-Nian Yang, Hong-Han Shuai, Wang-Chien Lee, Ming-Syan Chen

论文解读：现有的社会群体发现研究主要集中在社会网络中的稠密子图上。然而，寻找社会弱势群体也有许多重要的应用。该论文引入 k-三角形的概念来度量群的脆弱性，提出了最小 k-三角形断开群（MkTG），设计了 TERA 和 TERA-ADV 两种算法，利用图论方法有效地解决一般图上的 MkTG 问题。该算法在效率和求解质量上都优于现有的方法。

地址：

https://www.aminer.cn/pub/5c8bea0a4895d9cbc6b8c7c1/on-finding-socially-tenuous-groups-for-online-social-networks

标题：Unsupervised Feature Selection in Signed Social Networks

作者：Kewei Cheng, Jundong Li, Huan Liu

论文解读：社交媒体数据特征选择算法往往只关注链接实例之间的正向交互而忽略负相关交互。该论文研究了签名社交网络中的一个新的无监督特征选择问题，并提出了一个新的 SignedFS 框架，提供了一种原则性的方法来模拟用户潜在表征学习的正负链接；重新审视了符号社交网络中的同伦和平衡理论，

并将符号图正则化引入特征选择框架中，以捕捉签名社交网络中用户之间的一阶和二阶邻近性。在社交网络上的实验证明了该框架的有效性。

地址：

https://www.aminer.cn/pub/59ae3c262bbe271c4c71f4df/unsupervised-feature-selection-in-signed-social-networks

推荐系统（Recommender Systems）专题：

标 题 ： Collaborative Variational Autoencoder for Recommender Systems

作者：Xiaopeng Li, James She

论文解读：基于现代推荐系统基于协作方法存在的稀疏性、冷启动等缺点，该论文提出了一种在多媒体场景下同时考虑推荐内容和评分的贝叶斯生成模型。该模型以无监督的方式从内容数据中学习深层次的潜在表征，并从内容和评分两方面学习项目与用户之间的隐含关系。该模型通过推理网络在潜在空间而不是观察空间中学习内容的潜在分布，并且可以很容易地扩展到文本以外的其他多媒体形式。该模型能够显著地优于现有的推荐方法，具有更强的鲁棒性。

地址：

https://www.aminer.cn/pub/59ae3c262bbe271c4c71f4bc/collaborative-variational-autoencoder-for-recommender-systems

标题：HoORaYs: High-order Optimization of Rating Distance for Recommender Systems

作者：Jingwei Xu, Yuan Yao, Hanghang Tong, Xianping Tao, Jian Lu

论文解读：潜在因子模型已经成为推荐系统中一种流行的方法，现有的方法都是建立在一阶评分距离原则的基础上的。该论文提出了一种新的推荐系统潜在因子模型（HoORaYs）。该方法的核心思想是探索高阶评分距离，其目的不仅在于最小化同一用户、项目对的估计和实际评分之间的差异，也有同一用户在不同项目上的估计和实际评分差异之间的差异。该论文提出了一个有效且可

扩展的算法来求解，通过探索高阶等级距离，有助于减少估计量的方差，从而获得更好的泛化性能。

地址：

https://www.aminer.cn/pub/59ae3c262bbe271c4c71f500/hoorays-high-order-optimization-of-rating-distance-for-recommender-systems

标题：Post Processing Recommender Systems for Diversity

作者：Arda Antikacioglu, R. Ravi

论文解读：该论文解决了在推荐系统中增加多样性的问题，将推荐系统设计描述为从潜在推荐的候选超级图中选择子图的问题，其中多样性和评级质量都得到了明确的优化：该论文定义了一个新的灵活的多样性概念，允许系统设计者规定每个项目应该收到的建议数量，并顺利地惩罚偏离这个分布的情况；证明了最小成本网络流方法在理论和实践中产生了快速算法来设计优化这种多样性概念的推荐子图。在 Netflix 和 MovieLens 的标准评级数据集上展示了新模型和方法在保持高评级质量的同时增加多样性的有效性。

地址：

https://www.aminer.cn/pub/59a02e1db161e8ad1a7b6da9/post-processing-recommender-systems-for-diversity

此外，SIGKDD 2017 还评选了观众最欣赏奖，由亚马逊的 Daniel Hill 等人获得，具体如下：

标题：An efficient bandit algorithm for realtime multivariate optimization

作者：Daniel N. Hill, Houssam Nassif, Anand Iyer, Yi Liu, S.V.N. Vishwanathan.

论文解读：通常采用优化来确定网页的内容，然而随机实验无法很好地适应指数级大决策空间上的组合问题，因此，在实践中，通常仅限于一次优化网页的一个方面。这在实验的速度和布局决策之间可能的交互的利用方面都错失

了机会。该论文着重于交互式网页的多元优化制定了页面的不同组件之间的可能交互被显式建模的方法。该论文进一步应用算法来优化促进采用亚马逊服务的消息，实现了 21%的转化率增长。

地址：

https://www.aminer.cn/archive/an-efficient-bandit-algorithm-for-realtime-multivariate-optimization/59a02d63b161e8ad1a7b6d54

2.10.6 SIGKDD 2018

2018 年 8 月 19-23 日，第 24 届国际知识发现与数据挖掘大会（SIGKDD 2018）在英国伦敦召开。

本届大会主席由伦敦帝国理工学院计算机系计算科学教授、帝国理工学院数据科学研究所的创始主任、tranSMART 基金会的首席技术官 Yi-Ke Guo 教授，和 IBM Watson Watson Health 小组的首席科学家和高级经理 Faisal Farooq 博士共同担任，副主席由时任清华大学计算机科学与技术系教授唐杰担任；研究性程序委员会主席由国立台湾大学计算机科学与信息工程学院 Chih-Jen Lin 教授和新泽西州立大学管理科学与信息系统系的 Hui Xiong 教授共同担任；工业界程序委员会委员会主席由 Google 的杰出科学家 Andrei Broder 博士和马格德堡计算机科学系的 Myra Spiliopoulou 教授共同担任。共有来自 99 个国家的 3377 名学者注册了 KDD 2018。

KDD2018 共收到投稿论文 1479 篇，其中研究性论文 983 篇，创下新高，其中有 107 篇 oral，录用率 10.9%；74 篇 poster，录用率 7.5%。共收到应用数据科学论文 496 篇，创下新高，其中有 40 篇 oral，录用率 8.0%；72 篇 poster，录用率 14.5%。

KDD 2018 大会邀请了伦敦帝国理工学院数学的高级研究员和荣誉教授 David Hand，斯坦福大学的经济学教授、哈佛大学经济与工商管理荣誉教授、2012 年诺贝尔经济学奖得主之一的 Alvin E.Roth，牛津大学统计系的统计机器学习教授、DeepMind 的研究科学家 Yee Whye The，Avanessians 数据科学研究

所所长、哥伦比亚大学计算机科学系的 Jeannette M. Wing 教授，进行了四场大会主题报告。

SIGKDD 2018 具体获奖情况为：

- 最佳论文

✧ 最佳研究型论文

标题：Adversarial Attacks on Classification Models for Graphs（对图形分类模型的对抗性攻击）

作者：Daniel Zügner, Amir Akbarnejad, Stephan Günnemann

论文解读：为了解决离散域问题，提出了一种利用增量计算的高效算法。我们的实验研究表明，即使只进行少量的扰动，节点分类的准确率也会显著下降。更重要的是，我们的攻击是可转移的：学到的攻击方式可以推广到其他最先进的节点分类模型。

地址：

http://h5ip.cn/uy96

标题：XiaoIce Band: A Melody and Arrangement Generation Framework for Pop Music（XiaoIce Band：流行音乐的旋律与编曲生成框架）

作者：Hongyuan Zhu, Qi Liu, Nicholas Jing Yuan, Chuan Qin, Jiawei Li, Kun Zhang, Guang Zhou, Furu Wei, Yuanchun Xu, Enhong Chen

论文解读：音乐对人们的生活有着重要的影响。然而，创作音乐需要大量的专业知识和技能。近年来，如何利用机器学习技术自动进行音乐创作成为人工智能领域的热门话题。由于音乐元素的复杂性，如歌曲不同的和弦进行、乐段中结构鲜明的节奏型、不同特性的音轨（乐器）需要保持和谐一致等，使得高质量的单音轨作曲、多音轨编曲算法的设计充满了挑战性与特殊性。为此，论文基于深度神经网络和多任务学习等方法，从历史音乐数据（如十万多首歌曲）中学习音乐的音程关系、结构以及各种乐器的演绎特色，设计了一种基于和弦的节奏和旋律交叉的生成模型（CRMCG）来产生带有和弦进行的旋律；更进

一步，通过构建多个任务（即多个音轨，乐器序列）关联模型，为乐器的相互配合搭建了信息交互的桥梁，实现了一种多乐器联合编曲模型（MICA）。

地址：

https://www.aminer.cn/archive/xiaoice-band-a-melody-and-arrangement-generation-framework-for-pop-music/5b67b45517c44aac1c8607e9

✧ 最佳应用数据科学论文

标题： Real-time Personalization using Embeddings for Search Ranking at Airbnb（在 Airbnb 中使用嵌入式做搜索排名来实现实时的个性化）

作者： Mihajlo Grbovic（Airbnb）, Haibin Cheng（Airbnb）

论文解读： 嵌入式模型是为 Airbnb 量身定制的，能用于捕捉顾客的短期或长期兴趣，传递有效的住房推荐清单。在完全将他们部署到产品线上前，我们已经做了严谨的线上线下测试。

地址：

https://www.aminer.cn/archive/real-time-personalization-using-embeddings-for-search-ranking-at-airbnb/5b67b45517c44aac1c8607cb

标题： ActiveRemediation: The Search for Lead Pipes in Flint, Michigan（积极整治：密歇根弗林特主管道搜索）

作者： Jacob Abernethy（Georgia Institute of Technology）, Alex Chojnacki（University of Michigan）, Arya Farahi（University of Michigan - Ann Arbor）, Eric Schwartz（University of Michigan）和 Jared Webb（Brigham Young University）

论文解读： 除了统计与机器学习方法外，我们基于收入信息自适的统计模型，与政府就需要检验和替换的管道，提供推荐名单，进行沟通。最后，根据关于增加联邦政府基础建设投入的讨论，我们探索出了适用于全国范围内的方法。

地址：

https://www.aminer.cn/archive/activeremediation-the-search-for-lead-pipes-in-flint-michigan/5b67b45517c44aac1c860889

✧ 最佳博士论文

标题：Mining Entity and Relation Structures from Text: An Effort-Light Approach（从文本中挖掘实体和文本关系：一种轻松的方法）

作者：Xiang Ren 博士，其导师为韩家炜（Jiawei Han）教授

论文解读：理解信息社会大量不同形式的文本数据需要进行深入的内容分析。要将庞大的，非结构化的文本语料库转变为机器可读的知识，最大的挑战之一是要了解语料库中的类型化实体和关系结构。本文的重点是开发一种有原则的、可扩展的方法来提取类型化实体与人类注释的关系，以克服处理各种领域、流派和语言的文本语料库的障碍。

地址：

https://www.aminer.cn/pub/5ecfaaf49e795eb20a615022/mining-entity-and-relation-structures-from-text-an-effort-light-approach

标题：Probabilistic Models for Credibility Analysis in Evolving Online Communities（用于线上社区可信度分析的概率模型）

作者：Subhabrata Murherjee 博士，其导师为 Gerhard Weikum 教授

论文解读：我们提出了概率图形模型，它可以利用在线社区中多个因素之间的联合相互作用——比如用户交互、社区动态和文本内容——来自动评估用户贡献的在线内容的可信性，以及用户的专业知识及其演变。

地址：

https://d-nb.info/1137509414/34

标题：Characterization and Detection of Malicious Behavior on the Web（网页中恶意行为的特性描述和检测）

作者：Srijan Kumar 博士，其导师为 V.S. Subrahmanian 教授

论文解读：本文描述开发出的一套五种模型和算法，用于在多个 web 平台上准确地识别和预测几种不同类型的恶意行为——汪达尔人、恶作剧者、傀儡、巨魔和欺诈评论者。这些算法的分析对网络上的恶意行为形成了可解释的理解。

地址：

https://www.aminer.cn/pub/5c2c7a6717c44a4e7cf306d7/characterization-and-detection-of-malicious-behavior-on-the-web

✧ 时间检验奖（Test of Time Award）

标 题：Factorization meets the neighborhood: a multifaceted collaborative filtering model，KDD 2008

作者：Yehuda Koren

论文解读：这篇文章被奉为利用矩阵分解所推荐的教科书级论文，作者也是 Netflix 大赛得了 100 万美元的团队主力。论文提出了一个单一的框架，将邻域模型与潜在因子模型（现在通常称为“嵌入模型”）相结合，从而利用两种方法的优势：邻域模型在检测非常局部化的关系时最有效，而潜在因子模型则更多有效地估计与大多数或所有项目同时相关的整体结构。直到今天仍然有着它的实用意义。

地址：

https://www.aminer.cn/archive/factorization-meets-the-neighborhood-a-multifaceted-collaborative-filtering-model/53e9bc74b7602d97048f4169

✧ SIGKDD 创新和服务大奖

创新贡献奖：刘兵教授（伊利诺伊大学芝加哥分校）

创新奖（Innovation Award）由伊利诺伊大学芝加哥分校的刘兵教授获得。刘兵教授是数据挖掘重量级人物，是语义分析、观点挖掘研究领域的开创者之

一。他是伊利诺伊大学芝加哥分校（University of Illinois at Chicago，简称 UIC）计算机科学系优秀教授，于爱丁堡大学获得人工智能博士学位。他曾于 2013 年担任 ACM SIGKDD 的主席（2013 年 7 月 1 日至 2017 年 6 月 30 日）。曾担任许多其他数据挖掘领域会议主席（包括 ICDM，CIKM，WSDM，SDM 和 PAKDD），他担任 TKDE，TWEB，DMKD 等期刊的副主编；ACM、AAAI 和 IEEE 的会士（Fellow）。

服务奖：唐杰教授（清华大学）

2018 服务奖（Service Award）由清华大学的唐杰获得。唐杰教授是 AMiner 研发者。他是清华大学计算机科学与技术系教授。他发表了 200 多篇期刊/会议论文，拥有 20 项专利，引用超过 10,000。他曾担任 CIKM'16，WSDM'15，ASONAM'15，SocInfo'12，KDD 2018 副主席，组织 KDD'11-18 联合主席，IEEE TKDE / TBD 副主编的 PC 联合主席和 ACM TKDD / TIST。

✧ KDD CUP 竞赛

第一名：First floor to eat Latiao 队（中南大学的 Haoran Jiang 和 Binli Luo，北京邮电大学的 Jindong Han, Juan Liu, and Qianqian Zhang）

第二名：Getmax 队（微软的 Zhipeng Luo, 北京大学的 Jianqiang Huang, 阿里巴巴的 Ke Hu）

KDD CUP 2018 基于预测伦敦及北京空气质量问题，组委会在比赛中提供中国北京和英国伦敦的数据。比赛选手需要预测未来 48 小时内 PM2.5，PM10 和 O3 的浓度（伦敦只需要预测 PM2.5 和 PM10）。今年比赛共有来自 4180 支团队的 5687 位参赛者参与，覆盖超过 3000 所大学/机构的 49 个国家。First floor to eat Latiao 获得本次赛事的第一名。由罗志鹏（微软），胡可（阿里巴巴），黄坚强（北京大学）组成的 Getmax 团队今年 KDD Cup 获得两项第一，一项第二，是唯一包揽三项大奖的团队。去年该团队成员带领的 Convolution 团队也包揽了 KDD CUP 2017 的双料冠军。

除此之外，SIGKDD 2018 还颁发了最佳审稿人奖和初创团队奖。最佳审稿人奖由 George Valkanas 获得，这一奖项经由 SPC 提名，审稿质量及相应讨论表

现突出。初创团队奖（startup award）六支获奖团队分别是智能一点、ASTOUND、氪信科技、FactMATA、Infilect 及 Zugata。

- 代表性论文

2018 年的 KDD 论文还是分为两个 Track，分别是 Research Track 和 Applied Data Science （ADS） Track。Research Track 接收的论文中，Deep Learning、Representation、Embedding 相关的论文达到了 46 篇，是此次会议最受关注的主题，Supervised、 Unsupervised、 Transfer Learning 相关的论文有 41 篇紧随其后，Graph、Social、Temporal、Spatial 相关的论文也有 31 篇。下面介绍深度神经网络、深度学习和推荐系统专题的代表论文。

图 56 SIGKDD2018 论文研究热点的词云图

深度神经网络（Deep Neural Network）专题：

标题：StepDeep: A Novel Spatial-temporal Mobility Event Prediction Framework based on Deep Neural Network

作者：Bilong Shen, Xiaodan Liang, Yufeng Ouyang, Miaofeng Liu, Weimin Zheng, Kathleen M. Carley

论文解读：近年来，许多移动模式挖掘方法忽略了 POI 之间内在的空间和时间模式相关性，导致在不同的场景下泛化较差。该论文提出了一个基于深度神经网络（StepDeep）的时空移动事件预测框架，提出了一种可以自然地编码每个 POI 的空间和时间依赖性的新公式。StepDeep 因此通过将新的时间敏感卷积滤波器、空间敏感卷积滤波器和时空敏感卷积滤波器合并到一个网络中来预测时空事件。StepDeep 比现有的五个基线具有更高的预测精度，可以应用于许多时空事件预测场景。

地址：

https://www.aminer.cn/pub/5b67b45517c44aac1c860865/stepdeep-a-novel-spatial-temporal-mobility-event-prediction-framework-based-on-deep

深度学习（Deep Learning）专题：

标题：SHIELD: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression

作者：Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis, Duen Horng Chau

论文解读：基于深度神经网络（DNNs）对实用防御技术的迫切需要，该论文将 JPEG 压缩作为 SHIELD 防御框架的核心，利用其功能有效“压缩”此类像素操作。SHIELD 通过使用压缩图像对模型进行再训练来“接种”该模型，在压缩图像中应用不同的压缩级别以生成最终在整体防御中一起使用的多个接种模型。疫苗接种，组合和随机化的新颖组合使 SHIELD 成为了强化的多管齐下的防御系统。结果表明，该方法可消除高达 98%的由强大的对抗性技术提供的灰盒攻击。

地址：

https://www.aminer.cn/pub/5a9cb65d17c44a376ffb831b/shield-fast-practical-defense-and-vaccination-for-deep-learning-using-jpeg-compression

标题：Active Deep Learning to Tune Down the Noise in Labels

作者：Karan Samel, Xu Miao

论文解读：面对监督学习性能受到标签中错误限制的问题，该论文提出了一种新的主动深度去噪（ADD）方法，该方法首先建立 DNN 噪声模型，然后采用主动学习算法来确定最佳去噪函数。在低噪声条件下，只需要用 logn 示例查询 oracle，其中 n 是数据中的总数。企业应用实例表明，该方法能有效地降低预测误差的 1/3，且经 oracle 验证的实例仅为 0.1%。

地址：

https://www.aminer.cn/pub/5b67b45517c44aac1c8607ae/active-deep-learning-to-tune-down-the-noise-in-labels

推荐系统（Recommender System）专题：

标 题 ： Graph Convolutional Neural Networks for Web-Scale Recommender Systems

作者：Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, Jure Leskovec

论文解读：面对将图结构数据的深层神经网络扩展到 web 级的挑战。该论文开发了一种数据有效的图卷积网络（GCN）算法，该算法结合有效的随机游动和图卷积来生成包含图结构和节点特征信息的节点（即项目）的嵌入。与已有的 GCN 方法相比该算法可以训练和嵌入比一般 GCN 实现大四个数量级的图形，是迄今为止深度图嵌入的最大应用，为新一代基于图卷积体系结构的 web 级推荐系统铺平了道路。

地址：

https://www.aminer.cn/pub/5b67b45517c44aac1c860876/graph-convolutional-neural-networks-for-web-scale-recommender-systems

标题：Learning Tree-based Deep Model for Recommender Systems

作者：Han Zhu, Pengye Zhang, Guozheng Li, Jie He, Han Li, Kun Gai

论文解读：针对推荐系统的计算成本问题，该论文在提出了一种新的基于树的方法。主要思想是通过以自上而下的方式遍历树节点并为每个用户节点对做出决策，从粗到细地预测用户的兴趣。该论文发现树结构可以联合学习，以更好地适应用户的兴趣分布，从而便于训练和预测。实验表明，该方法明显优于传统方法。在淘宝展示广告平台上测试结果也证明了该方法的有效性。

地址：

https://www.aminer.cn/pub/5a9cb66717c44a376ffb8969/learning-tree-based-deep-model-for-recommender-systems

标 题 ： xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems

作者：Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, Guangzhong Sun

论文解读：近年来，学界已经提出了几种基于 DNN 的因子分解模型，但普通 DNN 存在在位级别上隐式地生成特征交互的问题。该论文提出了一个新的压缩交互网路（CIN），其目的是以明确的方式在向量层面产生特征交互。将一个 CIN 和一个经典的 DNN 结合成一个统一的模型 xDeepFM，能够显式地学习某些有界度的特征交互，可以隐式地学习任意的低阶和高阶特征交互。

地址：

https://www.aminer.cn/pub/5aed14d617c44a44381594f8/xdeepfm-combining-explicit-and-implicit-feature-interactions-for-recommender-systems

2.10.7 SIGKDD 2019

2019 年 8 月 4-8 日，第 25 届国际知识发现与数据挖掘大会（SIGKDD 2019）在美国阿拉斯加州市安克雷奇市召开。

本届大会主席由明尼苏达大学终身教授、知识发现与数据挖掘（KDD）领域的最高技术荣誉 ACM SIGKDD 创新奖得主 Vipin Kumar 教授和华盛顿塔科马大学工程技术学院的计算机科学与系统教授、数据科学中心的创始主任、前 ACM SIGKDD（知识发现和数据挖掘特别兴趣小组）信息官 Ankur Teredesai 教授共同担任。研究赛道主席由明尼苏达大学计算机科学与工程系教授 George Karypis 教授和波士顿大学计算机科学系教授 Evimaria Terzi 教授担任，应用赛道主席由美国波士顿大学计算机科学博士 Romer Rosales 和加拿大不列颠哥伦比亚大学计算机科学博士 Ying Li 担任。应用数据科学特邀讲座主席由新泽西州立大学罗格斯分校管理科学和信息系统系 Hui Xiong 教授、罗格斯大学计算机科学专业 Sofus Macskássy 博士、康奈尔大学计算机科学系 Johannes Gehrke 教授、和 Deepak Argwal 担任，应用数据科学特邀小组主席则由 IBM 研究部的 Chid Apte 担任。KDD 2019 共有 3150 位注册参会者，他们来自全世界 51 个国家，其中有 443 位学生。

相关数据统计显示，2019 年 KDD 的论文评审分为研究（Research track）和应用（Applied track）两大赛道，其中研究赛道共收到约 1200 篇投稿，最终约 110 篇被收录为 Oral 论文，60 篇被收为 Poster 论文，入选率约 14.2%。应用赛道则收到 700 余篇投稿，其中 45 篇被接收录为 Oral 论文，100 篇被收为 Poster 论文，入选率为 20.7%。

KDD 是数据挖掘领域国际最高级别的会议。和往年不同的是，2019 年的 KDD 采用的是双盲评审，且着重强调论文的“可复现性”，即论文之外还需额外提交内容展示可复现性，包括实验方法、经验评估和结果等，甚至在论文中展现研究代码和数据，所用的算法和资源。

SIGKDD2019 具体获奖情况为：

- 最佳论文

✧ 最佳研究型论文

标题：Network Density of States

作者：Kun Dong, Austin R. Benson, David Bindel

地址：

https://www.aminer.cn/pub/5cf48a1eda56291d5827f21c/network-density-of-states

论文解读：该篇论文获得“Research Track”Best Paper 奖。在本文中，研究者深入探索了真实世界图谱谱密度的核心。论文借用了凝聚态物理学中开发的工具，并添加了新的适应性来处理常见图形的谱特征。论文计算了单个计算节点上超过 10 亿个边的图的谱密度，证明所得到的方法非常高效。除了提供视觉上引人注目的图形指纹之外，研究者还展示了谱密度的估计如何简化许多常见的中心度量的计算，并使用谱密度估计关于图结构的有意义信息，这些信息不能仅从极值特征对推断出来。

标题：Optimizing Impression Counts for Outdoor Advertising

作者：Yipeng Zhang, Yuchen Li, Zhifeng Bao, Songsong Mo, and Ping Zhang

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb3f/optimizing-impression-counts-for-outdoor-advertising

论文解读：该篇论文获得“Research Track”Best Paper 奖。本文提出并研究了在考虑印象数量的情况下，如何优化户外广告的影响。给定一个广告牌数据库 U，其中每个广告牌都有一个位置和一个非统一成本，一个轨迹数据库 T 和一个预算 B，它的目标是找到在预算下影响最大的一组广告牌。根据广告消费者行为研究，我们在定义影响度量时，采用逻辑函数考虑广告（放置在不同广告牌上）对用户轨迹的印象计数。然而，这带来了两个挑战：a）我们的问题是 np 难在系数近似 O（T||1- ε ）对于任何 ε >0 在多项式时间内；b）影响度量是非子模的，不适用直接的贪婪方法。因此，我们提出一种基于切线的算法来计算子模函数来估计影响的上界。从今以后,我们引入一个与 θ -termination 条件和框架,实现 θ 2/（1-1/e）近似比。但是，当|U|很大时，这个框架非常耗时。因此,我们对其进行进一步优化进步修剪上界估计方法实现 θ 2/（1-1/e- ε ）

近似比例并显著降低了时间复杂度。我们在真实的广告牌和轨迹数据集上进行了实验，并表明所提出的方法的有效性比基线高出 95%。此外，优化的方法比原来的框架快了大约两个数量级。

✧ 最佳应用型论文

标题：Action Speaks Louder Than Goals: Value Player Actions in Soccer

作者：Tom Decroos, Lotte Bransen, Jan Van Haaren, Jesse Davi

地址：

https://www.aminer.cn/pub/5c8d2ff84895d9cbc643c2c4/actions-speak-louder-than-goals-valuing-player-actions-in-soccer

论文解读：该篇论文获得“Applied Data Science Track”Best Paper 奖。评估足球运动员比赛中个人行动所造成的影响是评估他们的重要指标。然而，大多数传统指标在解决此类任务时效果都不尽如人意，因为它们只关注整场比赛中仅有的几次特殊动作，比如射门和进球，而忽视了行动的背景。研究人员提出的方法包括：一种用于描述球场上各个球员动作的新语言；基于它对比赛结果的影响来评估任何类型球员动作的框架，同时考虑了动作发生的背景。

标题：Developing Measures of Cognitive Impairment in the Real World from Consumer-Grade Multimodal Sensor Streams

作者：Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, Marc Sunga, Han Hee Song, Hyun Joon Jung, Belle Tseng, Andrew Trister

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97dea9e/developing-measures-of-cognitive-impairment-in-the-real-world-from-consumer-grade

论文解读：该篇论文获得“Applied Data Science Track”Best Paper 奖。可穿戴消费设备和移动计算平台（智能手机、智能手表、平板电脑）的普遍性和显著的技术进步，以及大量可用的传感器模式，使患者及其日常活动的持续监测成为可能。这些丰富的纵向信息可以挖掘认知障碍的生理和行为特征，为及时、经济有效地检测 MCI 提供新的途径。在这项工作中，我们提出了一个平台，远程和不引人注目的监测症状相关的认知障碍使用几个消费级智能设备。我们演示了在 Lilly 探索性数字评估研究（12 周的可行性研究，监测了 31 名有认知障碍和 82 名无认知障碍的自由生活条件下的人）中，该平台如何收集了总计 16TB 的数据。我们描述了如何小心地统一数据、时间对齐和输入技术，以处理真实环境中固有的数据缺失率，并最终显示这些不同数据的效用，以区分症状与健康的控制，其基础是完全从设备数据计算出来的特征。

✧ 观众赞赏奖

标题：Temporal Probabilistic Profiles for Sepsis Prediction in the ICU

作者：Eitam Sheetrit, Nir Nissim, Denis Klimov, Yuval Shahar

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb62/temporal-probabilistic-profiles-for-sepsis-prediction-in-the-icu

论文解读：脓毒症很难预测，诊断和治疗，因为它涉及分析不同的多元时间序列集，通常存在数据丢失，采样频率不同和随机噪声等问题。在这里，我们提出了一个新的基于动态行为的模型，我们将其称为时间概率 proFile（TPF），用于多元时间序列的分类和预测任务。在 TPF 方法中，原始的带时间戳的数据首先被抽象为一系列更高级的有意义的概念，这些概念在表征时间段的时间间隔内保持不变。然后，我们发现了数据中频繁重复的时间模式。使用发现的模式，我们创建了整个实体总体，其中每个目标类别以及每个实体的时间模式的概率分布。然后，我们使用 TPF 作为元特征，通过测量它们与每个类别的汇总 TPF 或每个实体的单独 TPF 的 TPF 距离，对新实体的时间序列进行分

类或预测其结果。我们在大型基准临床数据集上的实验结果表明，TPF 改善了败血症的预测能力，并且比其他机器学习方法表现更好。

✧ SIGKDD 创新和服务大奖

ACM SIGKDD 创新奖是知识发现和数据挖掘（KDD）领域技术卓越的最高奖项。它授予个人或一组合作者，他们在 KDD 领域的杰出技术创新对推动该领域的理论和实践产生了持久的影响。这些贡献必须显著影响该领域的研究和开发方向，或者必须以重大和创新的方式转移到实践中或使商业系统得以发展。

Charu Aggarwal 博士是其 2019 年创新奖的获得者。他是 IBM TJ Watson 研究中心的杰出研究人员，并因其在高维数据、隐私、数据流、不确定数据、图形、文本挖掘和社交网络方面的研究贡献而受到认可。

ACM SIGKDD 服务奖是知识发现和数据挖掘（KDD）领域中最高的服务奖。它被授予个人或一组合作者，以表彰他们在知识发现和数据挖掘领域的杰出专业服务和贡献。

BM Watson Health 的董事兼杰出工程师 Balaji Krishnapuram 因通过开发机器学习产品以改善医疗保健为社会做出贡献而获得 2019 年服务奖。

✧ 最佳学生毕业论文奖

标题：Data Science for Human Well-Being

作者：Tim Althof

论文解读：可穿戴设备和移动设备（包括智能手机和智能手表）的普及已经产生了详细的行为数据。这些巨大的数字痕迹为我们提供了无与伦比的机会来实现新型的科学方法，从而使人们对我们的生活，健康和幸福有了新的见解。但是，要从这些数据中获得可行的见解，就需要新的计算方法，这些方法必须将观测的、科学上的“弱”数据转化为强大的科学结果，并可以通过计算来大规模地测试领域理论。在这篇论文中，我们描述了新颖的计算方法，该方法利用了数百万人采取的数十亿行动规模的数字活动轨迹。这些方法结合了来自数据挖掘，社交网络分析和自然语言处理的见解，以增进我们对身心健康的理解：

a）我们展示了大规模的数字活动痕迹如何揭示世界范围内未知的健康不平等现象；b）个性化的预测模型如何支持有针对性的干预措施以消除这种不平等；c）我们证明，对用户搜索引擎交互的速度进行建模可以提高我们对睡眠和认知表现的理解；d）最后，我们描述自然语言处理方法如何帮助改善数百万处于危机中的人们的咨询服务。

标题：Multidimensional Mining of Unstructured Data with Limited Supervision

作者：Chao Zhang

论文解读：非结构化文本数据是最重要的数据形式之一，在从社交网络和信息检索到医疗保健和科学研究等领域的数据驱动型决策中发挥着至关重要的作用。但是，将大量文本数据转换为多维知识仍然是现有数据挖掘技术无法轻易解决的挑战。在本文中，我们提出了在有限的监督下将非结构化文本数据转换为多维知识的算法。我们研究两个核心问题：a）如何在多个维度中通过声明性查询识别与任务相关的数据；b）如何从多维空间的数据中提取知识。

为了解决上述问题，我们提出了一个集成的多维数据集构建和开发框架。当将文本数据转换为多维知识时，提出的框架具有两个明显的优势：灵活性和标签效率。首先，它可以灵活地获取多维知识，因为多维数据集结构允许用户轻松地沿多维方向以不同的粒度识别与任务相关的数据，并进一步提取多维知识。其次，用于多维数据集构建和开发的算法几乎不需要监督。这使该框架吸引了许多应用程序，这些应用程序需要获取标记数据，而这些数据非常昂贵。

✧ KDD 杯创新奖

2019 年的 KDD CUP 比赛共有超过 2800 支注册队伍参赛。这些队伍来自 39 个国家和地区，包括了 230 个学术和研究机构，参赛人员在 5000 人以上，总共提交了超过 17000 次结果。今年的 KDD CUP 分为三个赛道，分别是：

常规机器学习竞赛（Regular Machine Learning Competition）

自动机器学习竞赛（Automated Machine Learning Competition）

「Research for Humanity」强化学习竞赛（「Research for Humanity」Reinforcement Learning Competition Track）

常规机器学习竞赛常规机器学习竞赛由百度赞助，分为两个任务：

任务一：最适合的交通方式推荐

任务二：开放研究/应用挑战

任务一的冠军团队成员为：Shiwen Cui, Changhua Meng, Can Yi, Weiqiang Wang, Xing Zhao, Long Guo，来自蚂蚁金服。

亚军团队成员为：Hengda Bao, Jie Zhang, Wenchao Xu, Qiang Wang, Jiayuan Xie, He Wang, Ceyuan Liang，来自上海微盟、趋势科技、滴滴、北京邮电大学、华南理工大学、京东等机构。

任务二的冠军团队成员为：Keiichi Ochiai, Tsukasa Demizu, Shin Ishiguro, Shohei Maruyama, Akihiro Kawana, 来自日本 NTT DOCOMO 公司。

亚军团队成员为：Yang Liu, Cheng Lyu, Zhiyuan Liu，来自东南大学。

PaddlePaddle 特别奖

获奖者为：Xianfeng Liang, Likang Wu, Joya Chen, Yang Liu, Runlong Yu, Min Hou, Han Wu, Yuyang Ye, Qi Liu, Enhong Chen，来自中国科学技术大学。

自动机器学习竞赛由第四范式赞助：

冠军团队成员为：Zhipeng Luo, Jianqiang Huang, Mingjian Chen, Bohang Zheng，来自 DeepBlueAI 和北京大学。

亚军团队成员为：Chengxi Xue, Shu Yao, Zeyi Wen, Bingsheng He，来自新加坡国立大学。

「Research for Humanity」强化学习竞赛由 IBM Africa 和 Hexagon-ML.com 赞助：

冠军团队成员为：Zi-Kuan Huang, Jing-Jing Xiao, Hung-Yu Kao，来自中国台湾国立成功大学。

亚军团队成员为：Lixin Zou, Long Xia, Zhuo Zhang, Dawei Yin，来自清华大学、京东和北京航空航天大学。

为促进创新和鼓励社会各界，2019 年 KDD 建立了 KDD 杯创新奖。在 2019 年，KDD 杯创新奖授予了强化学习竞赛平台赞助商 Hexagon-ML。

Hexagon-ML 获得此奖项的原因是：开创了一种新型竞赛，即强化学习竞赛，并在新颖的计算环境下为 KDD Cup 2019 实施该竞赛。与 IBM Research Africa 的成功合作为解决人类疟疾问题做出了贡献。在强化学习中收集、开发和发展数据科学社区的工作。

- 代表论文：

图 57 SIGKDD2019 论文研究热点的词云图

深度学习（Deep Learning）专题：

标题：ADMM for Efficient Deep Learning with Global Convergence

作者：Junxiang Wang, Fuxun Yu, Xiang Chen, Liang Zhao

论文解读：针对交替方向乘子法（ADMM）存在的缺乏全局收敛保证、收敛速度慢和特征维的立方时间复杂性等挑战。该论文提出了一种新的深度学习优

化框架 dlADMM，证明了一种基于 ADMM 的方法（dlADMM）在温和条件下处理深层神经网络问题的全局收敛性。实验表明，该算法优于大多数比较方法。

地址：

https://www.aminer.cn/pub/5d04e910da56295d08de1619/admm-for-efficient-deep-learning-with-global-convergence

标题：DeepGBM: A Deep Learning Framework Distilled by GBDT for Online Prediction Tasks

作者：Guolin Ke, Zhenghui Xu, Jia Zhang, Jiang Bian, Tie-Yan Liu

论文解读：针对梯度提升决策树（GBDT）和神经网络（NN）方法的不足，该论文提出了一个新的学习框架 DeepGBM，它通过使用 CatNN 和 GBDT2NN 两个 NN 组件来集成 NN 和 GBDT 的优点。在这两个组件的支持下，DeepGBM 可以同时利用分类和数字特性，同时保持高效的在线更新能力。实验表明，DeepGBM 在各种在线预测任务中的性能优于其他公认的基线。

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb82/deepgbm-a-deep-learning-framework-distilled-by-gbdt-for-online-prediction-tasks

标题：Sherlock: A Deep Learning Approach to Semantic Data Type Detection

作者：Madelon Hulsebos, Kevin Hu, Michiel Bakker, Emanuel Zgraggen, Arvind Satyanarayan, Tim Kraska, Çağatay Demiralp, César Hidalgo

论文解读：现有的数据准备和分析系统依赖字典查找和正则表达式匹配来检测语义类型，通常对脏数据不稳健，并且只检测有限数量的类型。该论文介绍了一种用于检测语义类型的多输入深层神经网络 Sherlock。该方法的支持加权超过了机器学习基线、字典和正则表达式基准以及众包注释的共识。

地址：

https://www.aminer.cn/pub/5d04e8d8da56295d08dafa5b/sherlock-a-deep-learning-approach-to-semantic-data-type-detection

异常检测（Anomaly Detection）专题：

标题：Deep Anomaly Detection with Deviation Networks

作者：Guansong Pang, Chunhua Shen, Anton van den Hengel

论文解读：深度学习对于异常检测的研究相对有限，存在数据学习效率低下和异常评分次优的问题，该论文介绍了一种新的异常检测框架及其实例，通过神经偏差学习来实现异常分数的端到端学习。结果表明，该方法可以有效地训练更多的数据，并且能获得更好的异常评分。

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb05/deep-anomaly-detection-with-deviation-networks

标题：Fast and Accurate Anomaly Detection in Dynamic Graphs with a Two-Pronged Approach

作者：Minji Yoon, Bryan Hooi, Kijung Shin, Christos Faloutsos

论文解读：针对异常类型分类并从理论上分析每种类型所产生的异常征兆的问题，该论文提出了一个用于动态图异常检测的在线算法 AnomRank。该算法使用了一种双管齐下的方法来定义两种新的异常性指标。每个度量、跟踪其自身版本的“节点得分”函数的导数，能够检测到任何节点重要性的突然变化。该方法快速准确，具有可伸缩性，同时在理论上十分可靠。

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb66/fast-and-accurate-anomaly-detection-in-dynamic-graphs-with-a-two-pronged

标题：Sequential Anomaly Detection using Inverse Reinforcement Learning

作者：Min-hwan Oh, Garud Iyengar

论文解读：针对异常检测中序列数据检测问题，该论文提出了一种基于反向强化学习（IRL）的端到端序列异常检测框架，其目的是确定决策主体触发其行为的潜在功能。该方法以目标代理的动作序列作为输入。然后，通过 IRL 推断出的奖励函数来理解agent的正常行为，并采用贝叶斯方法对IRL进行检测。实证研究表明，该方法在识别异常方面是有效的。

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb4f/sequential-anomaly-detection-using-inverse-reinforcement-learning

机器学习（Machine Learning）专题：

标题：Deep Uncertainty Quantification: A Machine Learning Approach for Weather Forecasting

作者：Bin Wang, Jie Lu, Zheng Yan, Huaishao Luo, Tianrui Li, Yu Zheng, Guangquan Zhang

论文解读：天气预报通常是通过数值天气预报（NWP）来解决的，由于初始状态设置不当，有时会导致预报效果不理想。该论文设计了一个数据驱动的方法，并辅以有效的信息融合机制来学习历史数据，并提出了一个新的负对数似然误差（NLE）损失函数。与数值预报相比，该方法的精度提高了 47.76%。

地址：

https://www.aminer.cn/pub/5ce2d0e4ced107d4c63cb461/deep-uncertainty-quantification-a-machine-learning-approach-for-weather-forecasting

标 题： FDML: A Collaborative Machine Learning Framework for Distributed Features

作者：Yaochen Hu, Di Niu, Jianming Yang, Shengping Zhou

论文解读：针对特征分布机器学习（FDML）问题，该论文提出了一种异步随机梯度下降（SGD）算法，以联合学习分布特征，在有界异步条件下具有理论收敛性保证。在参数服务器架构中实现了 FDML 系统。实验结果表明，所提出的 FDML 系统可以利用其他应用的用户和项目特征，显著增强腾讯 MyApp 中的应用推荐，同时在很大程度上保留了每个应用程序中功能的局部性和隐私性。

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deb32/fdml-a-collaborative-machine-learning-framework-for-distributed-features

标题：Towards Sustainable Dairy Management - A Machine Learning Enhanced Method for Estrus Detection

作者：Kevin Fauvel, Veronique Masson, Elisa Fromont, Philippe Farerdin

论文解读：该论文的研究目的是采用机器学习的方法来解决奶牛场牛奶生产资源利用效率的挑战。论文藉由机器学习的方法来处理行为性和无声性发情侦测，还提出了一种基于局部级联的发情检测算法，该算法的性能明显优于典型的商业解决方案。最后，该论文提出了一种基于全局和局部（行为与静默）算法可解释性（SHAP）的方法来减少发情检测方案中的不确定性。

地址：

https://www.aminer.cn/pub/5d3ed25a275ded87f97deab2/towards-sustainable-dairy-management-a-machine-learning-enhanced-method-for-estrus-detection

2.10.8 SIGKDD 2020

第 26 届 ACMSIGKDD 知识发现和数据挖掘会议（KDD2020）已于太平洋标准时间 8 月 23－27 日以虚拟线上方式召开。今年 KDD 收集了 338 篇论文（研究和应用轨道），34 个研讨会，45 个教程（讲座和实践），使其成为计算机科学中最大的应用研究会议之一。

尽管本次会议在线上举办，但仍然提供与往年相同的活动内容，包括主题演讲、专题小组、特邀报告、精选研究、应用数据科学论文、信息实践教程和workshop。2020 年 5 月 25 日，KDD2020 官方发布了接收论文，今年一共有 1279 篇论文提交至 Research Track（面向研究界的学术论文），共 216 篇被接受，接收率为 16.8%。而在去年，Researchtrack 共收到约 1200 篇论文投稿，其中约 110 篇被接收为 oral 论文，60 篇被接收为 poster 论文，接收率仅为 14%。今年的接收率有所提升。

SIGKDD2020 具体获奖情况为：

- 最佳论文

✧ 最佳研究型论文

标题：On Sampled Metrics for Item Recommendation

作者：Walid Krichene, Steffen Rendle

论文解读：这篇论文主要对抽样指标进行了详细的研究。在该项目中是使用依赖于相关项目位置的排名指标算法来进行评估，在任务中需要在给定的上下文情况下来对大量的项目进行排序。结果发现这些抽样指标与精确的度量值不一致，因为它们没有保留相关的语句。而论文证明了一种可行的方法就是通过应用一个修正项，即最小化不同的标准，如偏差或均方误差，来提高抽样指标的性能。最后通过对原始抽样指标及其修正变量实证评估，论文建议在度量计算中应避免抽样，但是如果实验研究需要抽样，那么所提出的修正项可以提高估计的质量。

地址：

https://www.aminer.cn/pub/5f03f3b611dc83056223202d/on-sampled-metrics-for-item-recommendation

标题：Malicious Attacks against Deep Reinforcement Learning Interpretations

作者： Mengdi Huai, Jianhui Sun, Renqin Cai, Liuyi Yao, Aidong Zhang

论文解读：该论文研究了 DRL 解释方法的脆弱性。具体来说，论文介绍了针对 DRL 解释的对抗攻击的第一项研究，并提出了一个优化框架，在此基础上可以得出最佳对抗攻击策略。此外，论文研究了 DRL 解释方法对模型中毒攻击的脆弱性，并提出了一种算法框架来严格地表述拟议的模型中毒攻击。最后，论文进行理论分析和广泛实验，以验证针对 DRL 解释提出的恶意攻击的有效性。

地址：

https://www.aminer.cn/pub/5f03f3b611dc830562232015/malicious-attacks-against-deep-reinforcement-learning-interpretations

标 题 ： TIPRDC: Task-Independent Privacy-Respecting Data Crowdsourcing Framework for Deep Learning with Anonymized Intermediate Representations

作者： Ang Li, Yixiao Duan, Huanrui Yang, Yiran Chen, Jianlei Yang

论文解读：该论文提出了 TIPRDC，这是一种与任务无关的尊重隐私的数据众包框架，带有匿名中间表示。该框架的目标是学习一个特征提取器，该特征提取器可以从中间表示中隐藏隐私信息。同时最大限度地保留嵌入在原始数据中的原始信息，以供数据收集器完成未知的学习任务。论文设计了一种混合训练方法来学习匿名中间表示：a）一种用于从特征中隐藏私人信息的对抗训练过程；b）使用基于神经网络的互信息估计器最大程度地保留原始信息。研究对 TIPRDC 进行了广泛评估，并使用两个图像数据集和一个文本数据集将其与现有方法进行了比较。论文的结果表明，TIPRDC 明显优于其他现有方法。其研究工作是第一个与任务无关的尊重隐私的数据众包框架。

地址：

https://www.aminer.cn/pub/5f03f3b611dc83056223205b/tiprdc-task-independent-privacy-respecting-data-crowdsourcing-framework-for-deep-learning-with

- 最佳应用型论文

标题：Temporal-Contextual Recommendation in Real-Time

作者： Ma Yifei, Murali Balakrishnan Narayanaswamy, Lin Bin 和 Hao Ding

论文解读：这篇论文的研究人员提出了一种黑匣子推荐系统，该系统可以适应各种场景，而无需手动调整。研究人员通过重要性抽样进一步提供了有效的培训技术，可以扩展到数百万个项目，而性能几乎没有损失。他们报告了对大量实际数据集的重大改进，并通过合成实验提供了对模型功能的直观了解。HRNN-meta 的一部分已大规模生产，供客户在 Amazon Web Services 上使用，并用作数千个网站的基础推荐引擎。

地址：

https://www.aminer.cn/pub/5f03f3b611dc8305622320dd/temporal-contextual-recommendation-in-real-time

- 代表性论文

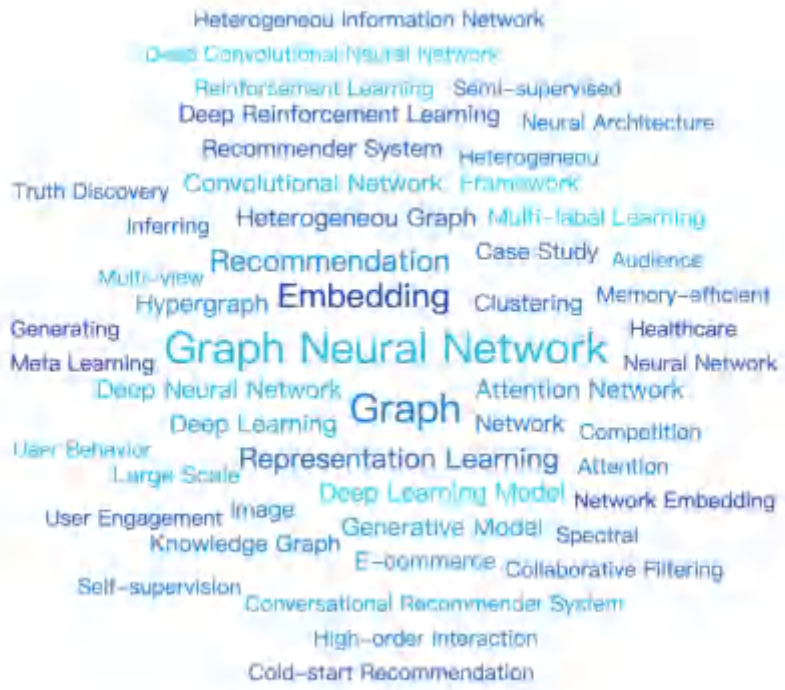


图 58 SIGKDD2020 论文研究热点的词云图

KDD2020 共录用论文 338 篇，论文作者 1479 人。2020 年 KDD 的论文主要集中在图神经网络、嵌入、推荐和表征学习几个主题中，下面共选出上述主题中的 10 篇论文进行介绍。

图神经网络（Graph Neural Network）：

标题：GPT-GNN: Generative Pre-Training of Graph Neural Networks

作者：Ziniu Hu, Yuxiao Dong, Kuansan Wang, Kai-Wei Chang, Yizhou Sun

论文解读：该论文提出了 GPT-GNN 框架，通过生成式预训练来初始化 GNN。GPT-GNN 引入了自我监督的属性图生成任务来预训练 GNN，以便它可以捕获图的结构和语义属性。该论文将图生成的可能性分解为两个部分：a）属性生成和 b）边缘生成。通过对两个组件进行建模，GPT-GNN 捕获了生成过程中节点属性和图结构之间的固有依赖性。在数十亿规模的开放式学术图谱和亚马逊推荐数据上进行的综合实验表明，GPT-GNN 的性能明显优于最新的 GNN 模型，而无需在各种下游任务中进行多达 9.1%的预训练。

地址：

https://aminer.cn/pub/5efb0d5691e011063336d354/gpt-gnn-generative-pre-training-of-graph-neural-networks?conf=kdd2020

标题：Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks

作 者 ： Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, Chengqi Zhang

论文解读：该论文提出了专门为多元时间序列数据设计的通用图神经网络框架。方法通过图形学习模块自动提取变量之间的单向关系，可以轻松地将诸如变量属性之类的外部知识集成到其中。进一步提出了一种新颖的混合跳跃传播层和一个扩张的起始层来捕获时间序列内的空间和时间依赖性。在端到端框架中共同学习图学习，图卷积和时间卷积模块。实验结果表明，论文提出的模型在 4 个基准数据集中的 3 个方面优于最新的基线方法，并且在提供额外结构信息的两个交通数据集上与其他方法相比，具有与众不同的性能。诸如变量属性之类的外部知识可以轻松地集成到其中。并为使用图神经网络处理各种非结构化数据打开了新的大门。

地址：

https://aminer.cn/pub/5ecce8ec91e0119170395ba4/connecting-the-dots-multivariate-time-series-forecasting-with-graph-neural-networks?conf=kdd2020

标题：Dynamic Heterogeneous Graph Neural Network for Real-time Event Prediction

作者：Wenjuan Luo, Han Zhang, Xiaodi Yang, Lin Bo, Xiaoqing Yang, Zang Li, Xiaohu Qie, Jieping Ye

论文解读：该论文建议对每个进行中的事件使用动态构造的异构图来对事件及其周围环境的属性进行编码。此外，论文提出了一种多层图神经网络模型，

以了解历史动作和周围环境对当前事件的影响，并生成有效的事件表示形式，以提高响应模型的准确性。论文针对 DiDi 平台上的两个实际应用研究了该框架。离线和在线实验表明，该框架可以显着提高预测性能。该框架已部署在在线生产环境中，每天处理数千万个事件预测请求。

地址：

https://aminer.cn/pub/5f03f3b611dc830562232099/dynamic-heterogeneous-graph-neural-network-for-real-time-event-prediction?conf=kdd2020

嵌入（Embedding）专题论文：

标题：Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding

作者：Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, Jiawei Han

论文解读：该论文提出了一个新任务 Hierarchical Topic Mining，该任务仅采用由类别名称描述的类别树，并旨在从文本语料库中挖掘每个类别的代表词集，以帮助用户理解他/她感兴趣的主题。论文中开发了一种新颖的联合树和文本嵌入方法，以及一种原则上的优化程序，该程序允许同时建模类别树结构和球形空间中的语料库生成过程，从而有效地进行类别代表术语的发现。综合实验表明，名为 JoSH 的模型可以高效挖掘高质量的层次结构主题集，并受益于弱监督的层次结构文本分类任务。

地址：

https://aminer.cn/pub/5f03f3b611dc830562231ff0/hierarchical-topic-mining-via-joint-spherical-tree-and-text-embedding?conf=kdd2020

标题：Adaptive Graph Encoder for Attributed Graph Embedding

作者： Ganqu Cui, Jie Zhou, Cheng Yang, Zhiyuan Liu

论文解读：该论文提出了一种自适应图形编码器（AGE），这是一种新颖的属性图嵌入框架。 AGE 由两个模块组成：a）为了更好地减轻节点特征中的高频噪声，AGE 首先应用了精心设计的拉普拉斯平滑滤波器；b）AGE 采用了自适应编码器，该编码器迭代地增强了滤波后的特征，以实现更好的节点嵌入。论文中使用四个公共基准数据集进行实验，以验证 AGE 在节点群集和链接预测任务上的作用。实验结果表明，AGE 在这些任务上始终优于现有的嵌入方法。

地址：

https://aminer.cn/pub/5f02f25491e011ee5e0258e0/adaptive-graph-encoder-for-attributed-graph-embedding?conf=kdd2020

标题：Personalized Prefix Embedding for POI Auto-Completion in the Search Engine of Baidu Maps

作者：Jizhou Huang, Haifeng Wang, Miao Fan, An Zhuo, Ying Li

论文解读：该论文提出了一种基于端到端基于神经的 POI-AC 框架，该框架最近已部署在百度地图的搜索引擎中，最大的 Web 映射应用程序之一，全球每月有数亿活跃用户。为了在用户之间建立联系，他们的个人输入习惯以及相应感兴趣的 POI，所提出的框架（简称 P3AC）由三个组件组成，即适应个性化前缀的多层 Bi-LSTM 网络，CNN 的网络对 POI 上的多源信息进行建模，并使用三元组排名损失函数来优化 POI 的个性化前缀嵌入和分布式表示。论文首先使用百度地图的大规模现实世界搜索日志来评估 P3AC 离线性能，该性能是通过多个指标（包括均值倒数排名（MRR），成功率（SR）和归一化折现累积收益（nDCG））来衡量的。大量的实验结果表明，它可以实现实质性的改进。然后，研究人员决定在线上启动它，并观察到其他一些有关用户满意度的关键指标，例如，POI-AC 会话中的平均击键次数和平均击键速度也明显下降。此外，研究人员已将 P3AC 的源代码和实验数据公开发布给公众，以进行可重复性测试。

地址：

https://aminer.cn/pub/5f03f3b611dc8305622320ce/personalized-prefix-embedding-for-poi-auto-completion-in-the-search-engine-of?conf=kdd2020

推荐（Recommendation）专题：

标题：Controllable Multi-Interest Framework for Recommendation

作者：Yukuo Cen, Jianwei Zhang, Xu Zou, Chang Zhou, Hongxia Yang, Jie Tang

论文解读：该论文为顺序推荐提出了一种新颖的可控多兴趣框架，称为ComiRec。论文的多兴趣模块从用户行为序列中捕获了多个兴趣，可将其用于从大型项目库中检索候选项目。然后将这些项目输入汇总模块以获取总体推荐。聚合模块利用可控因素来平衡推荐准确性和多样性。该研究对两个真实的数据集 Amazon 和 Taobao 进行顺序推荐实验。实验结果表明，该框架相对于最新模型取得了重大改进。框架也已成功部署在离线阿里巴巴分布式云平台上。

地址：

https://aminer.cn/pub/5ec48cc4da5629efe0884e02/controllable-multi-interest-framework-for-recommendation?conf=kdd2020

标题：Privileged Features Distillation at Taobao Recommendations

作者： Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, Wenwu Ou

论文解读：该论文从弥补训练与推理之间差距的蒸馏技术出发，提出了特征特征蒸馏（PFD）。论文训练两种模式，即与原始模型相同的学生模型和另外利用特权特性的教师模型。从更准确的教师中提取的知识被转移到学生身上，从而提高预测的准确性。在服务过程中，只提取学生部分，不依赖特权特征。论文在淘宝推荐中进行了两个基本预测任务的实验，即粗粒度排序的点击率（CTR），细粒度排名的 CVR。通过提取 CTR 服务期间禁止的交互功能和 CVR 事件后功能，对其强大基线进行了显著改进。在在线应付测试期间，CTR 任务中

的点击量值提高了+5.0%。在 CVR 任务中，转换度量提高了+2.3%。此外，通过解决 PFD 培训的几个问题，论文获得了与基线相比的训练速度，而不需要任何蒸馏。

地址：

https://aminer.cn/pub/5f03f3b611dc8305622320d5/privileged-features-distillation-at-taobao-recommendations?conf=kdd2020

表征学习（Representation Learning）专题：

标题：Understanding Negative Sampling in Graph Representation Learning

作者：Zhen Yang, Ming Ding, Chang Zhou, Hongxia Yang, Jingren Zhou, Jie Tang

论文解读：该论文从客观和风险两个角度系统地分析了负采样的作用，从理论上证明了负采样在确定优化目标和产生的方差方面与正采样同等重要。该论文是第一个推论并量化一个好的负采样分布为 pn（u|v）∝ pd（u|v）α，0＜α＜1 的人。理论上，论文提出 MCNS，通过自对比近似近似正分布，并通过 Metropolis-Hastings 加速负采样。论文在涵盖 19 个实验设置的 5 个数据集上评估了其方法，这些数据集涵盖了广泛的下游图形学习任务，包括链接预测、节点分类和推荐。这些相对全面的实验结果证明了其稳健性和优越性。

地址:

https://aminer.cn/pub/5ec7a32791e0118397f3ec20/understanding-negative-sampling-in-graph-representation-learning?conf=kdd2020

标题： Interpretable Deep Graph Generation with Node-edgeCo-disentanglement

作者： Xiaojie Guo, Liang Zhao, Zhao Qin, Lingfei Wu, Amarda Shehu, Yanfang Ye

论文解读：解纠缠表示学习近年来受到了广泛的关注，特别是在图像表示学习领域。然而，学习图背后的解纠缠表示在很大程度上仍未被探索，特别是对于同时具有节点和边缘特征的属性图。图生成的解纠缠学习面临着新的重大挑战，包括 a）缺少图的反褶积操作来联合解码节点和边缘属性；b）在各自影响的潜在因素之间的难解性：i）仅影响节点，ii）仅影响边缘，iii）它们之间的连接模式。为了解决这些问题，我们提出了一个新的属性图深层生成模型的解纠缠增强框架。特别地，提出了一种新的变分目标来解开上述三种潜在因素，并具有新的节点和边缘反褶积结构。此外，在每种类型中，个体因素的分离进一步增强，这被证明是对现有图像框架的一般化。在综合数据集和真实数据集上的定性和定量实验证明了该模型及其扩展的有效性。

地址:

https://aminer.cn/pub/5f03f3b611dc830562232029/interpretable-deep-graph-generation-with-node-edgeco-disentanglement?conf=kdd2020

3　人才篇

3 人才篇

依托清华大学自主研发的“科技情报大数据挖掘与服务平台[79]”（简称AMiner），采用大数据分析与挖掘技术，结合文献计量学方法和引文分析方法，基于1.3章节介绍的数据挖掘领域知识图谱和技术篇的技术关键词，以及附录1列出的代表性期刊和会议列表，获取这些期刊和会议上在2010年—2020年间收录的相关论文数据（共计14,582篇，论文引用量是336,281），通过分析和挖掘这些论文作者的信息，获取了21,018位研究学者。以这些学者以及他们发表的论文作为底层数据，对国家和机构的学术水平、学者合作和流动情况进行一个整体分析，并利用AMiner平台的学者画像功能，展示了国内外一些代表性学者的详细信息。通过以上数据的挖掘与分析，旨在为国家和机构在数据挖掘领域的学科建设和布局、人才政策制定等提供数据参考。

3.1 学者情况概览

本章节通过深度分析前述的论文和学者数据，从分布地图、学术水平、国际合作、学者流动等多种维度，详细介绍了全球和国内学者在数据挖掘领域的发展情况。

3.1.1 学者分布地图

人才的聚集可以推动一个城市相关产业的快速发展，人才地图的分布可以直观展示人才的地区分布，有利于调查和分析各地区人才竞争力的现况，对人才调查和引进尤为重要。

本节参考h-index作为筛选条件，选择TOP1000全球学者，以地图形式展示和分析他们的地区分布。每个学者分布地图，均是根据学者的当前单位地理位置进行绘制，其中颜色越红、圆圈越大，表示学者越集中。

（1） 全球学者分布地图

图 59展示了数据挖掘领域h-index排名前1000的高水平研究学者在全球范围的分布情况。从中可以看出，这些学者主要集中在以美国为首的北美洲、以英国、意大利为首的欧洲，和以中国为首的亚洲。表 1514 展示了数据挖掘

领域高水平学者数量排名前 10 的国家。美国的高水平学者约有 399 位，居全球首位，学者数量占比约 39.9%；其次，中国的高水平学者约有 132 位，数量占比约 13.2%；排名第三和第四的国家是欧洲的意大利和英国，高水平学者数量占比分别约是 4.5%和 4.4%，两国数量基本相同。中美两国的高水平学者数量远高于其他国家，在数据挖掘领域具有较强的学术实力。中国虽然仅次于美国，但是高水平学者数量约是美国的三分之一，差距较大。中国应继续在数据挖掘研究领域实施人才培养、引进和激励等策略，建设数据挖掘领域的高水平人才队伍。

图 59 数据挖掘领域 h-index 排名前 1000 学者的全球分布地图

表 14 h-index TOP1000 全球学者的国家统计

排名	国家名称	学者数量	排名	国家名称	学者数量
1	美国	399	6	加拿大	31
2	中国	132	7	新加坡	29
3	意大利	45	8	澳大利亚	29
4	英国	44	9	荷兰	27
5	德国	36	10	西班牙	24

(2) 中国学者分布地图

图 60 和表 15 展示了数据挖掘领域的 h-index 排名前 1000 的高水平研究学者在中国范围的分布情况。从地图中可以看出，这些高水平的研究学者主要

分布在中国东部发达地区，包括京津冀地区、长江三角洲地区、珠江三角洲、香港、台湾等地。其中，北京市数据挖掘领域高水平学者约有 32 名；其次是台湾、香港、上海和广东等东部地区；此外，陕西省和湖北省等中部地区也有少量学者分布，但与东部地区的人才数量相比差距明显。为更好促进我国数据挖掘相关研究的进展，应加大对中西部地区的科研投入和立项力度，通过科技政策协调各个地区的科研资源配置，改善高水平人才在小范围地区集中的情况。同时应加强地区之间的人才合作交流，使数据挖掘领域技术在各地区均衡发展。

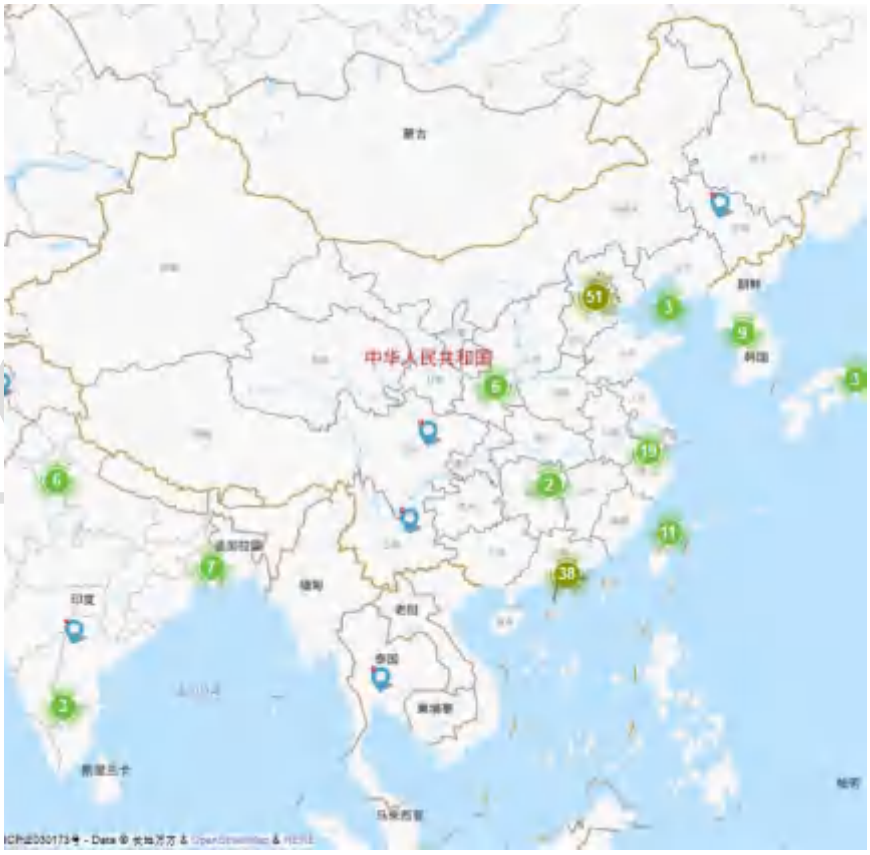


图 60 数据挖掘领域 h-index 排名前 1000 学者的中国分布地图

表 15 h-index TOP1000 学者的中国省市统计

排名	省市名称	学者数量	排名	省市名称	学者数量
1	北京市	32	6	江苏省	6
2	台湾省	12	7	河北省	5
3	香港特别行政区	10	8	陕西省	3
4	上海市	8	9	浙江省	2
5	广东省	7	10	湖北省	2

3.1.2 学术水平分析

本节通过分析指定期刊和会议上在 2010 年—2020 年间收录的相关论文数据

（共计 14,582 篇）和 21,018 位研究学者，统计分析了全球各个国家和机构的学术水平情况。

（1） 国家学术水平分析

表 16 显示了论文总被引频次排名前 10 的国家的论文发表量、学者数量等统计情况。这些国家主要分布在欧洲（5 个）、亚洲（2 个）、北美洲（2 个）、大洋洲（1 个），由此看出亚洲国家在数据挖掘相关领域发表论文的影响力较高。论文总被引频次最高的国家是中国，其次是美国，但两者相差不大。中国和美国在论文总被引频次、论文发表量、学者数量等指标上均远高于其他国家，由此看出中美两国在数据挖掘领域的科研实力和水平处于领头羊位置。结合前文所分析的高水平学者在各国的分布情况，中国的高水平学者数量是美国的约三分之一。中国在论文总被引频次上排名第一，与中国论文发表量和学者数量均较高有一定关联。中国仍需要重视高水平人才培养，在保证数量增长的同时，更加关注学术水平和论文质量的提升。

表 16 论文总被引频次排名前 10 的国家

国家名称	论文总被引频次	论文发表量	学者数量
中国	169998	4809	7773
美国	165478	4401	7299
德国	14886	486	496
澳大利亚	13519	586	1008
加拿大	11156	357	428
英国	10885	371	454
意大利	10834	329	364
新加坡	10078	330	404
西班牙	7961	247	316
法国	5732	270	366

（2） 机构学术水平分析

表 17 显示了论文总被引频次排名前 10 的全球各个机构的论文发表量、学者数量等统计情况。可以发现，排名前 10 的机构中大部分分布在美国，仅有 1 所分布在中国。9 所美国机构中既有企业，也有高校和研究中心，说明数据挖掘技术在美国的不同行业均有研究应用。中国唯一上榜的机构是清华大学，其论文总被引频次排名第 5，论文发表量排名第 3、学者数量排名第 1。对比中美高校情况，斯坦福大学的论文发表量虽只有清华大学的约 40%，但其论文总被

引频次是清华大学的 1.5 倍。类似地，清华大学的学者数量高于伊利诺伊大学香槟分校，但其论文发表量却仅有伊利诺伊大学香槟分校的约 80%。基于以上分析，不难发现在数据挖掘领域中国机构的影响力和竞争力还有待提高，特别是非高校科研机构在数据挖掘领域还未占据一席之地。其次，中国机构虽在论文发表量和学者数量上占优，但在论文总被引频次上没有体现出优势，这反映出中国机构在数据挖掘领域的学术影响力有待加强，学术质量和贡献度有待提高。

表 17 论文总被引频次排名前 10 的全球机构

机构名称	论文总被引频次	论文发表量	学者数量
微软 Microsoft	25934	301	389
伊利诺伊大学厄巴纳-香槟分校 University of Illinois at Urbana-Champaign	19864	370	517
斯坦福大学 Stanford University	19681	126	251
谷歌 Google	17101	235	279
清华大学 Tsinghua University	13446	298	591
卡内基梅隆大学 Carnegie Mellon University	9480	233	286
脸谱网 Facebook	7234	75	91
IBM 研究中心 IBM Research Center	5683	161	239
领英 Linkedin	5534	53	104
亚利桑那州立大学 Arizona State University	5365	118	166

表 18 显示了论文总被引频次排名前 10 的中国机构的论文发表量、学者数量等统计情况。这些机构主要分布在北京（4 个）、香港（2 个）、安徽（2 个）、浙江（1 个）、四川（1 个）。其中，论文总被引频次最高的中国机构是清华大学，其学者数量和论文发表量均远高于其他中国机构。此外，这 10 所机构中有 4 所是中国 C9 高校联盟成员，分别是清华大学、浙江大学、中国科学技术大学和北京大学。从以上数据分析结果可以看出，数据挖掘技术领先的中国科研机构主要分布在经济和政治发达地区，研究机构多为高校或研究院。经济发展水平较

高的地区，教育资源也较丰富，一定程度为这些地区的科技发展提供了优质人才资源，科技发展又会进一步促进地区经济发展，发达地区和欠发达地区在经济和科技方面的差距进一步拉大，形成马太效应。因此，合理规划和引导科研资源的分配有利于促进各地区均衡发展，也有益于前沿技术在全国范围内蓬勃发展。

表 18 论文总被引频次排名前 10 的中国机构

机构名称	论文总被引频次	论文发表量	学者数量
清华大学	13446	298	591
浙江大学	5047	139	165
香港科技大学	3833	133	223
中国科学技术大学	3770	86	97
中国科学院	3371	156	225
北京大学	3294	101	144
合肥工业大学	2709	40	77
香港中文大学	1770	70	94
北京航空航天大学	2088	52	54
电子科技大学	2047	51	57

3.1.3 国际合作分析

(1) 全球国家合作分析

通过统计全球不同国家在数据挖掘领域的合作论文发表情况，图 61 和表 19 展示了全球各国合作论文情况。从表中可以看出，排名前 10 的国家中大部分国家都倾向与中国、美国合作，与中国合作的国家数量为 6 个，与美国合作的国家数量为 5 个。国家之间合作论文数量排名第一的是中国与美国，比排在第二位的中国与澳大利亚合作论文数量多约 7 倍，论文总被引频次高约 10 倍。从图表中可以看出，中国和美国是全球论文合作网络中的重要节点国家，一方面反映两个大国之间的学术交流合作较为紧密，另一方面也反映出中美两国在数据挖掘领域的学术水平较为领先，是各国学术交流合作的热门选择。

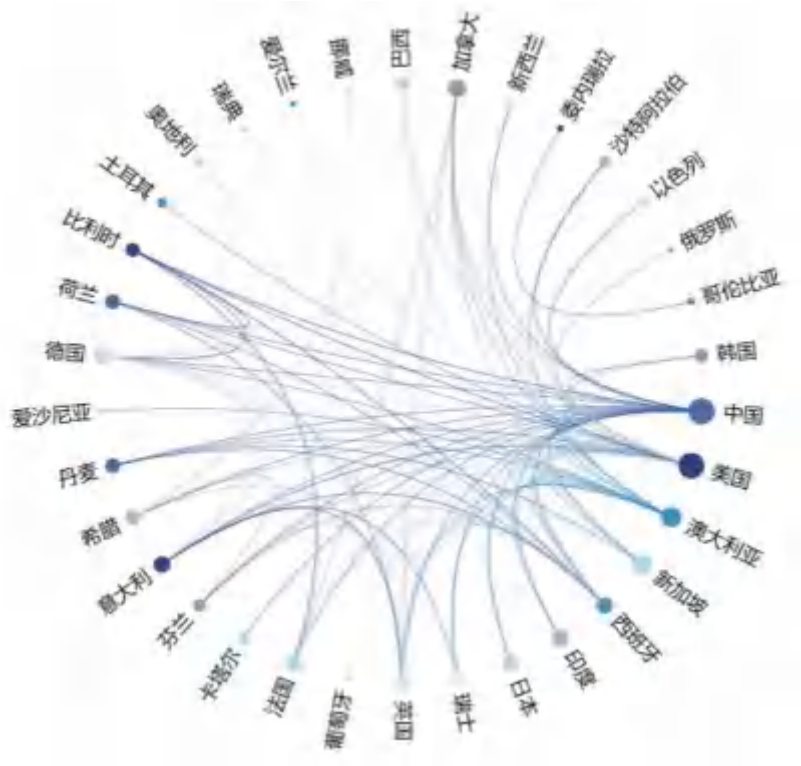


图 61 各国数据挖掘领域论文合作网络图

表 19 合作论文数量排名前 10 的国家列表

国家 1	国家 2	合作论文数量	论文总被引频次
中国	美国	1738	84917
中国	澳大利亚	261	8004
中国	新加坡	248	6958
中国	加拿大	137	5242
澳大利亚	美国	98	2722
印度	美国	96	1709
中国	日本	76	1369
德国	美国	73	1875
新加坡	美国	71	4226
中国	英国	69	1914

(2) 中国与其他国家合作分析

图 62 展示了中国与其他国家在数据挖掘领域的论文合作情况。从中可以看出，中国与美国合作最多，合作论文数量高达 1738 篇，占比超 50%；其次是澳大利亚、新加坡、加拿大。由此可以看出，中国与美国在数据挖掘领域的合作十分紧密。但考虑到中美关系日趋紧张，中国有必要加强与其他国家，如澳

大利亚、新加坡、加拿大的交流合作力度，以确保该领域对外学术交流合作可持续。

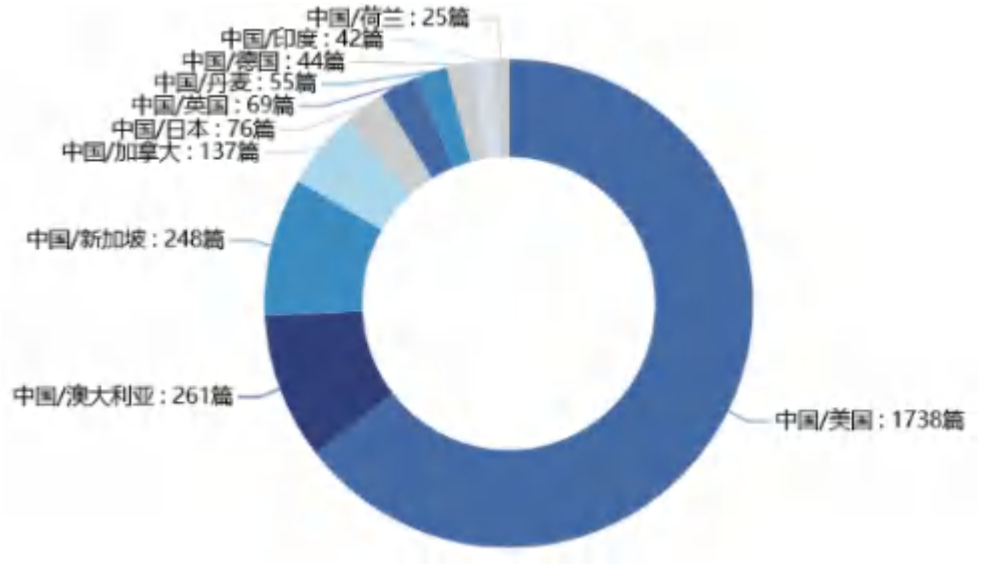


图 62 中国与其他国家的论文合作情况

3.1.4 学者流动情况

科技是第一生产力，人才是第一资源，高水平人才聚集可以促进一个国家或地区的快速发展，而人才的流动是实现人才聚集的重要途径。当下，人才迁徙已成为各地区人才竞争中不可忽视的社会现象。下文将聚焦学者所属单位地理位置的变化，统计全球各国和中国各个省、直辖市和特别行政区在数据挖掘领域的高水平学者流入、流出、流动差值等情况，为各个地区进行人才培养、人才交流和人才引进等工作提供有益参考。

（1） 全球学者流动分析

将学者流入和流出数量之和作为筛选条件，我们统计了数据挖掘领域的全球学者在 2010 年—2020 年间流动数量排名前 10 的国家，如图 63 所示。其中，美国和澳大利亚的学者流出数量大于流入数量，而中国、英国、印度、德国等国学者流入数量大于流出数量。无论是学者流入数量还是学者流出数量，美国均排在第一位，其次是中国。随着经济全球化深入发展，科技人才的跨国流动更加频繁，各国对科技人才的争夺更加激烈。中美两国的人才迁徙“流量”较大，但中国领域人才最终是净流入的，这与中国制定的一系列科技人才培养、引进和激励等政策和措施不无关系。除关系人才流动数量之外，仍需进一步关

注人才流动的质量，建设一支更富有国际竞争力的高水平科技人才队伍。

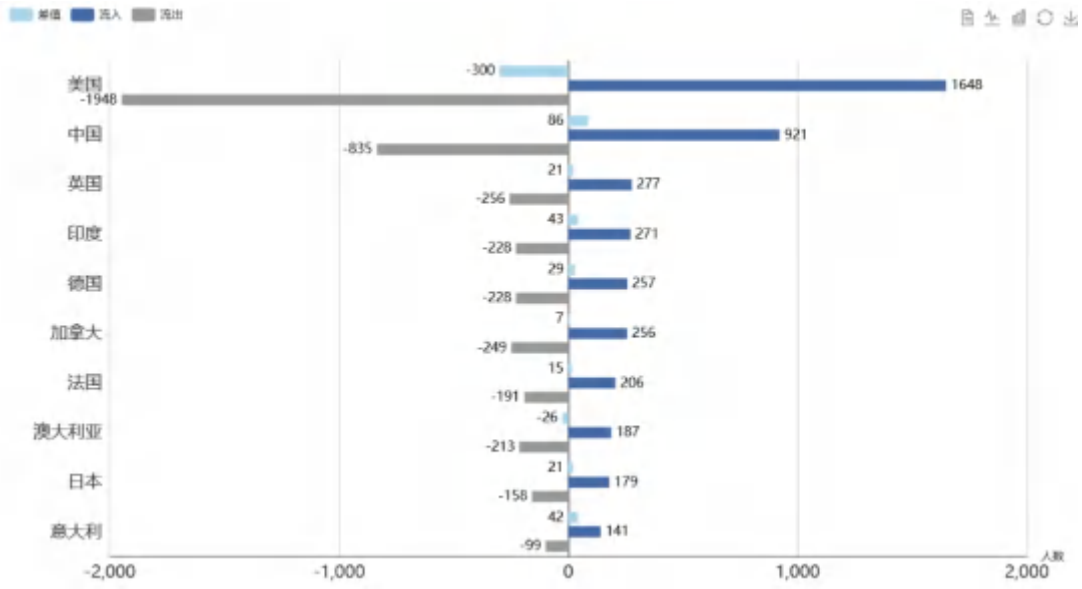


图 63 全球学者的流动情况

(2) 中国学者流动分析

本节分析了数据挖掘领域的中国学者于 2010 年—2020 年间在中国范围内不同地区的流动情况，根据学者流入和流出数量之和，展示了排名前 10 的中国省市，如图 64 所示。从图中不难发现，北京市学者流入和流出的数量明显高于其他地区，这可能与北京的高校和科研院所数量较多以及科研发展政策的倾斜有关。此外，北京、上海和香港的学者流入数量均高于流出数量，而南京、西安、武汉和杭州等地的学者流出数量高于流入数量。这表明，北京、上海和香港在数据挖掘技术发展方面具有竞争优势，能够吸引学者流入。而其余地区的竞争力相对较差，可以通过制定具有激励性的人才引进政策缩小地区差距，具体策略可以包括落户、补助和奖金等。

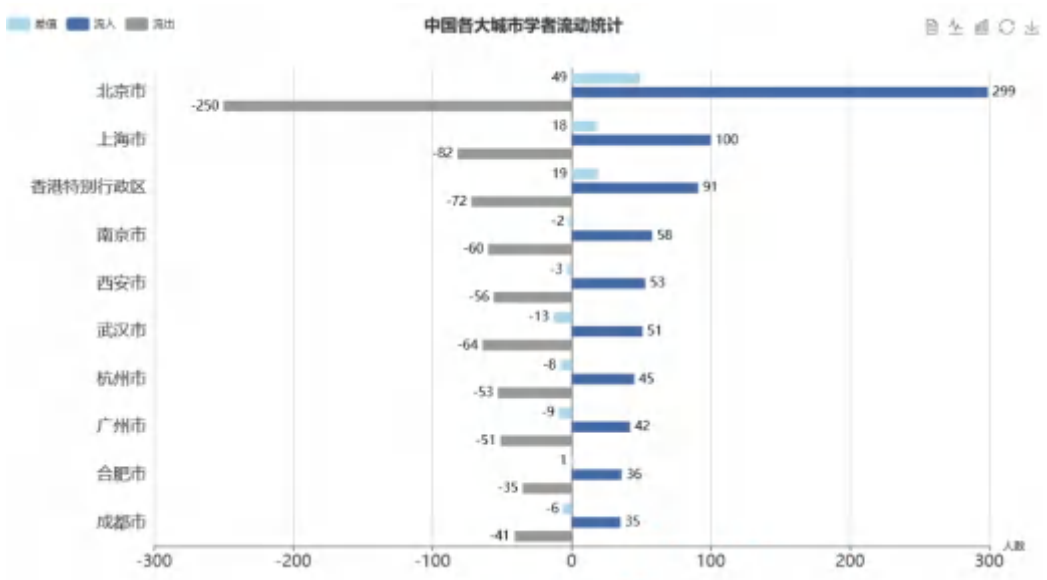


图 64 中国学者的流动情况

3.2 学者简介

学者选取来自近期将要发布的全球 AI 十年影响力人才发展报告，我们选取了 KDD 会议，对所涉学者及其论文关键信息进行抽取。依据各学者论文被引用的次数，来进行此次十年最具影响力人才的排名。我们依此排名对国内外学者进行介绍。

3.2.1 发展过程中代表学者简介

在数据挖掘发展过程中，有很多重要的研究学者在这一领域起到了推动作用，以下是有史以来在 SIGKDD 上发表论文的引用量排在前十名的学者。

Jon M. Kleinberg（乔恩·克莱因伯格） Follow | 66

h-index: 112 | #Paper: 387 | #Citation: 98441

Professor

Department of Computer Science, Cornell University

42579

Social Network | Computational Geometry | Social Networks | Complexity | Computer Algebra | Algorithmics

Similar Authors

Jiawei Han（韩家炜） Follow | 272

h-index: 175 | #Paper: 1252 | #Citation: 183472

Professor

Department of Computer Science, University of Illinois at Urbanan

161099

Data Mining | Data Cube | Data Analysis | Heterogeneous Information Network | Information Retrieval

Similar Authors

Bing Liu（刘兵） Follow | 32

h-index: 87 | #Paper: 395 | #Citation: 53733

Professor

Department of Computer Science, University of Illinois at Chicago

42181

Data Mining | Web Pages | Association Rule | Opinion Mining | Web Mining | Knowledge Discovery

Similar Authors

Jörg Sander Follow | 5

h-index: 40 | #Paper: 91 | #Citation: 30970

Professor

Department of Computing Science,University of Alberta

3919

Spatial Database | Clustering Algorithm | Sensor Network | Hierarchical Clustering | Data Mining | Indexation

Similar Authors

Philip S. Yu Follow | 131

h-index: 132 | #Paper: 855 | #Citation: 76031

Professor

University of Illinois at Chicago

53127

Data Mining | Transaction Processing | Social Network | Indexation | Distributed Databases

Similar Authors

Christos Faloutsos Follow | 68

h-index: 125 | #Paper: 796 | #Citation: 91458

Professor

Computer Science Department Carnegie Mellon University

47364

Data Mining | Social Network | Large Graph | Power Law | Anomaly Detection | Information Retrieval

Similar Authors

Pedro Domingos Follow | 20

h-index: 81 | #Paper: 213 | #Citation: 40949

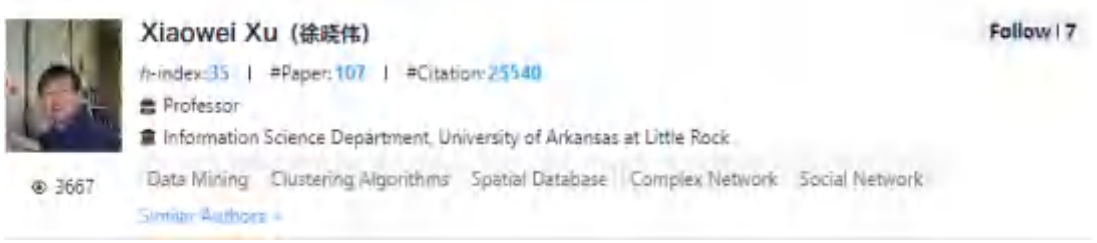
Professor

Department of Computer Science and Engineering,University of Washington

6200

Machine Learning | Markov Logic | First-order Logic | Markov Logic Network | First Order Logic

Similar Authors



● Martin Ester

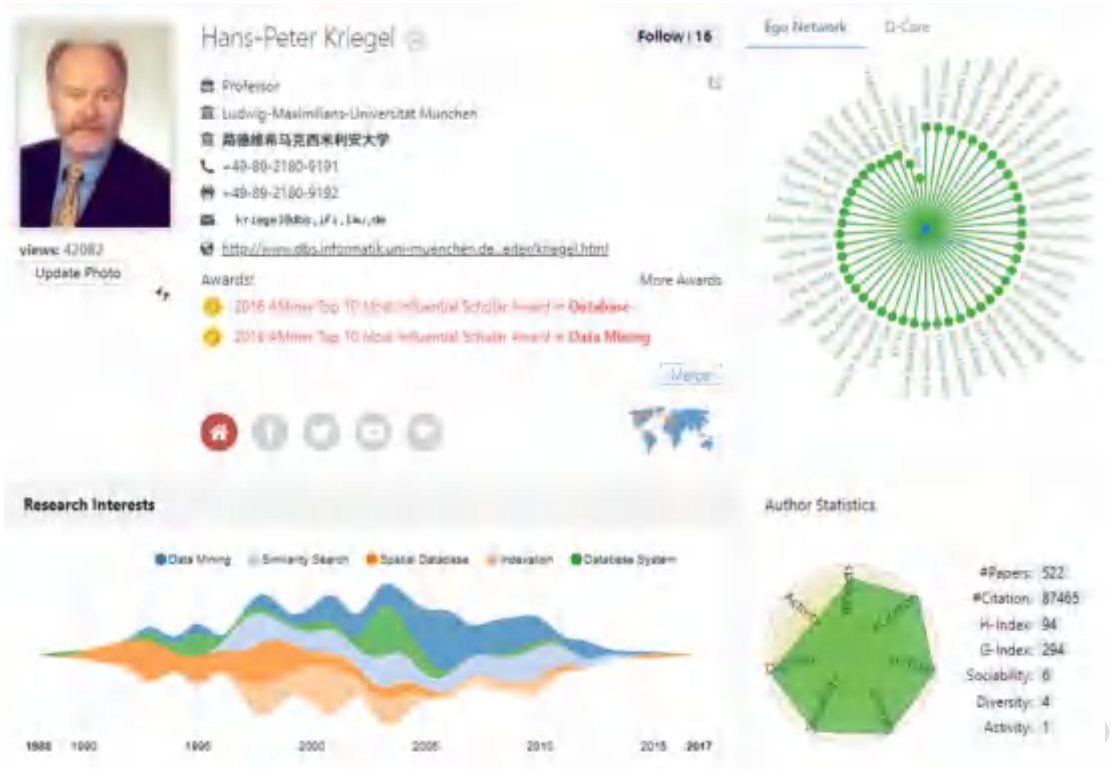
Martin Ester，在数据挖掘和机器学习领域拥有 20 多年的应用研究经验，探索了数据挖掘在支持精准医学方面的潜力，被认为是该领域的国际领导者。他是加拿大不列颠哥伦比亚省的西蒙弗雷泽大学（Simon Fraser University，简称 SFU）的计算机科学系教授，1990 年于瑞士苏黎世联邦理工学院获得计算机科学博士学位。他的研究领域为：社交媒体中的数据挖掘、生物网络中的数据挖掘、数据库系统、大数据在精准医学中的应用。

Martin Ester 博士是 SFU 数据库和数据挖掘实验室的联合主任、SFU 计算机科学学院院长、TKDE 副主编、NSERC 资金委员会成员。曾任 2014 ASONAM 的程序委员会主席

他致力于癌症标记物发现，患者分层和药物目标相互作用的预测，与 SFU 的 Fiona Brinkman 博士实验室共同开发的 PSORTb 仍然被广泛用作细菌中蛋白

质亚细胞定位预测的最先进工具。

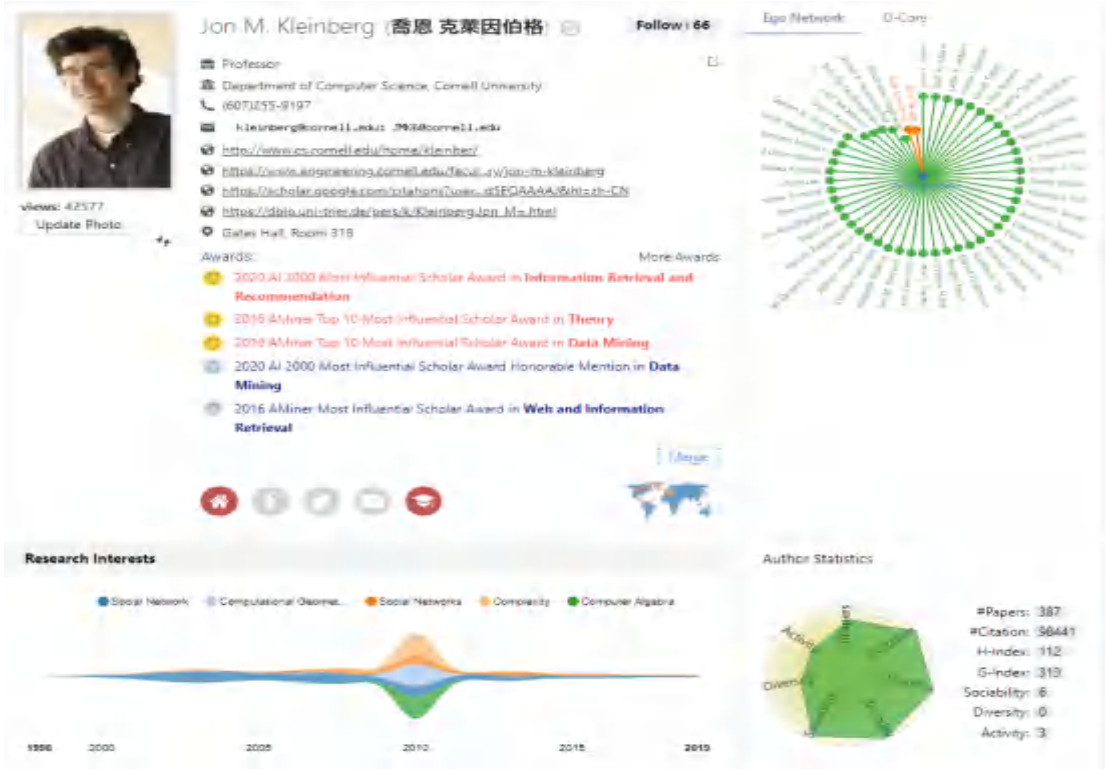
● Hans-Peter Kriegel



Hans-Peter Kriegel，因其在数据挖掘聚类，异常检测和高维数据分析（包括基于密度的方法）中的贡献而受到认可。他是德国慕尼黑大学（Ludwig Maximilian University of Munich）信息学教授，并领导计算机科学系数据库系统小组。研究领域为：相关聚类、高维数据索引和分析、空间数据挖掘和空间数据管理以及多媒体数据库。

Kriegel 教授是国际计算机协会（ACM）的研究员（2009）。因其在聚类，异常检测和高维数据分析（包括基于密度的方法）中的数据挖掘方面的影响，而获得了 SIGKDD 2015 创新奖（Innovation Award），这是数据挖掘领域的最高奖项。Kriegel 教授是“A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise （SIGKDD 1996）”的作者之一，该论文提出了 DBSCAN 算法，获 SIGKDD 2014 时间测试奖（Test of Time Award），对数据挖掘领域做出了突破性贡献。此外，他还曾获 2013 IEEE ICDM 研究贡献奖。

● Jon M. Kleinberg



Jon M. Kleinberg，计算机科学领域公认的天才，HITS 算法发明人，社会网络分析方向的知名学者。他是康奈尔大学（Cornell University）计算机教授， 1993 年于康奈尔大学获得了计算机科学学士学位，于 1996 年获得麻省理工计算机科学博士学位。研究方向为：社会网络与信息网络分析的基础理论、算法设计、万维网、计算机代数、近似算法。

Jon M. Kleinberg 教授是美国国家科学院（NAS）、美国国家工程院（NAE）、美国人文与科学院（AAAS）三料院士，是康奈尔大学的 University Professor（美国大学给教授的最高头衔）。曾获 NSF 职业奖，ONR 青年研究员奖，麦克阿瑟基金会奖学金，帕卡德基金会奖学金，斯隆基金会奖学金，2000 年 ACM 数据库系统原理研讨会最佳论文奖，2006 奈望林纳奖（Nevanlinna Prize）。

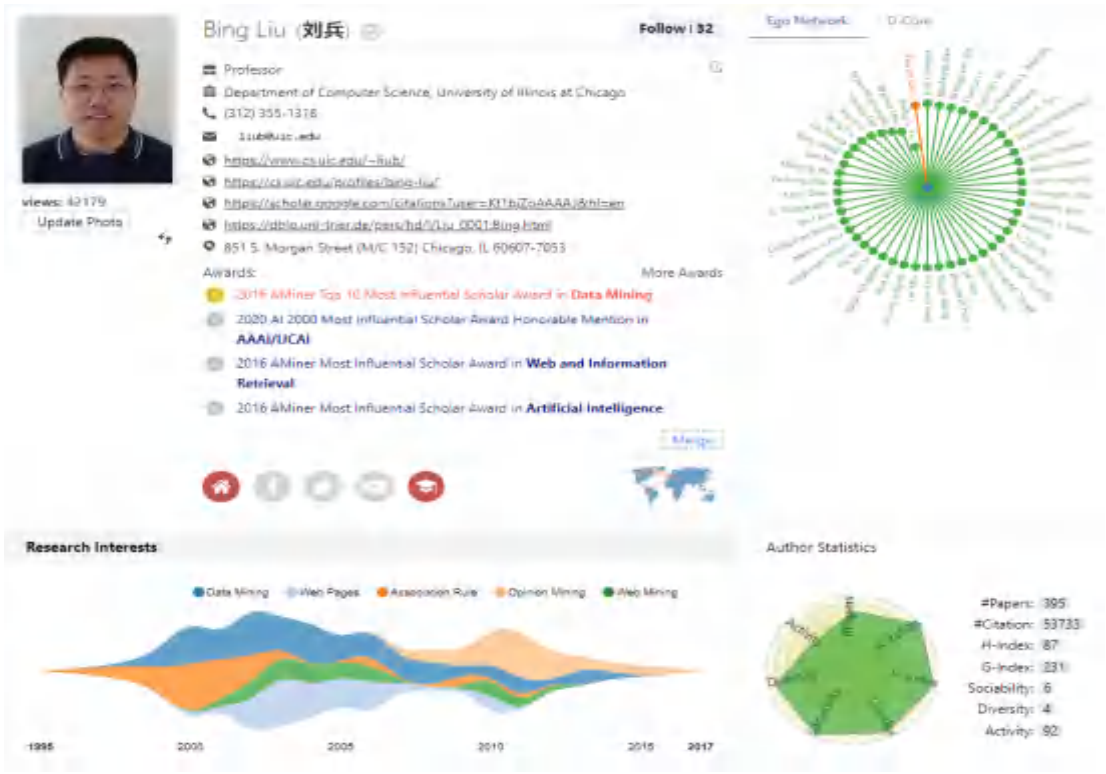
● 韩家炜

韩家炜，被称为“数据挖掘第一人”，是具有国际影响力的著名科学家。他是伊利诺伊大学香槟分校（University of Illinois at Urbana-Champaign，简称 UIUC）计算机系教授，1985 年于美国威斯康星大学（University of Wisconsin System）获计算机系博士学位。他的研究领域为：数据挖掘、数据库和信息网络；目前从事的项目有信息网络学术研究中心：建筑和采矿生物网络、社交媒体数据的多维结构的总结与挖掘、构建和挖掘结构、大型网络的嵌入式研究。

韩家炜教授是 IEEE 和 ACM 的会士（Fellow），曾任美国 ARL 资助的信息网络联合研究中心主任，曾担任 KDD、SDM 和 ICDM 等国际知名会议的程序委员会主席，创办了 ACM TKDD 学报并任主编。

他曾获 2009 年 the McDowell Award、ACM SIGKDD 2004 创新奖（2004 ACM SIGKDD Innovation Award）、ICDE 2002 杰出贡献奖。他出版的数据挖掘专著《Data Mining: Concepts and Techniques》，是国内外数据挖掘领域经典教材。在谷歌学术的 h-index 中，名列全球计算机科学领域高引作者前三。

●Bing Liu （刘兵）

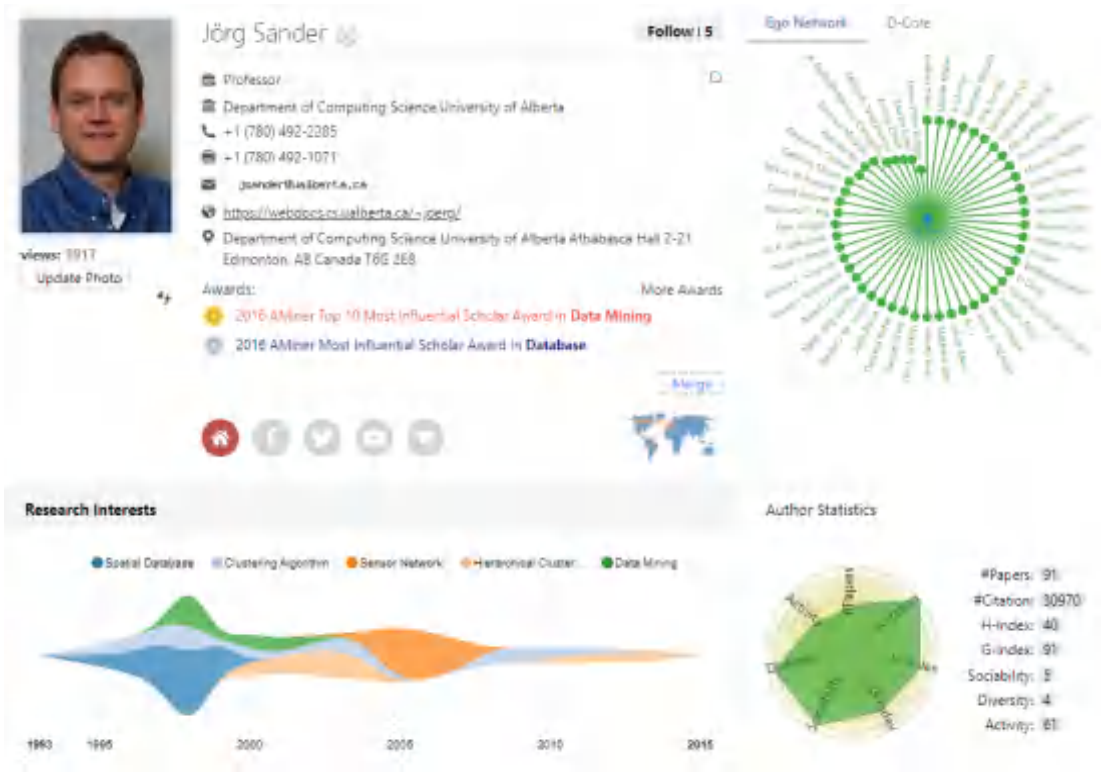


刘兵，数据挖掘重量级人物，是语义分析、观点挖掘研究领域的开创者之一。他是伊利诺伊大学芝加哥分校（University of Illinois at Chicago，简称 UIC）计算机科学系优秀教授，于爱丁堡大学获得人工智能博士学位。他的研究兴趣包括：语义分析和观点挖掘、机器学习、假/欺骗性意见检测、自然语言处理（NLP）。

他曾于 2013 年担任 ACM SIGKDD 的主席（2013 年 7 月 1 日至 2017 年 6 月 30 日）。曾担任许多其他数据挖掘领域会议主席（包括 ICDM，CIKM，WSDM，SDM 和 PAKDD），他担任 TKDE，TWEB，DMKD 等期刊的副主编；ACM、AAAI 和 IEEE 的会士（Fellow）。

刘兵教授于 2018 年获 SIGKDD 创新奖（Innovation Award），这是数据挖掘领域最高奖项。他的两篇论文（“Integrating Classification and Association Rule Mining（KDD 1998）”与“Mining and Summarizing Customer Reviews（SIGKDD 2003）”）分别获得 2014 年与 2015 年的时间测试奖（Test of Time）。

- Jörg Sander

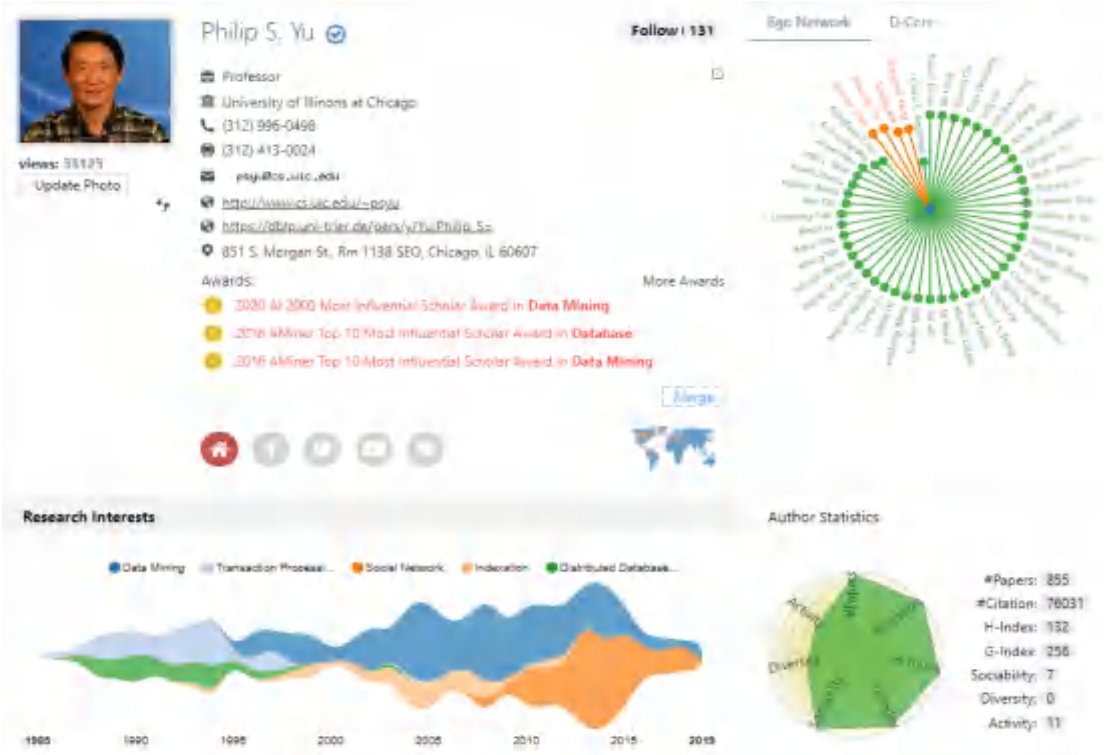


Jörg Sander，DBSCAN 算法创始人之一，聚类分析领域顶尖学者。他是阿尔伯塔大学（University of Alberta，简称 UA）计算机科学系教授，于 1998 年获得慕尼黑大学计算机科学博士学位。他的研究领域包括：数据库中的知识发现、空间数据挖掘、时空索引、生物信息学，目前关注聚类分析、生物数据库中的数据挖掘与动态位置数据。

Jörg Sander 教授是多个数据挖掘领域期刊的审稿人（包括 IEEE TKDE，The international Journal on Very Large Databases，DAMI，Information Systems，Transactions on Pattern Analysis and Machine Intelligence，KAIS，JAIR，JMLR， VLDB，Machine Learning，Systems Man and Cybernetic，PAA，Pattern Recognition Letters）。曾任第二届国际时空数据管理研讨会的研讨会联合主席和组织者，2002 ACM SIGKDD 知识发现与数据挖掘国际会议主席。

他曾获 2015 年 ACM 杰出会员奖，他在论文“A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise （SIGKDD 1996）”中提出的 DBSCAN 算法，对基于密度的聚类算法产生了巨大的影响，并已成为聚类算法中公认的重要算法之一，这篇论文获 2014 年 SIGKDD 时间测试奖（Test of Time）。

● 俞士纶

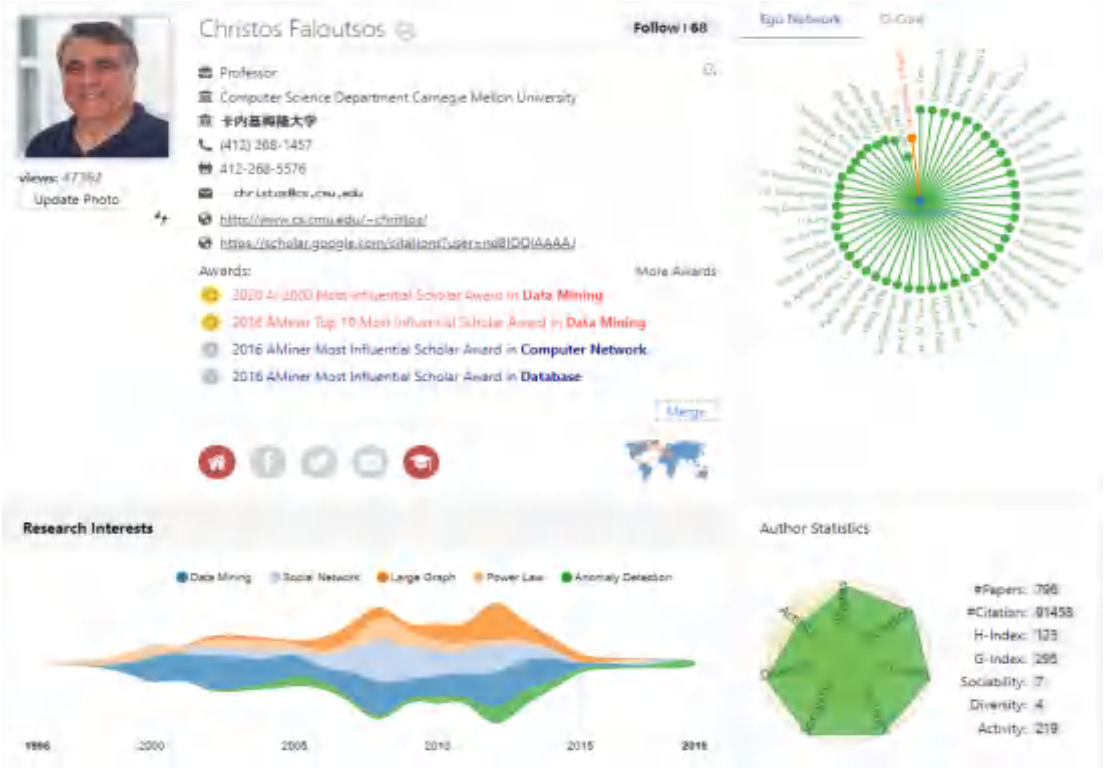


俞士纶，大数据挖掘领域领军人物。他是美国伊利诺伊大学芝加哥分校（UIC）计算机科学系教授，1978 年于斯坦福大学获得电气工程博士学位。他的研究领域为：数据挖掘、隐私保护发布和挖掘、数据流、数据库系统、互联网应用和技术、多媒体系统、并行和分布式处理以及性能建模。

俞士纶博士是 ACM 和 IEEE 院士（Fellow），现担任 ACM TKDD 主编、ACM CIKM 会议指导委员会成员、清华大学数据科学研究院院长，信息技术领域 Wexler 讲座教授。加入 UIC 之前，曾在美国 IBM Watson 研究中心工作多年，创建了世界知名的数据挖掘及数据管理部门。他还曾担任 IEEE ICDE 和 IEEE Data Mining 会议指导委员会成员，IEEE TKDE 主编。

因在大数据挖掘、融合以及匿名化方面深具影响力的研究和科学贡献，俞士纶博士曾获 2016 年 ACM SIGKDD 创新奖（2016 ACM SIGKDD Innovation Award）；因在大数据的可扩展性索引、查询、搜索、挖掘以及匿名化问题上开创性和基础性的创新贡献，获得 2013 年 IEEE CS 技术成就奖；因在数据挖掘领域开创性的贡献，获得 2003 年 IEEEICDM 研究贡献奖。此外他还曾获 2014 年 EDBT 时间检测奖（Test of Time Award）、2013 年 ICDM 十年最有影响论文奖。

● Christos Faloutsos

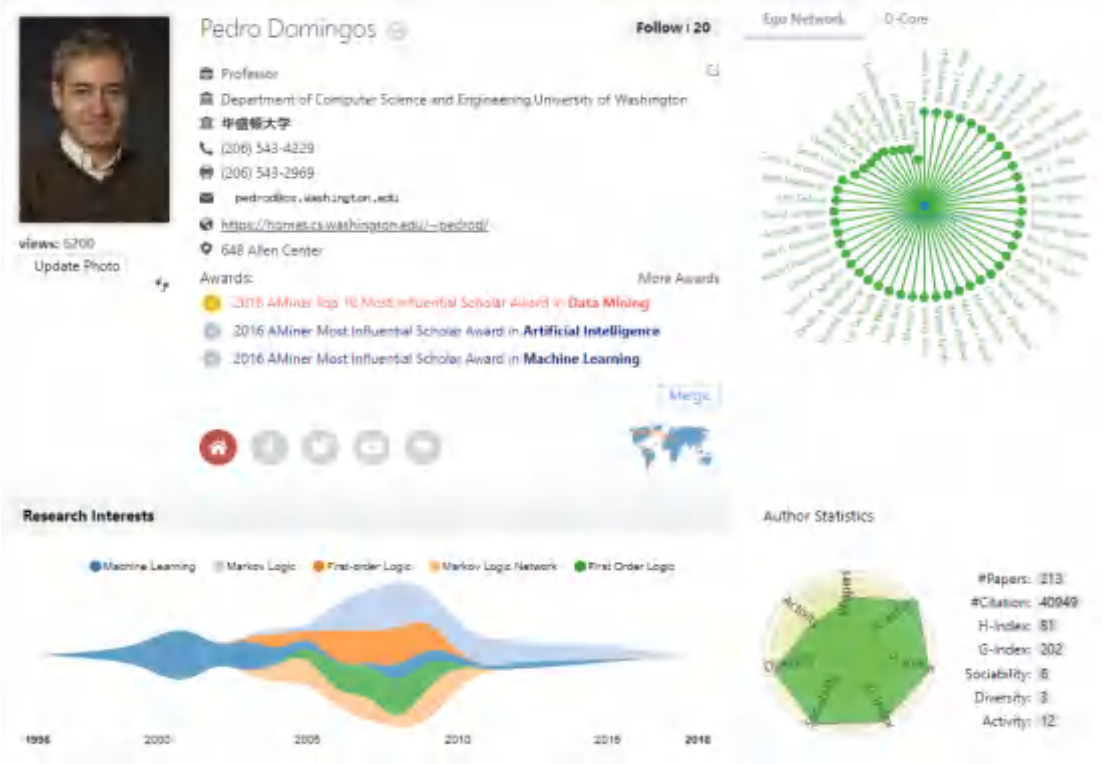


Christos Faloutsos，多媒体数据挖掘顶尖学者。他是美国卡耐基梅隆大学（Carnegie Mellon University，简称 CMU）计算机科学系教授，1989 年于牛津大学获得博士学位。研究领域为：数据挖掘、数据库性能、空间和多媒体数据库、图数据挖掘、生物与医学数据库索引。

Christos Faloutsos 教授是 SIGKDD 执行委员会成员、ACM（国际计算机学会）会员、IEEE TKDE 副主编，曾任 1999 年 ACM SIGMOD 主席、2003 年 ACM KDD 主席。

Christos Faloutsos 教授曾获 1997 年 VLDB 10 年论文奖、2010 年 SIGKDD 创新奖（Innovation Award，数据挖掘领域最高奖项）、2016 年 ICDM 10 年最高影响力论文奖。他是 SIGKDD 2016 的最佳研究论文“FRAUDAR： Bounding Graph Fraud in the Face of Camouflage”的作者之一，这篇论文研究如何在用户商品评论数据中识别假的评论，提出的方法能够识别伪装，比传统方法更加有效。除此之外，他曾获 1989 年“青年研究者奖”，这是美国国家科学基金会授予的青年科学家最高奖，1994 年 SIGMOD 最佳论文奖、2005 年 SIGKDD 最佳研究论文奖、2006 年 ICDM 研究贡献奖。

● Pedro Domingos



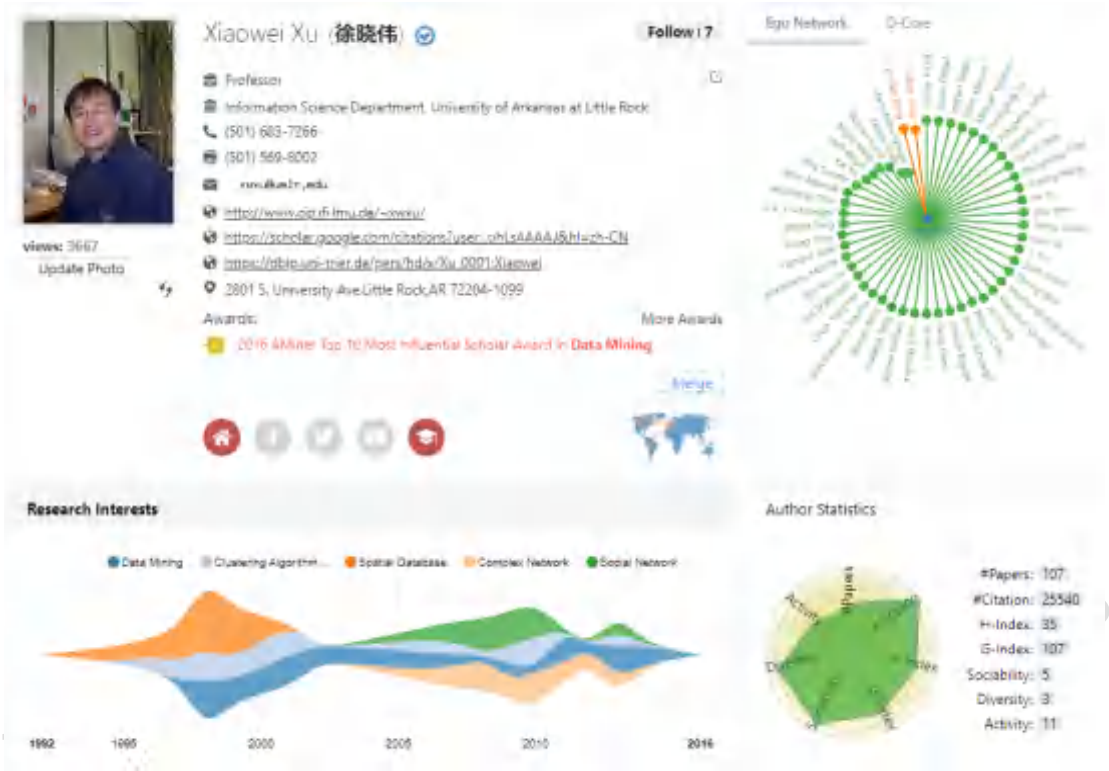
Pedro Domingos，VFDT 实时数据流挖掘系统创始人。他是华盛顿大学（University of Washington，简称 UW）计算机科学与工程系教授，1997 年于加州大学欧文分校（University of California, Irvine，简称 UCI）获信息与计算机科学博士学位。研究领域为：机器学习和数据挖掘、统计关系学习、可追溯的深度学习、机器阅读、集体知识基础、大规模机器学习。

Pedro Domingos 教授是 AAAI 会士（Fellow）、机器学习期刊（Machine Learning journal）的编辑委员会成员、国际机器学习协会（International Machine Learning Society）的联合创始人。曾任 SIGKDD 2003 联合主席、SRL-2009 联合主席、JAIR 副主编，曾在 AAAI，ICML，IJCAI，KDD，NIPS，SIGMOD，UAI，WWW 等项目委员会任职。

Pedro Domingos 教授曾获 SIGKDD 2014 创新奖（Innovation Award，数据挖掘领域最高奖项）、SIGKDD 2014 创新贡献奖、SIGKDD 1999 最佳基础研究论文奖，他与 Geoff Hulten 的论文“Mining High-Speed Data Streams （SIGKDD 2000）”曾获 SIGKDD 2015 时间测试奖（Test of Time），该论文提出了 VFDT 这一实时数据流挖掘系统，用于快速处理不断增长的数据记录。除此之外，他还

曾获 2015 年第二十四届国际人工智能联合会议杰出论文奖、2009 年自然语言处理经验方法会议最佳论文奖、2005 年第九届欧洲数据库知识发现原理与实践会议最佳论文。

● 徐晓伟



徐晓伟，基于密度的聚类算法 DBSCAN 创始人之一。他是美国阿肯色大学小石城分校（University of Arkansas at Little Rock，简称 UALR）信息科学系教授，1998 年于慕尼黑大学（Ludwig-Maximilians-Universität München，简称 LMU）获得计算机科学博士学位。研究领域为：图数据挖掘、潜在结构数据挖掘、非常大的图形数据管理和生物信息学。

徐晓伟教授现任国家毒理学研究中心（NCTR）的 ORISE 教师研究计划成员，曾任 SIDKDD 2015 计划委员会成员。

他是论文“A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise （SIGKDD 1996）”的作者之一，该论文获 SIGKDD 2014 时间测试奖，文中提出的 DBSCAN 算法对基于密度的聚类算法产生了巨大的影响，并已成为聚类算法中公认的重要算法之一，截止于 2018 年 9 月其 AMiner 引用次数为 12870 次。该算法还被 FDA 用于 atBioNet，一种基因

组学和蛋白质组学的知识扩展，网络建模和可视化的集成网络分析工具。

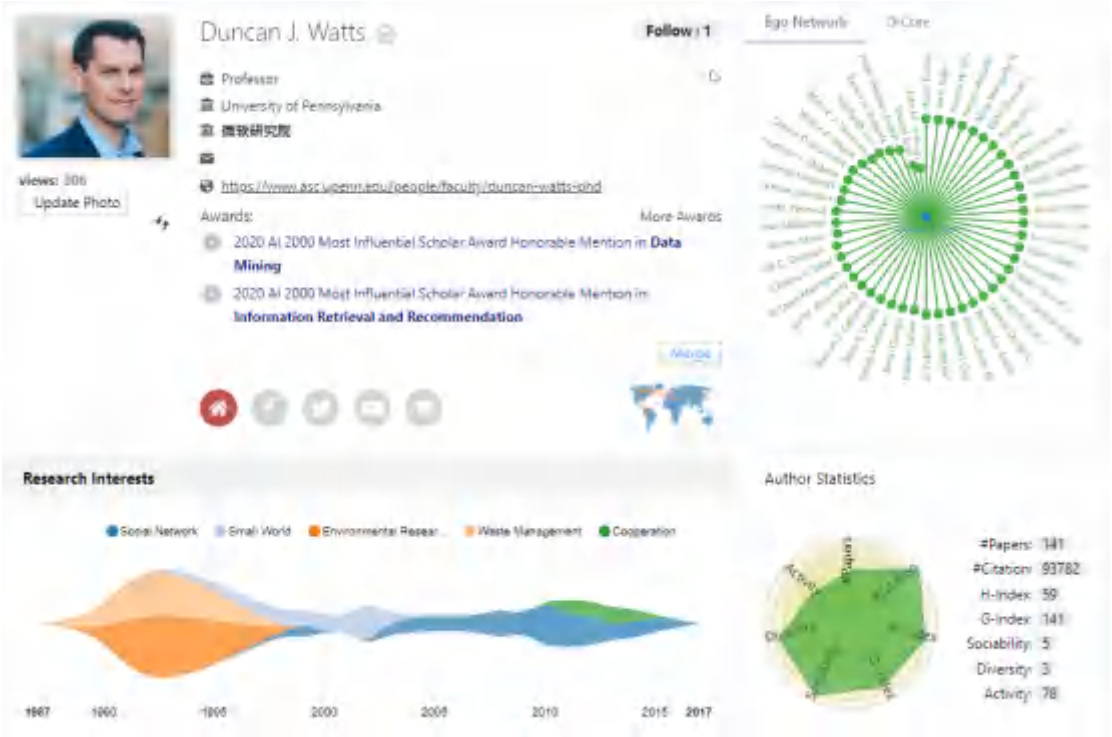
● Hector Garcia-Molina

Hector Garcia-Molina，网络数据挖掘领域顶尖学者，在学术领域与应用领域均享有很高声誉。他是斯坦福大学计算机科学和电气工程系的教授，1979年于斯坦福大学获得计算机科学博士学位。他的研究领域为：分布式计算系统、数字图书馆和数据库系统。

他是美国计算机协会（ACM）研究员、美国艺术与科学学院（AAAS）研究员，美国国家工程院（NAE）院士、苏黎世联邦理工学院的荣誉博士，现任总统信息技术咨询委员会（PITAC）的成员、Onset Ventures 的风险投资顾问、Oracle 董事会成员、State Farm Technical Advisory Council 的成员。他曾担任谷歌联合创始人谢尔盖布林的顾问。

他的论文“The Evolution of the Web and Implications for an Incremental Crawler，VLDB 2000”获 2010 VLDB 10 年最佳论文奖，研究了如何构建有效的增量爬虫，结合了最佳的设计选择，提出了增量爬虫的架构。此外他还曾获 2009 SIGMOD 最佳演示奖、2007 ICDE 影响力论文奖、1999 ACM SIGMOD 创新奖。

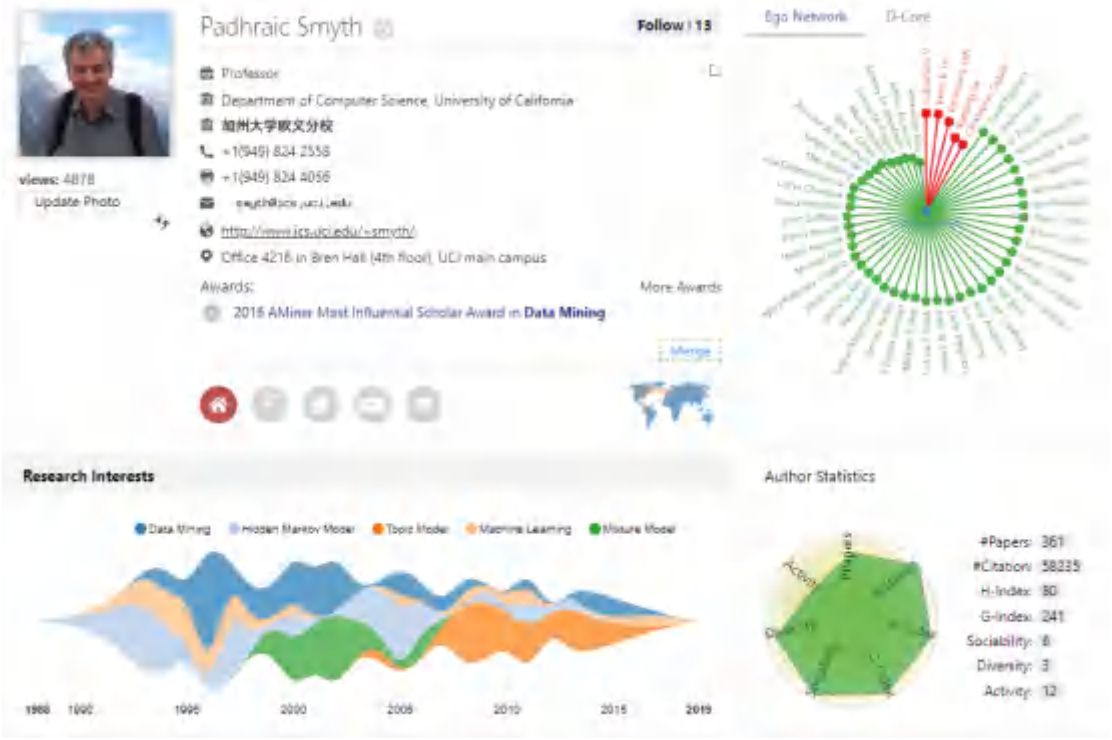
● Duncan J.Watts



Duncan J.Watts，卡内基梅隆大学计算机科学系教授，1997 年于康奈尔大学获应用力学博士学位。他的研究领域为：图形和流的挖掘、数据库性能评估、生物和医学数据库的索引、数据挖掘。

Duncan J.Watts 教授是 Microsoft Research 的首席研究员，也是 MSR-NYC 实验室的创始成员。他还是康奈尔大学的 AD 白人教授（AD White Professor）。在 2012 年加入 Microsoft Research 之前，他于 2000 年至 2007 年在哥伦比亚大学担任社会学教授，之后在雅虎担任首席研究科学家。他曾担任多个国际会议主席，如 ACM SIGMOD、SIGKDD 2013。他还曾获 2014 年埃弗雷特罗杰斯·罗杰斯奖、2013 年拉格朗日 - CRT 复杂性科学基金奖。

● Padhraic Smyth



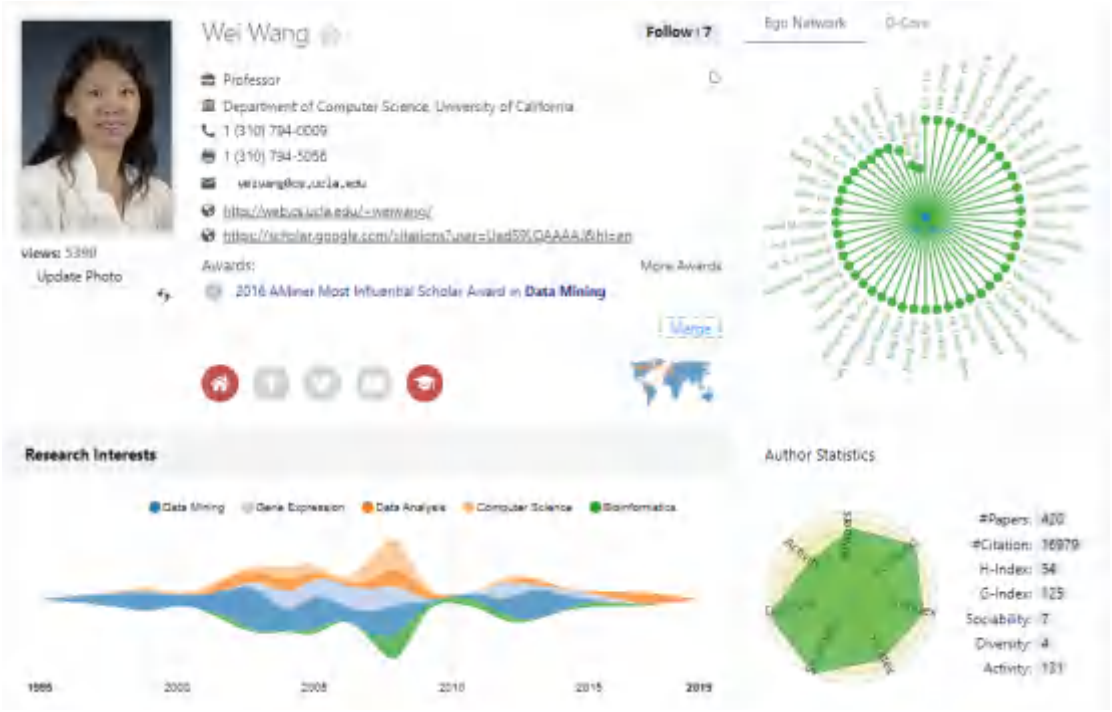
Padhraic Smyth，他是加州大学欧文分校（University of California,Irvine，简称 UCI）信息与计算机科学系教授，1988 年获得的加州理工学院博士学位。他的研究领域为：机器学习、数据挖掘、模式识别、应用统计。

Padhraic Smyth 教授是国际计算机学会（ACM，2013）与美国人工智能协会（AAAI，2010）的会士（Fellow），他曾任 2014 年 UCI 数据科学计划的创始董事（2007-2014）、Netflix 奖项竞赛的学术顾问（2006-2009）、帕萨迪纳喷气推进实验室（Jet Propulsion Laboratory）的技术组组长（1988-1996），曾担任 Machine Learning Research、ASA、TKDD 等期刊的编辑和顾问职位。他与 eBay，东芝，三星，甲骨文，诺基亚和 AT&T 等公司均有合作。

他曾获 SIGKDD 2009 创新奖（innovation award），这是数据挖掘领域的最高奖项。除此之外，他还曾获 SIGKDD 2002 最佳论文奖、2001 IBM 教员合作奖、SIGKDD 2000 最佳论文奖第二名、SIGKDD 1998 最佳论文奖第二名、SIGKDD 1997 最佳论文奖、1997 国家科学基金会职业奖（CAREER award）、1997 UCI ACM 教学奖。

排名靠前的其他学者简介：

- Wei Wang



Wei Wang，Leonard Kleinrock 加州大学洛杉矶分校（UCLA）计算机科学讲座教授，可扩展分析研究所（ScAi）主任。她于 1999 年获得加州大学洛杉矶分校计算机科学博士学位。2002 年到 2012 年，在北卡罗来纳大学教堂山分校担任计算机科学教授，卡罗莱纳基因组科学中心和 Lineberger 综合癌症中心的成员。且在 1999 年至 2002 年间担任 IBM TJ Watson 研究中心的研究人员。Wang 博士的研究兴趣包括大数据分析、数据挖掘、生物信息学和计算生物学以及数据库。

Wei Wang 博士曾获 2000 年和 2001 年 IBM 发明成就奖，2003 年 UNC 初级教师发展奖，2005 年 NSF 教师早期职业发展（CAREER）奖。2005 年被任命为微软研究院新任教师，2007 年 UNC Phillip 和 Ruth Hettleman 艺术与学术成就奖，2012 年 IEEE ICDM 杰出服务奖，2013 年大川基金会研究奖，2016 年 ACM SIGKDD 服务奖。

Wang 博士曾担任 IEEE TKDE（知识与数据工程交易）、IEEE Transactions on Big Data、ACM TKDD、KNOWL INF SYST、WIRES DATA MIN KNOWL、J COMPUT BIOL、IEEE ACM T COMPUT BI、IJKDB 的 associate editor（副主编）；INT J DATA MIN BIOIN 的 editorial board member（编委会成员）。她是国际会议组织和计划委员会的成员，委员会包括 ACM SIGMOD、ACM SIGKDD、ACM BCB、VLDB、ICDE、EDBT、ACM CIKM、IEEE ICDM、SIAM DM、SSDBM、ISMB、RECOMB、BIBM。

她于 2015 年被选为 ACM Bioinformatics、Computational Biology（计算生物学）和 SIGBio（生物医学信息学）特别兴趣小组的董事会成员。

● Thorsten Joachims

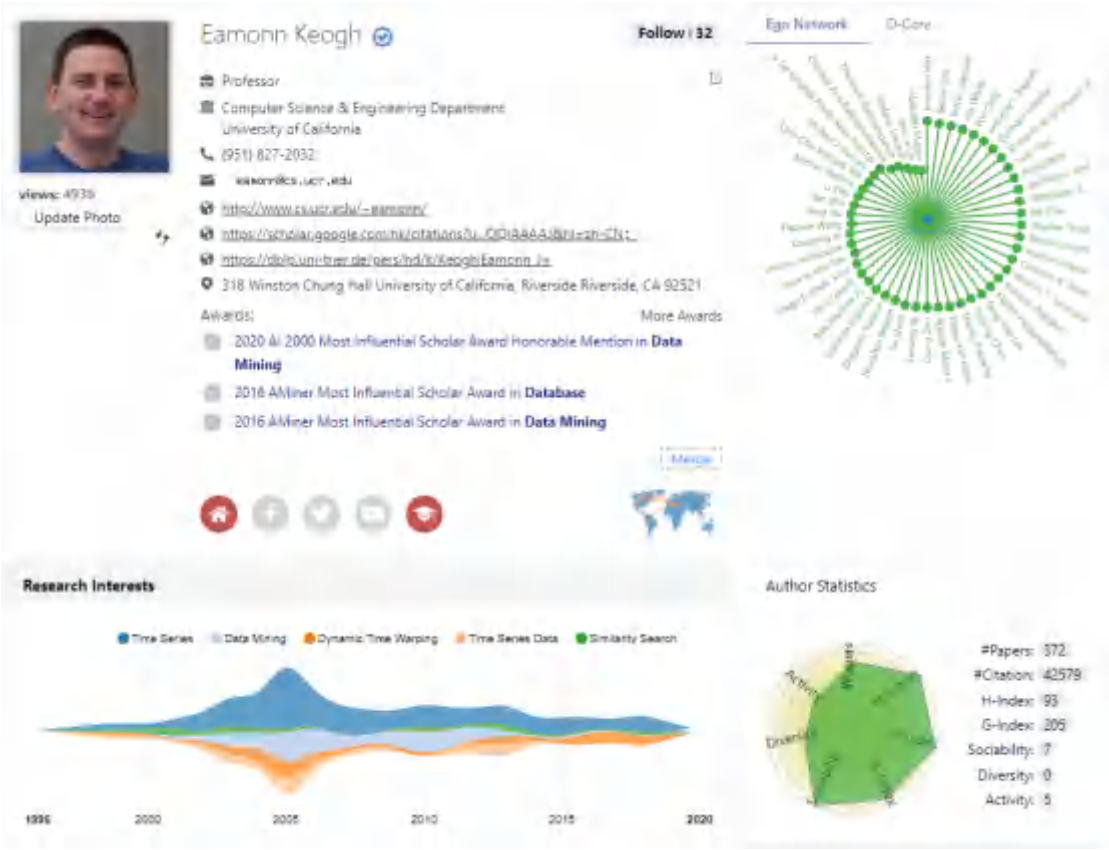
Thorsten Joachims，搜索引擎算法革新的重要推动者。他是美国康奈尔大学（Cornell University）计算机科学系教授，2001 年于多特蒙德工业大学获 AI-unit 博士学位。他的研究领域为：机器学习、人类行为学习数据和隐性反馈、搜索引擎算法、推荐算法、人与教育相关算法，目前从事的项目有 Propensity SVM-Rank、POEM、反事实评估与学习。

Thorsten Joachims 教授现任 ACM、AAAI 与 Humboldt 的会士（Fellow）。他曾与莫纳什大学的 Geoff Webb 教授共同担任 SIGKDD 2015 程序委员会主席，曾与 Johannes Fuernkranz 教授共同担任 2010 ICML 程序委员会主席，曾任 JMLR（2004 年-2009 年）、DMKD（2005 年-2008 年）MLJ 的行为编辑，JAIR 顾问委员会成员。

他曾获 2017WSDM 最佳论文奖、2009 ICML 最佳十年论文奖，他的论文“Optimizing Search Engines Using Click-Through Data （SIGKDD 2002）”

获 SIGKDD 2015 时间检测奖（Test of Time），论文率先提出了通过挖掘利用用户点击数据从而提高搜索引擎结果准确率的方法，该方法激发了一系列后继工作的发展并成为目前学术界与工业界搜索引擎研发的公认基础方法。此外他还曾获 SIGKDD 2016 最佳学生论文奖亚军、ISMIR 2014 最佳学生论文奖、ECML 2009 最佳论文奖、ICML 2005 最佳论文奖。

● Eamonn Keogh



Eamonn Keogh，时间序列领域顶尖学者。他是加州大学河滨分校（University of California, Riverside，简称 UCR）计算机科学与工程系教授，2001 年于加利福尼亚大学欧文分校（University of California, Irvine，简称 UCI）获计算机科学博士学位。研究领域为：时间序列、数据挖掘、指数化、动态时间扭曲、相似性搜索、异常检测。

Eamonn Keogh 教授曾获 2017 IEEE ICDM 最佳论文奖、2007 IEEE ICDM 最佳论文奖，他是论文“Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping”的作者之一，获 SIGKDD 2012 最佳论文奖。除此之外他还曾获 SIGMOD 2001 最佳论文奖、SIAM SDM 2010 最佳学

生论文，与 Bilson、JCDL 2009 最佳学生论文、SIGKDD 1997 最佳论文奖亚军。

● Rakesh Agrawal

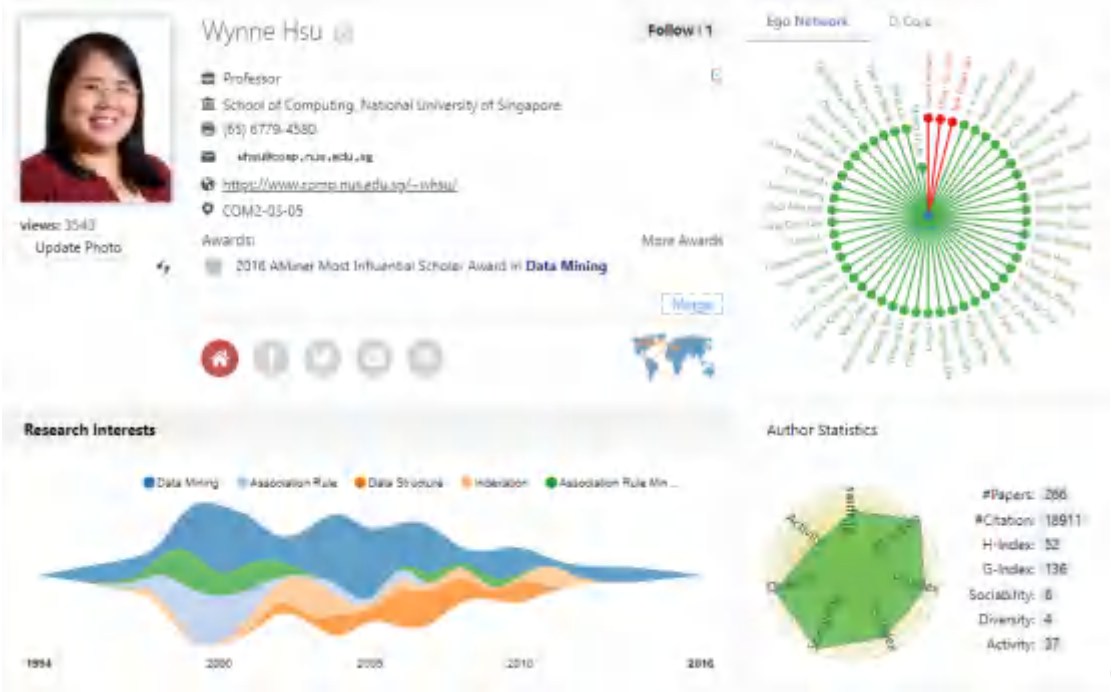
Rakesh Agrawal，是基础数据挖掘概念和技术开发的关键人物，开创了数据隐私的关键概念。他是 Data Insights 实验室的总裁兼创始人、微软研究实验室研究员，1982 年于威斯康星大学麦迪逊分校（University of Wisconsin-Madison，简称 UW-Madison）获计算机科学博士学位。研究领域为：数据挖掘、数据库系统、信息检索、关联法则、数据隐私。主要成就有：IBM 的商业数据挖掘产品 Intelligent Miner、DB2 Mining Extender、DB2 OLAP Server、WebSphere Commerce Server、Polyglot 面向对象类型系统、Alert 主动数据库系统、Ode（对象数据库和环境）、Alpha（具有通用传递闭包的关系数据库的扩展）、Nest 分布式系统。

Rakesh Agrawal 博士是美国国家工程院（NAE）院士、ACM 会士、IEEE 会士，曾任 IBM 院士，并在 IBM Almaden 研究中心领导 Quest 小组，《科学美国人（Scientific American）》将他列为首批 50 位顶尖科学家和技术专家。

Rakesh Agrawal 博士是因关联规则，挖掘序列等方面的开创性工作，获 2000 SIGKDD 创新奖（首届），他是“Order preserving encryption for numeric data”与“Mining association rules between sets of items in large databases”的第一作者，这两篇论文分别与 2014 年于 2013 年获 SIGMOD

时间测试奖。除此之外，他还曾获 2000 ACM-SIGMOD Edgar F. Codd 创新奖、2004 VLDB 10-Yr 最具影响力论文奖、ICDE 2008 最具影响力论文奖。

● Wynne Hsu

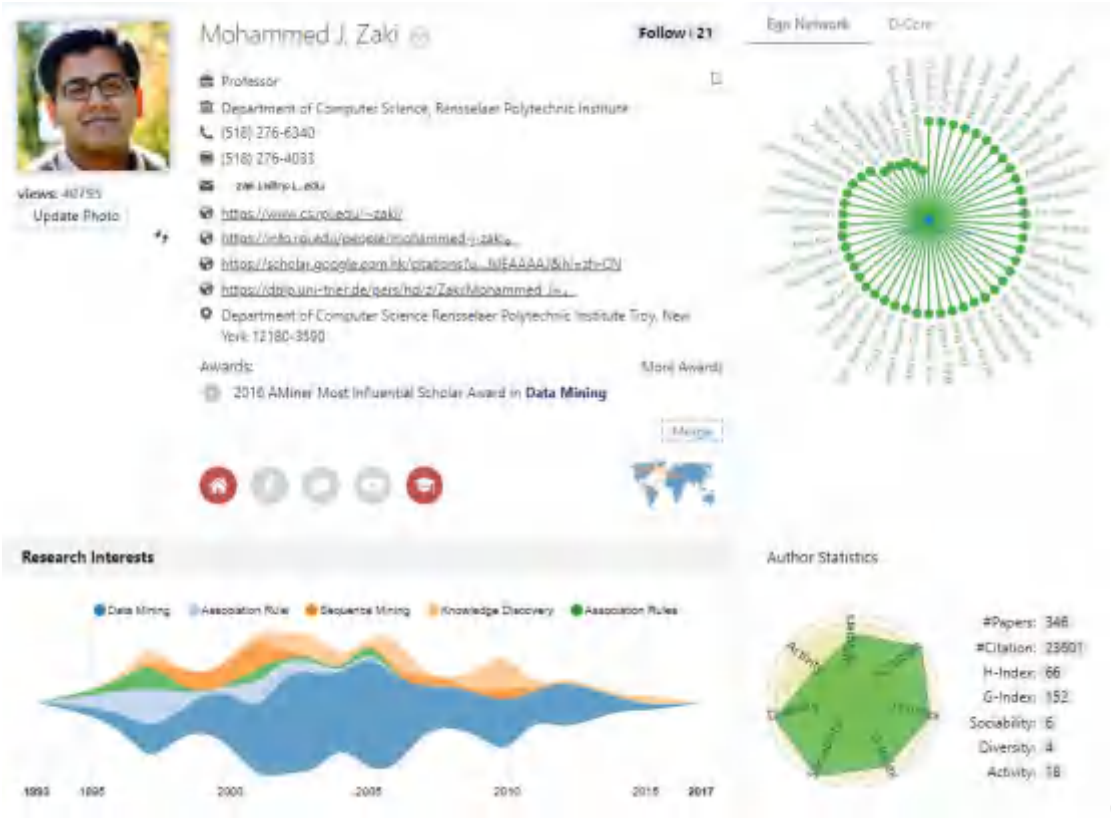


Wynne Hsu，图数据、医学数据挖掘领域的顶尖女性学者。她是新加坡国立大学（National University of Singapore，简称 NUS）计算机科学系教授，于普渡大学（Purdue University）获电气与计算机工程博士学位。研究领域为：数据挖掘、视网膜图像数据、生物医学数据挖掘。正在进行的项目为医学 AI、社交媒体分析。

Wynne Hsu 教授曾任 2013 VLDB 的副主编、2010 ICDM 的研究会副主席、2011 SIGMOD 的程序委员会成员、2010 COMAD 的程序委员会成员、2001-2009 年 SIGKDD 的程序委员会成员、2009 ICDM 的程序委员会成员。除此之外，她还曾担任多个数据挖掘领域会议的程序委员会成员，如 DASFAA、SAC、PRICAI、DEXA、ADMA、PAKDD。

Wynne Hsu 教授与伊利诺伊大学芝加哥分校的刘兵、LinkedIn 公司的 Yiming Ma 合作了论文“Integrating Classification and Association Rule Mining（KDD 1998）”，该文获 SIGKDD 2014 时间检测奖（Test of Time），该论文率先提出了整合关联规则和分类算法来帮助数据挖掘分类的思想，该方法激发了一系列后继工作的发展。

● Mohammed J. Zaki

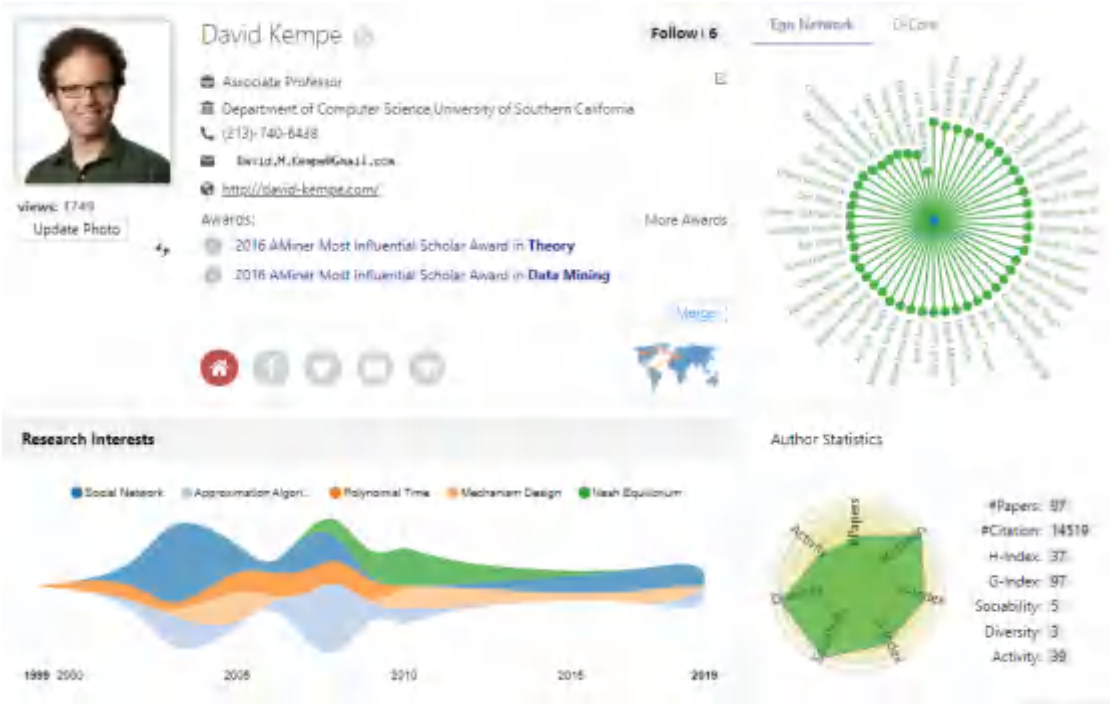


Mohammed J.Zaki，文本挖掘、社交网络与生物信息数据挖掘领域十分活跃的学者。他是伦斯勒理工大学（Rensselaer Polytechnic Institute，简称 RPI）计算机科学系的教授，1998 年于罗切斯特大学（University of Rochester）获计算机科学博士学位。他的研究领域为：新颖数据挖掘和机器学习技术的开发，特别是在文本挖掘、社交网络和生物信息学中的应用。

Mohammed J. Zaki 教授是 BIOKDD 系列研讨会的创始联合主席、Data Mining and Knowledge Discovery 的副主编、ACM SIGKDD 的董事会成员、ACM 杰出科学家和 IEEE 研究员。他曾任 CIKM 2018、IEEE Big Data 2015、ICDM 2012、CIKM 2012、BIBM 2011、SIGKDD 2009、PAKDD 2008、SDM 2008 的项目联合主席。他还曾担任 SADM 的区域编辑和 ACMTKDD、DMKD、SADM、IEEETKDE 等期刊的副主编，是 2014 年剑桥大学出版社出版的数据挖掘和分析教科书的作者。

他在 2010 年、2011 年和 2012 年三次获惠普创新研究奖，2003 年获得 ACM 服务奖，2002 年获得能源部早期职业首席研究员奖，2001 年获国家科学基金会职业奖。

● David Kempe

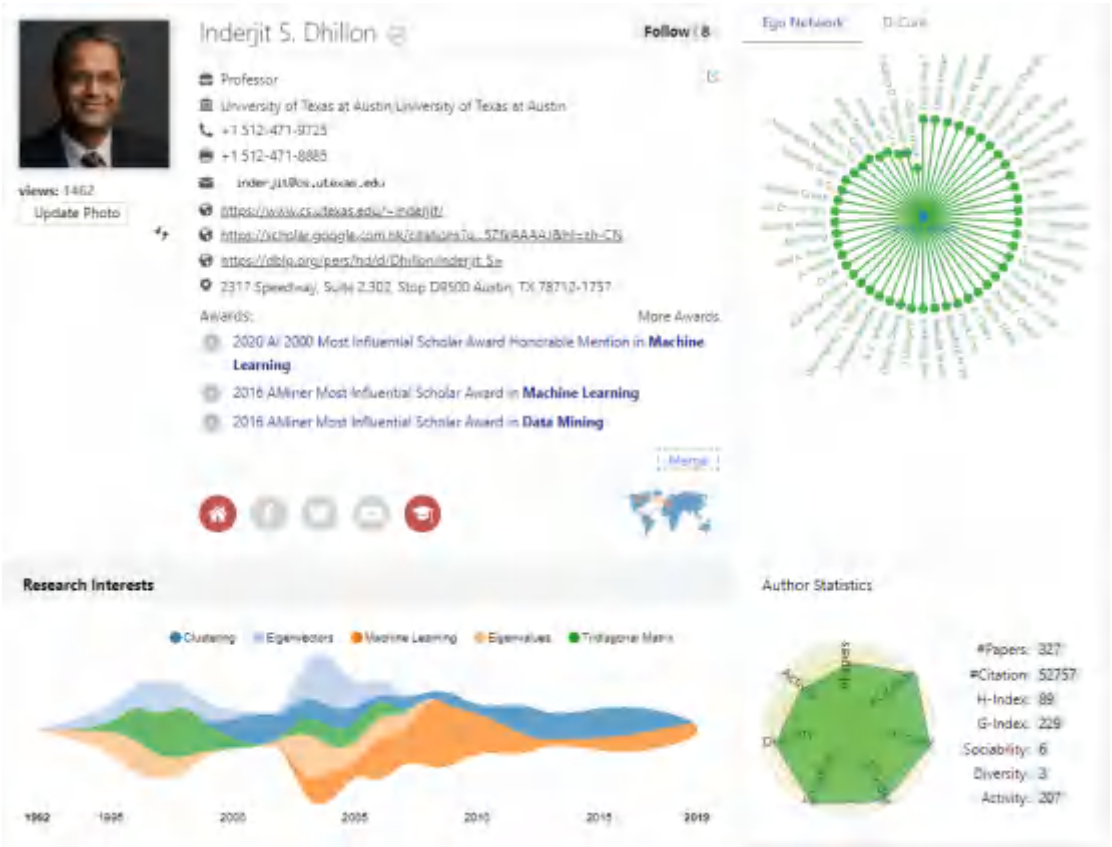


David Kempe，社交网络数据挖掘领域的学者，年轻有为。他是南加州大学（University of Southern California，简称 USC）计算机科学系副教授，2003 年于康奈尔大学（Cornell University）获计算机科学博士学位。他的研究领域为：计算机科学理论和算法的设计和分析，特别强调社会网络、经济学和计算交叉的主题，以及算法学习问题。

David Kempe 博士曾与 Ilias Diakonikolas 合作担任 2018 STOC 的区域主席，与 Shaddin Dughmi 和 Hamid Nazerzadeh 共同主持 2015 年 NEGT。

David Kempe 博士与康奈尔大学的 Jon Kleinberg、Eva Tardos 的“Maximizing the Spread of Influence through a Social Network （SIGKDD 2003）”获 SIGKDD 2014 时间检测奖（Test of Time），该论文是社交网络中影响力最大化问题研究的里程碑式工作，其在数学上证明了求解该类问题算法的性能指标界限，为此领域后继研究工作提供了理论支持。他曾获 ACM-EC 2014 最佳论文奖、SIGKDD 2003 最佳论文奖。除此之外他还曾获 2011 年 ICML 2011 年会议杰出论文奖、2009 年南加州大学梅隆分校优秀导师奖、2008 年海军研究办公室青年研究员奖、2005 年 NSF 职业奖等等奖项。

- Inderjit S. Dhillon



Inderjit S.Dhillon 是德克萨斯大学奥斯汀分校（University of Texas at Austin，简称 UT-Austin）计算机科学系教授，1997 年于伯克利加州大学（University of California, Berkeley）获计算机科学博士学位。他的研究领域为：机器学习、数据挖掘、数值线性代数、数值优化、网络分析、生物信息学；目前的一些研究课题为：高维数据分析、大数据分析的分治方法、社交网络分析与基因疾病预测。

Inderjit S. Dhillon 教授是 ACM、IEEE、SIAM、AAAS 的会士（Fellow），UT Austin 的 Gottesman 家族百年教授，同时担任 ICES 大数据分析中心的主任、A9 亚马逊的高级研究员。曾任 JMLR 编委、IEEE T PATTERN ANAL 编辑、SIMAX 编辑、《Foundations and Trends in Machine Learning》编辑。

Inderjit S. Dhillon 教授的“Scalable Coordinate Descent Approaches to Parallel Matrix Factorization for Recommender Systems”获 ICDM 2012 最佳论文奖。他还曾获 ICES 2013 杰出研究奖，SIAM 2011 杰出论文奖，2010-2011 Moncrief Grand Challenge 奖，SIAM 2006 线性代数奖，2001 大学研究卓越奖和 2001 NSF 职业奖。

3.2.2 近十年代表学者简介

以下是近十年在 SIGKDD 上发表论文的引用量排名靠前的学者。其中，俞士纶、Christos Faloutsos 和 Jiawei Han 三位学者在数据挖掘发展历史中做出了很大贡献，上一节已做介绍，本节不再赘述。

Jure Leskovec Follow | 64

h-index: 105 | #Paper: 309 | #Citation: 70120

Associate Professor

Computer Science Department, Stanford University

6223

Social Network | Social Networks | Machine Learning | Social Media | Data Mining | Information Network

Similar Authors

Philip S. Yu Follow | 131

h-index: 132 | #Paper: 855 | #Citation: 76031

Professor

University of Illinois at Chicago

53127

Data Mining | Transaction Processing | Social Network | Indexation | Distributed Databases

Similar Authors

Christos Faloutsos Follow | 68

h-index: 125 | #Paper: 796 | #Citation: 91458

Professor

Computer Science Department Carnegie Mellon University

47364

Data Mining | Social Network | Large Graph | Power Law | Anomaly Detection | Information Retrieval

Similar Authors

Yehuda Koren Follow | 3

h-index: 48 | #Paper: 112 | #Citation: 21634

Staff research scientist

Google

1284

Graph Drawing | Collaborative Filtering | Recommender System | Recommender Systems | Data Mining

Similar Authors

Jiawei Han (韩家炜) Follow | 272

h-index: 175 | #Paper: 1252 | #Citation: 183472

Professor

Department of Computer Science, University of Illinois at Urbanan

161099

Data Mining | Data Cube | Data Analysis | Heterogeneous Information Network | Information Retrieval

Similar Authors

Jie Tang (唐杰) Follow | 208

h-index: 61 | #Paper: 528 | #Citation: 15169

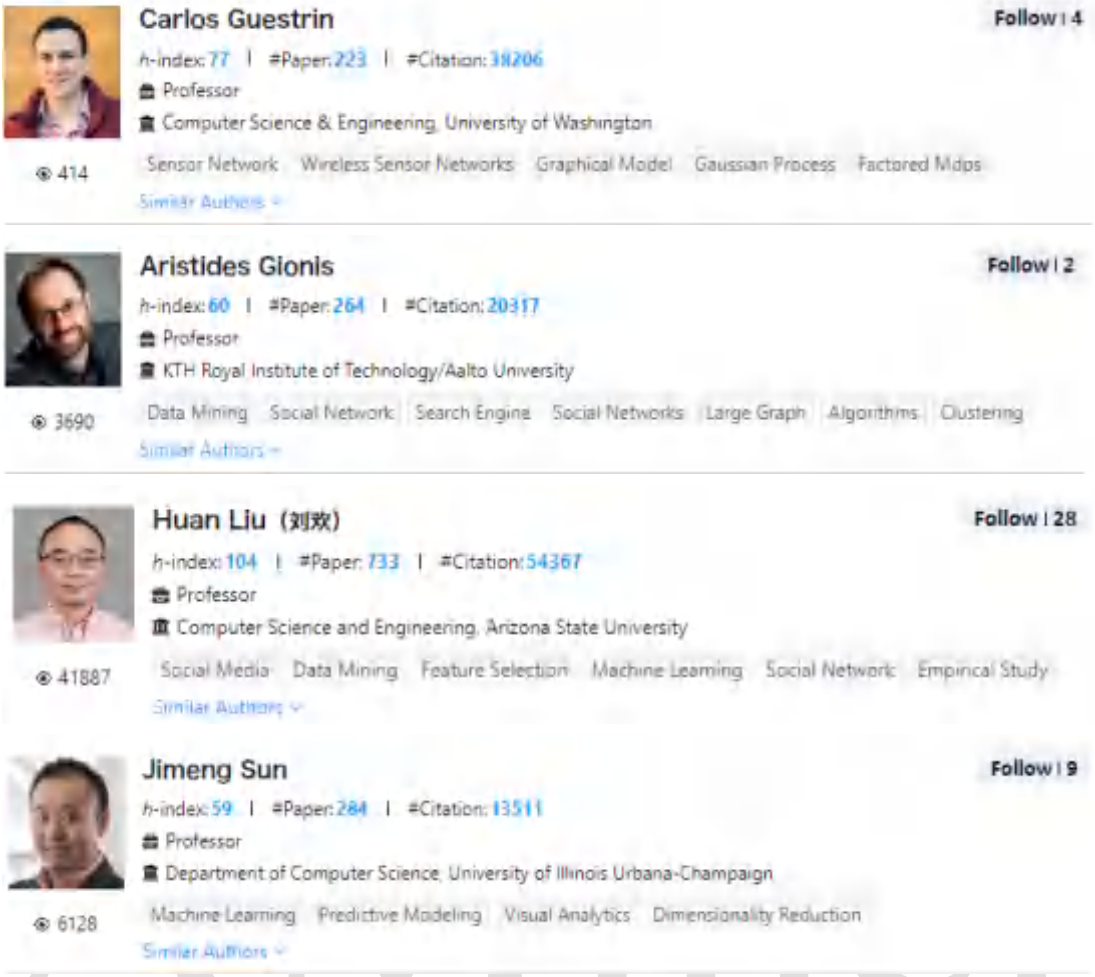
Professor

Department of Computer Science and Technology, Tsinghua University

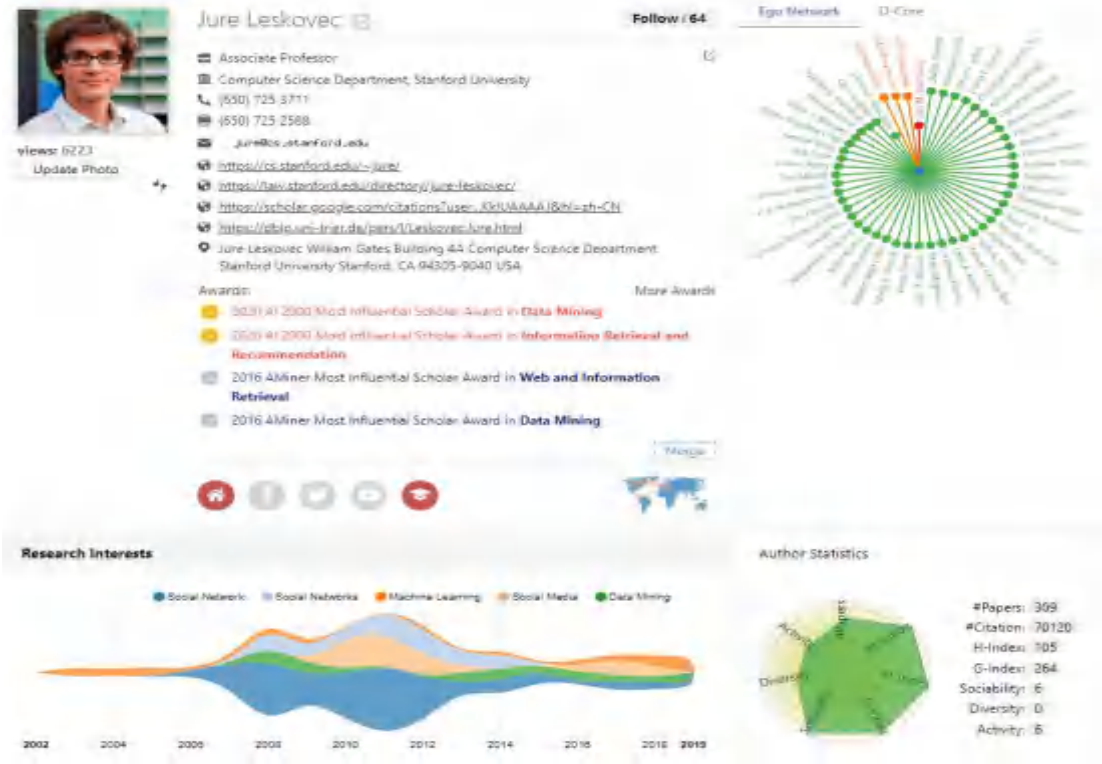
53981

Social Network | Social Influence | Semantic Web | Data Mining | Social Networks | Heterogeneous Network

Similar Authors



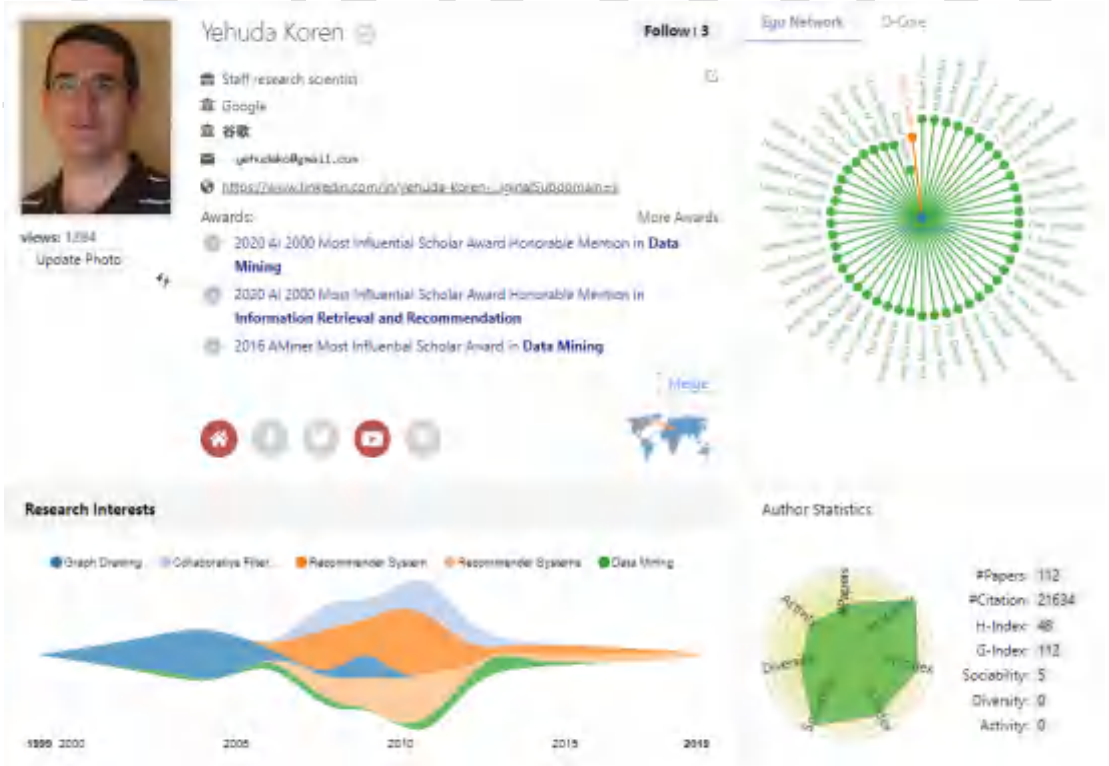
- Jure Leskovec



Jure Leskovec 是数据挖掘应用方面的顶尖学者。斯坦福大学计算机科学系副教授，2008 年于美国卡耐基梅隆大学（Carnegie Mellon University，简称 CMU）获得计算机科学学院机器学习系博士学位。研究领域为：大型社会和信息网络的数据挖掘和建模、大数据网络与媒体。

Jure Leskovec 副教授是 Chan Zuckerberg Biohub 的研究员。他是 SIGKDD 2017 最佳学生论文亚军的导师，他是“Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data”作者之一，获 SIGKDD 2017 最佳研究性论文亚军；他是“Modeling information diffusion in implicit networks”作者之一，获 ICDM 2010 最佳应用论文奖；他是“Graphs over time: densification laws, shrinking diameters and possible explanations”作者之一，获 SIGKDD 2005 最佳研究性论文。除此之外，他还在 JWRPM、SIGKDD、WSDM、ICDM、WWW 等多个期刊会议中多次获奖，如 SIGKDD 2010 最佳研究性论文奖亚军、JWRPM 2008 最佳研究性论文奖、SIGKDD 2007 最佳学生论文奖。

- Yehuda Koren



Yehuda Koren，推荐系统领域的顶尖科学家。于 Weizmann Institute（魏茨曼科技学院）取得计算机科学博士学位。他的研究领域为：推荐系统

（recommender system）、数据挖掘、机器学习、信息可视化。

Yehuda Koren 博士现任 Google 的高级研究员（staff research scientist），此前，他历任 Yahoo!的高级研究员（staff research scientist，2008-2012 年），AT&T Labs-Research 的研究员与主要成员（2003-2008）。他是 SIGKDD、ICDM、RecSys、WSDM、CIKM 的高级计划委员会（senior program committee）成员；IEEE TKDE 和 ACM TIST 的编委。

他的论文“Factorization meets the neighborhood: a multifaceted collaborative filtering model”曾获 SIGKDD 2018 时间检测奖（Test of Time Award），该论文提出了一个单一的框架，将邻域模型与潜在因子模型（现在通常称为“嵌入模型”）相结合，从而利用两种方法的优势：邻域模型在检测非常局部化的关系时最有效，而潜在因子模型则更多有效地估计与大多数或所有项目同时相关的整体结构，直到今天仍然有着它的实用意义。他曾带领团队在 Netflix Prize competition 两次赢得 progress awards，是 Netflix Grand Prize 获奖团队的成员。曾获 2011RecSys 最佳论文奖、SIGKDD 2009 最佳论文奖、INFOVIS 2005 最佳论文奖。

● 唐杰

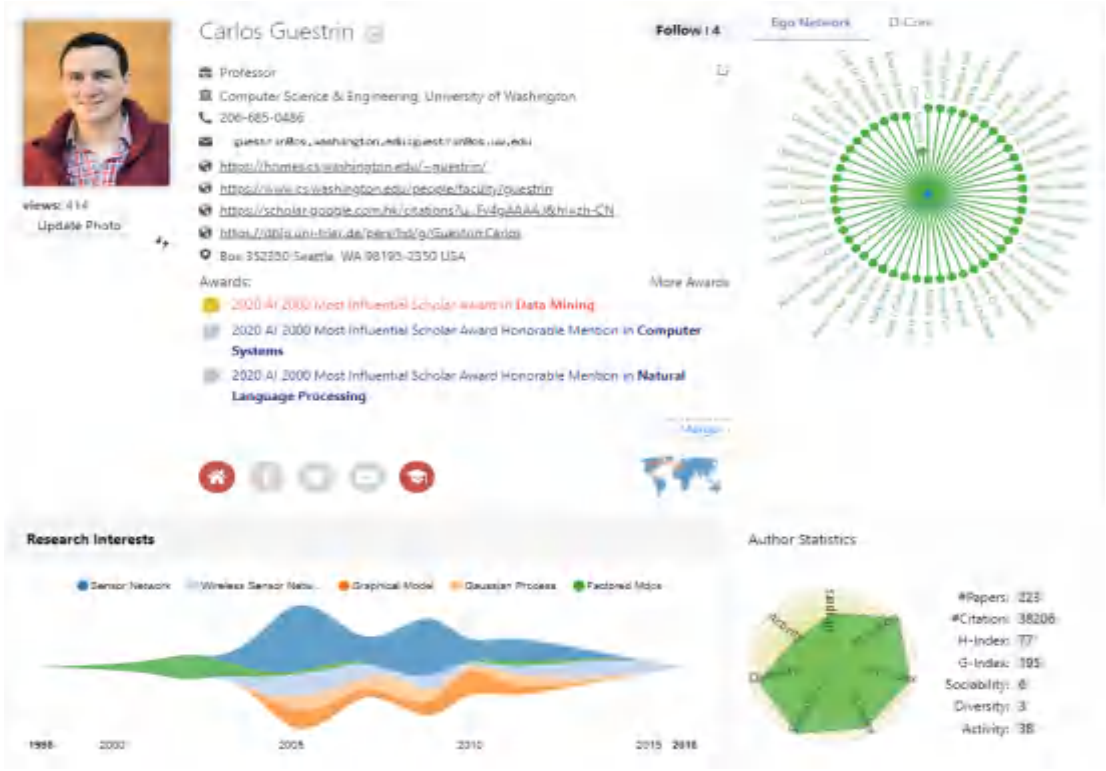
唐杰，AMiner 大数据平台的研发者，数据挖掘领域顶尖年轻学者。他是清华大学计算机科学与技术系教授、系科研办公室主任，2006 年于清华大学获计算机科学与技术工学博士学位。他的研究领域为：社会网络分析、数据挖掘和语义 Web。

他发表了 200 多篇期刊/会议论文，拥有 20 项专利，吸引了超过 10,000 项引用。他首次提出基于隐含话题的异构社会网络建模和社会网络影响力分析模型；针对 Web 信息的不同特点，提出多种有效的语义标注方法；提出的基于贝叶斯决策的多策略本体映射模型，在国际评测 OAEI 中获得多项第一的优秀成绩。研发的学术搜索系统 AMiner 的系统用户已覆盖 180 个国家。研究成果还在与 IBM、Google、Nokia、搜狐等多个国际合作和企业合作项目中得到推广应用。

唐杰教授曾担任 CIKM'16，WSDM'15，ASONAM'15，SocInfo'12，KDD 2018 副主席，组织 KDD'11-18 联合主席，IEEE TKDE / TBD 副主编的 PC 联合主席和 ACM TKDD / TIST。他曾参加国家中文新闻置标语言和分类标准的制定。

唐杰曾获数据挖掘顶会 KDD2018 杰出服务奖、2018 北京市科学技术奖一等奖、2016 AMiner 数据挖掘领域最具影响力学者（2016 AMiner Most Influential Scholar Award in Data Mining）、2012 年国家优秀青年科学基金、2011 北京市科技新星、2011 Scopus 全国青年科技之星、2010 年度清华大学学术新人奖（清华青年教师的最高学术荣誉）、2010 年 IBM 全球创新教师奖以及 ECML/PKDD'2011 Best Student Paper Runnerup、牛顿高级奖学金、CCF 青年科学家奖和 NSFC 优秀青年学者奖。

● Carlos Guestrin



Carlos Guestrin 是亚马逊机器学习专家，拥有斯坦福大学的博士学位，曾经在卡内基梅隆大学和华盛顿大学任职，主要授课内容为机器学习。他的研究方向主要有传感器网络、无线传感器网络、Graghical 模型、高斯过程和优化问题。

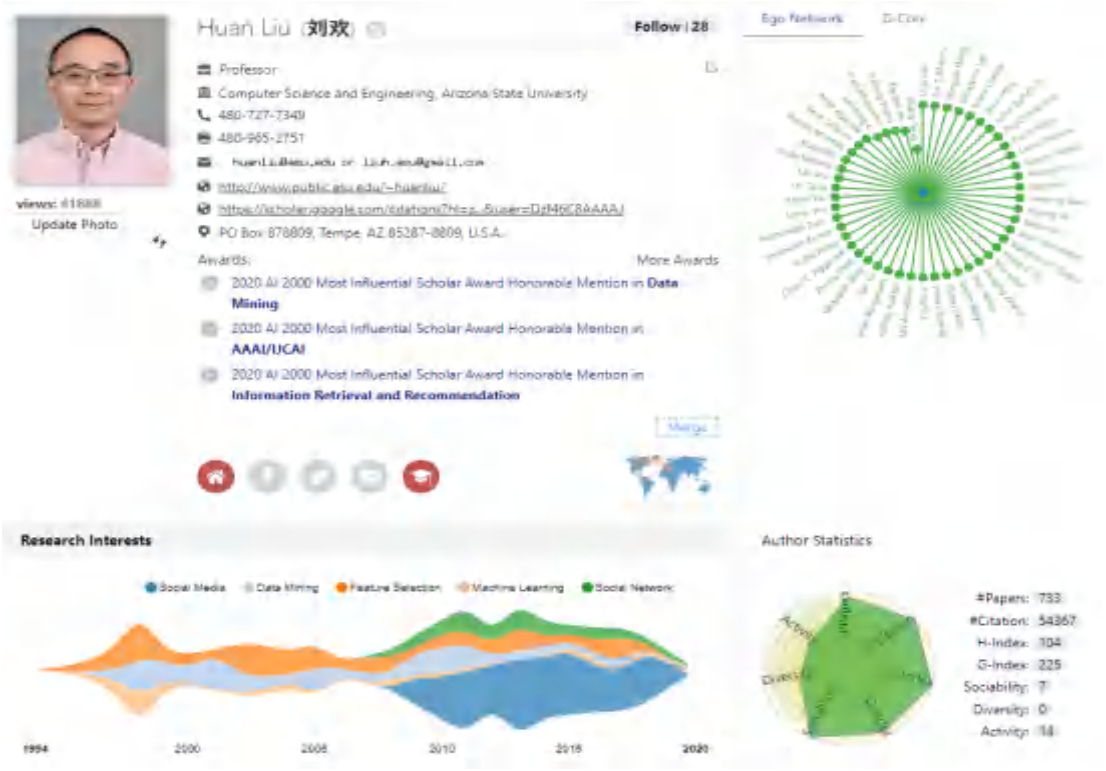
Carlos Guestrin 获得过许多会议和期刊的奖项，美国《大众科学》杂志曾将他评为“2008 年十大科学杰出人才之一”。他获得过两次 SIGKDD 的最佳论文奖，分别是，与 Dafna Shahaf 合作的论文“Connecting the Dots Between News”获得第 16 届 ACM SIGKDD 的最佳论文奖；论文“Cost-effective Outbreak Detection in Networks”获得第 13 届 ACM SIGKDD 获得的最佳论文奖。

● Gionis Aristides

Gionis Aristides 是阿尔托大学（Aalto University）计算机科学系的一名教授，是 EIT 数字数据科学专业的专员，也是 DMKD 和 TKDD 的编辑成员。他的研究领域是数据挖掘和算法数据分析，尤其是图表算法、社交网络分析和网络规模数据算法。

Gionis Aristides 教授的团队有五篇论文被 2018 年 IEEE 国际数据挖掘大会（ICDM）收录，他引用量最多的论文是“Reducing Controversy by Connecting Opposing Views”，在这篇论文中，Gionis Aristides 阐述了算法技巧在社交媒体上的应用。

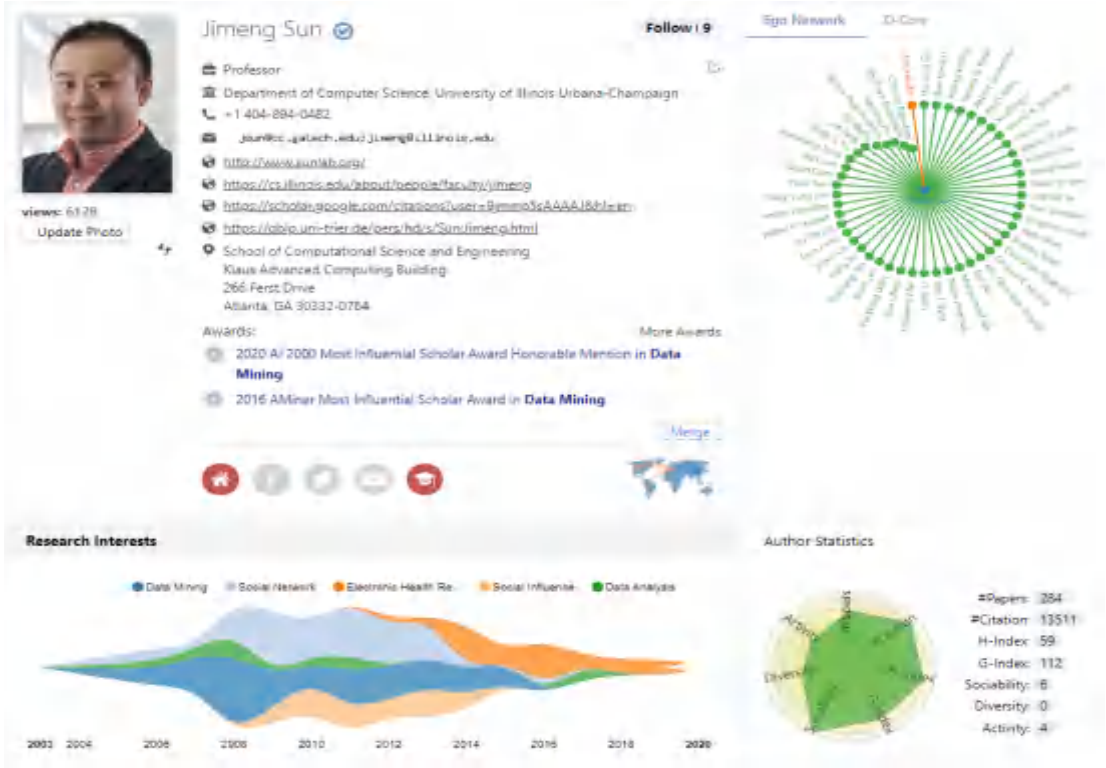
● 刘欢



刘欢，美国亚利桑那州立大学计算机科学与工程、信息学与决策系统工程系教授。本科毕业于上海交通大学，1985 年前往美国南加州大学（University of Southern California）学习，获得了南加州大学的计算机科学的硕士学位和博士学位。他目前的研究方向主要为机器学习、数据挖掘和社交网络的计算方法。

由于他在 Data Mining 和 Knowledge discovery 方面的突出贡献，在 2012 年被任命为 IEEE Fellow。在 2018 年 12 月初，被选为 ACM Fellow。2018 年 12 月中旬，被选为 AAAI Fellow。新入选 AAAI Fellow 的七位学者中有两位华人，刘欢教授便是其中一位。

- Jimeng Sun



Jimeng Sun，佐治亚理工学院计算机学院副教授，曾任 IBM TJ Watson 研究中心研究员。他的研究领域为数据挖掘和健康分析，尤其是在深度学习方法和大规模预测建模系统方面。

Jimeng Sun 发表过 120 多篇论文，提交了 20 多项专利。他获得了 SDM/IBM 2017 年度职业研究奖，在 2008 年获得了 ICDM 的最佳研究论文奖，在 2007 年获得了 SDM 最佳研究论文奖，在 2008 年 KDD 论文的亚军。

其他排名靠前的学者介绍：

- 林智仁

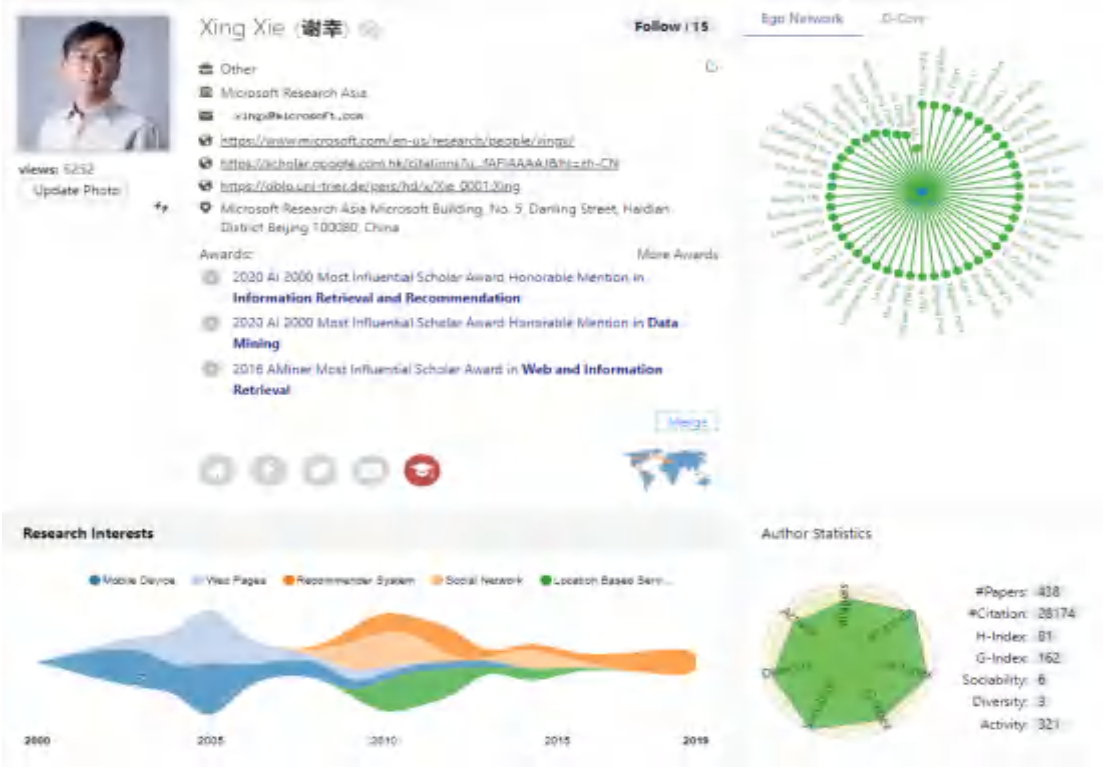
林智仁，在机器学习算法与软件领域有突出贡献。他是国立台湾大学计算机科学系的教授，1998 年于密歇根大学获工业和运营工程博士学位。他的研究领域为：机器学习、数据挖掘与数值最优化。

林智仁教授是电气和电子工程师协会（IEEE）会士（2011）、美国人工智能协会（AAAI）会士（2014）、国际计算机学会（ACM）会士（2015）、ACM 杰出科学家（2011）。

他曾获 2015 东元奖（Teco Award）、ACM RecSys 2013 最佳论文奖、SIGKDD 2010 最佳论文奖。他的 NTU 团队曾获 2013 年、2012 年、2010 年 SIGKDD CUP 冠军。

其开发设计的软件包 LIBSVM，是目前应用最广泛的支持向量机（SVM）软件，据 google scholar 统计已被引用超过 18000 次。

● 谢幸

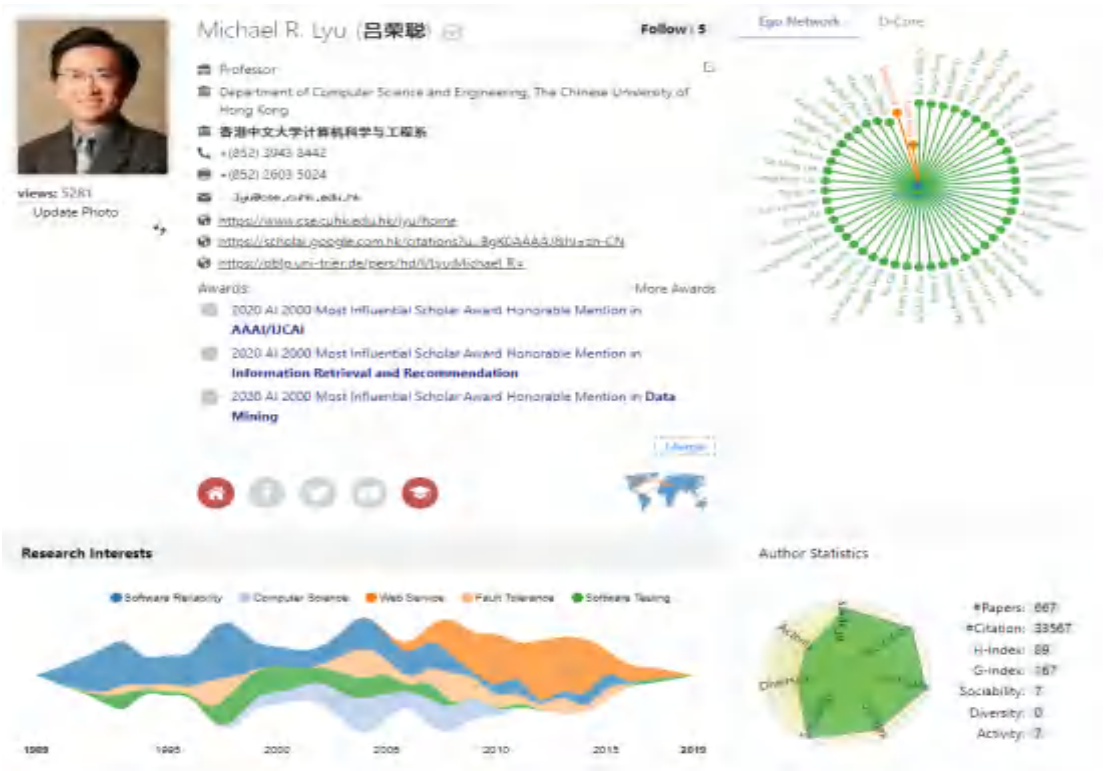


谢幸，数据挖掘与人工智能领域优秀年轻学者。他是微软亚洲研究院现任主管研究员，2001 年于中国科技大学获得计算机软件博士学位。他的研究领域为：空间数据挖掘、位置服务、社交网络、普适计算。

他在是 ACM Fellow、IEEE Fellow 和计算机学会高级会员，现任互联网搜索与挖掘组主管研究员，并任中国科技大学兼职博士生导师，多次担任 WWW、UbiComp、ACM SIGSPATIAL、KDD 等顶级国际会议程序委员会委员。他是 ACM TSC、ACM TIST、ACM IMWUT、GeoInformatica、Elsevier Pervasive and Mobile Computing、CCF TPCI 等杂志编委。他参与创立了 ACM SIGSPATIAL 中国分会，并曾担任 UbiComp 2011 大会程序委员会共同主席。

他多次在 KDD、ICDM 等顶级会议上获最佳论文奖，并被邀请在 ASONAM 2017、Mobiquitou 2016、SocInfo 2015、W2GIS 2011 等会议做大会主题报告。

- 吕荣聪



吕荣聪，国际知名软件可靠性工程及软件容错技术学者，是该领域最早的研究者之一。他是香港中文大学计算机科学与工程系教授，1988 年于加州大学圣洛杉矶分校获计算机科学博士学位。他的研究领域为：软件可靠性工程、分布式系统、容错计算、服务计算、多媒体信息检索和机器学习。

吕荣聪教授是 IEEE Fellow（2004）、AAAS Fellow（2007）、Croucher Senior Research Fellow（2008），国家教育部第八批“长江学者奖励计划讲座教授”。他创立了香港中文大学无线互联网视像技术实验室并出任主任。他曾在美国喷气式推进实验室、Murray Hill 贝尔实验室、Morristown 贝尔通讯研究公司工作。他还曾担任 IEEE Transactions on Reliability、IEEE Transactions on Knowledge and Data Engineering、Journal of Information Science and Engineering 的副主编。

● 裴健

裴建，加拿大西蒙弗雷泽大学计算科学学院教授，统计与精算学系和健康科学院兼职教授，现任京东集团副总裁。

主要研究兴趣为数据科学、大数据、数据挖掘和数据库系统等领域，尤其是数据密集型应用设计开发创新性的数据业务产品和高效的数据分析技术。

他是国际计算机协会（ACM）院士和国际电气电子工程师协会（IEEE）院士，ACM SIGKDD 现任主席，ACM 和 IEEE Fellow、加拿大研究讲席教授（Canada Research Chair, Tier I）。

因其在数据挖掘基础、方法和应用方面的杰出贡献，裴健教授获得数据科学领域技术成就最高奖 2017 ACM SIGKDD Innovation Award（ACM SIGKDD 2017 创新奖）和 IEEE ICDM Research Contributions Award（IEEE ICDM 研究贡献奖）。以及 2015 ACM SIGKDD Service Award（ACM SIGKDD 2015 杰出服务奖）。

- 杨强

杨强，人工智能研究的国际专家和领军人物，在学术界和工业界做出了杰出的服务和贡献，尤其近些年为中国人工智能（AI）和数据挖掘（KDD）的发展起了重要引导和推动作用。现任香港科技大学新明工程学讲席教授、计算机科学和工程学系主任，大数据研究所所长，1989 年于马里兰大学获得计算机科学博士学位。他的研究领域为：机器学习、数据挖掘和自动规划。

杨强教授是 IEEE、AAAI、AAAS 和 IAPR 的 Fellow，是诺亚方舟研究实验室创始人。他于 2013 年 7 月当选为国际人工智能协会（AAAI）院士，是第一位获此殊荣的华人，之后又于 2016 年 5 月当选为 AAAI 执行委员会委员，是首位也是至今为止唯一的 AAAI 华人执委，2015 年 IJCAI 大会、2010 年 ACM KDD 大会的学术委员会主席（PC Chair），以及 2012 年 ACM KDD 大会主席。2017 年 8 月他当选为国际人工智能联合会（IJCAI），国际人工智能领域创立最早的顶级国际会议）理事会主席，是第一位担任 IJCAI 理事会主席的华人科学家，《ACM TIST》期刊和《IEEE 大数据》期刊的创始主编。

他曾获 Nokia MDC 2012 两项冠军、ACM CIKM 2011 最佳跨学科论文奖、ACM KDDCUP 2011 评级预测第三位、2005 ACM KDDCUP Competition 冠军队教练、2005 第一届 ICAPS 合作冠军、2004 ACM KDDCUP Competition 冠军队教练，获 SIGKDD 2016 杰出服务奖（ACM SIGKDD 2016 Service Award）。

● 吴信东

吴信东，国际著名的数据挖掘研究与应用学者，对多数据源的知识发现、不同种类的噪音对知识发现精度的影响，以及数据挖掘在生物、影像数据库中的应用等方面有自己的突出贡献。他是合肥工业大学教授、博士生导师，1993 年于英国爱丁堡大学（University of Edinburgh）大学获人工智能博士学位。他的研究领域为：大数据分析、数据挖掘、机器学习、人工智能等。

吴信东教授是 IEEE、AAAS 的会士（Fellow）、国家“千人计划”入选者、长江学者、明略数据公司首席科学家、美国路易斯安那大学计算机科学终身教授、兼任该校计算与信息学院院长，他还是 KAIS 主编、IEEE ICDM 指导委员会主席、AI&KP 主编。他曾任 IEEE TKDE 主编，IEEE TCII 专委会主任，IEEE/ACM ASONAM’14、ACM CIKM 2010、ACM SIGKDD 2007、ACM/WIC/IEEE WI 2006、IEEE ICDM 2003、PAKDD 1998 等国际会议的程序委员会主席/联合主席。

吴信东教授曾获 2012 IEEE 计算机学会技术进步奖（Technical Achievement Award）、2012 IEEE/WIC/ACM WI 最佳论文奖、2006 IEEE ICDM 杰出服务奖（Outstanding Service Award）、2011 与 2005 IEEE ICTAI 最佳论文奖、2004 ACM SIGKDD 杰出服务奖（Service Award）。

3.3 部分国内学者的研究成果

3.3.1 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[80]。

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[81]。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[82]。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[83]。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[84]。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[85]。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[86]，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[87]。

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[88]。

浙江大学的蔡登等人研究了非监督学习中特征选择问题[89]。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[90]。

浙江大学的何晓飞和亚利桑那州立大学的叶杰平，共同研究了无监督的数据补齐这一极具挑战的问题，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/N2)（N 是算法中的迭代次数）[91]。

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[92]。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[93]。

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[94]。

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[95]。

3.3.2 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之一。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。

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[96]。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[97]，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[98]。

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[99]。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[100]。

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[101]。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[102]。

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。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[103]。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）[104]，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[105]。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）[106]。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[107]。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[108]。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[109]，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[110]。

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[111]，他们在国内豆瓣网络（www.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。

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。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。

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。

3.3.3 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。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[112]。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[113]。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[114]。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[115]。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[116]。北京大学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[117]。

4 应用篇

4 应用篇

数据挖掘技术从一开始就是面向应用的。目前，在很多领域数据挖掘都是一个很时髦的词，尤其是在如零售业、旅游业、物流业、医学等领域，数据挖掘技术的使用，可以大大提高行业效率和行业质量。

4.1 零售业

数据挖掘技术源于商业的直接需求，虽然它在各种领域都存在广泛的使用价值，但零售领域是数据挖掘的主要应用领域之一。这是因为由于条形码技术的发展，零售业的销售部门可以利用前端收款机系统收集存储大量的售货数据、顾客购买历史记录、货物进出状况和消费与服务记录等等。零售业也同其他数据密集型企业一样积累了大量的数据，这些数据正是数据挖掘的基础。数据挖掘技术有助于识别顾客购买行为，发现顾客购买模式和趋势，改进服务质量，取得更高的顾客保持力和满意程度，减少零售业成本。

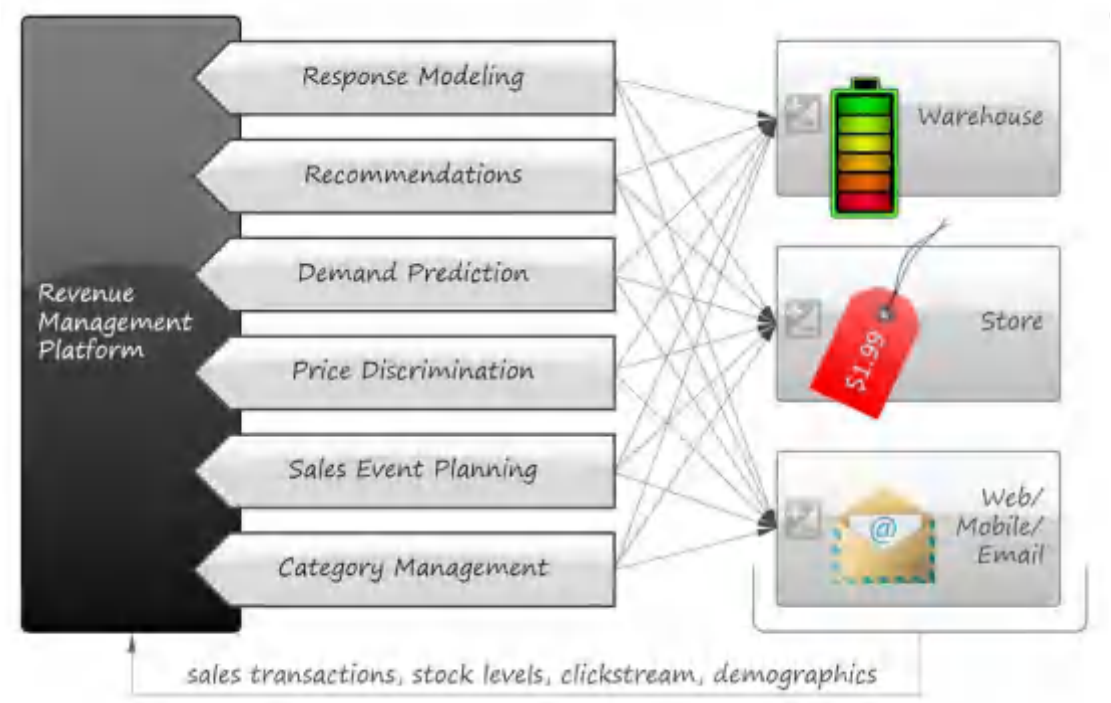


图 65 数据挖掘方法在零售业中的应用[118]

数据挖掘在零售业中的具体应用有以下：

（1） 了解销售全局。通过分类信息按商品种类、销售数量、商店地点、

价格和日期等了解每天的运营和财政情况，对销售的每一点增长、库存的变化以及通过促销而提高的销售额都可了如指掌。零售商店在销售商品时，随时检查商品结构是否合理十分重要，如每类商品的经营比例是否大体相当。调整商品结构时需考虑季节变化导致的需求变化、同行竞争对手的商品结构调整等因素。

（2） 降低库存成本。通过数据挖掘系统，将销售数据和库存数据集中起来，通过数据分析，以决定对各个商品、各色货物进行增减，确保正确的库存。数据挖掘系统还可以将库存信息和商品销售预测信息，通过电子数据交换直接送到供应商那里，这样省去零售业中介，而且供应商负责定期补充库存，零售商可减少自身负担。

（3） 商品分组布局、购买推荐和商品参照分析。通过从代销记录中挖掘相关信息，可以发现购买某一种商品的顾客可能购买其他商品。这类信息可用于形成一定的购买推荐，或者保持一定的最佳商品分组布局，以帮助客户选择商品，刺激顾客的购买欲望从而达到增加销售额，节省顾客购买时间的目的。

（4） 促销活动的有效性分析。零售业常常通过广告、优惠券、各种折扣和让利的方式搞促销活动，以达到促销产品，吸引顾客的目的。但只有充分了解客户，才能定位促销活动，提高客户响应率，降低促销活动成木。利用数据挖掘技术可以分析出应该在什么时间、在什么地点、以何种方式和对什么样的人搞促销活动，能真正达到促销目的，避免企业资源的不必要浪费。同时，数据挖掘也可以使用过去有关促销的数据来寻找未来投资中回报最大的用户。

（5） 市场和趋势分析。利用数据挖掘工具和统计模型对数据库的数据仔细研究，以分析顾客的购买习惯、广告成功率和其它战略性信息。利用数据库通过检索数据库中近年来的销售数据，作分析和数据挖掘，可预测出季节性、月销售量，对商品品种和库存的趋势进行分析。还可确定降价商品，并对数量和运作做出决策。

（6） 顾客忠诚度分析。各个零售企业往往通过办理会员卡的方式，建立了顾客会员制度来跟踪顾客的消费行为。通过对顾客会员的信息进行数据挖掘，可以记录一个顾客的购买系列，顾客的忠诚和购买趋势可以按系统的方式加以

分析。由同一顾客在不同时期购买的商品可以分组为序列。序列模式挖掘可用于分析顾客的消费或忠诚度的变化，据此对价格和商品的花样加以调整和更新，以便留住老客户，吸引新客户[119]。

4.2 旅游业

旅游大数据及挖掘在旅游业中的广泛应用，不仅仅为现代化旅游企业的飞速发展提供了有利的促进作用，同时对于人们对旅游信息的科学化搜集和掌握也提供了一定的便利，不仅仅对客流的趋向有着准确的预知性，同时对于游客的喜好也有着直接性的掌握，并对现代化旅游公共服务的改善有着极其有利的作用[120]。

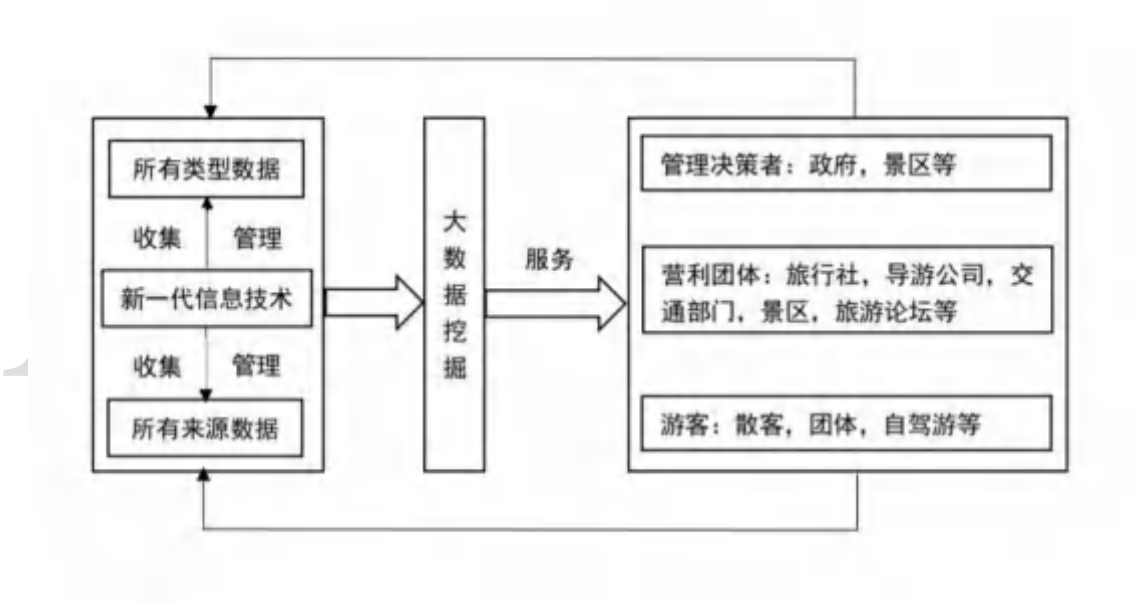


图 66 数据挖掘应用于智慧旅游的概念结构[121]

数据挖掘在旅游业中的具体作用如下：

（1） 对有价值的旅游信息加以挖掘。挖掘有价值的旅游信息的过程中，通过对游客对旅游网站日志的点击率进行分析，并分析用户较为常见的一种浏览行为，进而对游客较为感兴趣的旅游目标进行掌握和搜集。结合大数据挖掘的信息，旅游管理相关部门就要对旅游信息网站进行优化设计，并尽可能地保证网站的优化设计和游客的实际需求有着一定的吻合度，并保证现行的旅游服务更优质和更全面。

（2） 对潜在旅游客户挖掘。在对潜在旅游客户进行收集的过程中，通过对数据进行聚类性的分析，并结合游客的一些访问记录，对游客的相关知识进行综合性的分析和总结，及时地关注和搜集潜在客户的喜好，并将客户最感兴趣的旅游信息加以推荐。

（3） 旅游路线的优化。旅游大数据及挖掘在旅游行业应用的过程中，通过挖掘游客对旅游线路和目的地的访问情况，并进行综合性的分析，进而对最具有市场潜力的旅游路线加以选择，进而合理地规划好相关性的旅游路线，对旅游网站的结构进行动态性的处理，并将网站的粘性有效增加，进而将网站的访问量全面提高。

（4） 旅游项目和目的地的推荐。旅游大数据及挖掘在旅游行业的引用中，借助于数据挖掘，对旅游项目和目的地有着一定的推荐作用，通过对涵盖所有目的地的旅游数据库创建，并借助于数据挖掘工具，对客户的行为和兴趣爱好进行分析，并形成不同类型的旅游服务方案，集合游客的实际旅游爱好，对最佳旅游项目和其目的地进行推荐，尽可能地将旅游客户的满意度全面提高[122]。

4.3 物流业

数据挖掘技术能够帮助企业在物流信息系统管理中，及时、准确地收集和分析客户、市场、销售及整个企业内部的各种信息，对客户的行为及市场趋势进行有效的分析，了解不同客户的爱好，从而可以为客户提供有针对性的产品和服务，大大提高各类客户对企业和产品的满意度。

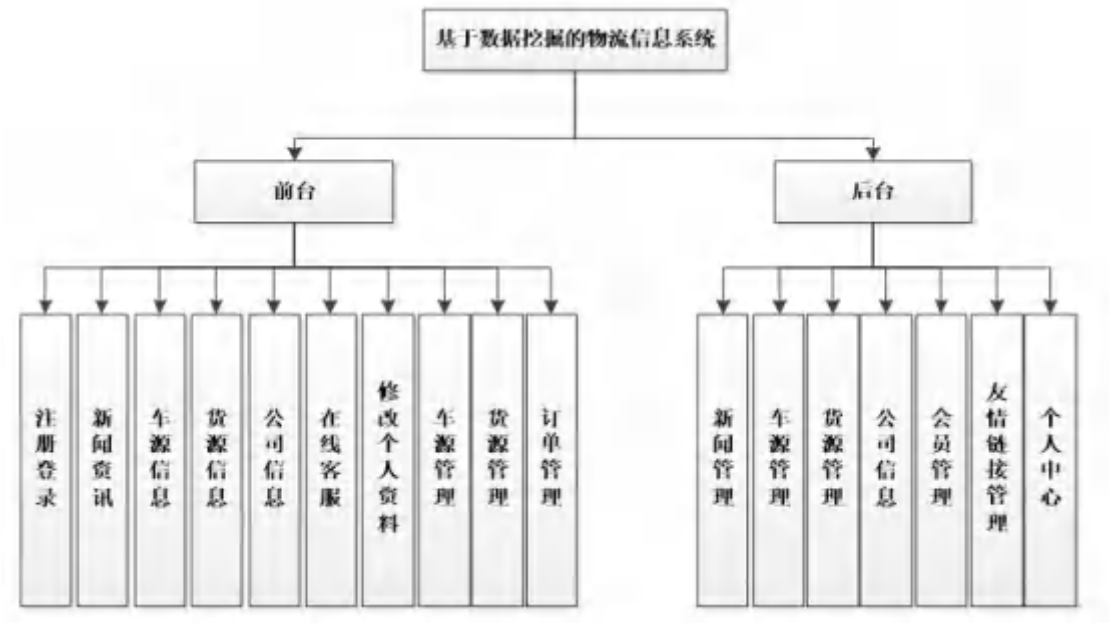


图 67 基于数据挖掘的物流信息系统[123]

数据挖掘在物流业中的具体应用有：

（1） 数据挖掘在客户关系管理中的作用。利用数据挖据可以找出未来潜在的客户，利用数据挖掘整理出资料的特点，找出最有兴趣的客户群，让他们有机会接触到该项产品和服务，并最终成为真正的客户。面对真正的客户，数据挖掘可以发现客户的消费嗜好，通过刺激客户的消费嗜好显然可以提高企业收入，通过关联规则挖掘还可以增加交叉销售，促使客户购买尚未使用的产品和服务，并且数据挖掘还可以判断哪些客户对促使他们将现有产品和服务升级的增量销售有意义。面对历史客户，数据挖掘可以通过建立流失模型，发现客户离开的原因，预测什么样的客户有离开的意向，找到解决方法，从而避免将来类似的客户再次流失。

（2） 数据挖掘在物流配送中的应用。数据挖掘可以提高车辆的利用率，如何安排车辆路线和进行车辆调度既能满足配送任务，又使车辆运行总里程最短。利用一定的算法，可以得出一个最优解，节省物流配送中的成本。

4.4 医学界

随着信息时代的到来，数据挖掘被越来越多地应用于临床实践。利用信息技术，医疗记录和随访数据可以更有效地被存储和提取。同时，从医学数据中

寻找潜在的关系或规律，可以获得对病人进行诊断、治疗的有效知识；增加对疾病预测的准确性，及早发现疾病，提高治愈率。

S.No	疾病类型	数据挖掘工具	技术	算法	传统方法	DM应用的精度等级%
1.	Tuberculosis	WEKA	朴素贝叶斯分类器	KNN	概率统计	78 Percent
2.	Heart Disease	ODND, NCC2	Classification	Naive	Probability	60%
3.	肾脏透析	RST	Classification	做决定	Statistics	76%
4.	糖尿病	ANN	Classification	C4.5算法	神经网络	82%
5.	血库部门	WEKA	Classification	J48		90 Percent
6.	Dengue	SPSS建模器		C5.0	Statistics	80 Percent
7.	丙型肝炎	SNP	Information	Gain	决策规则	74 Percent

图 68 医疗领域数据挖掘工具的准确性对比[124]

数据挖掘对医学实践的重要性主要体现在四个方面：

（1） 医学数据挖掘会大大增加产生新知识的速度，利用计算机技术分析电子病历中包含的非结构化数据能够更好地进行自动化的数据采集。

（2） 医学数据挖掘可以帮助进行知识传播，大量的研究使得知识的转化非常困难，大部分临床医生难以跟上最新的证据来指导临床实践。这个问题可以通过分析现有的电子病历产生一个仪表板来指导临床决策。如 IBM 的沃森超级计算机与斯隆凯特林癌症中心合作，利用这个方法来辅助临床医生对癌症患者做出诊断和提出治疗方案[125]。

（3） 通过整合系统生物学与电子病历数据，医学数据挖掘可以为个性化医疗计划转变为临床实践提供机会。

（4） 通过直接向病人提供信息来转化卫生保健知识，让病人在治疗过程中发挥更积极的作用[126]。

4.5 金融业

在当今信息科技迅猛发展的时代，数据挖掘技术由于其极好的数据信息提取能力能够有效助力于信息决策、信息管理、科学研究等，在金融行业具有良好的应用前景。目前，数据挖掘技术已经在金融领域里得到了广泛的应用，主

要体现在多维数据分析、贷款偿还预测、客户信用政策、目标市场客户分类和聚类以及对金融犯罪防范等方面。

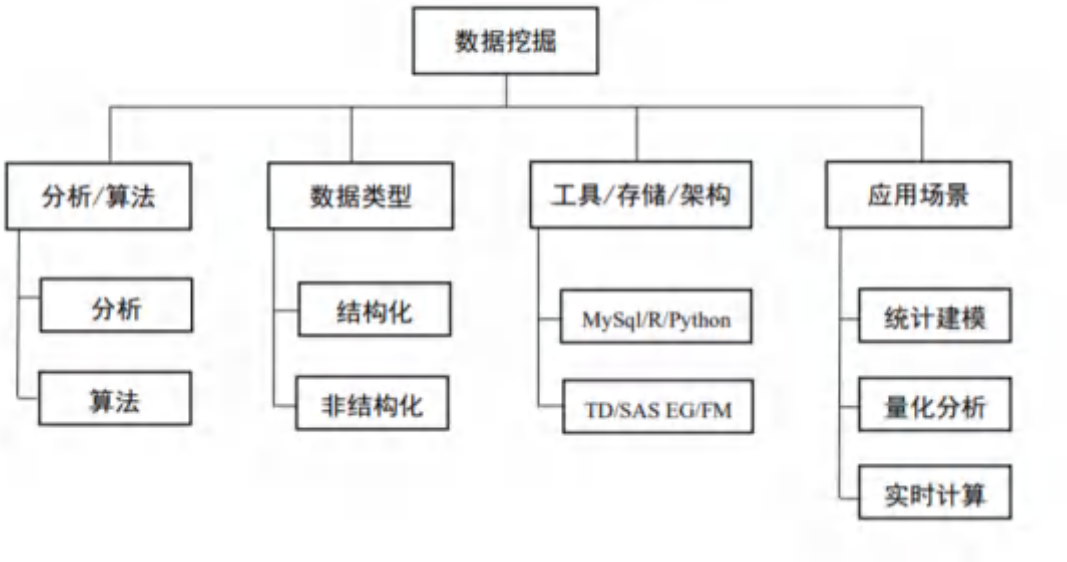


图 69 互联网数据挖掘与金融数据挖掘对比[127]

金融行业中的数据挖掘应用主要体现在这五个方面：

（1） 多维数据分析。数据挖掘在金融领域中的多维数据分析主要包括对金融数据进行分析后，对相关的数据挖掘设计以及为相关金融机构建构对应的数据仓库，从而掌握相关数据的一般性，助力于金融机构相关业务的开设和开展。

（2） 贷款偿还预测。放贷是银行最主要的业务之一，因此做好贷款偿还预测有助于银行相关业务的顺利展开。数据挖掘能够充分用到特征选择等方法对各种因素进行有效识别，并选取重要因素，排除非相关性因素，从而实现最为准确的贷款偿还预测。

（3） 客户信用政策。做好客户信用政策有助于银行相关业务的顺利展开。数据挖掘能够充分用到特征选择等方法对各种因素进行有效识别，并选取重要因素，排除非相关性因素，从而制定最为准确的客户信用政策。

（4） 目标市场客户分类和聚类。数据挖掘在对金融领域目标市场客户的分类和聚类上，主要是对目标客户进行有效识别和对目标市场进行有效分析，

并通过分类和聚类将目标客户分划到相应的客户群体中，从而推动目标市场的建立。

（5） 防范金融犯罪。金融犯罪是金融业面临的问题之一，包括洗钱、恶意透支、伪造信用卡等。利用数据挖掘技术，可以将数据库信息进行有效集成，应用多种分析工具识别出其中出现异常的模式，并对其进行定点追踪。例如在数据链上识别出在某一段时间内，某一对象频繁出现了大量现金流量，且资金流向存疑等。面对这样的异常模式，相关人员可以通过可视化、分类、联接等多种工具进行系统化地分析，获取相关犯罪事实[128]。

4.6 电信业

信息时代，数据挖掘技术对通信行业具有不可估量的作用，是通信行业制定科学经营发展战略的技术支撑，对营销效果的增强以及营销效益的提升意义重大。随着社会经济的不断发展，移动通信技术水平的日益提高，各通信网络运营商对数据挖掘技术的重视程度也在逐步提升。通信行业中数据挖掘技术应用的领域主要是客户流失的调查分析、客户关系处理与维系、客户行为的了解以及经营战略与数据的调整等方面。

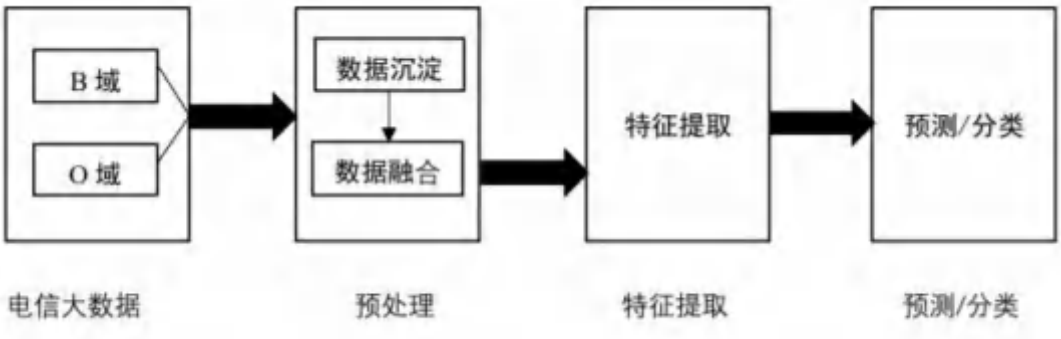


图 70 电信大数据的数据挖掘流程[129]

数据挖掘在电信行业中的应用主要集中在以下几个方面：

（1） 数据挖掘技术强化实施对客户关系的维护与管理。在电信行业内部，从传统角度进行分析，其客户关系管理应用的是 CRM 系统，主要是为了提高自身核心竞争力，发挥信息技术优势，以互联网为依托，强化企业与客户关系的协调，实现在营销、服务等多领域的和谐发展。借助对不同意见的分析与处理，

促使管理方式更具科学性，在根本上为客户提供独具特色的个性化交互与服务，有效促进科学性与合理性。

（2） 数据挖掘系统的应用强化对客户资源准确与全面的审核分析。在数据挖掘系统的应用下，通信行业能够实现对客户资源的准确、全面与快速的评估，明确彼此之间的关系状态。同时，借助对不同客户偏好、使用频率等的分析能够有针对性地制定服务模式与类型，增强客户的归属感，更显被尊重感，这对于提升整个通信行业整体竞争力具有突出作用，为经济与社会效益的获取创造有利条件。

（3） 数据挖掘技术实现了对用户行为的综合分析，以便构建个性化营销规划。对于客户行为的分析需要采取聚类分析法。在通信行业，聚类分析应用的领域为移动电话业务的分析工作，涉及漫游、通话、呼叫等。针对差异化的人群，其在费用选择、产品使用规划等方面采取不同的方法，需要结合其收入、家庭组成、知识水平以及社交量等因素进行全面与综合的分析，以此为依据，构建更具优质性的系统性评估。

（4） 数据挖掘技术满足不同区域经营运行数据的分析需求，明确通信企业经营发展目标。针对经营性数据，数据挖掘技术主要依托关联规则开展应用，善于对隐性数据关系进行分析。具体而言，是选择指定的数据集合，对数据记录组合进行分析，构建相互关联性，将处于通信行业不同领域的内容进行有机联系，做出科学性分析[130]。

5　趋势篇

5 趋势篇

本节依托 AMiner 平台技术分析系统（http://trend.aminer.cn），搜索相关研究论文并进行深入挖掘，探索分析了数据挖掘技术的研究发展趋势和研究热点等，为科技工作者了解数据挖掘领域相关技术的研究历史和现状、快速识别前沿热点研究问题提供信息窗口。

5.1 技术研究发展趋势

图 71 为数据挖掘技术研究发展趋势图，展示了数据挖掘领域技术研究的时间发展轨迹和研究热度变化情况，对于学科领域的布局和发展具有重要的意义和参考价值。其中展示了数据挖掘领域的部分关键词，包括 Data Mining（数据挖掘）、Social Network（社交网络）、Big Data（大数据）、Time Series Mining（时间序列挖掘）、Anomaly Detection（异常检测）、Support Vector Machine（支持向量机）、Decision Tree（决策树）、Text Mining（文本挖掘）、Clustering（聚类）、Classification（分类），其研究发展趋势如图所示。图中每条色带表示一个研究话题，色带宽度表示该话题在当年的研究热度，与当年该话题的论文数量呈现正相关关系，不同色带的高低排序是由当年这些话题的研究热度决定。

从图中可以看出，大部分话题的研究热度随着时间推进呈增长趋势。其中数据挖掘和社交网络的研究热度居高不下，而近五年来大数据、异常检测和时间序列挖掘的研究热度呈现明显的上升趋势。另外，趋势图也显示，聚类和分类的研究热度一直维持在较高水平，但在近五年内呈现比较明显的下滑趋势。

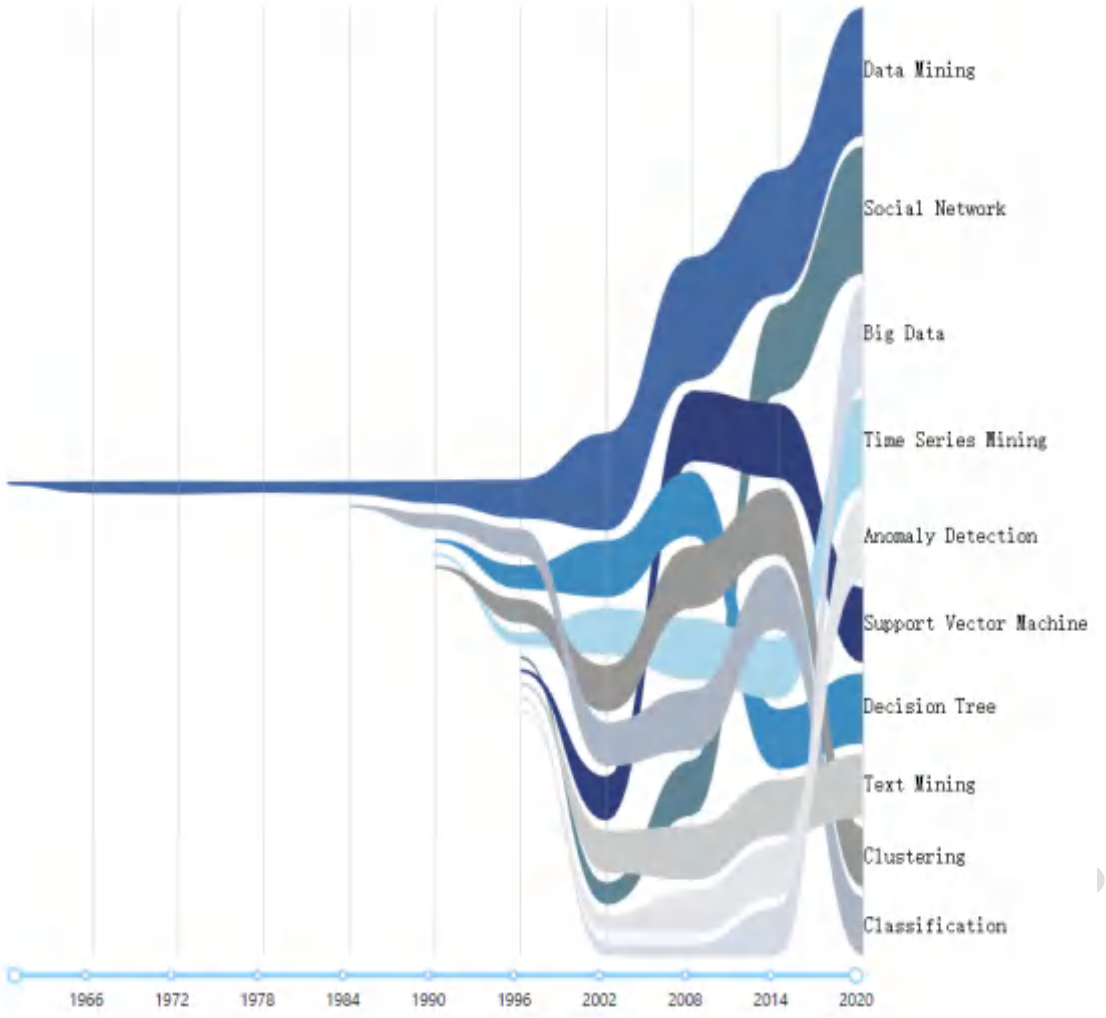


图 71 数据挖掘领域的技术研究发展趋势

5.2 技术研究创新热点

图 72 展示了数据挖掘领域研究热度较高的话题分布，其中每个关键词的大小表示研究热度高低，与研究的论文数量成正比。从中可以看出，该领域研究热度较高的是 Data Mining、Social Network、Information Retrieval、Computer Science、Machine Learning、Support Vector Machine 等。

将 h-index 指标作为筛选条件，从 AMiner 平台上获取了以上几个研究热度较高的领域的代表性学者信息，其中 Data Mining 研究领域的代表性学者是伊利诺伊大学厄巴南分校计算机科学系韩家炜教授，他获得了清华大学 AMiner 团队发布的“AI 2000 人工智能全球最具影响力学者榜单 ”（简称 AI 2000）的数据挖掘领域最具影响力学者奖。Social Network 领域的代表性学者是麻省理

工学院数据、系统和社会研究所Daron Acemoglu教授，他获得了2018年AMiner经济学十大最具影响力学者奖。Information Retrieval领域的代表性学者是伊利诺斯大学电子与计算机工程系黄煦涛教授，他获得了2020 AI 2000计算机视觉领域最具影响力学者荣誉奖。Computer Science领域的代表性学者是马里兰大学计算机科学系Ben Shneiderman教授，他获得了AI 2000可视化领域最具影响力学者荣誉奖。Machine Learning领域的代表性学者是加州大学统计学系Michael I. Jordan教授，他获得了AI 2000年度机器学习领域最具影响力学者荣誉奖。Support Vector Machine领域的代表性学者是麻省理工学院大脑和认知科学系Tomaso A. Poggio教授，他获得了2016年AMiner计算机视觉最具影响力学者奖。

图 72 数据挖掘领域的研究热点词云图

表 20 数据挖掘领域关键词的论文数统计

关键词	论文数	关键词	论文数
Data Mining	5937	Text Analysis	383
Social Network	1277	Probability	381
Information Retrieval	1163	Database Theory	375
Computer Science	1141	Social Media	373
Machine Learning	1074	Data Structure	360
Support Vector Machine	886	Web Pages	354
Data Analysis	839	New Algorithm	352
Decision Tree	675	Experimental Evaluation	343
Time Series Mining	638	Application Software	343
Clustering	622	Knowledge Base	330
Computational Complexity	613	Knowledge Based Systems	329
Data Structures	593	Search Space	327
Graph Theory	584	Transaction Processing	315
Efficient Algorithm	580	Privacy	298
Data Set	570	Big Data	297
Classification	560	Pattern Clustering	291
Feature Extraction	554	Query Optimization	290
Xml	504	Anomaly Detection	285
Database System	489	Data Integrity	285
Data Engineering	486	Scalability	281
Feature Selection	482	Pattern Recognition	278
Optimization	471	Artificial Intelligence	276
Training Data	427	Query Language	273
Relational Database	417	Knowledge Representation	270
Semantics	397	Search Engines	269
Recommender System	397	Recommender Systems	268
Text Mining	393	Optimization Problem	263

Satisfiability	393	Concurrency Control	256
Data Privacy	393	Uncertainty	256
Data Stream	393	Information Extraction	256
Synthetic Data	392	Probability Density Function	252
Real Data	388	Database Indexing	249

将 h-index 指标作为筛选条件，从 AMiner 平台上获取了以上几个研究热度较高的领域的代表性学者信息，其中 Data Mining 研究领域的代表性学者是伊利诺伊大学厄巴南分校计算机科学系韩家炜教授，他获得了清华大学 AMiner 团队发布的“AI 2000 人工智能全球最具影响力学者榜单 ”（简称 AI 2000）的数据挖掘领域最具影响力学者奖。Social Network 领域的代表性学者是麻省理工学院数据、系统和社会研究所Daron Acemoglu 教授，他获得了 2018 年 AMiner 经济学十大最具影响力学者奖。Information Retrieval 领域的代表性学者是伊利诺斯大学电子与计算机工程系黄煦涛教授，他获得了 2020 AI 2000 计算机视觉领域最具影响力学者荣誉奖。Computer Science 领域的代表性学者是马里兰大学计算机科学系 Ben Shneiderman 教授，他获得了 AI 2000 可视化领域最具影响力学者荣誉奖。Machine Learning 领域的代表性学者是加州大学统计学系 Michael I. Jordan 教授，他获得了 AI 2000 年度机器学习领域最具影响力学者荣誉奖。Support Vector Machine 领域的代表性学者是麻省理工学院大脑和认知科学系 Tomaso A. Poggio 教授，他获得了 2016 年 AMiner 计算机视觉最具影响力学者奖。

根据 AMiner 平台的不完全统计，表 21 展示了这些学者的学术指标。

表 21 数据挖掘研究热点子领域的代表性学者的学术指标统计

学者姓名	论文发表量	论文总被引频次	H-index	G-index	Sociability	Diversity	Activity
韩家炜	1252	183472	175	416	9	0	11
Daron Acemoglu	501	130496	133	361	7	2	3
黄煦涛	1441	103589	147	300	8	4	261
Ben Shneiderman	900	118368	126	339	7	4	27
Michael I.	821	190243	153	435	7	0	7

Jordan							
Tomaso A. Poggio	711	96325	136	307	7	0	1

5.3 数据挖掘专利数据分析

本章节根据 1.3 章节介绍的数据挖掘知识图谱，并结合第 2 章节中的数据挖掘相关技术作为查询关键词，从 AMiner 平台上检索到 2010-2019 年期间的专利数据，共计 286,769 篇。

图 73 展示了中国在 2010-2019 年间的历年专利数量的分布情况。从图中可以看出，中国近 10 年的专利数量整体上呈现逐年递增趋势，在 2012 年和 2015 年专利数量增长率达到了高峰（约为 35.5%和 35.7%）。

图 73 中国历年的专利数量分布（2010-2019 年）

图 74 展示了中国在 2010-2019 年间专利数量最多的前 10 个机构。这 10 个机构中包含 3 所公司企业和 7 所高校，主要分布在北京（3 所）、广东（2 所）、四川（1 所）、陕西（1 所）、浙江（1 所）、江苏（1 所）、天津（1 所）等地区。

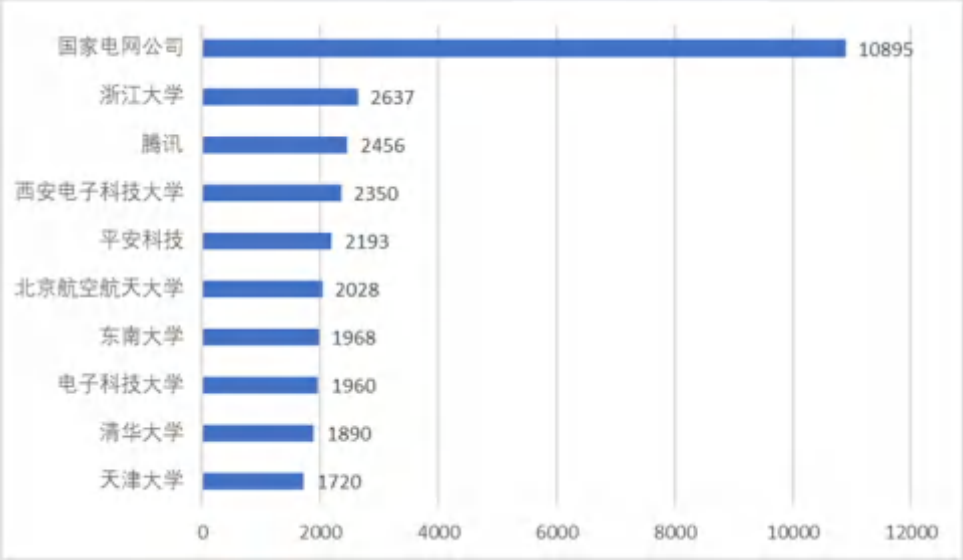


图 74 2010-2019 年中国专利数量 TOP 10 机构

5.4 国家自然科学基金支持情况

本报告根据附录 1 中展示的数据挖掘领域关键词列表，从 AMiner 数据库中查找到 593 个国家自然科学基金在 2010 至 2020 年支持的相关项目（包含未结题的项目），如表 22 所示。593 个项目主要分为 14 类，其中面上项目和青年科学基金项目占据绝大多数，占比约 81.7%。

表 22 数据挖掘相关国家自然科学基金项目分类情况（2010-2020 年）

项目类别	项目数量	数量占比
面上项目	266	44.9%
青年科学基金项目	218	36.8%
地区科学基金项目	44	7.4%
重点项目	14	2.4%
联合基金项目	9	1.5%
重大研究计划	8	1.3%
国际(地区)合作与交流项目	7	1.2%
国家杰出青年科学基金	7	1.2%
应急管理项目	6	1.0%
优秀青年科学基金项目	6	1.0%
专项基金项目	5	0.8%
创新研究群体项目	1	0.2%
海外及港澳台学者合作研究基金	1	0.2%
专用基金项目	1	0.2%

图 75 展示了数据挖掘领域国家自然科学基金支持项目在 2010 年到 2020 年的分布情况。从中可以看出，每年的支持项目数量相差不大，2013-2018 年间的支持项目整体数量较高。2012 年，美国奥巴马政府在白宫网站上发布了《大数据研究和发展倡议》，旨在提升利用大量复杂数据集合获取知识和洞见的能力，六大联邦政府机构达成一致，宣布将为此投入 2 亿美元以上经费，支持大力发展对数字化数据的接入、组织和挖掘的工具和技术，并进一步扩展，形成了包括联邦政府 12 个部门和机构的多项研究计划。这一倡议掀起了全球范围内政府推动大数据分析和研究的热潮，进而提升了数据挖掘的研究热度。

图 75 数据挖掘领域国家自然科学基金项目支持历年分布情况

图 76 展示了获得国家自然科学基金支持项目（数据挖掘领域）数量 TOP 15 的中国机构列表。这些机构主要分布在东部地区的北京市（3 个）、上海市（2 个）、广东省（1 个），中部地区的湖南省（2 个）和湖北省（2 个），西北地区的陕西省（1 个），西南地区的四川省（1 个），以及东北地区的黑龙江省、吉林省、辽宁省各 1 个。这些机构的地区分布比较均匀，但是数量上比较偏向中国政治和经济发达地区。

图 76 数据挖掘领域国家自然科学基金项目支持数量 TOP 15 机构统计

6　总结与展望

6 总结与展望

数据挖掘成为了日常生活中的一部分，影响日常生活的方方面面，从当地超市供应的商品、网上冲浪看到的广告，到犯罪预防。数据挖掘技术是一把双刃剑，通过改进服务和提高顾客满意度，以及生活方式，可以为人类生活带来很多好处，但是同时也会严重威胁到个人隐私权和数据安全。

本报告设计了概述篇、技术篇、人才篇、应用篇、趋势篇等 5 个章节，详细介绍了数据挖掘的概念内涵和关键技术，以及该领域的科研学者发展情况和行业应用情况，并通过分析论文、专利和国家自然科学基金等数据，挖掘数据挖掘领域的研究热点，旨在为未来发展趋势分析提供参考。

近年来，随着互联网、物联网、云计算、三网融合等 IT 与通信技术的迅猛发展，数据的快速增长成为了许多行业共同面对的严峻挑战和宝贵机遇，因而信息社会已经进入了大数据时代。网络大数据的复杂性、不确定性和涌现性，给数据挖掘技术带来很大挑战。其中，复杂性主要体现在数据类型的复杂性、数据结构的复杂性和数据内在模式的复杂性。不确定性主要包括数据本身的不确定性、模型的不确定性和学习的不确定性。涌现性主要表现为模式的涌现性、行为的涌现性和智慧的涌现性。有效的数据挖掘方法、系统和服务的开发，交互的和集成的数据挖掘环境的构建是关键的研究领域。使用数据挖掘技术解决大型或复杂的应用问题是数据挖掘研究人员、数据挖掘系统和应用的开发人员面临的重要任务。比如：现有的数据挖掘系统或技术主要是面向特定的问题场景，没有一个统一通用的理论框架来处理所有领域问题。为了实现低成本处理不同行业领域的任务需求，有必要研究构建统一的数据挖掘理论框架。由于收集的数据量不断剧增，与传统的数据分析方法相比，数据挖掘必须能够有效地处理大量数据，并且尽可能是交互的。在增加用户交互的同时，全面提高挖掘过程的总体效率。传统的数据挖掘挖掘方法是集中式的，在当今很多分布式环境（例如，互联网、局域网、传感器网络、云计算等）下不能很好工作，因此期待未来分布式数据挖掘方法有所进展，并且在流数据应用方面建立实时动态数据挖掘模型。

参考文献

[1] JiaweiHan, MichelineKamber, JianPei,等. 数据挖掘:概念与技术[M]. 机械工业出版社, 2012.

[2] 琚会婧.数据挖掘技术在客户关系管理(CRM)中的应用研究[D]. 华北理工大学, 2019.

[3] 熊菊霞,吴尽昭.异构复杂信息网络敏感数据流动态挖掘[J].计算机工程与科学,2020,42(04):628-633.

[4] 王丹丹,刘同明.复杂类型数据挖掘技术的研究现状[J].华东船舶工业学院学报(自然科学版),2003(01):72-76.

[5] 百度文库. 网络数据挖掘 [EB/OL]. https://wenku.baidu.com/view/e047e912ce84b9d528ea81c758f5f61fb73628e9.html

[6] Gan W, Lin J C W, Chao H C, et al. Data mining in distributed environment: a survey[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2017, 7(6): e1216.

[7] 刘滨.分布式数据挖掘综述[J].河北科技大学学报,2014,35(01):80-90.

[8] Chaimontree S, Atkinson K, Coenen F. Multi-agent based clustering: Towards generic multi-agent data mining[C]//Industrial Conference on Data Mining. Springer, Berlin, Heidelberg, 2010: 115-127.

[9] Mateo R M A, Lee J. Data mining model based on multi-agent for the intelligent distributed framework[J]. International journal of intelligent information and database Systems, 2010, 4(4): 322-336.

[10] 刘嬰雯.多模态生物数据分析与挖掘研究[D].西安电子科技大学,2018.

[11] 靳秀英.环境大数据挖掘应用浅析[C].中国环境科学学会环境信息化分会,2016:51-53.

[12] Pandey K K, Shukla D. Challenges of Big Data to Big Data Mining with their Processing Framework[C]. international conference on communication systems and network technologies, 2018.

[13] Kaisler S, Armour F, Espinosa J A, et al. Big data: Issues and

challenges moving forward[C]//2013 46th Hawaii International Conference on System Sciences. IEEE, 2013: 995-1004.

[14] 张博.基于隐私保护的数据挖掘分析[J].信息通信,2018(11):171-174.

[15] 黄晓斌,张兴旺.网络动态数据挖掘研究进展与展望[J].图书情报工作,2015,59(10):6-13.

[16] 向鸿鑫,杨云.不平衡数据挖掘方法综述[J].计算机工程与应用,2019,55(04):1-16.

[17] 郑恩辉. 基于支持向量机的代价敏感数据挖掘研究与应用[D]. 浙江大学

[18] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowledge and information systems, 2008, 14(1):1-37.

[19] 博客园 .C4.5 算法总结 [EB/OL]. https://www.cnblogs.com/sumuncle/p/5610877.html

[20] CSDN. 数据挖掘十大经典算法 [EB/OL]. https://blog.csdn.net/qq_36523839/article/details/82383597?utm_medium=distribute.pc_relevant_download.none-task-blog-BlogCommendFromBaidu-4.nonecase&depth_1-utm_source=distribute.pc_relevant_download.none-task-blog-BlogCommendFromBaidu-4.nonecas.

[21] 天赋好书 . 机器学习原理 [EB/OL]. https://www.cntofu.com/book/85/ml/cluster/kmeans.md.

[22] Wu X, Kumar V, Quinlan J R, et al. Top 10 algorithms in data mining[J]. Knowledge and information systems, 2008, 14(1):1-37.

[23] CSDN. 机器学习复习笔记 [EB/OL]. https://blog.csdn.net/qq_41282377/article/details/104425399.

[24] CSDN. 数据挖掘十大算法（四）：Apriori（关联分析算法）[EB/OL]. https://blog.csdn.net/qq_36523839/article/details/82191677.

[25] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the royal statistical society. Series B (methodological), 1977: 1-38.

[26] 知乎专栏. EM 算法详解[EB/OL]. https://zhuanlan.zhihu.com/p/40991784.

[27] CSDN. Adaboost 算法总结[EB/OL]. https://blog.csdn.net/u014465639/article/details/73609978?utm_source=blogxgwz6.

[28] 知乎专栏. KNN 分类算法[EB/OL]. https://zhuanlan.zhihu.com/p/43570890.

[29] cnblog. kNN 算法：K 最近邻(kNN, k-Nearest Neighbor)分类算法[EB/OL]. https://www.cnblogs.com/jyroy/p/9427977.html#idx_0.

[30] CSDN. 机器学习十大算法之五：朴素贝叶斯算法[EB/OL]. https://blog.csdn.net/u010700066/article/details/81298781.

[31] 马刚. 朴素贝叶斯算法的改进与应用[D].安徽大学,2018.

[32] CSDN. 决策树的生成与剪枝 CART[EB/OL]. https://blog.csdn.net/The_lastest/article/details/106487469.

[33] 李航.统计学习方法.北京:清华大学出版社,2012.

[34] 颜雪松，蔡之华.一种基于 Apriori 的高效规则挖掘算法的研究[J]计算机工程与应用,2002（10）：208-212

[35] 罗可，吴杰.一种基于 Apriori 的改进算法[J]计算机工程与应用,2001（22）：20-22

[36] 何童.不确定性目标的 CLARANS 聚类算法[J].计算机工程,2012,38(11):56-58.

[37] 百度百科. CLARANS[EB/OL]. https://baike.baidu.com/item/CLARANS/2332995?fr=aladdin

[38] (美)韩家炜(Han, J.)等著,范明等译.数据挖掘:概念与技术[M].北京:机械工业出版社,2012.

[39] 海沫.大数据聚类算法综述[J].计算机科学,2016,43(S1):380-383.

[40] CSDN. CURE 算法详解[EB/OL]. https://blog.csdn.net/qq_40793975/article/details/83574309

[41] CSDN. CLIQUE 算法 ---- 子空间聚类算法[EB/OL].

https://blog.csdn.net/zhinanpolang/article/details/84331510

[42] 博客园. ROCK 聚类算法[EB/OL]. https://www.cnblogs.com/1zhk/p/4539645.html

[43] 知乎专栏. 数据挖掘入门笔记——ROCK 聚类（东邻西舍）[EB/OL]. https://zhuanlan.zhihu.com/p/55081137

[44] 百度百科. DBSCAN[EB/OL]. https://baike.baidu.com/item/DBSCAN/4864716?fr=aladdin

[45] CSDN. OPTICS 算法[EB/OL]. https://blog.csdn.net/zanghui426/article/details/50370783

[46] CSDN. DENCLUE — 基于密度分布函数的聚类[EB/OL]. https://blog.csdn.net/weixin_41521681/article/details/90484937

[47] 于洋.一种改进的 COBWEB 算法研究[D]. 哈尔滨工程大学, 2010.

[48] 博客园. SOM 聚类算法[EB/OL]. https://www.cnblogs.com/carreyBlog/p/11984336.htmlhttps://www.cnblogs.com/carreyBlog/p/11984336.html

[49] 陈红宇. 基于网络疾病传播模型的溯源算法及其应用综述[J].电脑编程技巧与维护, 2019(07).

[50] 廖岭.基于文本挖掘和主路径分析的技术趋势预测方法及案例研究[M]. 华中科技大学.2017.07.

[51] 廖君华, 孙克迎, 钟丽霞. 一种基于时序主题模型的网络热点话题演化分析系统[J]. 图书情报工作, 2013, 57(9):96-102.

[52] Venugopalan S , Rai V . Topic based classification and pattern identification in patents[J]. Technological Forecasting & Social Change, 2015, 94:236-250.

[53] 李晓松,雷帅,刘天.基于 IRD 的前沿技术预测总体思路研究[J].情报理论与实践,2020,43(01):56-60.

[54] 尹忠博,罗威,罗准辰.数据驱动的技术预测:现状、技术与趋势[J].情报理论与实践,2018,41(12):35-39.

[55] Zhang Y, Zhang F, Yao P, et al. Name Disambiguation in AMiner:

Clustering, Maintenance, and Human in the Loop[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1002-1011.

[56] 廖剑岚. 决策支持系统中的数据挖掘与 OLAP——数据仓库环境下的信息分析[D].华东师范大学,2002.

[57] 赵伟, 林芬芬, 彭洁等. 创新型科技人才评价理论模型的构建[J]. 科技管理研究, 2012, 032(024):131-135.

[58] 陈苏超, 薛晔. 基于模糊神经网络的高层次创新型科技人才的评价[J]. 太原理工大学学报, 2014, 45(3): 420-424.

[59] 盛楠,孟凡祥,姜滨,等. 创新驱动战略下科技人才评价体系建设研究[J]. 科研管理,2016,37(S1):602-606.

[60] 李瑞, 吴孟珊, 吴殿廷. 工程技术类高层次创新型科技人才评价指标体系研究[J]. 科技管理研究, 2017（18）.

[61] Tang J , Zhang J , Yao L , et al. ArnetMiner: Extraction and Mining of Academic Social Networks[C]// Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2008.

[62] Kumar A , Sangwan S R , Nayyar A . Multimedia Social Big Data: Mining[M]// Multimedia Big Data Computing for IoT Applications. 2020.

[63] 蒋良孝, 蔡之华. 多媒体数据挖掘技术研究[J]. 微型机与应用, 2003(12):46-48+57.

[64] 百度百科. 语音识别 [EB/OL]. https://baike.baidu.com/item/%E8%AF%AD%E9%9F%B3%E8%AF%86%E5%88%AB/10927133?fr=aladdin

[65] 百度百科. 语音合成 [EB/OL]. https://baike.baidu.com/item/%E8%AF%AD%E9%9F%B3%E5%90%88%E6%88%90/9790227?fr=aladdin

[66] 刘钊, 蒋良孝. 图像数据挖掘之研究[J]. 计算机工程与应用, 2003, 039(033):202-204.

[67] 李向伟, 康毓秀. 基于内容的视频检索与挖掘关键技术研究[J]. 软件, 2014(08):26-31.

[68] 杨洋,陈红军.隐私保护数据挖掘技术研究综述[J].微型电脑应用,2020,36(08):41-44+54.

[69] Sweeney L. k-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 557-570.

[70] Homayoun S, Ahmadzadeh M, Hashemi S, et al. BoTShark: A deep learning approach for botnet traffic detection[M]//Cyber Threat Intelligence. Springer, Cham, 2018: 137-153.

[71] Agrawal D, Aggarwal C C. On the design and quantification of privacy preserving data mining algorithms[C]//Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. 2001: 247-255.

[72] Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-anonymity and l-diversity[C]//2007 IEEE 23rd International Conference on Data Engineering. IEEE, 2007: 106-115.

[73] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of cryptography conference. Springer, Berlin, Heidelberg, 2006: 265-284.

[74] Sirole T, Choudhary J. A survey of various methodologies for hiding sensitive association rules[J]. International Journal of Computer Applications, 2014, 96(18).

[75] Wang Q, Wang C, Li J, et al. Enabling public verifiability and data dynamics for storage security in cloud computing[C]//European symposium on research in computer security. Springer, Berlin, Heidelberg, 2009: 355-370.

[76] Ge W, Wang W, Li X, et al. A privacy-preserving classification mining algorithm[C]//Pacific-Asia Conference on Knowledge Discovery

and Data Mining. Springer, Berlin, Heidelberg, 2005: 256-261.

[77] Vaidya J, Clifton C. Privacy-preserving k-means clustering over vertically partitioned data[C]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. 2003: 206-215.

[78] 张锋军,杨永刚,李庆华,许杰,牛作元,石凯.大数据安全研究综述[J].通信技术,2020,53(05):1063-1076.

[79] 简称 AMiner（https://aminer.cn/），2006 年上线，已建立起全领域学术论文文献及专家学者库，收录论文\专利文献超过 3 亿篇，学者信息 1.36 亿份。

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

[81] L. Li, X. Jin, S. Pan et al. Multi-domain Active Learning for Text Classification[C].Proc. KDD 2012. pp. 1086-1094

[82] X. Liu, Z. Zhou. Learning with Cost Intervals[C]. Proc. KDD 2010. pp. 403-412.

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

[84] M. Zhang, K. Zhang. Multi-Label Learning by Exploiting Label Dependency[C]. Proc. KDD 2010. pp. 999-1008.

[85] P. Zhang, J. Li, P. Wang et al. Enabling Fast Prediction for Ensemble Models on Data Streams[C]. Proc. KDD 2011. pp.177-185

[86] L. Han, G. Song, G. Cong et al. Overlapping Decomposition for Causal Graphical Modeling[C]. Proc. KDD 2012. pp.114-122.

[87] X. Liu, Z. Nie, N. Yu et al. BioSnowball: Automated Population of Wikis[C]. Proc. KDD 2010. pp. 969-978

[88] J. Zhang, Y. Song, C. Zhang et al. Evolutionary Hierarchical Dirichlet Processes for Multiple Correlated Time-Varying Corpora[C]. Proc. KDD 2010. pp.1079-1088

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

[90] Z. Qi, M. Yang, Z. Zhang et al. Mining Partially Annotated Images[C]. Proc. KDD 2011. pp. 1199-1207

[91] Y. Hu, D. Zhang, J. Liu et al. Accelerated Singular Value Thresholding for Matrix Completion[C]. Proc. KDD 2012. pp. 298-306

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

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

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

[95] Z. Zou, H. Gao, J. Li. Discovering Frequent Subgraphs Over Uncertain Graph Databases Under Probabilistic Semantics[C]. Proc. KDD 2010. pp. 633-643

[96] C. Tan, J. Tang, J. Sun et al. Social Action Tracking via Noise Tolerant Time-Varying Factor Graphs[C]. Proc. KDD 2010. pp. 1049-1058.

[97] J. Tang, J. Sun, C. Wang et al. Social Influence Analysis in Large-scale Networks[C]. Proc. KDD 2009. pp. 807-816.

[98] C. Tan, J. Tang, J. Sun et al. Social Action Tracking via Noise Tolerant Time-Varying Factor Graphs[C]. Proc. KDD 2010. pp. 1049-1058.

[99] Y. Wang, G. Cong, G. Song et al. Community-Based Greedy Algorithm for Mining Top-K Influential Nodes in Mobile Social Networks[C]. Proc. KDD 2010.. pp. 1039-1048.

[100] L. Xiang, Q. Yuan, S. Zhao et al. Temporal Recommendation on Graphs Via Long- and Short-Term Preference Fusion[C]. Proc. KDD 2010. pp. 723-732.

[101] J. Huang, H. Sun, Q. Song et al. Revealing Density-Based

Clustering Structure from the Core-Connected Tree of a Network[J]. IEEE Trans. Knowledge and Data Engineering, 25(8): 1876-1889, 2012.

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

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

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

[105] P. Cui, S. Jin, L. Yu, F. Wang et al. Cascading Outbreak Prediction in Networks: A Data-Driven Approach[C]. Proc. KDD 2013. pp. 901-909.

[106] J. Tang, S. Wu, J. Sun et al. Cross-domain Collaboration Recommendation[C]. Proc. KDD 2012. pp. 1285-1293.

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

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

[109] J. Zhang, X. Fan, J. Wang et al. Keyword-Propagation-Based Information Enriching and Noise Removal for Web News Videos[C]. Proc. KDD 2012. pp. 561-569.

[110] G. Li, Y. Yang, T. Wang et al. Location-Aware Publish/Subscribe[C]. Proc. KDD 2013. pp. 802-810.

[111] H. Yin, Y. Sun, B. Cui et al. LCARS: A Location-Content-Aware Recommender System[C]. Proc. KDD 2013. pp. 221-230.

[112] V. Mayer-Schonberger, K. Cukier. Big Data: A Revolution That

Will Transform How We Live, Work, and Think[M]. Eamon Dolan/Houghton Mifflin Harcourt. 1 edition

[113] G. He, H. Feng, C. Li et al. Parallel SimRank Computation on Large Graphs with Iterative Aggregation[C]. Proc. KDD 2010. pp. 543-552.

[114] W. Zhang, Y. Zhang, B. Gao et al. Joint Optimization of Bid and Budget Allocation in Sponsored Search[C]. Proc. KDD 2012. pp. 1177-1185

[115] J. Zhu, X. Zheng, L. Zhou et al. Scalable Inference in Max-Margin Topic Models[C]. Proc. KDD 2013. pp 964-972.

[116] M. Ou, P. Cui, F. Wang et al. Comparing Apples to Oranges: A Scalable Solution with Heterogeneous Hashing[C]. Proc. KDD 2013. pp. 230-238.

[117] S. Jiang, L. Bing, B. Sun et al. Ontology Enhancement and Concept Granularity Learning: Keeping Yourself Current and Adaptive[C]. Proc. KDD 2011.. pp. 1244-1252.

[118] 搜狐. 几组模型教你看清零售业的数据挖掘问题[EB/OL]. https://www.sohu.com/a/223452821_400678

[119] 安建华.数据挖掘技术在零售业中的应用研究[D].东北财经大学,2005.

[120] 徐蓉艳.旅游大数据与挖掘及其在旅游行业的应用方向[J].中国市场,2014(51):204-205+208.

[121] 梁昌勇,马银超,路彩红.大数据挖掘:智慧旅游的核心[J].开发研究,2015(05):134-139.

[122] 罗成奎.大数据技术在智慧旅游中的应用[J].旅游纵览(下半月),2013(08):59-60.

[123] 计算机毕业设计. asp.net 基于数据挖掘的物流信息系统[EB/OL]. http://www.360bysj.com/a/asp_net/2018/1117/486.html

[124] snc.mimi. 医疗保健中的数据挖掘[EB/OL]. http://www.srcmini.com/12138.html

[125] Jensen PB, Jensen LJ, Brunak S Mining electronic health records:towards better research applications and clinical care Nat Rev Genet, 2012, 13(6):395-405

[126] 秦文哲,陈进,董力.大数据背景下医学数据挖掘的研究进展及应用[J].中国胸心血管外科临床杂志,2016,23(01):55-60.

[127] 新金融. 互联网和金融，在数据挖掘上究竟存在什么区别？[EB/OL].https://www.sohu.com/a/200137473_116182

[128] 朱永佳,宋俊典,蔡浩淼.基于大数据背景下数据挖掘在金融行业的应用[J].科技视界,2020(09):121-122.

[129] 朱成,刘海强,朱峰,孙启新.电信大数据的数据挖掘关键技术分析与探讨[J].电信快报,2018(06):22-24.

[130] 张军.试分析数据挖掘在通信行业营销中的应用[J].信息通信,2018(07):254-255.

附录 1 数据挖掘领域关键词

本报告根据 1.3 章节介绍的数据挖掘知识图谱，结合第 2 章介绍的技术领域整理了如下所示的数据挖掘领域的关键词列表。其中考虑到一些关键词（如支持向量机、决策树等）不只属于数据挖掘领域，为了保证数据获取准确性，针对此类关键词增加了条件约束，即这些关键词出现的文本中必须同时出现“数据（Data）”或“挖掘（Mining）”等关键词。

表 23 数据挖掘领域关键词列表

英文关键词	中文关键词
Data Mining	数据挖掘
Social Network	社交网络
Clustering Algorithm	聚类算法
Text Mining	文本挖掘
Data Analysis	数据分析
Anomaly Detection	异常检测/异常发现
Association Rule	关联规则
Big Data	大数据
Web Mining	web 挖掘/网络挖掘
Outlier Detection	离群点检测
Classification Algorithm	分类算法
Decision Tree	决策树
Data Warehouse	数据仓库
Data Cube	数据立方体
Frequent Pattern	频繁模式/频繁项集
Apriori Algorithm	Apriori 算法
C4.5 Algorithm	C4.5 算法
K-Means Algorithm	K-Means 算法
Support Vector Machine	支持向量机
Expectation Maximization	最大期望算法/EM 算法
PageRank	PageRank
Adaboost Algorithm	Adaboost 算法
K-Nearest Neighbor	k 近邻算法
Naive Bayes	朴素贝叶斯
Classification And Regression Trees	CART 算法
Graph Mining	图挖掘
Time Series Analysis	时间序列分析
Knowledge Discovery	知识发现
Multimedia Mining	多媒体挖掘

附录 2 期刊和会议列表

表 24 数据挖掘领域代表性期刊和会议列表

序号	刊物全称	刊物简称
1	IEEE Transactions on Knowledge and Data Engineering	TKDE
2	ACM Transactions on Knowledge Discovery from Data	TKDD
3	Data and Knowledge Engineering	DKE
4	Data Mining and Knowledge Discovery	DMKD
5	ACM Knowledge Discovery and Data Mining	SIGKDD
6	IEEE International Conference on Data Engineering	ICDE
7	ACM International Conference on Web Search and Data Mining	WSDM
8	International Conference on Data Mining	ICDM
9	SIAM International Conference on Data Mining	SDM

附录 3 国家自然科学基金 NSFC 项目

表 25 数据挖掘相关国家自然科学基金项目列表（2010-2020 年）

项目类别	项目标题	依托单位
专用基金项目	用于空间环境分析的自主航天测控数据挖掘技术研究	北京航天飞行控制中心
	数据挖掘：原理与方法研讨	贵州大学
	面向大数据的媒体内容分析与关联语义挖掘研究	南京大学
	基于协同管理的道路交通安全风险分析的知识发现与推理研究	大连海事大学
	汉语文本数据挖掘的统计方法	东北师范大学
	Web 文本意见挖掘关键技术研究	北京理工大学
重点项目	在线社会关系网络的挖掘与分析	中国科学院计算技术研究所
	生物网络数据分析与挖掘中相关理论与关键技术	西安电子科技大学
	面向网络事件的跨平台异质媒体语义协同与挖掘	中国科学院大学
	面向脑疾病的人体肠道微生物组学数据挖掘方法研究	复旦大学
	基于云计算的海量数据挖掘关键技术研究	清华大学
	基于云计算的海量数据挖掘	中国科学院计算技术研究所
	基于移动群智感知的物联网大数据挖掘与应用	北京航空航天大学
	基于图与组合优化的生物数据和网络数据挖掘算法研究	山东大学
	基于数据挖掘与机器学习的重要能源转化过程电催化剂的设计与创新	南开大学

	基于数据挖掘的南海岛礁演变机制及多尺度模拟预测技术研究	长沙理工大学
	大数据理解与知识发现的若干方法	中南大学
	大数据环境下地理关联模式挖掘的理论与方法	中南大学
	WEB 搜索与挖掘的新理论与方法	北京大学
	Web 搜索与挖掘的新理论和新方法—支持舆情监控的 Web 搜索与挖掘的理论与方法研究	中国科学院计算技术研究所
重大研究计划	证券管理决策大数据挖掘云服务平台研究	中国科学院计算技术研究所
	一年期滚动项目——证券管理决策大数据挖掘云服务平台研究	中国科学院计算技术研究所
	网络舆情大数据的地理空间情报价值发现与挖掘	中国科学院自动化研究所
	基于多源时空大数据群体时空移动规律挖掘与动力学建模	深圳大学
	基于大数据分析的犯罪模式挖掘与犯罪预测研究	中国电子科技集团公司第三十八研究所
	基于大数据的中国资本市场全息社会网络深度挖掘及研究	上海财经大学
	基于大数据的灾难信息挖掘与决策支持的关键技术研究	南京邮电大学
	管理与决策大数据的模式识别与敏感内容挖掘研究	合肥工业大学
优秀青年科学基金项目	知识发现与知识工程	清华大学
	数据挖掘与商务智能	北京航空航天大学
	基于模式识别理论的生物信息数据挖掘	中国科学院新疆理化技术研究所
	管理信息与数据挖掘（研究领域）	西南财经大学
	多模态数据知识发现	天津大学
	大数据挖掘与应急管理	中国人民解放军国防科学技术大学
应急管理项目	面向复杂海洋大数据的特征挖掘与建模及其应用研究	中国海洋大学
	基于点对点网络架构的分布式协同数据聚类的关键技术研究	济南大学
	基于大数据的国家自然科学基金成果挖掘与统计分析研究	东北师范大学
	多源异构数据的决策特征提取与知识发现	华南理工大学
	单细胞测序数据计算方法开发和数据挖掘	南方科技大学
	大规模属性图中的高效模式挖掘方法研究	西安交通大学
	中医药知识发现可靠性研究	浙江工商大学

青年科学基金项目	中药橘叶抗乳腺炎的药效机制研究及基于多策略数据挖掘系统的"谱－效"关系构建	成都医学院
	智能感知的盾构地铁运营风险知识发现与安全状态时变精准评价研究	华中科技大学
	知识驱动下基于数据挖掘的滑坡预报研究	中国地质大学（武汉）
	知识驱动的多目标决策数据挖掘理论框架及应用实验系统研究	电子科技大学
	整合猪 miRNA 和功能基因表达谱芯片元数据挖掘肌肉生长发育新的调控通路	华中农业大学
	整合基因组和表型组数据挖掘遗传病潜在致病变异的方法研究	复旦大学
	针对农残检测数据中隐性知识发现的可视分析模型与算法研究	北京工商大学
	用数据挖掘及机器学习算法进行伽玛暴分类研究	河北师范大学
	影像遗传学中海量数据挖掘算法研究及其在老年痴呆症中的应用	西北工业大学
	以患者为中心视角下基于数据挖掘的慢性病精准预防与管理模型构建	中国人民大学
	移动社交网络中用户签到位置轨迹挖掘及信息推荐策略研究	中南财经政法大学
	移动社交电商平台 UGC 对用户购买意愿的影响研究：基于大数据文本挖掘	上海外国语大学
	一种基于知识发现方法的临床检验决策支持新模式研究	中国人民解放军第四军医大学
	岩石高边坡安全监控模型的数据挖掘研究及应用	河海大学
	享乐性与实用性商品的个性化推荐策略与技术研究——基于用户评论大数据驱动的商品特征挖掘	华南理工大学
	物联网大数据中价值富集离群点检测方法的研究	大连海事大学
	无指导汉语文本挖掘的统计模型和统计推断	清华大学
	微生物漆酶数据挖掘及底物杂泛性分析	安徽大学
	网络重要节点及链路挖掘方法研究	杭州师范大学
	网络意见挖掘若干前沿问题研究	山东大学
	探索雾霾形成的模糊认知图构建及其数据挖掘方法	北京石油化工学院
	台湾海峡船舶 AIS 轨迹数据挖掘及其航路规划优化	集美大学
	数据聚类问题中的一类张量优化方法研究	广东工业大学
	时空轨迹数据挖掘及其隐私保护方法研究	安徽师范大学
	时间演化尺度下大规模社会网络特征分析与社区结构挖掘	山东科技大学
	生物医学文本大数据中的疾病关系并行挖掘模型研究	长沙理工大学
	社交网络开放平台漏洞挖掘及威胁评估方法研究	中国科学院大学

	社交网络互动中用户“信息窄化”机理分析：基于微博的数据挖掘	同济大学
	社交-推荐网络中的隐式朋友挖掘	北京航空航天大学
	上下文感知的移动社交网络社会化挖掘与推荐技术研究	电子科技大学
	区间值时序数据挖掘中聚类与预测的研究	山东青年政治学院
	破坏性创新环境下的未来创新路径研究：构建数据挖掘与技术预测桥梁的视角	北京理工大学
	纳米分辨率显微视频的像素级数据挖掘研究	衢州学院
	面向智慧交通的社交媒体大数据挖掘关键技术研究	东北林业大学
	面向隐私数据保护的支持向量机新方法及其抗攻击模型研究	许昌学院
	面向隐私保护的分布式数据挖掘关键问题研究	河北工程大学
	面向循证医学的大规模医学信息关联模式挖掘技术研究	首都医科大学
	面向新型隐私保护的海量图数据挖掘	哈尔滨工程大学
	面向问题模式挖掘的复杂芯片制造过程近全局优化动态调度方法研究	同济大学
	面向糖尿病电子病历大数据的可解释性时序信息挖掘研究	暨南大学
	面向顺式调控元件及模块识别的近似序列模式挖掘	北京交通大学
	面向数据流的无限论域动态粒的不确定性分析与知识发现研究	安徽大学
	面向事件预测的因果知识发现、验证和推理研究	中国科学院信息工程研究所
	面向社交位置大数据的用户潜在兴趣地点挖掘	北京邮电大学
	面向全基因组关联研究的动态数据挖掘与深度查询方法	中南大学
	面向技战术分析的足球视频中的运动模式挖掘	中国农业大学
	面向候鸟迁徙行为的数据挖掘算法研究	中国科学院计算机网络信息中心
	面向个性化推荐服务的社交网络数据深挖掘关键技术研究	湖州师范学院
	面向感应网络的移动现实挖掘及复杂行为模式分析研究	中国科学院沈阳自动化研究所
	面向访问控制的数据挖掘分类外包技术研究	中国人民解放军国防科学技术大学
	面向多示例数据标注的隐变量支持向量机研究	广东工业大学
	面向动态复杂数据的粒化模型与知识发现研究	江西农业大学
	面向大型社会网络融合的关联用户挖掘模型与方法	北京建筑大学

	面向大数据的非平行超平面支持向量机理论、算法与应用研究	深圳大学
	面向大量数据的半监督支持向量机的优化方法研究	华中农业大学
	面向大规模数据挖掘的隐私保护支持向量机增量与并行学习算法研究	山东科技大学
	面向大规模基因表达谱的数据挖掘及并行分析方法研究	哈尔滨工业大学
	面向城市公租自行车优化管理的大数据挖掘方法	北京建筑大学
	满足差分隐私的频繁模式挖掘研究	北京邮电大学
	领域知识驱动的演化行为模式挖掘研究	同济大学
	矿床领域文本数据挖掘与知识图谱构建	中国地质大学（武汉）
	空间目标特性数据挖掘的可微分编程方法研究	中国人民解放军战略支援部队航天工程大学
	可降低和评价不确定性影响的可拓时空关联规则挖掘方法研究	浙江大学
	精细风场要素驱动下的数据挖掘林火蔓延元胞自动机模型研究	成都信息工程大学
	结核病基因转录组学数据挖掘方法的研究	西南大学
	结合情境感知的移动互联网高维、多源、异构用户数据挖掘方法研究	西北工业大学
	基于众数的函数型数据聚类方法	浙江工业大学
	基于中医临床大数据的名老中医诊治慢性肾脏病经验挖掘方法的应用研究	南京中医药大学
	基于中小规模样本数量的半导体材料基因组数据挖掘技术	中国农业大学
	基于中国之星小望远镜阵 CSTAR 的数据挖掘及所探测变源的后续研究	中国科学院国家天文台
	基于运动视频的针刺手法参数采集及神经网络数据挖掘研究	上海中医药大学
	基于云计算平台 Hadoop 的海量数据聚类研究	湘潭大学
	基于元基因组相似度计算的海量微生物群落数据挖掘	中国科学院青岛生物能源与过程研究所
	基于语义的医学领域前沿知识发现及演化机制研究	中国医学科学院
	基于语义的突发危机事件知识发现与决策支持研究	上海交通大学
	基于有限样本的高分五号高光谱数据分类方法研究	自然资源部国土卫星遥感应用中心
	基于引文网络图数据挖掘的热点技术领域预测研究	东北林业大学
	基于意见传播原理的社交网络数据挖掘理论和应用研究	深圳大学

	基于已存知识重用的大数据分布式递进分类挖掘方法研究	江苏大学
	基于循证评价与数据挖掘构建我国肺癌筛查高风险人群决策树模型的研究	中国医学科学院肿瘤医院
	基于穴位皮区模型和体表空间分析的中医皮肤科临床数据挖掘探索性研究	中国中医科学院中医药信息研究所
	基于序列模式挖掘的火电厂制粉系统节能优化技术研究	西安交通大学
	基于稀疏松弛匹配与图聚类分析的共同视觉模式挖掘方法与应用研究	安徽大学
	基于吸引子传播的半监督文本挖掘方法研究	吉林大学
	基于无监督学习的单分子电导数据挖掘	厦门大学
	基于文本挖掘的网络信息与股票市场关联机制研究	浙江大学
	基于文本挖掘的社会媒体药品不良反应抽取研究	天津财经大学
	基于数据挖掘与运气学说的香港地区气象疫病监测预警系统研究	香港浸会大学深圳研究院
	基于数据挖掘与即时反馈的医疗建筑循证设计信息技术开发研究	青岛理工大学
	基于数据挖掘技术实现早期乳腺癌的个体化预后预测	中山大学
	基于数据挖掘技术的疏散星团研究	常州大学
	基于数据挖掘技术的焦虑抑郁共病中医证候学规律研究	北京中医药大学
	基于数据挖掘和实用性随机对照试验的五运六气临床价值评价	北京中医药大学
	基于数据挖掘和感知分析的非对称失真视觉质量评价模型研究	浙江科技学院
	基于数据挖掘和代谢组学技术的小青龙汤证证本质研究	山西中医学院
	基于数据挖掘方法及不确定性分析对公共建筑冷水机组群控策略的识别、评估和优化方法研究	同济大学
	基于数据挖掘的组蛋白修饰模式发现及转录调控功能研究	东华大学
	基于数据挖掘的住宅建筑用户行为及相应能耗预测模型研究	湖南大学
	基于数据挖掘的隧道施工全过程安全风险动态评估与管控方法	山东大学
	基于数据挖掘的水上交通肇事逃逸船舶自动追踪方法研究	大连海事大学
	基于数据挖掘的数字地价模型及其可靠性理论	北京建筑大学
	基于数据挖掘的区域综合能源系统能效状态评价与运行优化策略研究	太原理工大学

	基于数据挖掘的脑卒中就医延迟风险预测模型建立与风险管理机制研究	哈尔滨医科大学
	基于数据挖掘的空调负荷预测方法研究	中原工学院
	基于数据挖掘的结直肠癌临界点的预警算法	华南理工大学
	基于数据挖掘的共享航次计划用户行为分析研究	国家海洋局第一海洋研究所
	基于数据挖掘的高血压肝火亢盛证的生物学机制研究	北京中医药大学
	基于数据挖掘的第三方构件安全性测试方法研究	江苏大学
	基于数据挖掘的超高分辨率影像海岸带地物二次分类方法研究	国家海洋局第一海洋研究所
	基于数据挖掘的阿尔茨海默症风险因素的动态模式探测研究	广东工业大学
	基于数据挖掘的 Sepsis 患儿集束化治疗时间窗预测及内环境变化模式的研究	广州医科大学
	基于数据挖掘-Copula 理论的银行信贷中观组合管理研究	浙江大学宁波理工学院
	基于树结构模式挖掘的 Web 信息抽取研究	合肥工业大学
	基于属性偏序可视化表示原理的脾虚证方剂配伍规律知识发现方法	燕山大学
	基于时空数据挖掘的国人伤害死亡风险分布特征及影响因素研究	华中科技大学
	基于生物数据挖掘技术的脓毒症早期诊断新模型的构建及应用研究	苏州大学
	基于生物大数据挖掘技术的华法林个体化用药预测新模型的构建及应用	中南大学
	基于深度学习的乳腺癌分子生物信息的文本挖掘研究	南京邮电大学
	基于深度交通事故调查的驾驶人应急行为数据库建设及数据挖掘研究	广东技术师范学院
	基于熵的公共卫生大数据信息挖掘方法研究	南京医科大学
	基于认知语言学原理的工艺知识发现方法研究	西北工业大学
	基于缺失属性值区间型描述的不完备数据聚类方法及应用研究	大连理工大学
	基于迁移学习的企业不完备数据中的知识发现方法研究	大连理工大学
	基于偏最小二乘和贝叶斯理论的代谢组学数据挖掘的新算法研究	上海交通大学
	基于模糊推理的复杂形状物体不完整点云数据分类策略及模糊估值研究	齐齐哈尔大学
	基于模糊粗糙集的概率数据挖掘方法研究	中国人民大学

	基于免疫蚁群优化模型的多源遥感湿地精细分类知识发现	中国科学院东北地理与农业生态研究所
	基于扩展模糊积分的生物信息数据挖掘研究	华南农业大学
	基于跨媒体信息挖掘的网络舆情分析研究	中国科学院自动化研究所
	基于跨媒体数据挖掘的社会图像事件分析与标注	北京航空航天大学
	基于空间数据挖掘的高分辨率遥感图像水上桥梁目标识别与损毁评估	浙江海洋大学
	基于空间数据挖掘的复杂型面制造质量特征获取方法研究	天津大学
	基于进化组合支持向量机（EC-SVM）的代谢组学数据分析方法研究	复旦大学
	基于结构知识发现与推理的高适应性室内三维重建	武汉理工大学
	基于监测数据挖掘的库区滑坡表面变形模式及时空规律研究	三峡大学
	基于集成异构网络的民航旅客-航班关联挖掘研究	中国民航大学
	基于集成学习的商务智能中非均衡数据分类方法研究	合肥工业大学
	基于集成学习的不平衡流数据分类问题研究	浙江工商大学
	基于宏观症状和微观指标的肝炎后肝硬化证候分类、演变规律及生物学机制的数据挖掘研究	北京中医药大学
	基于果蝇 piRNA 大数据挖掘的基因调控机制研究	上海交通大学
	基于轨迹数据的景区游客时空运动模式挖掘及内在机理研究	厦门大学
	基于关系 Markov 网的多关系数据聚类分析方法研究	吉林大学
	基于工业大数据挖掘的复杂产品总完工时间动态预测	西安电子科技大学
	基于高维数据聚类的算法交易策略若干关键问题研究	武汉大学
	基于复杂网络理论的肺系病中医古籍本体构建与知识发现研究	广州中医药大学
	基于复杂网络的大众生产信息挖掘：模型、算法与应用	华南理工大学
	基于复杂电磁大数据的辐射源联合参数模型挖掘和识别	中国电子科技集团公司第二十八研究所
	基于蜂窝基站轨迹数据挖掘的语义化位置感知计算研究	浙江工业大学
	基于分子信标微流控芯片的大数据存储与挖掘	安徽理工大学
	基于分形与数据流挖掘技术的动态数据挖掘方法及其应用研究	安徽大学

	基于分布式序列模式挖掘物联网服务用户隐私保护关键技术研究	青岛科技大学
	基于二值化的网络视觉大数据高效挖掘与分析	中国科学院信息工程研究所
	基于多组学数据挖掘阿尔茨海默病生物网络功能模块研究	常州工学院
	基于多组学数据的癌症驱动模块网络挖掘方法研究	曲阜师范大学
	基于多重分形和文本数据流技术的网络金融信息动态挖掘研究	合肥工业大学
	基于多源信息融合与知识发现的复杂工业过程工况识别方法研究	湖南师范大学
	基于多源软件数据挖掘的修改分析与预测关键技术研究	扬州大学
	基于多维潜变量模型和数据挖掘的溃疡性结肠炎中医证候量化方法学研究	广州中医药大学
	基于多视图学习的癌症多组学数据聚类方法研究	中国矿业大学
	基于多目标优化的约束模式挖掘方法研究	安徽大学
	基于多粒度知识发现的人群复杂行为模式分析及预测模型研究	西安科技大学
	基于多关系数据挖掘的跨媒体推荐关键技术研究	北方工业大学
	基于第二代测序技术的 DNA 甲基化数据挖掘分析新方法的研究及网络服务器的搭建	中国科学院北京基因组研究所
	基于大数据整合挖掘的肾细胞癌分子进化机制研究	深圳大学
	基于大数据协同感知的学习行为时间模式挖掘研究	西南大学
	基于大数据挖掘技术的盾构刀具磨损分析与识别研究	中国人民解放军理工大学
	基于大数据挖掘的数控机床多工况载荷谱系研究	华中科技大学
	基于大数据挖掘的桥梁结构健康状态分析与评估研究	重庆大学
	基于粗糙概念格理论的多尺度模具知识发现及重用方法研究	广东工业大学
	基于变分原理的数据聚类模型的数学理论与计算方法	华中科技大学
	基于本体方法的我国上市甲乙肝疫苗相关不良反应数据挖掘与建模分析	重庆邮电大学
	基于本地化差分隐私的高效用频繁模式挖掘技术研究	中国海洋大学
	基于 microRNA 原位杂交和数据挖掘技术的三阴性乳腺癌预后预测模型研究	复旦大学
	基于 GIS 与数据挖掘的裂隙岩质隧道三维模拟与灾害预测	吉林大学
	混合型生物医学知识网络构建及隐性知识发现方法研究	吉林大学

	关联规则集上的知识发现	河南理工大学
	顾及多变量关联特性的时空异常模式挖掘模型与方法	武汉大学
	功能性前噬菌体数据挖掘及数据库的建立	中国人民解放军军事科学院军事医学研究院
	公安视频大数据中的步态模式挖掘与表达	中国人民公安大学
	高维数据聚类信息核保存与隐藏技术研究	东南大学
	干旱区径流过程的数据挖掘研究——以塔里木河流域为例	中国科学院地理科学与资源研究所
	复杂工况下基于数据挖掘的资源消耗会计分摊方法研究	东北财经大学
	复合关系粗糙集模型及高效知识发现算法研究	西南交通大学
	多元时间序列数据挖掘中的特征表示和相似性度量方法研究	华侨大学
	多维视角下基于文献实体单元共现网络分析的药物关系知识发现研究	山西医科大学
	多率含噪时滞非线性系统基于数据挖掘的自适应控制	中国计量大学
	多粒度视角下大规模数据聚类算法研究	山西大学
	多采样策略下的强精度近似频繁模式挖掘	中国人民解放军国防科技大学
	多标记数据分类及其特征选择算法研究	浙江师范大学
	盾构施工诱发建筑物损坏粒计算知识发现与动态风险评价研究	华中科技大学
	地理环境多因子作用下的点集异常丛聚模式挖掘	武汉大学
	当机器智能遇到人类计算—基于众包的分类数据挖掘技术研究	南京大学
	大数据中的多粒度知识发现模型与方法研究	重庆邮电大学
	大数据挖掘在科技项目查重中的应用研究	中国科学技术信息研究所
	大数据环境下稀有类数据挖掘研究	武汉大学
	大数据环境下面向互联网金融的个人信用深度挖掘与评价研究	长沙理工大学
	大数据环境下基于同步原理的数据流挖掘算法研究	电子科技大学
	大数据环境下轨迹破案中的移动对象数据挖掘关键技术研究	北京工业大学
	大数据环境下的文本挖掘并行处理技术研究	中南林业科技大学
	大数据环境下的土石坝病险挖掘和诊断	水利部交通运输部国家能源局南

		京水利科学研究院
	大规模网络子图模式快速挖掘方法研究	西安交通大学
	大规模社交网络中基于复杂社会行为的社会信任关系挖掘	苏州大学
	大规模高分辨质谱数据挖掘新方法研究	中南大学
	尺度驱动的时空集聚模式挖掘模型与方法	中南大学
	城市交通出行时空分布模式挖掘与个性化路径规划	中国科学院地理科学与资源研究所
	不完备多视图数据聚类分析关键技术研究	西安电子科技大学
	不确定性数据流上的频繁项集挖掘关键技术研究	中央财经大学
	不确定数据分类学习的支持向量机算法研究	哈尔滨工业大学
	不确定环境下基于数据挖掘的群体偏好行为评估	电子科技大学
	病证结合证候特征模式挖掘的方法学研究	北京中医药大学
	标准信息计量方法与数据挖掘研究	中国科学院武汉文献情报中心
	保护隐私的海量数据挖掘	中山大学
	Web 图像视觉模式挖掘及其应用	中国科学院上海高等研究院
	WEB2.0 环境下基于本体学习的观点挖掘研究	中南财经政法大学
	PageRank 问题的研究及其在基因芯片数据挖掘中的应用	江苏师范大学
	MapReduce 集群环境下的恒星光谱关联规则挖掘及性能优化	太原科技大学
	FOXO3 调控脐血干细胞红系分化的系统组学数据挖掘	中国科学院北京基因组研究所
面上项目	宗教、外部监督与审计质量：数据挖掘与经验证据	厦门大学
	自适应的中文网络意见挖掘关键技术研究	黑龙江大学
	准同型相界原理结合数据挖掘技术设计窄热滞后形状记忆合金	西安交通大学
	专家学术活动轨迹识别的时序关联规则挖掘与知识发现	中南大学
	注意力机制下的名老中医诊疗经验知识发现与问答模型构建方法研究	广州中医药大学
	主动配电网态势图优化建模及大数据态势图片挖掘理论与方法研究	杭州电子科技大学
	重大疾病多组学与医学大数据挖掘基础理论及关键技术	中山大学
	中医药治疗肺癌核心处方的数据挖掘研究	上海中医药大学
	智能电网知识可视化模型及知识发现策略研究	东北电力大学

	支持网络舆情分析的观点挖掘及观点社群发现关键技术研究	东北大学
	整合大数据挖掘与路网脆弱性评估的疏散交通管理研究	哈尔滨工业大学
	整合 GWAS 和 eQTL 数据挖掘绵羊肉质性状关键基因	兰州大学
	针对来自众包的大数据支持向量机研究	南京信息工程大学
	长链烷基糖苷酶基因的数据挖掘、克隆表达及催化性能研究	华东理工大学
	长非编码 RNA 与复杂疾病关联异常模式挖掘方法研究	西安电子科技大学
	噪声数据的非凸损失函数支持向量机最优化模型与算法研究	海南大学
	在线社交网络跨媒体大数据挖掘与个性化智能搜索研究	北京邮电大学
	在线多分辨率时间演化图聚类及簇演变模式挖掘	西安电子科技大学
	云计算环境下基于分形数据挖掘技术的商务智能系统的研究	合肥工业大学
	云计算环境下海量不平衡数据分类研究	辽宁师范大学
	云计算环境下顾及用户关系的手机用户时空轨迹模式挖掘方法研究	南京师范大学
	云服务环境下知识发现与数据挖掘的安全与隐私保护关键技术	浙江工商大学
	语义支撑下以原油市场为例的 Web 知识发现关键方法及实证研究	中国科学院科技战略咨询研究院
	应用数据挖掘技术研究中医药治疗再生障碍性贫血的组方规律	中国人民解放军第二一 0 医院
	信息网络中基于结构及属性的社区挖掘研究	中国科学技术大学
	消费者投诉的短文本挖掘及新消保法的影响研究	上海交通大学
	文件流型大数据的分析模型构建与知识发现研究	南开大学
	文本挖掘的统计建模	北京大学
	微生物组大数据整合与数据挖掘方法研究	华中科技大学
	外包软件项目风险智能决策支持系统研究-基于因果分析和可行动知识发现集成框架	暨南大学
	图最大化问题的近似算法及其在金融数据挖掘中的应用	西安交通大学
	提高支持向量机处理复杂数据效能的方法研究	山西大学
	数据挖掘中的稀疏张量优化方法研究	曲阜师范大学
	数据挖掘与找矿靶区资源潜力模拟非线性技术研究	中国地质大学（北京）

	数据挖掘驱动的水沙模型时序耦合关键理论与技术研究	黄河水利委员会黄河水利科学研究院
	数据挖掘和静态分析相结合的重复代码缺陷检测及重构方法	哈尔滨工业大学
	数据仓库中行列混合存储引擎的优化模型	东华大学
	时滞复杂系统的集群特征及其在数据挖掘中的应用研究	中国人民解放军国防科学技术大学
	时空异步关联规则挖掘的模型和算法研究	武汉大学
	生物分子网络的统计构建与分析及其在油菜组学数据挖掘中的应用	河南大学
	射电干涉数据自动化处理管线程序 SAND 与 VLBI 监测数据挖掘	中国科学院新疆天文台
	社团发现的图挖掘算法及其并行化以及应用研究	电子科技大学
	社交网络中基于图挖掘的用户行为分析与异常检测技术研究	北京邮电大学
	社交媒体健康知识发现与个性化诊疗方法研究	北京理工大学
	社会网络分析与挖掘若干关键技术研究	中山大学
	软件系统复杂网络层次化实体挖掘方法及关键技术研究	燕山大学
	人机融合智能数据挖掘系统中众包策略优化的研究	大连理工大学
	群组决策动态过程的建模和数据挖掘方法研究	清华大学
	迁移学习及其在气象雷达数据分类中的应用	中国人民解放军国防科学技术大学
	南京民国建筑修缮 BIM 模型实例库的构建及其数据挖掘与知识发现研究	东南大学
	目标导向的网络挖掘与推荐关键问题研究	中国科学院大学
	模糊动态多目标优化及在演化数据聚类中的应用研究	淮北师范大学
	面向中医临床大数据的现代名老中医肺癌辨治规律并行挖掘策略及方法学研究	南京中医药大学
	面向智慧城市交通系统的时空大数据挖掘技术研究及应用	湖南大学
	面向知识发现的网络对齐方法及其在生物网络中的应用	中南大学
	面向在线教育考试评估的数据挖掘技术研究与应用	中国科学技术大学
	面向移动用户个性化服务的情境数据挖掘方法及应用研究	中国科学技术大学
	面向微博数据的位置相关事件检测和时空异常聚类模式挖掘研究	武汉大学

	面向网络异构信息源的问答资源挖掘	哈尔滨工业大学
	面向图文混合的网络舆情新事件发现及其关联挖掘	西安邮电大学
	面向数据挖掘的轨迹数据结构化表示研究	复旦大学
	面向社交网络的药物不良反应的隐含知识发现	大连理工大学
	面向热门争议话题的基于社交网络文本与结构的层次观点挖掘研究	东南大学
	面向群体协同开发的软件工程关联数据挖掘	上海交通大学
	面向流行病防控的大规模人口动态接触网络建模与挖掘方法	吉林大学
	面向领域用户知识发现的数据结构化建模与多粒度融合	重庆邮电大学
	面向领域的多粒度动态海量数据挖掘理论模型与方法	重庆邮电大学
	面向客户全生命周期分析与决策的数据挖掘关键算法研究	哈尔滨工业大学
	面向可拓建筑策划与设计的可拓数据挖掘理论及其方法研究	哈尔滨工业大学
	面向金融大数据的半监督聚类集成挖掘关键技术研究	吉林财经大学
	面向交叉领域文本挖掘的主题模型研究	山东工商学院
	面向基于图的数据挖掘的 FPGA 加速方法研究	华中科技大学
	面向互联网大数据的用户兴趣挖掘及预测研究	香港城市大学深圳研究院
	面向海量数据的语境离群点检测技术研究	南京大学
	面向过程的海洋异常变化时空关联模式挖掘方法	中国科学院遥感与数字地球研究所
	面向轨迹大数据的语义标注与语义模式挖掘算法研究	烟台大学
	面向关系数据库知识发现的概率逻辑贝叶斯网络研究	吉林大学
	面向高维数据挖掘的非负矩阵分解关键问题研究	北京交通大学
	面向高光谱数据分类的深度学习方法研究	南京信息工程大学
	面向复杂数据的多粒度知识发现建模与三支决策分析	西南大学
	面向动态数据认知的知识发现理论模型与方法	重庆邮电大学
	面向大众生产的软件信息网络挖掘及其应用研究	中国人民解放军国防科学技术大学
	面向大数据的位置信息挖掘与推荐方法研究——以南京市出租车 GPS 数据为例	南京信息工程大学

	面向大规模复杂数据的多粒度知识发现关键理论与方法研究	中国矿业大学
	面向大规模复杂数据的地铁施工安全多粒度知识发现与动态风险感知研究	华中科技大学
	面向船型优化的数据挖掘方法及应用研究	武汉理工大学
	面向产品质量特征需求基于数据挖掘的公差再设计建模与优化	湖南科技大学
	面向财经应用的文本挖掘若干关键技术研究	江西财经大学
	面向 Web 主观性文本意见挖掘研究	北京理工大学
	面向 LAMOST 天文光谱特征线的数据挖掘方法研究	太原科技大学
	面向 GML 的时空关联规则及序列模式挖掘研究	江西理工大学
	密态数据的隐私计算及在深度学习中的外包数据挖掘研究	上海海洋大学
	蒙汉双语网络挖掘层次关联分析方法研究	中国科学院合肥物质科学研究院
	流程工业生产系统非稳定工况下基于数据挖掘的节能调控方案生成理论与方法	浙江大学
	领域知识驱动的深层知识发现研究	中国科学院大学
	领域驱动空间 co-location 模式挖掘技术研究	云南大学
	利用知识图谱进行生命组学数据知识发现的关键技术研究	中国人民解放军军事医学科学院
	利用数据挖掘技术探析微针系统疗法的优势病种及用穴规律	河北中医学院
	利用计算机数据挖掘技术研究分子筛材料的结构化学	吉林大学
	利用大数据信息挖掘和基因进化方法研究禽流感病毒的跨地域传播	中国科学院微生物研究所
	利用大数据分析与挖掘技术探析五输穴主治优势病症和配伍规律	山东中医药大学
	跨模态课堂大数据分析挖掘研究	西安交通大学
	空间数据挖掘的数据各向异性研究	武汉大学
	空间关联规则挖掘尺度优化的模型和算法	武汉大学
	空间大数据挖掘的关键技术研究	北京理工大学
	聚类导向的字典学习及基于稀疏表示的高维数据聚类研究	北京科技大学
	具有迟滞特征对象的多空间关联和多值映射数据挖掘与聚类分析	上海师范大学
	局部空间关联规则挖掘模型与算法	武汉大学
	结合样本优化和核学习子空间的多源异质遥感数据分类及城市应用	上海师范大学
	健康大数据的建立及其在知识发现和个体化移动健康管理的研究	北京大学
	监控视频数据中知识发现方法研究	江苏理工学院

	间歇过程数据扩散映射本征维度空间的 k 近邻故障诊断研究	沈阳化工大学
	价值模式挖掘及应用研究	北京航空航天大学
	几何演变驱动的高维机加工艺知识发现与重用研究	西北工业大学
	集群环境下的天体光谱离群数据挖掘与性能优化	太原科技大学
	集成建模和数据挖掘方法的乳腺癌网络随机动力学研究	复旦大学
	基于众源轨迹大数据的行为模式挖掘与定量空间优化	武汉大学
	基于知识指导和模糊信息粒化的时序大数据分析和挖掘	北京师范大学
	基于知识图谱的农业大数据碎片化知识发现方法研究	安徽农业大学
	基于知识发现和数据挖掘的建筑实践谱系认知与演变机制研究	同济大学
	基于在线健康社区的病患知识发现和个性化诊疗推荐方法研究	广东工业大学
	基于约束的高维数据聚类	大连理工大学
	基于语义距离的分布式数据挖掘理论与方法	河北科技大学
	基于用户搜索意图理解的在线社交网络跨媒体精准搜索与挖掘研究	北京邮电大学
	基于异构二部图模型的多源大规模数据聚类集成算法研究	华南农业大学
	基于移动端虚拟病人的护理学专业本科生问诊能力培养模式的构建及多视角数据挖掘研究	中山大学
	基于医案与机器学习信息融合的名中医诊疗糖尿病知识发现方法研究	广州中医药大学
	基于药物大数据的潜在不良药物反应挖掘研究	武汉大学
	基于信息系统同态理论的混合数据挖掘理论与方法研究	渤海大学
	基于相似度学习的异构数据聚类算法研究及其应用	中山大学
	基于相关性的大数据分类理论与方法研究	华北电力大学
	基于网树的无重叠多维对比序列模式挖掘及其在序列分类中的应用	河北工业大学
	基于网上用户行为的 Web 使用挖掘模式发现关键技术研究	哈尔滨工业大学
	基于图学习的学术合作网络挖掘技术	大连理工大学
	基于图模型的多源异构在线产品评论数据融合与知识发现研究	吉林大学
	基于数学属性偏序表示原理的《伤寒论》方证群结构知识发现方法研究	广州中医药大学

	基于数据挖掘与运气学说的中医肺系疫病预测与辨治规律研究	南京中医药大学
	基于数据挖掘近似函数的智能制造装备生产决策参数稳健优化方法研究	重庆科技学院
	基于数据挖掘技术的一次性医用材料费用控制研究	天津医科大学
	基于数据挖掘技术的信号通路识别模型与算法研究	广西大学
	基于数据挖掘技术的肺癌危险度评价与早期预警系统研究	郑州大学
	基于数据挖掘技术的方剂数据库构建及筛选代表方的药效学验证研究	黑龙江大学
	基于数据挖掘技术的蛋白质标志物与基因突变联合检测的分子诊断系统构建及应用	郑州大学
	基于数据挖掘和分布式智能体的产品平台逆向建模与变异性分析	武汉大学
	基于数据挖掘构建症状间及症状与证候间关联模式的研究	河南中医药大学
	基于数据挖掘方法的软件安全特性建模与分析	燕山大学
	基于数据挖掘方法的区域 CO2 通量时空分布格局及不确定性研究-以青藏高原草地样带为例	中国科学院地理科学与资源研究所
	基于数据挖掘的我国大学生高质量就业研究：评价体系、影响因素和实现路径	中国政法大学
	基于数据挖掘的煤矿灾害预测研究	辽宁工程技术大学
	基于数据挖掘的煤矿安全可视化管理模型及图元体系研究	中国矿业大学（北京）
	基于数据挖掘的煤矿安全风险评价体系研究	辽宁工程技术大学
	基于数据挖掘的货币市场与资本市场联接途径研究	山东科技大学
	基于数据挖掘的行业景气预测预警研究	华中师范大学
	基于数据挖掘的地下流体及热红外异常信息相关性分析	太原科技大学
	基于数据挖掘的大型燃煤发电机组节能诊断理论与方法研究	华北电力大学
	基于数据挖掘的刺灸法效应特异性基本规律与特点的研究- - -毫针法、灸法及刺络放血法部分	河北中医学院
	基于数据挖掘的刺灸法效应特异性基本规律及特点的研究	河北医科大学
	基于数据挖掘的城镇存量建设用地信息获取与动态监管研究	浙江大学
	基于数据挖掘、风险对冲与经济激励的配电网规划方法研究	清华大学

	基于数据驱动知识发现的智能故障诊断方法与专家系统关键技术	燕山大学
	基于数据分布评估和支持向量机方法的分布式数据流挖掘模型和算法研究	中央财经大学
	基于输出纠错编码的开集多类数据挖掘算法研究	厦门大学
	基于时态交通网络的移动对象时空统计分析、数据挖掘及交通敏感导航技术	中国科学院软件研究所
	基于生物医学文献的隐含知识发现方法研究	大连理工大学
	基于深度学习和迁移学习的非结构化临床文本挖掘的方法探索	北京大学
	基于深度学习的文本和语音多模态数据挖掘研究	内蒙古民族大学
	基于深度学习的城市场景点云数据分类识别与建模	南京航空航天大学
	基于深度学习的槟榔成瘾多中心多模态数据分类研究	中南大学
	基于深度神经网络的社会媒体用户情感及兴趣挖掘方法研究	中国科学院自动化研究所
	基于三维立体数据库和多维数据挖掘的中药注射剂致过敏反应关键影响因素研究	北京中医药大学
	基于认知计算的大数据挖掘理论与技术	广西师范学院
	基于模糊信息粒化和知识指导的大规模区间值属性图序列的数据挖掘	北京师范大学
	基于旅游攻略数据挖掘的城市内部游客流动特征与形成机制研究——以沪宁杭为例	南京师范大学
	基于粒计算的多源异构动态数据挖掘关键技术研究	西南交通大学
	基于粒计算的多模态多标记数据分类建模研究	闽南师范大学
	基于空间数据挖掘的点-轴开发与生态环境协调发展研究-以哈大齐工业走廊为例	哈尔滨师范大学
	基于可能世界的不确定数据聚类	大连理工大学
	基于进化模糊机制的 Web 新闻挖掘关键技术	河南科技大学
	基于结构化大数据深度挖掘的呼吸道症候群监测与早期预警机制研究	重庆大学
	基于间隙约束的序列模式挖掘关键技术及其在特征提取中的应用	河北工业大学
	基于基因组和化学大数据的植物内生真菌“隐形天然产物”挖掘	吉林大学
	基于机器学习的网络结构知识发现和隐匿研究	浙江工业大学
	基于核学习的大型复杂数据挖掘理论与方法研究	哈尔滨工业大学
	基于海量、多域、高维数据挖掘的中医疗效分析方法研究	中国中医科学院西苑医院
	基于轨迹数据的用户意图挖掘关键技术研究	山东大学
	基于构造型最优结构前馈神经网络的时间序列数据挖掘研究	中南大学

	基于公理模糊集和粒计算的多元动态数据的知识发现与语义表示	大连理工大学
	基于高通量数据挖掘揭示染色质调控因子新的作用机制	同济大学
	基于高通量测序对骆驼抗体库的多样性分析和数据挖掘及新纳米抗体的发现	新疆大学
	基于高斯过程动态系统的多输出时序数据分类与回归	华东师范大学
	基于非显式隐私保护的大规模高维数据聚类方法研究	哈尔滨工业大学
	基于非参数概率混合模型的方向数据聚类算法研究	华侨大学
	基于多智能体一致性理论的分布式数据聚类算法	中国科学技术大学
	基于多源异构大数据的风电机组本征挖掘与状态异常辨识研究	江南大学
	基于多源信息融合与数据挖掘的土压平衡盾构机-岩状态识别研究	鲁东大学
	基于多模态网络数据挖掘的景区游客流量预测与预警研究	北京联合大学
	基于多模态关联图模型的医学媒体数据挖掘关键技术研究	哈尔滨工程大学
	基于多核表示和模糊近似的混合数据分类方法研究	北京建筑大学
	基于多关系的模糊认知图挖掘模型、算法与评价机制研究	北京科技大学
	基于动态数据挖掘的物流信息智能分析研究	武汉大学
	基于动态决策和数据挖掘的集装箱翻倒问题研究	清华大学
	基于电子病历数据挖掘的急性冠脉综合征风险评估及干预效果分析方法研究	浙江大学
	基于大数据挖掘与深度学习的中国书法仿写与智能创作的算法研究	西安交通大学
	基于大数据挖掘的颞下颌关节动态匹配参数模型的建立及致病机制研究	中山大学
	基于大数据挖掘的煤矿安全管理决策模型及仿真研究	中国矿业大学
	基于大数据的公共价值挖掘与公共决策支持研究	华中科技大学
	基于粗集理论的多准则决策分析及其知识发现过程研究	南开大学
	基于超网络的大众协同创新社区用户知识主体挖掘方法研究	华南理工大学
	基于不确定图模型的医学图像数据挖掘关键技术的研究	哈尔滨工程大学
	基于贝叶斯本体的 Web 知识发现研究	北京航空航天大学

	基于 Web 知识挖掘与融合的命名实体消歧技术研究	中国科学院自动化研究所
	基于 Web2.0 的社会网络媒体信息的组织与挖掘方法研究	中国科学院自动化研究所
	基于 SCADA 数据挖掘的风电机组状态在线识别与预警	湖南科技大学
	基于 QFD 和数据挖掘的卷烟产品叶组配方优化关键技术研究	东北大学
	基于 n 阶引力场理论的数据分类研究	厦门理工学院
	基于 DEM 的洞庭湖流域河网特征分析与数据挖掘问题研究	中南大学
	基于 CER 模式的中医真实世界数据挖掘关键技术研究	广州中医药大学
	基于 Agent 的跨媒体数据挖掘和旅游信息导航研究	北京邮电大学
	会计政策选择策略系数：基于数据挖掘的计量改进和因素分析	郑州航空工业管理学院
	海量高维天体光谱数据挖掘及其并行化研究	太原科技大学
	海量不确定图挖掘算法研究	哈尔滨工业大学
	海岸带空间利用模式挖掘与时空分异分析	中国科学院地理科学与资源研究所
	关于 ELMs 技术的理论评价及其在数据挖掘中的应用研究	西北大学
	高性能计算环境下社交媒体地理大数据热点挖掘与智能推荐	北京大学
	高维时间序列数据聚类分析及应用研究	华侨大学
	高维混合型数据聚类及应用研究	汕头大学
	高维混合数据异常知识发现的粒计算模型关键问题研究	哈尔滨工程大学
	高复杂性脑电数据的数据挖掘及其在癫痫性发作自动检测中的应用	西北大学
	高度可扩展的数据仓库数据编码方法及查询处理新技术研究	中国人民大学
	复杂信息系统的多粒度知识发现与不确定性分析	浙江海洋大学
	复杂网络的局域同步及其在数据聚类和网络路由中的应用	中国科学技术大学
	复杂生物网络集的频繁模式挖掘算法研究	中国人民解放军空军工程大学
	复杂社会网络跨组织业务过程挖掘及其动态优化模型研究	上海第二工业大学
	复杂空间对象的轨迹建模与模式挖掘	中国科学院地理科学与资源研究所

	复杂疾病全基因组 SNP 高阶互作模式的网络挖掘方法研究	曲阜师范大学
	复杂场景下监控视频目标的运动模式挖掘方法研究	中国人民解放军国防科学技术大学
	负序列模式挖掘关键技术及其在医保欺诈检测中的应用研究	齐鲁工业大学
	分布式计算环境下的并行数据挖掘算法与理论研究	中国科学院计算技术研究所
	非均衡概念漂移网络舆情大数据流挖掘模型、算法与评价机制研究	国家行政学院
	方剂药物组配规律的有向图挖掘技术研究	闽南师范大学
	多源直觉模糊数据集知识发现的粒计算方法研究	重庆理工大学
	多源数据挖掘的关键技术研究	浙江师范大学
	多态异构机器学习及其在大数据挖掘中的应用	华南理工大学
	多粒度框架下带偏好直觉模糊数据集的信息融合与知识发现研究	重庆理工大学
	多尺度数据的粒计算与知识发现研究	浙江海洋大学
	多尺度概念格的构造与知识发现方法研究	中国石油大学（华东）
	多靶标生物标志物多模式检测体系的构建及其联合数据挖掘技术的应用	郑州大学
	动态系统分段模型及其在视频运动模式挖掘中的应用	中国科学院自动化研究所
	动态数据挖掘中的演化聚类模型与算法研究	大连理工大学
	动态数据挖掘的构造性机器学习方法研究	中国人民解放军电子工程学院
	叠前数据挖掘与储层参数非线性预测	中国石油大学（华东）
	电子商务和电子服务的信任网络结构模式挖掘研究	浙江万里学院
	电子健康档案中的大数据非均衡挖掘研究	河北大学
	单细胞 RNA 测序数据聚类方法研究	复旦大学
	代谢组学数据挖掘方法研究	郑州大学
	大数据驱动的城市交通流机理研究和语义挖掘	复旦大学
	大数据驱动的产品用户需求挖掘模型及产品创新优化策略研究	同济大学
	大数据环境下基于视觉主题模型的视觉数据分类方法研究	湖北工程学院
	大数据环境下高维数据流挖掘算法及应用研究	大连理工大学
	大规模时空数据集的高效知识发现核心问题研究	厦门理工学院
	大规模多源异质蛋白质数据挖掘中的若干关键问题研究	南京理工大学

	大规模 SNP 数据挖掘及其在复杂疾病分析中的应用研究	湖南大学
	超高层建筑动态位移全天候摄像测量及数据挖掘研究	温州大学
	部分相关的多任务数据聚类	大连理工大学
	P2P 网贷中投资者决策的变量优先级——基于前景理论的数据挖掘	西南财经大学
	O2O 商务模式下多源异构大数据的挖掘、融合与应用研究	中央财经大学
联合基金项目	面向三旧改造的多源异构大数据管理分析与挖掘研究	华南农业大学
	漏洞相关数据集中的知识发现及在漏洞检测中的应用	中国人民大学
	基于移动大数据的特异群组挖掘与行为预测	复旦大学
	基于数据挖掘的煤运重载铁路列车运行参数动态偏移规律及调度优化决策支持研究	北京交通大学
	基于潜在出行主题模型的民航旅客大数据挖掘与分析	南开大学
	基于机器学习的生物医学大数据挖掘理论与方法研究	中山大学
	基于短文本大数据的关联分析和挖掘	南开大学
	基于 POLAR 的太阳高能物理相关数据挖掘、处理和分析	中国科学院紫金山天文台
	LAMOST 低质量光谱的分析处理与数据挖掘	山东大学
海外及港澳台学者合作研究基金	医疗保健社会化媒体上的参与者行为分析及健康知识发现方法研究	哈尔滨工业大学
国家杰出青年科学基金	知识发现与知识工程	清华大学
	数据挖掘与多目标决策（研究领域）	电子科技大学
	时空数据挖掘	中国科学院地理科学与资源研究所
	社会化媒体大数据挖掘与应用	北京航空航天大学
	情境数据挖掘及应用	中国科学技术大学
	多媒体信息处理与分析	西安电子科技大学
	Web 信息检索与数据挖掘	中国科学院计算技术研究所
	最优化数据挖掘的商业智能方法以及在金融与银行管理中的应用	中国科学院大学

国际(地区)合作与交流项目	面向天文光谱的数据挖掘算法性能分析与并行化研究	太原科技大学
	面向海上移动作业的海洋大数据挖掘与仿真重构	中国海洋大学
	基于数据挖掘的软件管理国际大会	电子科技大学
	基于从头设计、分子模拟及数据挖掘的高活性头孢菌素酰化酶的设计研究	清华大学
	多模态迁移学习及其在网络数据挖掘中的应用	上海交通大学
	第一届空间数据挖掘与地理知识服务国际会议暨第八届北京地理信息系统国际研讨会	福州大学
地区科学基金项目	整合基因组和转录组数据挖掘影响绵羊尾脂重的关键基因	甘肃农业大学
	云平台上基于海量医学图像并行数据挖掘的计算机辅助诊断技术研究	贵州大学
	云南跨境民族网络舆情信息挖掘关键技术研究	云南民族大学
	应用数据挖掘技术研究水环境中化学污染物的解析和辨识	内蒙古大学
	新疆维、哈、汉族高血压流行病和中医证候调查及与 BMI 关系研究——基于数据挖掘和 HLM 模型构建与分析	石河子大学
	维吾尔文 WEB 舆情挖掘的关键理论及技术研究	新疆大学
	社会化学习网络社区中基于数据挖掘视角的用户行为研究	昆明理工大学
	融合关联规则数据挖掘和基于案例推理的经方“方证相应”证治系统构建研究	广西中医药大学
	南海台风客观预报的流形学习数据挖掘和 Boosting 集成预报方法	广西壮族自治区气象减灾研究所
	面向智能农业的山地果园监测与数据挖掘技术研究及其应用	华东交通大学
	面向知识发现的协同进化算法及在镍基高温合金组织结构超声评价中的应用	南昌航空大学
	面向文本挖掘的特征选择关键问题研究	内蒙古民族大学
	面向复杂数据的粒计算知识发现方法研究	南昌工程学院
	面向标记分布数据的粒计算模型与知识发现研究	江西农业大学
	蒙医方剂数据挖掘关键技术研究	内蒙古民族大学
	结合外部资源的地方志文本挖掘模型研究	江西师范大学
	极限学习机不平衡数据分类研究	宁夏大学
	基于质性数据挖掘的农产品市场供给-价格交互下多重稳定机制设计与动态模拟	云南财经大学
	基于知识发现与动力统计集成的暴雨短期预报建模研究	广西师范学院
	基于位置大数据的城市热点区域和居民出行模式的挖掘研究	云南大学

	基于数据挖掘技术和 Delphi 法的小儿肺炎郁热辨证研究	广西中医药大学
	基于数据挖掘技术的抑制调控模式发现与算法研究	广西大学
	基于数据挖掘技术的 RA 方证构效规律及与脏腑相关性研究	新疆医科大学
	基于数据挖掘和复杂网络的 UML 类图复杂性度量研究	江西财经大学
	基于数据挖掘的跨区域网络情报智能分析研究---以东盟十国为例	广西大学
	基于苗医药特色理论与适应性数据挖掘技术的苗医方剂释理技术研究暨苗医特色方剂学理论框架构建	贵阳中医学院
	基于结构模型的 miRNA 协同作用模式数据挖掘研究	广西大学
	基于节点相似度的大规模融合情感社交网络表示、分析与挖掘	华东交通大学
	基于健康体检大数据挖掘下的新疆不同民族城镇居民健康管理模式研究——以代谢综合征为例	新疆医科大学
	基于复杂网络的商务大数据聚类与关联应用研究	兰州交通大学
	基于分子网络数据挖掘的高血压与 2 型糖尿病共享遗传机制的研究	广西医科大学
	基于多源异构大数据挖掘融合的高原公路交通风险多因素时空耦合识别及其控制技术研究——以云南高原地区为例	昆明理工大学
	基于大规模短文本的自动知识发现关键技术研究	延安大学
	基于大尺度网络的轻度创伤性脑损伤(mTBI)功能磁共振成像数据挖掘	南昌大学
	基于策略模式的中药性效数据挖掘方法研究	江西中医药大学
	基于不确定数据挖掘的滑坡区域地质灾害危险性评价方法研究	江西理工大学
	基于 GEP 和亚复杂系统的跨媒体时空数据挖掘关键技术研究	广西师范学院
	高维数据分析中的支持向量机核学习及其应用研究	北方民族大学
	复杂系统多粒度故障知识发现及维修决策机制研究	内蒙古大学
	大宗农产品价格内涵属性之效应分解研究--基于自变量扰动循环的数据挖掘集成技术	江西财经大学
	大规模数据挖掘中嵌入式数据归约的稀疏模型与算法研究	海南大学
	大规模数据聚类的并行进化算法骨架研究	江西师范大学
	大规模动态社交网络的骨架挖掘及其应用研究	江西师范大学
	不确定数据的空间 co-location 模式挖掘技术研究	云南大学
创新研究群体项目	数据挖掘与智能知识管理:理论与应用研究	中国科学院大学

版权声明