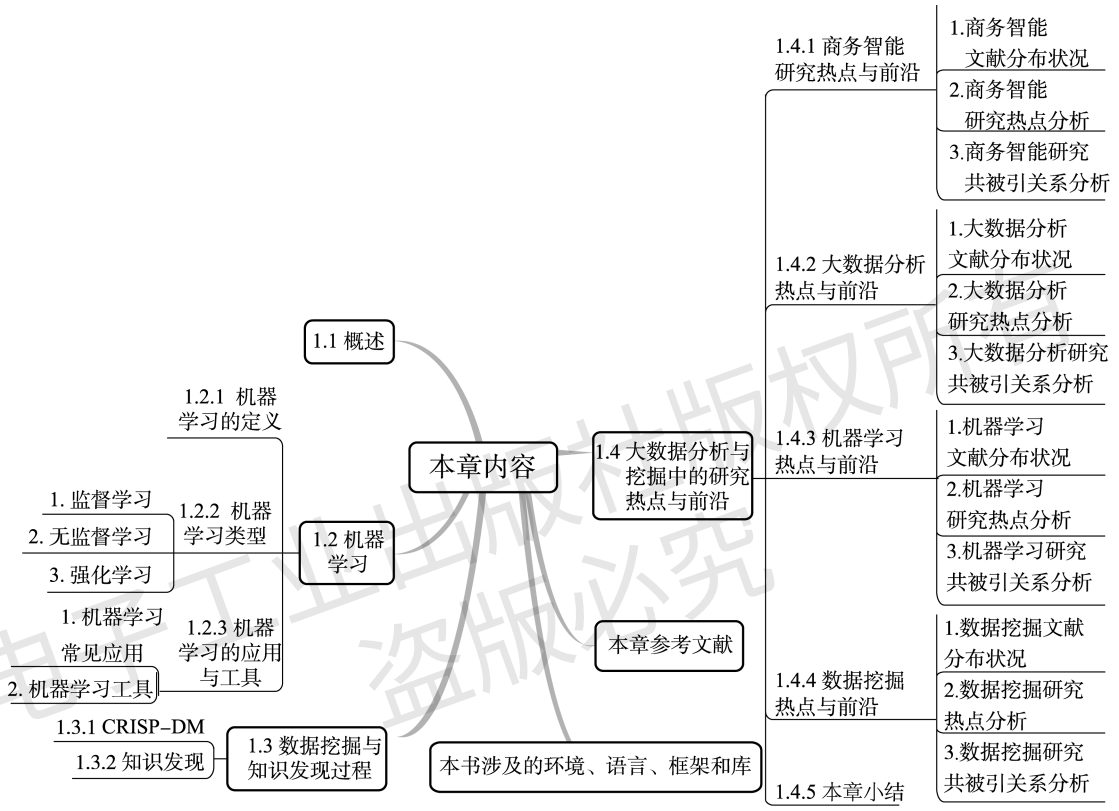
第1篇

绪　论

第 1 章　大数据分析与挖掘的概念与理论



“数据挖掘是从数据中提取隐含的、以前未知的和潜在有用的信息。这个想法是建立自动筛选数据库的计算机程序，寻找规律或模式。如果发现强大的模式，可能会推广到对未来数据做出准确预测。…… 机器学习为数据挖掘提供了技术基础。它用于从数据库中的原始数据中提取信息……”。“数据挖掘被定义为发现数据模式的过程。该过程必须是自动的或(更常见的）半自动的。发现的模式必须是有意义的，因为它们会带来一些优势，通常是经济优势。数据总是大量存在。”

——《数据挖掘：实用的机器学习工具和技术》

“数据挖掘，通常也称为从数据中发现知识（KDD），是自动或方便地提取表示隐式存储或捕获在大型数据库、数据仓库、Web、其他海量信息存储库或数据流中的知识的模式。”

——《数据挖掘：概念和技术》

针对“大数据”（Big Data），研究机构 Gartner 给出的定义：“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力来适应海量、高增长率和多样化的信息资产。麦肯锡全球研究所给出的定义：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、

多样的数据类型和价值密度低四大特征。当前大数据已经渗透到了所有的学科和研究领域（包括计算机科学、医学和金融等），因为它在所有这些领域都具有潜力。数据生成和数据收集的变化也导致了数据处理的变化。

1.1 概述

大数据分析与挖掘是知识发现过程的核心阶段，旨在从数据中提取有趣和潜在的有用信息。数据挖掘可以作为人工智能和机器学习的基础。这个方向的许多技术可以归入人工智能（AI）、机器学习（ML）和深度学习（DL）等领域。

人工智能（Artificial Intelligence，AI）是一个广泛而复杂的概念，通常用于描述一个模仿人脑认知功能的概念或系统，它可以从经验中学习，可以通过使用知识来执行任务、推理和做出决策。它包括机器学习、自然语言处理、语言合成、计算机视觉、机器人学、传感器分析、优化和模拟。人工智能的类型有很多，如专家系统、神经网络和模糊逻辑等。

机器学习（Machine Learning，ML）是人工智能技术的一个子集，它使计算机系统能够基于过往经验（即数据观察）学习，并改善其在特定任务中的行为。ML 技术包括支持向量机、决策树、贝叶斯学习、K－means 聚类、关联规则学习、回归和神经网络等。

深度学习（Deep Learning，DL）是使用人工神经网络的机器学习的一个子集。人工神经网络是受人脑结构启发而设计出的计算模型。典型的 DL 架构是深度神经网络（DNNs）、卷积神经网络（CNNs）、循环神经网络（RNNs）和生成对抗网络（GAN）等。深度学习适用于执行复杂的任务，如对象识别、语音识别和翻译，是一种特别流行的机器学习类型。

人工智能、机器学习、深度学习的关系如图 1－1 所示。

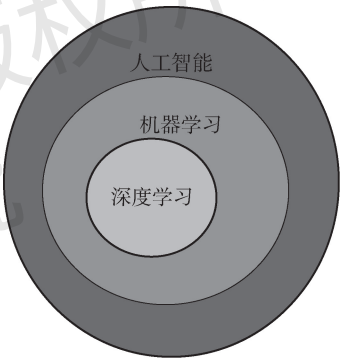


图 1－1　人工智能、机器学习、深度学习的关系

1.2 机器学习

1.2.1 机器学习的定义

机器学习是一门属于人工智能范畴的多领域交叉的学科（见图 1－2），涉及概率论、统计学、逼近论、凸分析及算法复杂度理论等多门学科。机器学习的主要研究对象是人工智能，它是人工智能的核心之一，主要研究计算机怎样模拟或实现人类的学习行为，特别是如何在经验学习中提高具体算法的性能，获取新的知识或技能，重新组织已有的知识结构。

机器学习的定义有如下两种：

（1）机器学习是对能通过经验自动改进的计算机算法的研究。

（2）机器学习是根据数据或以往的经验，优化计算机程序的性能标准的方法。

一种经常引用的英文定义：A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

如果一个程序在使用既有的经验（E）执行某类任务（T）的过程中被认定为是“具备

学习能力的”，那么它一定需要展现出：利用现有的经验（E），不断改善其完成既定任务（T）的性能（P）的特质。

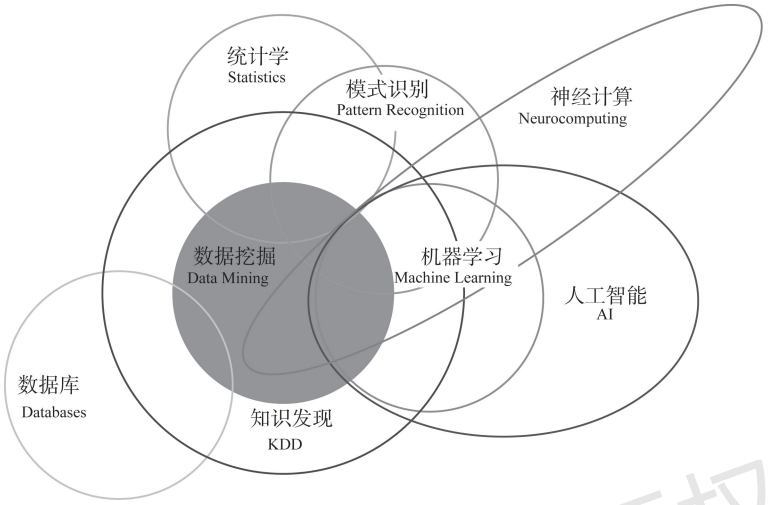


图 1－2　各领域之间的关系

（扫码看彩图）

三个关键术语：任务 T（Task）、经验 E（Experience）、性能 P（Performance）。

所谓学习就是针对经验 E、一系列的任务 T 和一定表现的性能 P，如果随着经验 E 的积累，针对定义好的任务 T 可以提高表现 P，那么就说计算机具有学习能力。

机器学习大量的应用都与大数据高度耦合，几乎可以认为大数据是机器学习应用的最佳场景，它常用于分析大型数据集并在其中找到规则和模式。

1.2.2　机器学习类型

机器学习类型由需要解决的问题定义，并且要分析目标的内在因素。

- 首先有一个要预测的目标、值或类，例如，想要根据不同的输入（星期几、广告、促销）预测商家的收入，模型将根据历史数据进行训练，并使用该训练结果来预测未来的收入。那么该模型是有监督的，因为它知道要学习什么。
- 如果有未标记的数据，并想要在这些数据中查找模式和组，例如，希望根据客户订购的产品类型、购买产品的频率、上次访问等要素进行聚类，无监督机器学习将自动区分不同的客户。
- 如果想达到一个目标，例如，想找到在指定规则下赢得某游戏的最佳策略，一旦指定了这些规则，强化学习技术将多次玩此游戏以找到最佳策略。

1. 监督学习

监督学习是最常见的机器学习类型之一。它用于发现数据中的模式，并根据过去的经验预测未来的行为。在监督学习中，数据被分成两部分，称为训练集和测试集。训练集用于训练模型，测试集用于评估模型的准确性。

监督学习任务的基本架构和流程如图 1－3 所示。首先，准备训练数据，可以是文本、图像和音频等；其次，抽取所需要的特征，形成特征向量；接着，把这些特征向量连同对应的标记/目标（Labels）送入机器学习算法中，训练出一个预测模型；然后，采用同样的特征抽取方法作用于新测试数据，得到用于测试的特征向量；最后，使用预测模型对这些测试

的特征向量进行预测并得到结果。

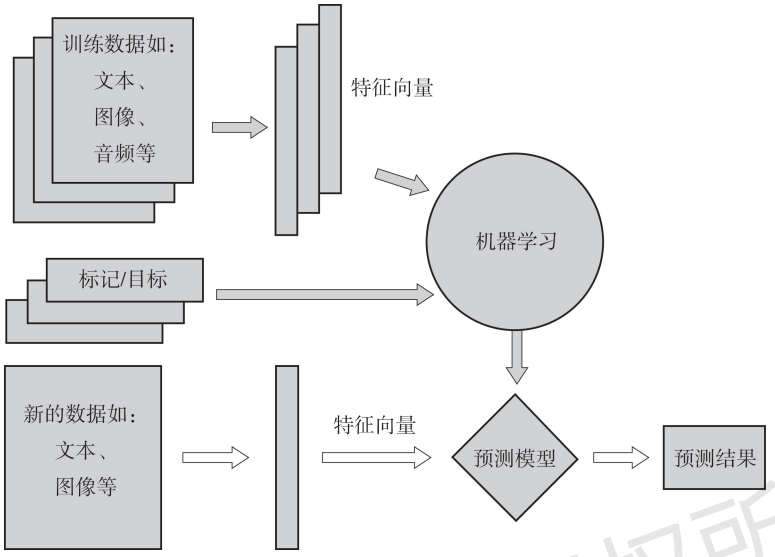


图 1－3　监督学习任务的基本架构和流程

监督学习的原则：第一，训练数据集包含输入数据（预测变量）和预测的值（可以是数字也可以不是）；第二，该模型将使用训练数据来学习输入和输出之间的联系。基本思想是训练数据可以泛化，并且模型可以以一定的准确性用于新数据。

常用的监督学习算法：线性和逻辑回归、支持向量机、朴素贝叶斯、神经网络、梯度提升、分类树和随机森林等。

监督学习通常用于图像识别、语音识别、预测和某些特定业务领域（目标、财务分析等）中的专家系统。

2. 无监督学习

无监督学习着重于发现数据本身的分布特点、数据中的结构。与监督学习不同，无监督学习不需要对数据进行标记。它还可用于查找数据中的组、集群或识别数据中的异常。

无监督学习算法可以分为三个不同的类别：

- 聚类算法，如 K－means、层次聚类或混合模型。这些算法试图区分和分离不同组中的观察结果。
- 降维算法（大多是无监督的），如 PCA、ICA 或自动编码器。这些算法以较少的维度找到数据的最佳表示。
- 异常检测，用来发现数据中的异常值，即不遵循数据集模式的记录。

无监督学习可用于查找相似客户群。

3. 强化学习

强化学习是机器学习的一种，如图 1－4 所示，它用于寻找可以最大化奖励的最佳行动或决策，也可用于寻找问题的最佳解决方案，最优解取决于奖励函数。

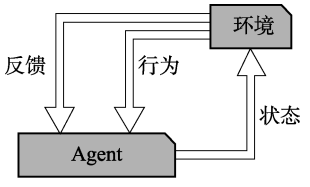


图 1－4　强化学习

强化学习可用于优化不同类型的问题。例如，它可用于优化非线性函数或查找网络中的最短路径。

强化学习是一种无须人工干预即可训练模型的方式，模型从环境中交互学习。当模型获得事件或对象时，它会尝试预测

我们想要的结果。如果结果正确，那么模型就会得到奖励；如果结果错误，那么模型就会受到惩罚。这样，模型就知道下次应该做什么。

1.2.3　机器学习的应用与工具

1. 机器学习常见应用

机器学习应用程序包括降维、自然语言处理、计算机视觉、异常检测、时间序列、分析和推荐系统等。机器学习应用程序如图 1－5 所示。

图 1－5　机器学习应用程序

降维（Dimensionality Reduction，DR）在保留最相关信息的同时减少数据维度。它用于图像和音频压缩及机器学习模型创建过程中的特征工程。

自然语言处理（Natural Language Processing，NLP）是一个广泛的领域，与其他机器学习应用程序越来越不同，许多专家认为 NLP 是一门独立的学科。ML 在 NLP 中的应用包括主题建模、文本分类、情感分析、机器翻译、自然语言生成、语音识别、文本到语音、文本分析、摘要、实体识别和关键词提取。

与 NLP 一样，计算机视觉（Computer Vision，CV）正在成为一个巨大的独立主题。最著名的 CV 应用是图像分类、图像分割和对象检测。

异常检测（Anomaly Detection）属于一种应用程序，其目的是识别数据中意外的、非典型的东西，对不匹配预期模式或数据集中其他项目的项目、事件或观测值的识别。异常检测分为新颖性检测、异常值检测和欺诈检测等。异常检测应用包括银行欺诈、结构缺陷、医疗问题和文本错误等类型的问题。异常也被称为离群值、新奇、噪声、偏差和例外。

时间序列（Time Series）是将同一统计指标的数值按其发生的时间先后顺序排列而成的数列，如证券交易所价格、天气数据及物联网传感器数据等。我们可以根据已有的历史数据

对未来进行预测，可以分析时间序列来预测可能的未来值。

分析（Analysis）是探索数据性质和模式的经典领域。它包括预测分析（预测未来或未见数据可能发生的事情）、当前状态分析（我们可以从当前数据中获得哪些见解，而无须构建预测模型）和优化问题（如探索如何从以消耗最少资源的方式从 A 点到 B 点）。

最后，推荐系统（Recommender System），解决信息过载问题，能够根据用户的兴趣和爱好将相关内容推荐给用户。此类系统包含各种 ML 推荐技术，利用用户和内容项的已知数据，实现个性化服务。

2. 机器学习工具

Python 是一种方便调试的解释型语言，能进行跨平台作业，拥有广泛的应用编程接口。在软件工程中有一个非常重要的概念，便是代码与程序的重用性。为了构建功能强大的机器学习系统，如果没有特殊的开发需求，通常情况下我们都不会从零开始编程。Python 自身免费开源的特性使得大量专业的编程人员，参与到 Python 第三方开源工具包（程序库）的构建中，并且大多数的工具包（程序库）都允许个人免费使用或商用，如很多用于机器学习的第三方程序库，便于向量、矩阵和复杂科学计算的 NumPy 与 SciPy；各种样式绘图的 Matplotlib；包含大量经典机器学习模型的 Scikit – learn；对数据进行快捷分析和处理的 Pandas；集成了上述所有第三方程序库的综合实践平台 Anaconda。

Python 有很多用于开发具有不同功能和优势的机器学习工具，如以下三种。

- 机器学习框架：Scikit – learn、PyTorch、TensorFlow、Keras。
- 辅助框架：Pandas、Numpy、Matplotlib、Seaborn、OpenCV。
- 编程语言：Python、R、C + + 。

Nguyen（2019）等对机器学习工具的总结如图 1 – 6 所示。

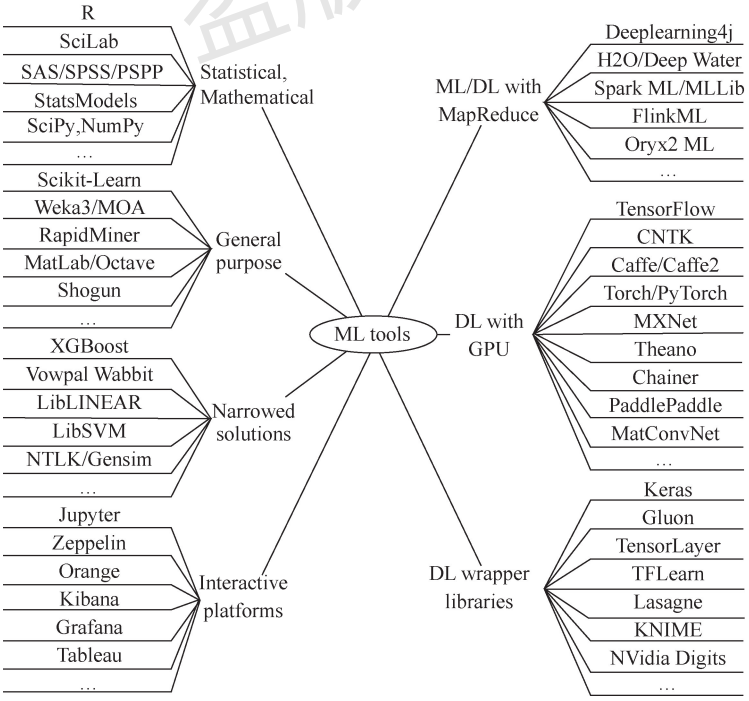


图 1 – 6　机器学习工具

1.3　数据挖掘与知识发现过程

1.3.1　CRISP－DM

数据挖掘多生活领域的实现导致了数据挖掘周期的跨行业标准流程（CRISP－DM 1999），它现在是数据挖掘的主要事实标准。CRISP－DM 周期（见图 1－7）由六个阶段组成：

（1）业务理解通常基于所提供的探求公式和数据描述。

（2）数据理解基于所提供的数据及其文档。

（3）数据准备包括数据转换、探索性数据分析（EDA）和特征工程。每个步骤都可以进一步划分为更小的子步骤。例如，特征工程包括特征提取、特征选择。

（4）建立模型阶段，各种 ML 算法都可以应用不同的参数校准。数据和参数可变性之间的结合会导致模型训练－测试－评估周期的大量重复。如果数据是大规模的，那么建立模型阶段将有耗时和计算密集的要求。

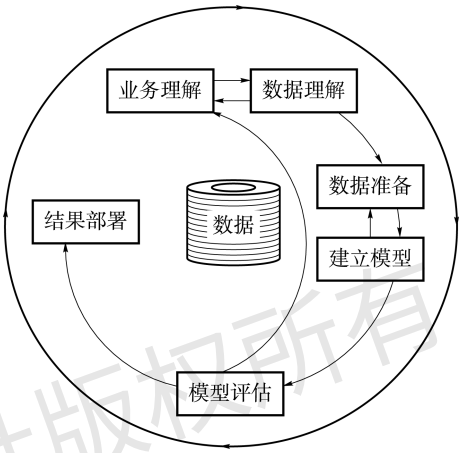


图 1－7　CRISP－DM 周期

（5）模型评估阶段，可以在各种标准下进行，对 ML 模型进行彻底测试，以便为结果部署阶段选择最佳模型。

（6）结果部署阶段，也称为生产阶段，包括使用经过训练的 ML 模型来利用其功能，创建一个数据管道进入生产。

埃森哲（Accenture）对 CRISP－DM 分析框架的总结如图 1－8 所示。

整个 CRISP－DM 周期是重复的。前五个阶段中的一组，称为开发阶段，可以根据评估结果以不同的设置重复进行。结果部署阶段对重复性要求下的实际生产至关重要，它意味着在线评估、监测、模型维护、诊断和再培训。需要强调的是，由于机器学习需要从数据中学习，因此，在实践中预计数据理解和数据准备阶段会消耗每个数据挖掘的大部分时间。

1.3.2　知识发现

知识发现（Knowledge Discovery in Database，KDD），是“数据挖掘”的一种更广义的说法，即从各种媒体表示的信息中，根据不同的需求获得知识。《数据挖掘：概念和技术》一书第 1 章中，作者总结了知识发现 KDD 过程：

（1）数据清洗，去除噪音和不一致的数据；

（2）数据集成，可以组合多个数据源；

（3）数据选择，从数据库中检索与分析任务相关的数据；

（4）数据转换，通过执行汇总或聚合操作将数据转换并合并为适合挖掘的形式；

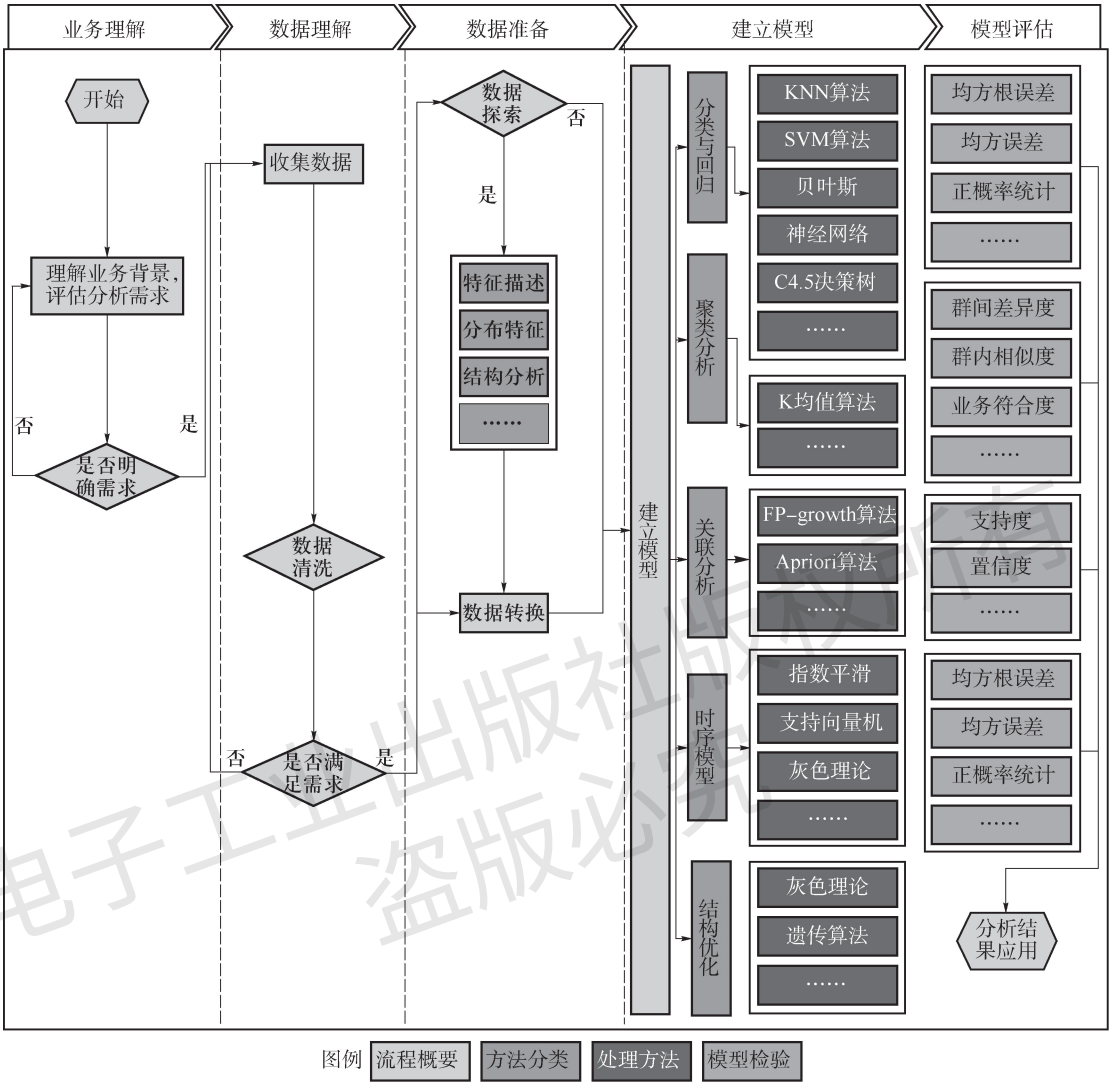


图 1－8　CRISP－DM 分析框架

（5）数据挖掘，这是应用智能方法提取数据模式的基本过程；

（6）模式评估，以基于有趣的度量来识别代表知识的真正有趣的模式；

（7）知识呈现，使用可视化和知识表示技术将挖掘出的知识呈现给用户。

从数据挖掘到数据库中的知识发现，Usama Fayyad、Gregory Piatetsky－Shapiro 和 Padhraic Smyth 于 1996 年在《人工智能》杂志上发表的一篇文章中将 KDD 定义为数据库中的知识发现，“…… KDD 领域关注的是用于理解数据的方法和技术的开发。……该过程的核心是将特定的数据挖掘方法应用于模式发现和提取。”和“…… KDD 是指从数据中发现有用知识的整个过程，而数据挖掘是指这个过程中的特定步骤。数据挖掘是应用特定算法从数据中提取模式。”描述总结如下。

第 1 步：选择（数据转化为目标数据）；

第 2 步：预处理（目标数据转化为处理后的数据）；

第 3 步：转换（将处理后的数据转换或统一成适合挖掘的形式）；

第 4 步：数据挖掘（将数据转换为模式）；

第 5 步：模式评估（根据某种兴趣度度量，识别表示知识的真正有趣模式）；

第 6 步：知识展现（将挖掘学习出来的知识展示出来）。

这个过程很简单，是处理问题时常用的模型。如果对 KDD 过程进行更详细的展开，描述解释如下：

（1）了解应用领域和流程目标；

（2）创建目标数据集作为所有可用数据的子集；

（3）数据清理和预处理进行去除噪声、处理丢失的数据和异常值；

（4）数据缩减和投影，专注于与问题相关的特征；

（5）将过程目标与数据挖掘方法相匹配。确定模型的目的，如汇总或分类；

（6）选择数据挖掘算法以匹配模型的目的（来自第 5 步）；

（7）数据挖掘，即对数据运行算法；

（8）解释挖掘的模式，使用户可以理解它们，如通过总结和可视化；

（9）根据发现的知识采取行动，如报告或做出决定。

数据驱动的问题解决方法如图 1－9 所示。

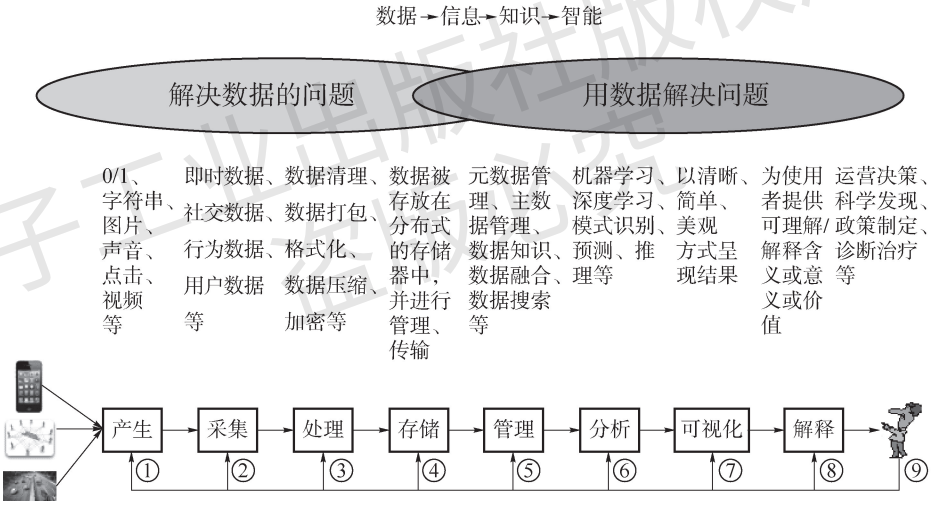


图 1－9　数据驱动的问题解决方法

1.4　大数据分析与挖掘中的研究热点与前沿

我们使用美国 Drexel University 教授陈超美基于 Java 开发的信息可视化研究工具——Citespace 软件，绘制知识图谱，发现大数据分析与挖掘领域的研究热点和前沿，预测前沿领域的发展趋势。Citespace 是一款引文可视化的分析软件，着眼于分析科学中蕴含的潜在知识，是在科学计量学、数据可视化背景下逐渐发展起来的，该软件的所用数据来自 Web of Science，时间范围为 2000 年至 2021 年 5 月。

WOS 中检索条件为：选择文献类型为“article”，主题词为“business intelligent”或“business intelligence”，经人工筛选后得到文献 1490 篇；主题词为“big data analysis”，经人工筛选后得到文献 1460 篇；主题词为“machine learning”，经人工筛选后得到文献 1121 篇；

主题词为“data mining”，经人工筛选后得到文献 1158 篇。

1.4.1 商务智能研究热点与前沿

1. 商务智能文献分布状况

（1）时间分布

图 1－10 所示的时间分布图展示了 WOS 中以“business intelligence”为主题，2008—2021 年按年份的分布情况。从图中可以看出，随着时间的变化，文献数量逐步增多，总体上呈递增的趋势。根据现状及发展趋势，预测在未来几年，相关研究也会不断增多。

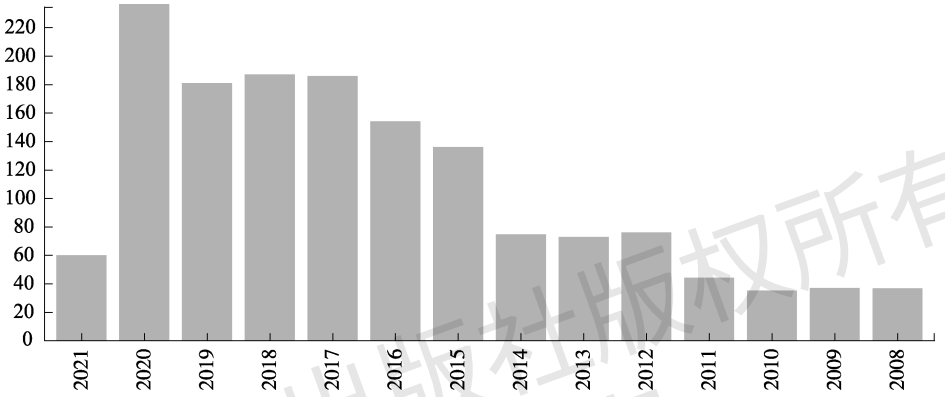


图 1－10 商务智能研究文献的时间分布图

（2）国家和地区分布

通过对 WOS 中商务智能文献分析（见图 1－11 和表 1－1）可知，截至 2021 年 5 月，美国在商务智能领域发文最多，占 23.221%；第二为我国大陆地区，发文占比为 11.208%；第三为澳大利亚，发文占比为 6.980%。由此可知，我国近年由于科学技术水平提升，带动了商务智能研究走在世界前列。

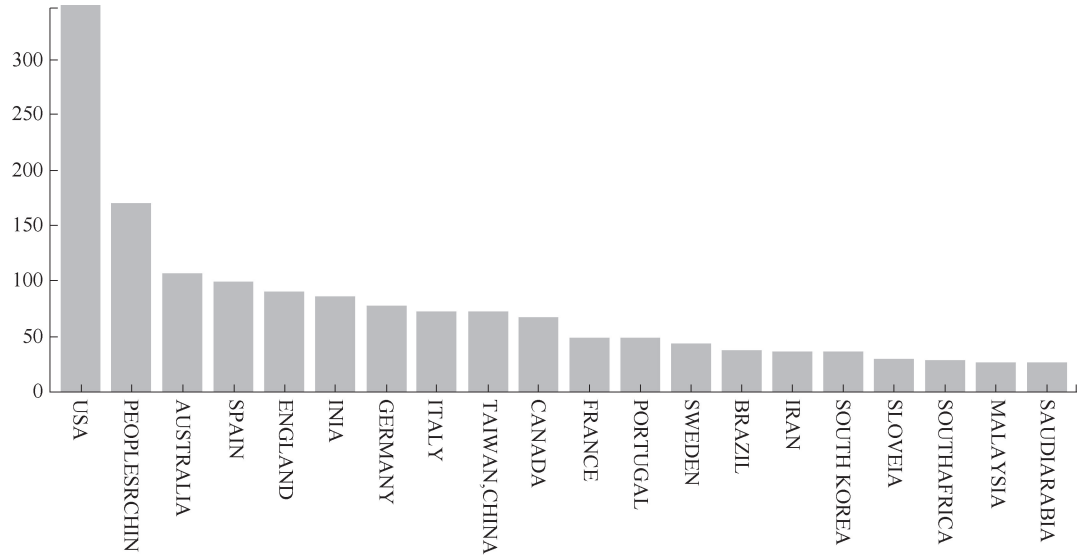


图 1－11 商务智能研究文献的国家和地区分布图

表 1 – 1　商务智能研究文献的国家和地区分布表

字段：国家/地区	记录数	%/1,490	柱状图
USA	346	23.221%	
PEOPLES R CHINA	167	11.208%	
AUSTRALIA	104	6.980%	
SPAIN	97	6.510%	
ENGLAND	88	5.906%	
INDIA	84	5.638%	
GERMANY	75	5.034%	
ITALY	70	4.698%	
TAIWAN, CHINA	69	4.631%	
CANADA	65	4.362%	

2. 商务智能研究热点分析

通过 Citespace 关键词共现分析，合并相同概念关键词，选择生成最小树 MST 剪枝策略，得到商务智能研究关键词共现统计表，如表 1 – 2 所示。结果显示，除检索条件“business intelligence”以外，最常出现的关键词体现了商务智能数据、信息、知识和智能的价值。

表 1 – 2　商务智能研究关键词共现统计表

Count	Centrality	Year	Keywords
766	0.05	2008	business intelligence
203	0.02	2013	big data
153	0.04	2008	management
140	0.14	2009	system
132	0.11	2009	model
114	0.01	2013	analytics
98	0.07	2012	performance
83	0.05	2009	impact
82	0.07	2008	data mining
81	0.06	2009	framework
80	0.07	2008	information
74	0.06	2009	technology
64	0.05	2008	data warehouse
61	0.03	2009	information system
57	0.04	2010	information technology

续表

Count	Centrality	Year	Keywords
56	0. 04	2008	knowledge
55	0. 07	2008	design
53	0. 02	2015	big data analytics
50	0. 11	2011	strategy

得到关键词共现可视化图，如图 1 – 12 所示，知识管理、技术、策略和质量等关键词均处于中心位置。将关键词聚类后（见图 1 – 13），研究热点较为集中，时间上变化差异不大。主成分分析、组织敏捷性、数据挖掘、机器学习和技术验收模型是商务智能的前沿领域。

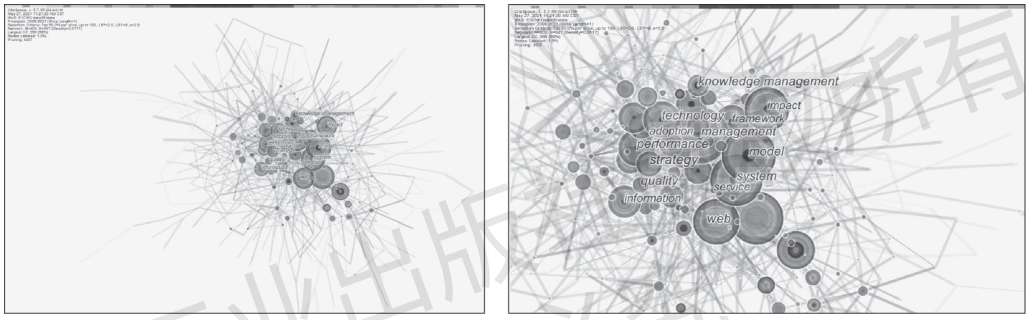


图 1 – 12　商务智能研究关键词共现可视化图

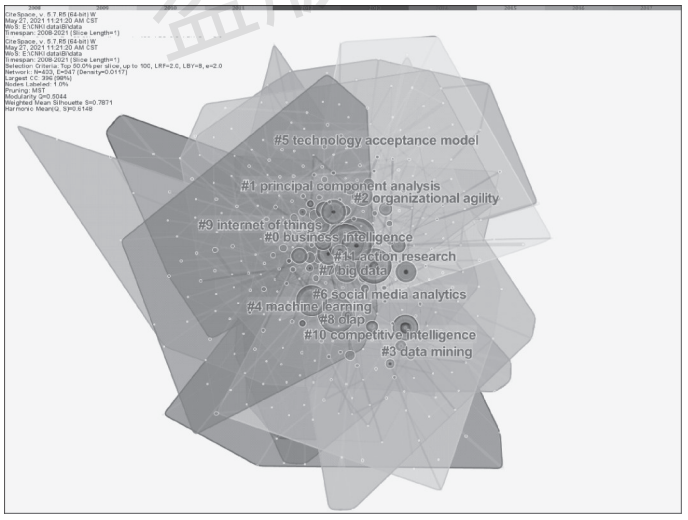


图 1 – 13　商务智能研究关键词聚类图

在前 20 关键词突现表（见表 1 – 3）中，我们可以看到具体研究热点随时间的演变。2008—2012 年研究主要集中于数据挖掘、数据仓库、知识管理、系统和框架等技术及优化方面。近三年，数据分析、机器学习成为新的研究热点，反映了当下研究已逐步由技术转向对数据的商业分析。

表 1 – 3　商务智能研究前 20 关键词突现表

Keywords	Year	Strength	Begin	End	2008—2021 年
data mining	2008	16. 83	2008	2013	
web	2008	7. 13	2008	2015	
business intelligence	2008	6. 5	2008	2011	
knowledge management	2008	5. 61	2008	2013	
data warehouse	2008	4. 13	2008	2013	
ontology	2008	3. 66	2008	2013	
Integration	2008	2. 96	2008	2011	
system	2008	9. 25	2009	2012	
framework	2008	4. 24	2009	2011	
management	2008	3. 9	2009	2011	
model	2008	3. 57	2010	2013	
business intelligence system	2008	3. 03	2010	2017	
optimization	2008	4. 01	2011	2016	
olap	2008	5. 59	2012	2013	
design science	2008	3. 75	2012	2015	
support	2008	3. 58	2012	2014	
acceptance	2008	3. 55	2012	2015	
data analytics	2008	7. 14	2019	2021	
machine learning	2008	5. 24	2019	2021	
organizational performance	2008	3. 28	2019	2021	

3. 商务智能研究共被引关系分析

选择网络节点为“reference”，阈值设置为（2，2，10）、（2，2，10）、（2，2，10），得到商务智能研究文献共被引网络图，如图 1 – 14 所示，得到其领域文献之间的被共引关系统计表，如表 1 – 4 所示。从图 1 – 14 及表 1 – 4 中可以看出，最突出的是 Chen HC 在 2012 年发表于 *MIS QUART* 的文章。从共被引关系突现表（见表 1 – 5）中可知近年在商务智能研究上具有突出贡献、具有转折意义的作者及文献，如 Hevner AR、Wamba SF 等。

表 1 – 4　商务智能研究文献共被引关系统计表

Count	Centrality	Year	Cited Refernce
225	0. 00	2012	Chen HC, 2012, MIS QUART, V36, P1165
70	0. 03	2012	Popovic A, 2012, DECIS SUPPORT SYST, V54, P729, DOI 10. 1016/ j. dss. 2012. 08. 017
70	0. 05	2012	McAfee A, 2012, HARVARD BUS REV, V90, P60
58	0. 04	2013	Isik O, 2013, INFORM MANAGE-AMSTER, V50, P13, DOI 10. 1016/j. im. 2012. 12. 001
56	0. 04	2015	Gandomi A, 2015, INT J INFORM MANAGE, V35, P137, DOI 10. 1016/ j. ijinfomgt. 2014. 10. 007

续表

Count	Centrality	Year	Cited Refernce
51	0.03	2011	Chaudhuri S, 2011, COMMUN ACM, V54, P88, DOI 10.1145/1978542.1978562
39	0.02	2010	Yeoh W, 2010, J COMPUT INFORM SYST, V50, P23
38	0.04	2011	Lavalle S, 2011, MIT SLOAN MANAGE REV, V52, P21
38	0.04	2015	Wamba SF, 2015, INT J PROD ECON, V165, P234, DOI 10.1016/j.ijpe.2014.12.031
35	0.00	2011	Manyika J, 2011, BIG DATA NEXT FRONTI, V0, P0

图 1-14　商务智能研究文献共被引网络图

表 1-5　商务智能研究文献共被引关系突现表（前 10 名）

References	Year	Strength	Begin	End	2008—2021 年
Hevner AR, 2004, MIS QUART, V28, P75	2004	10.17	2008	2012	
Davenport T H, 2007, COMPETING ANAL NEW S, V0, P0	2007	8.55	2009	02015	
Elbashir Mohamed Z, 2008, Inte …… Information Systems, v9, p135, DOI	2008	12.52	2011	2016	
Jourclan Z, 2008, INFORM SYSTMANAGE, V25, P121, DOI 10.1080/10580530801941512, DOI	2008	9.33	2012	2016	
Watson HJ, 2007, COMPUTER, V40, P96, DOI 10.1109/MC.2007.331, DOI	2007	6.7	2012	2015	
Bo Pang, 2008, Foundations and Trends in Information Retnieval, V2, P1, DOI 10.1561/1500000001, DOI	2008	5.92	2012	2016	
Yeoh W, 2010, J COMPUT INFORM SYST, V50, P23	2010	7.69	2014	2017	
Wixom Barbara, 2010, Internati …… Intelligence Research, V1, P13, DOI	2010	7.37	2014	2016	
Watson HJ, 2009, COMMUN ASSOC INF SYS, V25, P487	2009	6.18	2015	2017	
Wamba SF, 2017, J BUS RES, V70, P356, DOI 10.1016/j.jbusres.2016.08.009, DOI	2017	6.89	2019	2021	

1.4.2　大数据分析热点与前沿

1. 大数据分析文献分布状况

（1）时间分布

图 1－15 所示的时间分布图展示了 WOS 中以“big data analysis”为主题，2013—2021 年按年份的分布情况。从图中可以看出，随着时间的变化，文献数量逐步增多，呈递增的趋势。根据现状及发展趋势，预测在未来几年，相关研究也会不断增多。

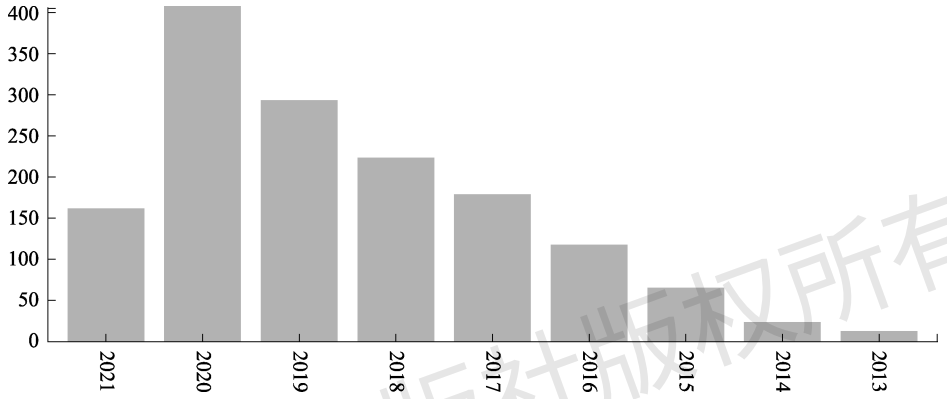


图 1－15　大数据分析研究文献的时间分布图

（2）国家和地区分布

通过对 WOS 中大数据分析文献分析（见图 1－16 和表 1－6）可知，截至 2021 年 5 月，我国大陆地区在大数据分析领域发文最多，占 38.288%；第二为美国，发文占比为 17.260%；第三为韩国，发文占比为 11.438%。国外研究虽早于国内，但国内文献数量始终高于国外，并且二者差距越来越大，由此可知近年来大数据分析领域应用研究热度的增长，国内明显高于国外。

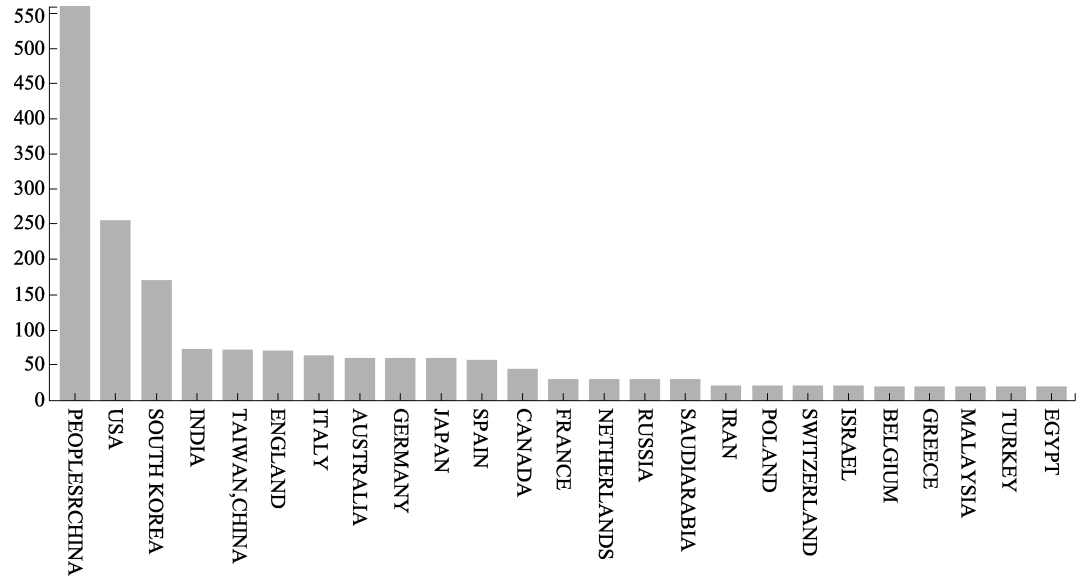


图 1－16　大数据分析研究文献的国家和地区分布图

表 1－6　大数据分析研究文献的国家和地区分布表

字段：国家/地区	记录数	%/1,460	柱状图
PEOPLES R CHINA	559	38.288%	
USA	252	17.260%	
SOUTH KOREA	167	11.438%	
INDIA	69	4.726%	
TAIWAN，CHINA	68	4.658%	
ENGLAND	66	4.521%	
ITALY	60	4.110%	
AUSTRALIA	56	3.836%	
GERMANY	56	3.836%	
JAPAN	56	3.836%	

2. 大数据分析研究热点分析

通过 Citespace 关键词共现分析，合并相同概念关键词，选择生成最小树 MST 剪枝策略，得到大数据分析研究关键词共现统计表，如表 1－7 所示。结果显示，除检索条件“big data analysis”及“big data”以外，出现频次最高的是 model，第二是 system；中心性最高的是 system、data mining、machine learning 和 prediction，说明模型、系统、机器学习和预测性是国际学者研究大数据分析领域的热点。

表 1－7　大数据分析研究关键词共现统计表

Count	Centrality	Year	Keywords
203	0.04	2016	big data
132	0.05	2016	big data analysis
54	0.04	2016	model
51	0.15	2016	system
41	0.07	2016	algorithm
38	0.15	2016	machine learning
35	0.08	2016	classification
32	0.06	2016	network
31	0.03	2016	cloud computing
29	0.10	2016	mapreduce
26	0.03	2016	impact
26	0.15	2016	prediction
25	0.05	2016	framework
24	0.15	2016	data mining
22	0.05	2016	performance
21	0.13	2016	optimization

续表

Count	Centrality	Year	Keywords
20	0. 10	2016	challenge
20	0. 14	2016	management
20	0. 08	2016	internet

生成国内外大数据分析研究前沿知识图谱，如图 1 – 17 所示。将关键词聚类后（见图 1 – 18），研究热点在时间上交叉重合较明显。物联网、社交媒体、数据分类是大数据分析的前沿领域。

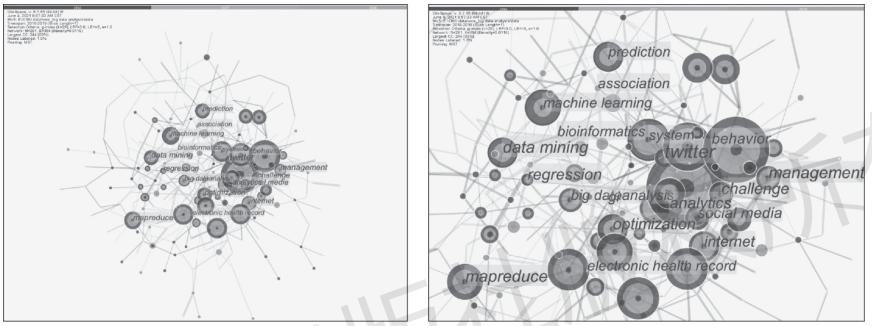


图 1 – 17　国内外大数据分析研究前沿知识图谱

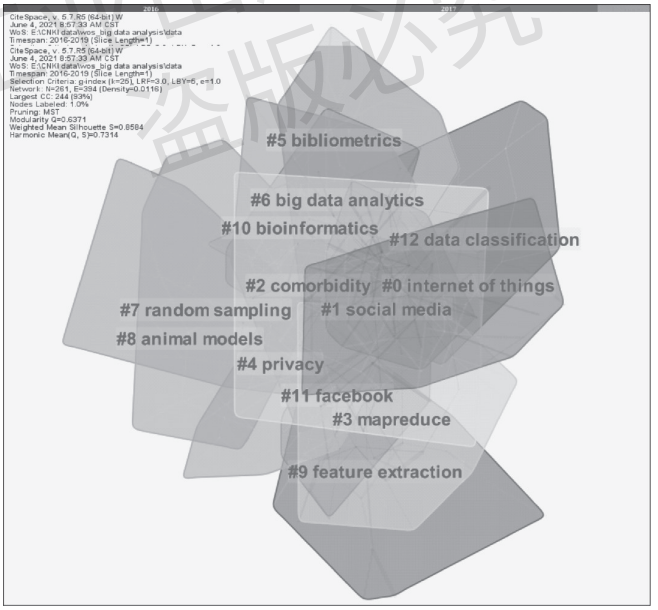


图 1 – 18　大数据分析研究关键词聚类图

3. 大数据分析研究共被引关系分析

选择网络节点为“reference”，阈值设置为（2，2，10）、（2，2，10）、（2，2，10），得到大数据分析研究文献共被引网络图，如图 1 – 19 所示，得到其领域文献之间的被共引关系统计表，如表 1 – 8 所示。从图 1 – 19 及表 1 – 8 中可以看出，最突出的是 Manyika J 在 2011 年发表于 *BIG DATA NEXT FRONTI* 的文章，且其中心性也最高，远远超过了被认为是

中心度节点的 0.1 的标准。从共被引关系突现表（见表 1－9）中可知近年在大数据分析研究上具有突出贡献、具有转折意义的作者及文献，如 Manyika J、Chen HC、LeCun Y 等。

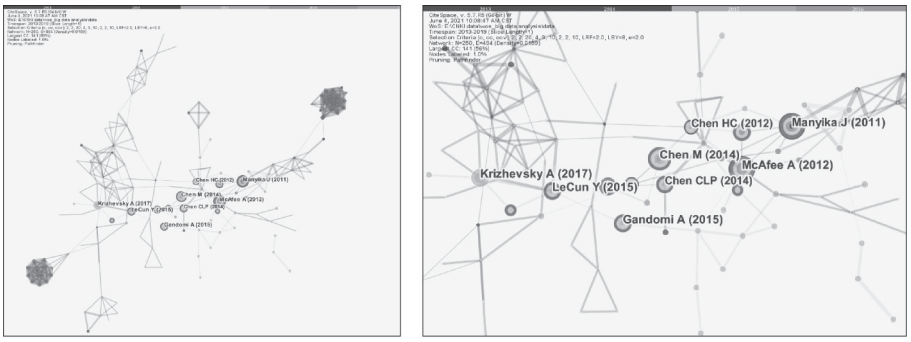


图 1－19　大数据分析研究文献共被引网络图

表 1－8　大数据分析研究文献共被引关系统计表

Count	Centrality	Year	Cited Refernce
18	0.18	2011	Manyika J, 2011, BIG DATA NEXT FRONTI, V0, P0
17	0.12	2012	McAfee A, 2012, HARVARD BUS REV, V90, P60
16	0.02	2017	Krizhevsky A, 2017, COMMUN ACM, V60, P84, DOI 10.1145/3065386
15	0.06	2015	LeCun Y, 2015, NATURE, V521, P436, DOI 10.1038/nature14539
15	0.04	2015	Gandomi A, 2015, INT J INFORM MANAGE, V35, P137, DOI 10.1016/j. ijinfomgt.2014.10.007
15	0.04	2014	Chen M, 2014, MOBILE NETW APPL, V19, P171, DOI 10.1007/s11036－013－0489－0
14	0.01	2014	Chen CLP, 2014, INFORM SCIENCES, V275, P314, DOI 10.1016/j. ins.2014.01.015
13	0.17	2012	Chen HC, 2012, MIS QUART, V36, P1165
10	0.03	2013	Mayer－Schonberger V, 2013, BIG DATA REVOLUTION, V0, P0
10	0.01	2012	Han J, 2012, MOR KAUF D, V0, P1

表 1－9　大数据分析研究文献共被引关系突现表（前 7 名）

References	Year	Strength	Begin	End	2013—2019 年
Manyika J, 2011, BIG DATA NEXT FRONTI, V0, P0	2011	3.76	2013	2016	
Chang F, 2008, ACM T COMPUTSYST, V26, P0, DOI 10.1007/s11036－013－0489－0, DOI	2008	3.59	2014	2016	
McAree A, 2012, HARVARD BUS REV, V90, P60	2012	5.27	2015	2016	
Chen M, 2014, MOBILE NETW APPL, V19, P171, DOI 10.1007/s11036－013－0489－0, DOI	2014	3.15	2016	2017	
Krizhevsky A, 2017, COMMUN ACM, V60, P84, DOI 10.1145/3065386, DOI	2017	4.86	2017	2019	
Chen HC, 2012, MIS QUART, V36, P1165	2012	3.66	2017	2019	
LeCUN Y, 2015, NATURE, V521, P436, DOI 10.1038/nature14539, DOI	2015	3.03	2017	2019	

1.4.3　机器学习热点与前沿

1. 机器学习文献分布状况

（1）时间分布

图 1－20 所示的时间分布图展示了 WOS 中以“machine learning”为主题，2008—2021 年按年份的分布情况。从图中可以看出随着时间的变化，文献数量逐步增多，递增趋势变化明显。根据现状及发展趋势，预测在未来几年，相关研究也会不断增多。

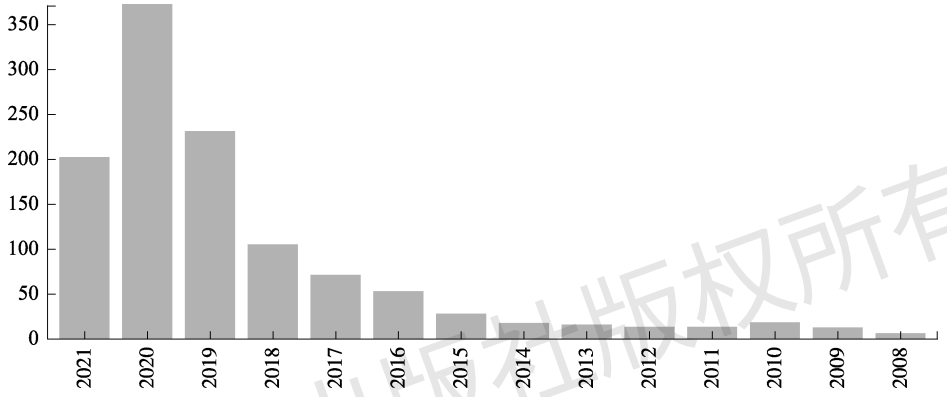


图 1－20　机器学习研究文献的时间分布图

（2）国家和地区分布

通过对 WOS 中机器学习文献分析（见图 1－21 和表 1－10）可知，截至 2021 年 5 月，美国在机器学习领域发文最多，占 39.697%；第二为我国大陆地区，发文占比为 12.489%；第三为英国，发文占比为 9.991%。由此可知，与其他相关研究发展相似，我国的机器学习领域研究也同时走在世界前列。

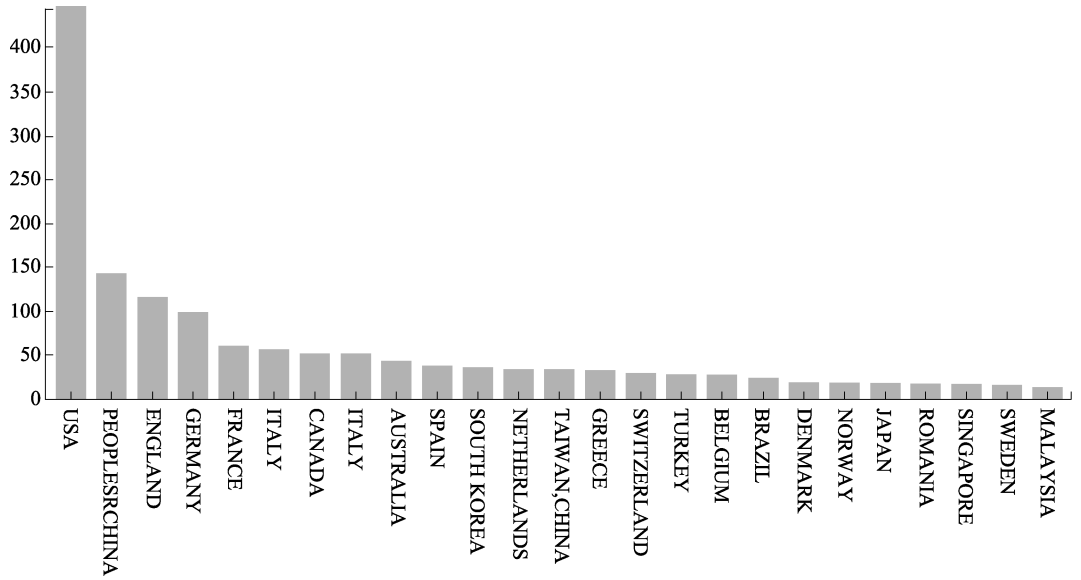


图 1－21　机器学习研究文献的国家和地区分布图

表 1－10　机器学习研究文献的国家和地区分布表

字段：国家/地区	记录数	%/1,121	柱状图
USA	445	39.697%	
PEOPLES R CHINA	140	12.489%	
ENGLAND	112	9.991%	
GERMANY	96	8.564%	
FRANCE	57	5.085%	
ITALY	53	4.728%	
CANADA	48	4.282%	
INDIA	48	4.282%	
AUSTRALIA	40	3.568%	
SPAIN	35	3.122%	

2. 机器学习研究热点分析

通过 Citespace 关键词共现分析，合并相同概念关键词，选择生成最小树 MST 剪枝策略，得到机器学习研究关键词共现统计表，如表 1－11 所示，生成机器学习研究前沿关键词共现知识图谱，如图 1－22 所示。由可视化结果可知，除检索条件“machine learning”以外，频次从小到大依次为 model（模型）、neural network（神经网络）、learning technique（学习技法）、artificial intelligence（人工智能）等，组成了机器学习研究近年来的研究热点。根据关键词中心性，text mining、stock market、propensity score、variable importance 位于前列且中心性超过 0.1，说明该关键词在近年的研究中起到了不可或缺的作用。

表 1－11　机器学习研究关键词共现统计表

Count	Centrality	Year	Keywords
622	0.04	2008	machine learning
191	0.03	2010	model
120	0.08	2008	neural network
98	0.00	2013	learning technique
96	0.01	2013	artificial intelligence
94	0.06	2008	learning method
87	0.01	2011	learning algorithm
84	0.02	2017	big data
81	0.04	2016	random forest
76	0.01	2014	learning approach
23	0.10	2017	text mining
13	0.13	2017	stock market
11	0.12	2017	propensity score
6	0.16	2019	variable importance

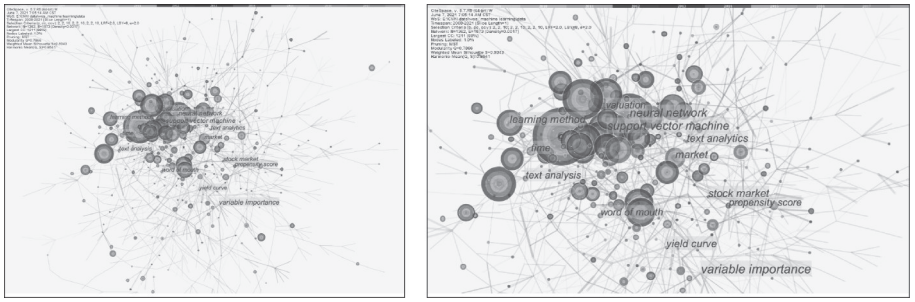


图 1－22　机器学习研究前沿关键词共现知识图谱

将关键词聚类后（见图 1－23），研究热点较为集中，时间上变化差异不大。higher education（高校教育）、predictability（可预测性）和 text data（文本数据）等是机器学习的前沿研究领域。

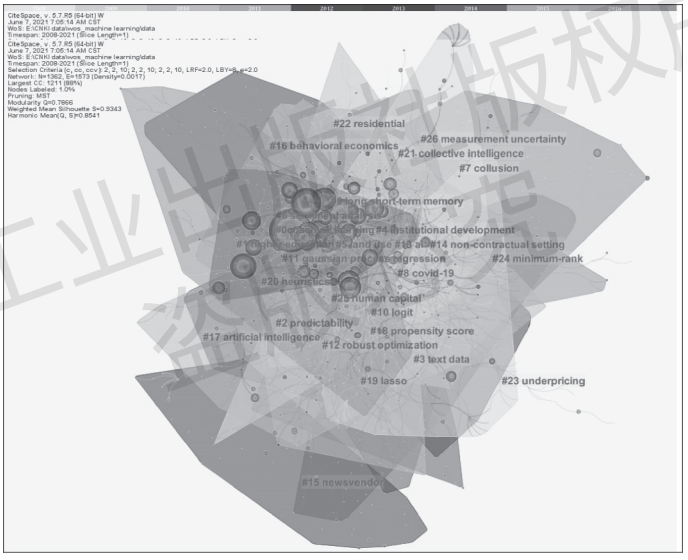


图 1－23　机器学习研究关键词聚类图

在机器学习研究前 10 关键词突现表（见表 1－12）中，我们可以看到具体研究热点随时间的演变。研究主要集中于 machine learning（机器学习）、neural network（神经网络）、support vector machine（支持向量机）、classification（分类）、social network（社会网络）和 data analysis（数据分析）等，共同组成了近 20 年的机器学习领域的研究前沿和研究新兴领域。

表 1－12　机器学习研究前 10 关键词突现表

Keywords	Year	Strength	Begin	End	2008—2021 年
machine learning	2008	19. 55	2008	2015	
neural network	2008	9. 7	2008	2016	
support vector machine	2008	13. 62	2009	2017	
classification	2008	5. 79	2009	2017	

续表

Keywords	Year	Strength	Begin	End	2008—2021 年
genetic algorithm	2008	5. 62	2009	2017	
system	2008	4. 21	2009	2017	
model	2008	10. 46	2010	2017	
support vector regression	2008	6. 08	2015	2017	
social network	2008	4. 16	2015	2018	
data analysis	2008	3. 71	2015	2018	

3. 机器学习研究共被引关系分析

同样选择网络节点为“reference”，阈值设置为（2，2，10）、（2，2，10）、（2，2，10），得到机器学习研究文献共被引网络图，如图 1 – 24 所示，得到其领域文献之间的被共引关系统计表，如表 1 – 13 所示。从图 1 – 24 及表 1 – 13 中可以看出，最突出的是 Mullainathan S 在 2017 年发表于 *J ECON PERSPECT* 的文章。从共被引关系突现表（见表 1 – 14）中可知近年在机器学习研究上具有突出贡献、具有转折意义的作者及文献，如 Hardle W、Hastie TJ 等。

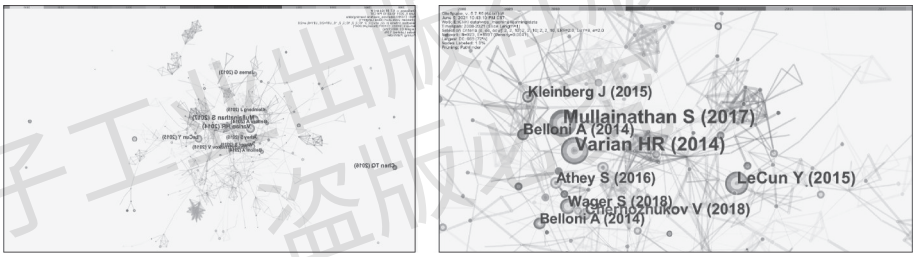


图 1 – 24　机器学习研究文献共被引网络图

表 1 – 13　机器学习研究文献共被引关系统计表

Count	Centrality	Year	Cited Refernce
50	0. 00	2017	Mullainathan S, 2017, J ECON PERSPECT, V31, P87, DOI 10. 1257/jep. 31. 2. 87
48	0. 01	2014	Varian HR, 2014, J ECON PERSPECT, V28, P3, DOI 10. 1257/jep. 28. 2. 3
35	0. 00	2016	Chen TQ, 2016, KDD16; PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, V0, P785, DOI 10. 1145/2939672. 2939785
29	0. 02	2015	LeCun Y, 2015, NATURE, V521, P436, DOI 10. 1038/nature14539
23	0. 00	2013	James G, 2013, INTRO STAT LEARNING, V0, P0
21	0. 02	2014	Belloni A, 2014, REV ECON STUD, V81, P608, DOI 10. 1093/restud/rdt044
6	0. 10	2017	Krizhevsky A, 2017, COMMUN ACM, V60, P84, DOI 10. 1145/3065386
4	0. 13	2019	Carmona P, 2019, INT REV ECON FINANC, V61, P304, DOI 10. 1016/j. iref. 2018. 03. 008
4	0. 12	2014	Nassirtoussi AK, 2014, EXPERT SYST APPL, V41, P7653, DOI 10. 1016/j. eswa. 2014. 06. 009
4	0. 11	2015	Geng RB, 2015, EUR J OPER RES, V241, P236, DOI 10. 1016/j. ejor. 2014. 08. 016
2	0. 10	2014	Agarwal R, 2014, INFORM SYST RES, V25, P443, DOI 10. 1287/isre. 2014. 0546

表 1－14　机器学习研究文献共被引关系突现表（前 8 名）

Keywords	Year	Strength	Begin	End	2008—2021 年
Hardle W，2009，J FORECASTING，V28，P512，DOI 10. 1002/for. 1109，DOI	2009	3. 06	2015	2017	
Hastie TJ，2009，ELEMENTS STAT LEARNI，V0，P0，DOI 10. 1007/978－0－387－84858－7，DOI	2009	7. 55	2016	2017	
Friedman J，2010，J STAT SOFTW，V33，P1，DOI 10. 18637/jss. v033. i01，DOI	2010	3. 76	2016	2017	
Loughran T，2011，J FINANC，V66，P35，DOI 10. 1111/j. 1540－6261. 2010. 01625. x，DOI	2011	3. 69	2016	2019	
Varian HR，2014，J ECON PERSPECT，V28，P3，DOI 10. 1257/jep. 28. 2. 3，DOI	2014	4. 2	2017	2018	
Chang CC，2011，ACM T INTEL SYST TEC，V2，P0，DOI 10. 1145/1961189. 1961199，DOI	2011	2. 74	2017	2019	
Pedregosa F，2011，J MACH LEARN RES，V12，P2825	2011	4. 81	2018	2019	
Mullainathan S，2017，J ECON PERSPECT，V31，P87，DOI 10. 1257/jep. 31. 2. 87，DOI	2017	3. 1	2018	2019	

1. 4. 4　数据挖掘热点与前沿

1. 数据挖掘文献分布状况

（1）时间分布

图 1－25 所示的时间分布图展示了 WOS 中以“data mining”为主题，2008—2021 年按年份的分布情况。随着时间的变化，文献数量逐步增多，总体上递增趋势变化明显，2011 年、2014 年、2018 年和 2020 年略有下降。根据现状及发展趋势，预测在未来几年，相关研究会稳定增多。

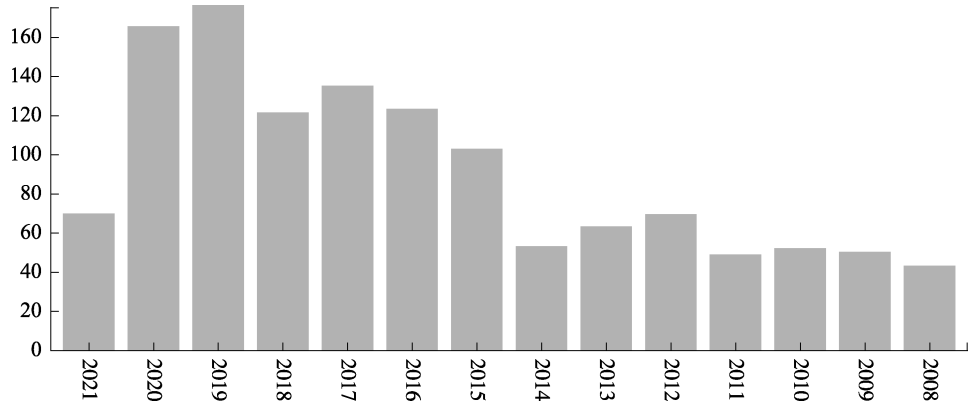


图 1－25　数据挖掘研究文献的时间分布图

（2）国家和地区分布

通过对 WOS 中数据挖掘文献分析（见图 1－26 和表 1－15）可知，截至 2021 年 5 月，美国在数据挖掘领域发文最多，占 26. 153%；第二为我国大陆地区，发文占比为

14.194%；第三为英国，发文占比为8.426%。从发文量上看，我国的数据挖掘领域研究也同时走在世界前列。

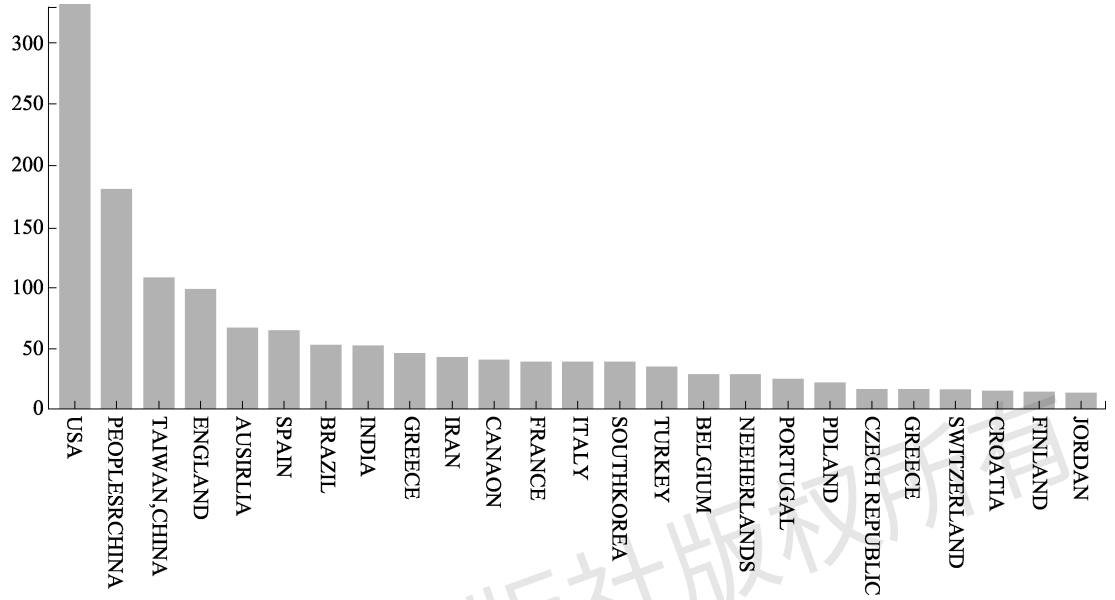


图1－26　数据挖掘研究文献的国家和地区分布图

表1－15　数据挖掘研究文献的国家和地区分布表

字段：国家地区	记录数	%/1，258	柱状图
USA	329	26.153%	
PEOPLES R CHINA	178	14.149%	
TAIWAN，CHINA	106	8.426%	
ENGLAND	98	7.790%	
AUSTRALIA	65	5.167%	
SPAIN	63	5.008%	
BRAZIL	51	4.054%	
INDIA	50	3.975%	
GERMANY	44	3.498%	

2. 数据挖掘研究热点分析

通过Citespace关键词共现分析，合并相同概念关键词，选择生成最小树MST剪枝策略，得到数据挖掘研究关键词共现统计表，如表1－16所示，生成数据挖掘研究关键词共现可视化图，如图1－27所示。其中，除检索条件“data mining”以外，频次从小到大依次为model（模型）、classification（分类）、machine learning（机器学习）等，都是近年来数据挖掘相关的研究热点。根据关键词中心性，neural network、service、learning algorithm位于前列且中心性超过0.1，说明该关键词在近年的研究中起到了不可或缺的作用。

表 1－16　数据挖掘研究关键词共现统计表

Count	Centrality	Year	Keywords
871	0. 00	2008	data mining
164	0. 04	2008	model
137	0. 09	2008	classification
104	0. 02	2008	machine learning
85	0. 06	2008	algorithm
78	0. 09	2008	data mining technique
77	0. 02	2008	system
76	0. 00	2014	big data
75	0. 05	2009	support vector machine
70	0. 10	2008	neural network
18	0. 13	2009	service
19	0. 10	2008	learning algorithm

图 1－27　数据挖掘研究关键词共现可视化图

通过将关键词聚类（见图 1－28），可知研究热点较为集中，时间上变化差异不大。分类、大数据、网络和情感分析等是当今数据挖掘的前沿领域。

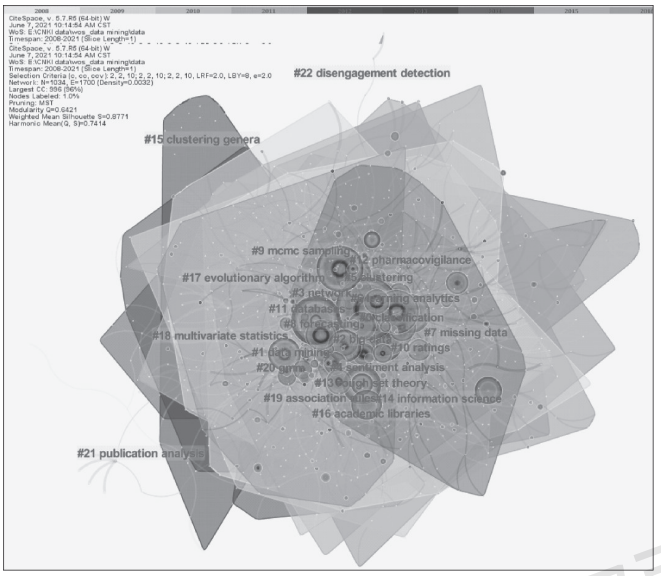


图 1－28　数据挖掘研究关键词聚类图

在前 10 关键词突现表（见表 1－17）中，我们可以看到具体研究热点随时间的演变。2008—2015 年研究主要集中于大数据、预测、商务智能、可视化和信息等。近三年，人工智能成为新的研究热点。

表 1－17　数据挖掘研究前 10 关键词突现表

Keywords	Year	Strength	Begin	End	2008—2021 年
large database	2008	3. 57	2008	2009	
prediction	2008	3. 72	2009	2011	
business intelligence	2008	4. 56	2010	2012	
visualization	2008	4. 35	2010	2015	
information	2008	6. 05	2012	2014	
statistical analysis	2008	15. 34	2013	2016	
knowledge	2008	4. 02	2013	2015	
high－dimensional data	2008	3. 89	2013	2015	
asa data science journal	2008	9. 68	2016	2016	
artificial intelligence	2008	6. 2	2019	2021	

3. 数据挖掘研究共被引关系分析

选择网络节点为“reference”，阈值设置为（2，2，10）、（2，2，10）、（2，2，10），得到数据挖掘研究文献共被引网络图，如图 1－29 所示，得到其领域文献之间的被共引关系统计表，如表 1－18 所示。从图 1－29 及表 1－18 中可以看出，最突出的是 Han J 在 2012 年发表于 *MOR KAUF D* 的文章，其次是 Hastie Trevor 在 2009 年发表于 *ELEMENTS STAT LEARNI* 的文章，且其中心度远超 0. 1 标准，高达 0. 27。从共被引关系突现表（见表 1－19）中可知近年在数据挖掘研究上具有突出贡献、具有转折意义的作者及文献，如 Han J、Chen HC、Pedregosa F 等。

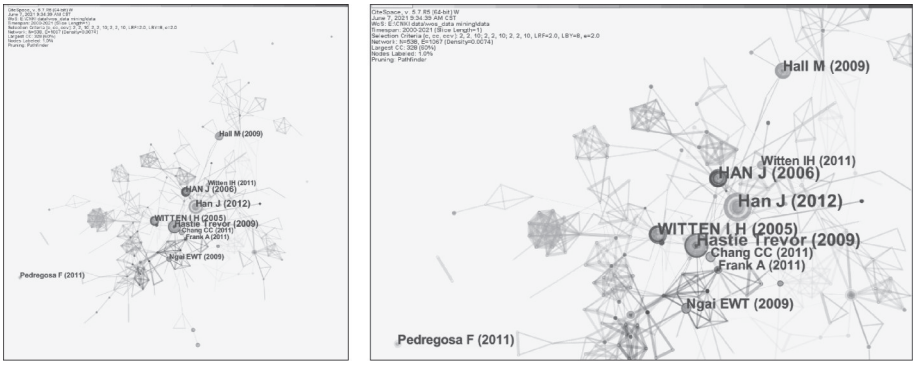


图 1－29　数据挖掘研究文献共被引网络图

表 1－18　数据挖掘研究文献共被引关系统计表

Count	Centrality	Year	Cited Refernce
33	0. 09	2012	Han J, 2012, MOR KAUF D, V0, P1
28	0. 27	2009	Hastie Trevor, 2009, ELEMENTS STAT LEARNI, V2nd ed., P0, DOI 10. 1007/978－0－387－84858－7]
23	0. 10	2006	HAN J, 2006, DATA MINING CONCEPTS, V0, P0
20	0. 11	2005	WITTEN I H, 2005, DATA MINING PRACTICA, V0, P0
15	0. 03	2009	Hall M, 2009, ACM SIGKDD EXPLOR NE, V11, P1
6	0. 24	2006	Neslin SA, 2006, J MARKETING RES, V43, P204, DOI 10. 1509/jmkr. 43. 2. 204
5	0. 20	2013	He W, 2013, INT J INFORM MANAGE, V33, P464, DOI 10. 1016/j. ijinfomgt. 2013. 01. 001
9	0. 14	2006	Tan P－N, 2006, INTRO DATA MINING, V0, P0
4	0. 13	2015	Lessmann S, 2015, EUR J OPER RES, V247, P124, DOI 10. 1016/j. ejor. 2015. 05. 030
10	0. 12	2011	Chang CC, 2011, ACM T INTEL SYST TEC, V2, P0, DOI 10. 1145/1961189. 1961199
4	0. 12	2001	Hand DJ, 2001, ADAP COMP MACH LEARN, V0, P0
2	0. 12	2012	Kim AJ, 2012, J BUS RES, V65, P1480, DOI 10. 1016/j. jbusres. 2011. 10. 014
2	0. 12	2013	Cortez P, 2013, INFORM SCIENCES, V225, P1, DOI 10. 1016/j. ins. 2012. 10. 039
4	0. 11	2018	Vu HQ, 2018, J TRAVEL RES, V57, P883, DOI 10. 1177/0047287517722232

表 1－19　数据挖掘研究文献共被引关系突现表（前 10 名）

References	Year	Strength	Begin	End	2000—2021 年
WITEEN I H, 2005, DATA MINING PRATICA, V0, P0	2005	7. 75	2008	2012	
Friedman J, 2001, ELEMENTS STATLEARNI, V1, P0	2001	5. 16	2008	2009	
Han Jiawei, 2001, DATA MINING CONCEPTS, V0, P0	2001	4. 58	2008	2009	

续表

References	Year	Strength	Begin	End	2000—2021 年
HAN J，2006，DATA MINING CONCEPTS，V0，P02017	2006	7.95	2009	2014	
Hastie Trevor，2009，ELEMENTS STAT LEARNI，V2nd ed.，P0，DOI 10.1007/978－0－387－84858－7]，DOI	2009	6.81	2013	2017	
Hall M，2009，ACM SIGKDD EXPLOR NE，V11，P1	2009	6.02	2014	2017	
Witten IH，2011，MOR KAUF D，V0，P1	2011	3.61	2015	2018	
Han J，2012，MOR KAUF D，V0，P1	2012	5.81	2017	2021	
Chen HC，2012，MIS QUART，V36，P1165	2012	3.6	2017	2019	
Pedregosa F，2011，J MACH LEARN RES，V12，P2825	2011	5.42	2018	2019	

1.4.5 本章小结

通过文献的时间分布、国家和地区分布情况可知，20 世纪末到 21 世纪初，大数据分析与挖掘及相关领域的文献不断发表，总体呈现逐步上升的趋势。总体上来看，美国在商务智能领域的发文量位居首位，第二是中国，中国的商务智能及相关研究在全球范围内处于领先地位。其中，在大数据分析领域中，中国位居首位。

运用 Citespace 进行可视化工具的关键词聚类、突现功能，可以发现数据挖掘、数据仓库、知识管理、系统、框架、数据分析和机器学习成为新的研究热点，反映了当下研究已逐步由技术转向对数据的商业分析。运用 Citespace 对与商务智能领域及相关的大数据分析、机器学习、数据挖掘进行进一步分析可知，各个领域间的研究热点交叉度也较高，算法、模型等是近些年关注的热点。热点算法有分类、支持向量机、回归算法、神经网络和随机森林等。

运用 Citespace 工具生成被共引网络关系图，得到 Hevner AR 等人在 2004 年于 *MIS QUART* 上发表的文章、Elbashir Mohamed Z 在 2008 年发表的文章等都是商务智能领域研究的关键性节点文件；大数据分析研究中心最重要的文献是 Manyika J 在 2011 年发表于 *BIG DATA NEXT FRONTI* 的文章；Mullainathan S 在 2017 年发表于 *J ECON PERSPECT* 的文章是机器学习研究的重要文献；Han J 在 2012 年发表于 *MOR KAUF D* 的文章、Hastie Trevor 在 2009 年发表于 *ELEMENTS STAT LEARNI* 的文章是数据挖掘研究的关键性节点文件。这些文献对未来的商务智能研究都会有重要的参考引用价值。

本章参考文献

[1] LanH. Witten，EibeFrank，MarkA. Hall，等. 数据挖掘：实用机器学习工具与技术 [M]. 北京：机械工业出版社，2014.

[2] 韩家炜,坎伯,裴健,等. 数据挖掘概念与技术 [M]. 北京：机械工业出版社，2012.
[3] MITCHELL T M. Does Machine Learning Really Work? [J]. Ai Magazine,1997,18(3):11 – 20.
[4] FAYYAD U, PIATETSKY – SHAPIRO G, SMYTH P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. AAAI Press,2000.
[5] FAYYAD U, PIATETSKY – SHAPIRO G,SMYTH P. From Data Mining to Knowledge Discovery in Databases[C]. Ai Magazine. 1996:37 – 54.
[6] NGUYEN G, DLUGOLINSKY S, M BOBÁK, et al. Machine Learning and Deep Learning frameworks and libraries for large – scale data mining: a survey[J]. Artificial Intelligence Review,2019.
[7] 张昭. 基于 Citespace 的商务智能研究热点与前沿可视化分析 [J]. 情报探索，2012(12)：6 – 9.
[8] 萧文龙，王镇豪，陈豪，徐瑀婧 . 国内外商务智能及大数据分析研究动态和发展趋势分析 [J]. 科技与经济，2020，33 (06)：66 – 70.
[9] 埃森哲大数据分析方法论及工具 [R]. 北京：埃森哲公司，2014：1 – 65.
[10] JEANNETTE M. Wing. The Data Life Cycle. Harvard Data Science Review,Iss 1,Jan. 2019.

本书涉及的环境、语言、框架和库

（1）语言、环境

- Python programming language. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Anaconda—the most popular python data science platform. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Anaconda for Cloudera—data science with python made easy for big data. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Project jupyter. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e

（2）数据挖掘与机器学习 Python 库

- NumPy—the fundamental package for scientific computing with Python. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- SciPy—Scientific computing tools for Python. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Pandas—Python Data Analysis Library. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Matplotlib—Visualization with Python. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Scikit – learn machine learning in Python. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e

- Natural language toolkit. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- SciLab—open source software for numerical computation. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e

（3）深度学习框架

- TensorFlow—an open - source software library for machine intelligence. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- PyTorch—deep learning framework that puts python first. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e

（4）数据分析与挖掘的工具

- SPSS. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Tableau software: business intelligence and analytics. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- RapidMiner open source predictive analytics platform. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- Weka3: data mining software in Java. https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e
- SAS (previously statistical analysis system). https://www.hxedu.com.cn/hxedu/w/inputVideo.do?qid=5a79a0187deba829017dfa80f95a4d5e

注：如无特别解释，本书所有案例都是基于 Python3.7 开发。