知识图谱技术与应用

Knowledge Graph
Technology and Application

闫树 魏凯 洪万福 等◎著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

知识图谱技术与应用

闫树 魏凯 洪万福等　著

人民邮电出版社

北京

图书在版编目（CIP）数据

知识图谱技术与应用 / 闫树等著.--北京：人民邮电出版社，2019.11

ISBN 978-7-115-51966-5

Ⅰ.①知… Ⅱ.①闫… Ⅲ.①知识管理 Ⅳ.①G302

中国版本图书馆CIP数据核字（2019）第192137号

◆著 闫树 魏凯 洪万福 等

责任编辑 唐名威

责任印制 彭志环

◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路11号

邮编 100164 电子邮件 315@ptpress.com.cn

网址 http://www.ptpress.com.cn

北京市艺辉印刷有限公司印刷

◆ 开本：700×1000 1/16

印张：10 2019年11月第1版

字数：124千字 2019年11月北京第1次印刷

定价：59.00元

读者服务热线：(010)81055493 印装质量热线：(010)81055316

目录

参考文献

内容提要

本书系统地介绍了知识图谱的相关概念、技术要素与应用，不仅涵盖了知识图谱技术的发展历程与特点，也涵盖了当前阶段知识图谱的主要应用，并分析了未来的发展趋势与挑战。本书从理论综述、技术解读、应用场景、实战分析等多个角度进行了阐述，内容全面且易于理解。

本书是一本入门级图书，面向具备一定计算机知识但没有知识图谱构建经验的读者，旨在帮助他们掌握知识图谱构建的专业知识。同时，本书还面向渴望了解知识图谱应用的各行业人员，旨在帮助他们拓展视野、开阔思路。相信所有对知识图谱感兴趣的读者通过阅读本书都能有所收获。

前言

在互联网飞速发展的今天，万物互联成为可能，智能分析由只专注于个体转开始变为更关注个体之间的关系。伴随着数据处理技术（Data Technology，DT）时代的到来，数据量呈爆发式的增长。在这些海量的非结构化文本数据、大量的半结构化表格和网页以及生产系统的结构化数据中，蕴含着大量的关系信息。利用知识图谱技术，人们可以对这些关系信息进行结构化、语义化的智能处理，形成大规模的知识库，并支撑业务应用，使得机器能够更好地理解网络、理解用户、理解资源，为用户提供新型智能化服务。

然而，市面上知识图谱的相关书籍，要么聚焦于科学引文网络或其他行业应用，要么主要介绍相关工具的使用，专门讲述知识图谱全面理论的书籍还比较少。作者希望以此书的出版弥补这一空白。

本书的主要特色包括：①系统性，从知识图谱的起源发展入手，层层推进，让读者对知识图谱这一技术工具建立系统的印象；②全面性，既包含通用知识图谱，也包含行业知识图谱，内容涉及较广；③基础性，本书面向对知识图谱有兴趣的读者，力求内容通俗易懂；④实用性，理论与实践相结合，通过案例让读者对知识图谱的应用有直观的了解。

全书共6章。第一章从定义、发展历程、架构、特点等方面对知识图谱进行了基础性的概念解读；第二章以知识表示与建模、知识抽取与挖掘、知识存储与融合、知识检索与推理4个过程为主线，对搭建通用知识图谱的技术要素进行了介绍；第三章介绍了行业知识图谱的特点，并重点研究了知识图谱在公安、金融、教育、电信、工业领域中的应用场景；第四章对知识图谱的发展趋势和挑战进行了分析；

第五章通过医疗决策辅助系统、“虚拟生命”、股份制银行、禁毒大数据分析平台4个案例，从痛点难点、实现路径和应用效果3个方面对知识图谱的实战应用进行了解读；在第六章中，作者列举了Pajek、CiteSpace等8种国内外较为常用的知识图谱构建工具，并对各工具的主要功能和特点进行了介绍。

从最初的搜索引擎到现在的聊天机器人、大数据风控、证券投资、智能医疗、自适应教育、推荐系统等，知识图谱的应用越来越多，它在技术领域的热度逐年上升。大规模构建并应用知识图谱，对于互联网行业、传统行业甚至社会治理具有重要的意义。随着理论和技术的不断发展，学术界和产业界对知识图谱的认识在不断地变化与更新。未来，相关的研究和应用的边界将不断扩展。

本书的编写成员包括闫树、魏凯、洪万福、钱智毅、王彬、符山、姜春宇。本书在编写过程中得到了中国信息通信研究院何宝宏所长、张雪丽副所长、刘寒、刘成成、马鹏玮、王妙琼、李雨霏、王卓、李俊逸、吕艾临等同事的大力支持。厦门渊亭信息科技有限公司、北京明略软件系统有限公司、中移（苏州）软件技术有限公司、中软国际有限公司、深圳狗尾草智能科技有限公司等企业的专家对本书提出了建议或提供了相关案例，在此对他们一并表示感谢。

由于作者水平有限、编写时间仓促，书中难免会出现一些错误或有争议的地方，恳请读者批评指正。如果您有任何建议或遇到了任何问题，欢迎发送邮件至yanshu@caict.ac.cn，期待得到您的反馈。

第一章 知识图谱概述

在互联网时代，信息量呈爆炸式增长，这给人们有效地获取信息和知识带来了巨大的挑战。知识图谱（Knowledge Graph，KG）以其强大的语义处理功能和快速分析能力，迅速成为互联网用户信赖的，可以快速、准确地获取信息资源的智能化搜索工具。特别是随着人工智能的逐步发展与应用，知识图谱已成为一门关键技术，被广泛应用于智能问答、大数据分析、个性化推荐等领域。知识图谱同深度学习一起，成为推动人工智能发展的核心驱动力之一。本章将从定义、发展历程、架构、特点等方面对知识图谱进行介绍。

第一节 什么是知识图谱

一、知识图谱的定义

作为一种智能、高效的知识组织方式，知识图谱能够帮助用户迅速、准确地查询到自己需要的信息，近年来得到了飞速发展。尽管产业界对其内涵有了基本共识，但实际上目前尚没有一个公认的定义。

知识图谱由Google公司在2012年提出，但发布时Google公司并没有对这一概念做出清晰的定义。维基百科上知识图谱的词条实际是对Google公司搜索引擎使用的知识库功能的描述，即知识图谱是Google公司使用的一个知识库及服务，它利用从多种来源收集的信息提升搜索引擎返回的结果的质量。

百度百科将知识图谱定义为“通过将应用数学、图形学、信息可视化技术、信息科学等学科的理论和方法与计量学引文分析、共现分析

等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构，达到多学科融合目的的现代理论。”但从该词条的详细内容可以看出，百度百科的定义仍是一种对知识图谱的早期理解和对Google公司提出的知识图谱功能的复述。

国内外学术机构围绕知识图谱进行了大量研究，近年来我国高校学者也在知识图谱领域发表了许多优秀的论文，并对知识图谱做出了比较完整和全面的定义。如华东理工大学教授王昊奋认为：“知识图谱旨在描述真实世界中存在的各种实体或概念。其中，每个实体或概念用一个全局唯一确定的ID来标识，这个ID被称为它们的标识符。‘属性-值’对（Attribute-Value Pair，AVP）用来刻画实体的内在特性，而关系用来连接两个实体，刻画它们之间的关联。”而电子科技大学的刘峤等人认为：“知识图谱是结构化的语义知识库，用于以符号形式描述物理世界中的概念及其相互关系，其基本组成单位是‘实体-关系-实体’三元组以及实体及其属性-值对，实体间通过关系相互联结，构成网状的知识结构。”

在互联网上有大量对知识图谱的讨论，在知乎等社交问答网站上存在多种对知识图谱的表述，内容大同小异，如“本质上，知识图谱旨在描述真实世界中存在的各种实体或概念及其关系，其构成一张巨大的语义网络图，节点表示实体或概念，边则由属性或关系构成。现在的知识图谱已被用来泛指各种大规模的知识库”。

技术厂商和用户对知识图谱有着不同的理解，但大多没有提出明确的定义。

从以上研究中可以看到知识图谱的起源和在中国的传播发展。综合其中的共识，作者对本书探讨的知识图谱给出以下定义：知识图谱本质上是一种语义网络，是新一代的知识库技术，通过结构化、语义化的处理将信息转化为知识，并加以应用。

二、对知识图谱定义的解读

对于上述知识图谱的定义，可以从以下几个方面进行解读。

1.表现形式

知识图谱的抽象表现形式是以语义互相连接的实体，是把人对实体世界的认知通过结构化的方式转化为计算机可理解和计算的语义信息。我们可以将知识图谱理解成一个网状知识库，这个知识库反映的是一个实体及与其相关的其他实体或事件，不同的实体之间通过不同属性的关系相互连接，从而形成了网。由此，知识图谱可以被看成对物理世界的一种符号表达。

2.涵盖范围

知识图谱由传统的知识库演变而来，可以说狭义的知识图谱就是知识库，但广义的知识图谱应涵盖知识库、从信息到知识的知识库构建以及高效定位正确的知识、发现隐含的知识的知识库运用等方面，目标是解决信息过载和信息缺失的问题。

3.技术表现

知识图谱在技术上表现为技术栈，通常被认为是由知识提取、知识融合、知识加工、知识呈现4层技术组合而成的。知识图谱在知识

库的构建方面具备接入多数据源的能力，比传统的人工方式更加高效。除了知识库部分外，知识图谱技术还包括可以生成新知识的推理引擎，被视为自动化、智能化的新一代知识库技术。

4.研究价值

知识图谱是人工智能的关键技术之一，人工智能追求的目标是利用机器快速、便捷地获得高质量的数据信息，进而辅助人们进行更多智能化的应用。在实现这一目标的过程中，知识就是核心力量。知识对于人工智能的价值在于让机器具备对数据的认知能力和理解能力。构建知识图谱的目的就是让机器形成这种认知能力，使其能够理解这个世界。

知识的获取，特别是自动获取本身就很难，知识的来源广泛，且种类多样，形成知识的数据可能是结构化的，也可能是非结构化的。获取知识后的理解和推理是对知识的选择和应用，是将零散的数据整合到一起并梳理出脉络的过程，更为复杂。

这个时候，可以通过“图”这样一种直观、清晰的基础性通用“语言”，清楚地还原各个数据之间的复杂关系。知识图谱的构建就是在Web网页的基础上增加一层覆盖的网状的图，将Web网页上的概念相互链接起来，用最小的成本将互联网中大量的信息组织成可以被利用的知识。

5.应用价值

知识图谱提供了一种从海量数据中抽取结构化知识的手段，快速便捷，拥有广阔的应用前景。

对于使用知识图谱的人来说，相比文字，图更加直观、有条理，因此知识图谱可以帮助人们更好地理解和记忆知识。很多人应用思维导图对知识进行记忆和梳理，在这个过程中应用的是使用者本身的记忆习惯和技巧。知识图谱是从知识本身出发，保留了知识原来的组织，引导使用者理解知识。

对于使用知识图谱的软件、服务、系统来说，知识图谱提供了结构化的数据存储格式，降低了软件、服务、系统在数据挖掘和管理过程中的难度。同时，知识图谱可以在较好地保存数据及数据之间关联的基础上，挖掘出更多的有效信息，开发更多的应用场景。在使用知识图谱服务进行搜索时，人们可以直接获得与数据关联的答案，而不是可能包含答案的网页。

知识图谱由复杂多层次的技术栈构成，内涵覆盖构建、应用等多个生命周期环节，知识图谱技术的供需双方对于知识图谱的理解和着眼点实际是不同的。需求方企业往往倾向于简单化理解，或者将其等同于传统的专家库，或者认为其就是图可视化的炫酷展现形式；而技术厂商可能基于自身在技术栈不同层面的优势宣传和解读这一技术。透过复杂的技术栈和纷繁的技术术语来看，知识图谱的本质是运用新的技术在知识结构化和分析洞察两个方面提升信息转化为知识并且被利用的效率，具体如下。

- 知识结构化：与传统知识库相比，知识图谱在知识构建部分除了专家人工的方式，还利用机器学习算法等手段进行文本挖掘和自然语言处理，从大量的非结构化和半结构化数据中抽取知识。

- 分析洞察：在人、企业、产品、兴趣、想法、事实存在交织的关联关系时，使用图分析这些复杂的关系效率更高，也更加有可扩展性。如应用图遍历、最短路径、三角计数、连通分量、类中心等算法

进行目标实体搜寻、实体关联识别、关联程度评价、关键人物和特殊关系群体发现等工作时，可得到较好的效果。

从企业级信息管理的全局视角来看，知识图谱无疑是企业信息管理的一种方式和手段。知识图谱的主要功能（如文本分析、语义计算等）与传统的数据采集、清洗、整合等数据处理功能在处理方法和流程上有一定的相似性，在技术上也有互通或重合的内容。知识图谱的建设横跨企业级数据建设和应用的多个环节，在技术的整合方面复杂度较高，因此应用知识图谱的用户企业需要具备一定的数据基础和数据技术能力基础，比如持续的数据治理和知识管理机制、较好的基础数据质量、对数据技术能力和团队的积累等。

三、知识图谱的通用表示

从本质上来看，可以将知识图谱理解成一张由不同知识点相互连接形成的语义网络。任何一种网络都是由节点和边构成的，因此，知识图谱也是由节点和边构成的。节点表示实体或概念，边表示实体的属性或实体间的关系。

知识图谱中的节点分为以下两种。

● 实体：指具有可区别性且独立存在的某种事物，如一个人、一座城市、一种商品等。某个时刻、某个地点、某个数值也可以作为实体。实体是一个知识图谱中最基本的元素，每个实体可以用一个全局唯一的ID进行标识。

● 语义类/概念：语义类指具有某种共同属性的实体的集合，如国家、民族、性别等；而概念则反映一组实体的种类或对象类型，如人物、气候、地理等。

知识图谱中的边分为以下两种。

● 属性（值）：指某个实体可能具有的特征、特性、特点以及参数，是从某个实体指向它的属性值的“边”，不同的属性对应不同的边，而属性值是实体在某一个特定属性下的值。例如，图1所示的“类别”“首都”是不同的属性，“北京”是中国在“首都”这一属性下的属性值。

● 关系：是连接不同实体的“边”，可以是因果关系、相近关系、推论关系、组成关系等。在知识图谱中，将关系形式化为一个函数。这个函数把若干个节点映射到布尔值，其取值反映实体间是否具有某种关系。

基于以上定义，可以更好地理解三元组。三元组是知识图谱的一种直观、简洁的通用表示方式，可以方便计算机对实体关系进行处理。

用三元组$G=(E,R,S)$表示知识图谱，其中，$E=\{e_1,e_2,\ldots\ldots,e_E\}$是知识图谱中的实体集合，包含$|E|$种不同的实体；$R=\{r_1,r_2,\ldots\ldots,r_E\}$是知识图谱中的关系集合，共包含$|R|$种不同的关系；$S\subseteq E\times R\times E$是知识图谱中的三元组集合。三元组的基本形式主要包括（实体1，关系，实体2）以及（概念属性，属性值）等。（实体1，关系，实体2）、（实体，属性，属性值）都是典型的三元组。如图1所示，方块是实体，椭圆是属性值，实线是两个实体之间的关系，虚线是实体的属性。中国的首都是北京就可以用（中国，首都，北京）表示。

图1 知识图谱示例

第二节 知识图谱的发展历程

虽然知识图谱这一命名是在2012年才出现的，但是它的发展历程却可以追溯到20世纪的引文网络、语义Web、描述逻辑和专家系统

等。在这一技术的历史演变过程中，出现了多次发展瓶颈，人们也多次通过技术的发展突破了这些瓶颈。本节对知识图谱的发展历程进行简要回溯。

一、起源：科学知识图谱

1955年，尤金·加菲尔德（Eugene Garfield）在《科学》（*Science*）杂志发表了一篇题为《Citation Indexes for Science: A New Dimension in Documentation Through Association of Ideas》的论文，提出了“引文索引”的设想，即提供一种文献计量学的工具，帮助科学家识别其感兴趣的文献。这一引文技术的概念开创了从引文角度研究文献及科学发展动态的新方法。

1965年，普莱斯发表了《Networks of Scientific Papers》一文，提出了用引证网络表示科学文献之间印证关系的方法。这相当于为当代科学发展绘制了一张地形图，由此引文网络开始成为研究科学发展脉络的方法，进而形成了科学知识图谱（Mapping Knowledge Domain）的概念。但在这一阶段，科学知识图谱主要应用于研究科学发展的历程，更多地被用在科学计量学科和情报学科，致力于发展科学文献引用网络的可视化。

1968年，奎林（J.R.Quillian）提出了语义网络（Semantic Network）的概念，为人类联想记忆提供了一个明显的公理模型。这一模型的本质是一种用图表示知识的结构化方式，可以看成一种用于存储知识的图的数据结构。但在语义网络被提出之后，有人认为自然语言比语义网络更适合表示人类的知识，于是展开了对语义网络和自然语言谓词逻辑之间联系的讨论。在20世纪70年代的研究成果中，Bertram C.Bruce提供了一种将语义网络转化成谓词逻辑的算法，且该

算法在计算上具有一定优势；B.Kaiser给出了用语义网络表示连接词的方法。在此之后，语义网络可以方便地将自然语言的句子用图进行表达和存储，此技术可被广泛应用于机器翻译、问答系统和自然语言理解等任务。

二、发展：知识库

1977年，美国斯坦福大学的计算机科学家费根·鲍姆教授在第五届国际人工智能大会上提出了知识工程（Knowledge Engineering）的概念。知识工程是通过存储现有的专家知识对用户的提问进行求解的系统，本质上是一个通过智能软件建立的专家系统，研究如何由计算机进行问题的自动求解。知识工程的提出使人工智能的研究从基于推理的模型转向基于知识的模型，从理论转向了应用。随后，作为知识工程的一个重要组成部分，知识库（Knowledge Base，KB）应运而生，并成为知识图谱技术发展史上的重要阶段。

知识库来自于人工智能-知识工程领域和数据库领域两方面技术的有机融合。它经过分类和有序化，根据一定格式将相互关联的各种知识存储在计算机中。相比于一般的数据库，知识库可以对知识结构进行分析，根据知识的各方面特征将其编构成便于利用的、有结构的组织形式。相比于一般的应用程序只能把问题求解的知识隐含地编码在程序中，知识库则可以将问题的答案显式地表达，并单独组成一个相对独立的程序实体。

对于知识库的研究，核心在于对知识的组织和表达，因此逻辑基础十分重要。在此后的一段时期，对语义网络的研究方向逐渐转变为具有严格逻辑语义的表示和推理。从20世纪80年代末到20世纪90年代，语义网络的工作集中在对概念之间关系的建模，有人提出了术语

逻辑（Terminological Logic）以及描述逻辑的概念。这一时期比较有代表性的工作是Brachman等人提出的CLASSIC 语言和Horrock实现的FaCT 推理机。

进入21世纪，语义网（Semantic Web）和链接数据（Linked Data）的出现开启了语义网络应用的新场景。语义网和链接数据是万维网之父Tim Berners Lee分别在1998年和2006年提出的。相对于语义网络，语义网和链接数据倾向于描述万维网中资源、数据之间的关系。

语义网中的“Web”希望将数据相互链接，组成一个庞大的信息网络，正如互联网中相互链接的网页，只不过基本单位变为粒度更小的数据。在万维网诞生之初，网络上的内容只有人类可读，计算机无法理解和处理。在用户浏览网页时，计算机只能判断这是一个网页，网页里面有图片、有链接，但并不知道图片描述的是什么，也不清楚链接指向的页面与当前页面有何关系。语义网是对Web的一个扩展，其核心是给Web上的文档添加能够被计算机理解的“元数据”，使网络上的数据对于机器可读，进而使整个互联网成为一个通用的信息交换媒介。

语义网与传统Web的最显著区别是用户可以上传各种图结构的数据，并且数据之间可以建立链接，从而形成链接数据。链接数据产生的目的是定义如何利用语义网技术在网上发布数据，强调在不同的数据集间创建链接。链接数据项目汇集了很多高质量的知识库，如FreeBase、DBpedia和YAGO，这些知识库都来源于人工编辑的大规模知识库——维基百科，随后出现的知识图谱就是对链接数据这一概念的进一步包装。

在这一阶段，由于技术发展程度的限制，知识库更多以机构知识库的形式出现。对于特定的机构，由于该机构所在领域的知识规模通常相对较小，因此容易通过知识库的理论和方法进行有效的组织和管

理。有了机构知识库，对机构内容知识的保存、管理、访问更加方便，人们甚至可以利用机构知识库进行预测和决策支持。

三、形成：知识图谱

随着互联网的发展，知识与信息呈现爆发式增长，搜索引擎的使用越来越广泛。但海量的信息使得传统万维网并不能满足人们快速、准确地获取高质量信息的需求，于是，知识图谱出现了。

2012年11月，Google公司率先提出知识图谱的概念，表示将在其搜索结果中加入知识图谱的功能。此时的知识图谱与最初在引文网络中出现的科学知识图谱有很大的区别，但与知识库在理论和方法上还比较相近，只是由于建立在互联网搜索引擎的发展之上，知识图谱的含义更加宽泛。从发展愿景来看，知识图谱里的知识应该包含人们生活中的万事万物，涵盖人类文明发现和创造的所有知识。

知识图谱由知识及知识之间的关系组成，知识（实体）的内部特征使用属性-值对表示；知识（实体）之间的关系通过相互连接的边表示。从机构知识库到互联网搜索引擎，面向知识图谱的研究不断深入。传统的搜索引擎是基于关键词匹配的，而知识图谱是利用知识（实体或概念）之间的匹配度建立一个有序的知识组织，为用户提供智能化的访问接口，使用户在搜索时可以更加快速、准确地获得一个全面的信息体系。其工作原理如图2所示。

图2 知识图谱工作原理

Google公司拥有数量众多的互联网用户，有需求和资本建立一个庞大的知识图谱。Google公司采用多种语言对知识图谱中的实体、属性和实体间的关系进行描述。根据2015年统计的数据，Google公司构

建的知识图谱拥有5亿个实体、约35亿条实体关系信息，已被广泛用于提高搜索引擎的搜索质量。

在Google知识图谱中，一个大规模的、协同合作的知识库——FreeBase起到了重要作用。FreeBase即链接数据的一个数据集，采用“图”的数据结构，把知识库绘制成一个有向图。这种数据模型相对于传统数据库的优势在于其可以处理更复杂的数据以及方便数据的插入。Google知识图谱的模式（Schema）是由Google公司的专业团队在FreeBase的基础上开发和设计的。在Google知识图谱中，所有的对象都有属于它的类型（Type），类型的数量是不固定的。

在Google之后，微软、百度、搜狗等互联网公司纷纷开始构建自己的知识图谱。随着探索研究的不断深入，知识图谱作为一种新的知识管理思路，不再局限于搜索引擎的拓展应用中，开始在各类智能系统（如IBM Watson）以及数据存储等领域发挥关键作用。但是目前的知识图谱构建尚不完善，期待知识图谱在实体之间更加复杂的关系推理等方面有更多的突破。

第三节 知识图谱的架构

知识图谱的架构包括知识图谱自身具备的逻辑架构和构建知识图谱采用的技术架构两部分。

一、逻辑架构

知识图谱的逻辑架构可分为两个层次：数据层和模式层。数据层是知识图谱的基础，由一系列的事实（Fact）组成。知识以事实为单位存储在图数据库中，例如Google的Graphd和微软的Trinity都是典型

的图数据库。采用（实体，关系，实体）或（实体，属性，属性值）这样的三元组作为事实的基本表达方式，可以将存储在图数据库中的所有数据构建成庞大的实体关系网络，形成一个知识的“图谱”。知识图谱的逻辑架构如图3所示。

图3 知识图谱的逻辑架构

知识图谱的模式层在数据层之上，是知识图谱的核心。模式层存储的是经过提炼的知识。通常采用本体库管理模式层，借助本体库对公理、规则和约束条件的支持能力规范实体之间的联系。

这里提到的“本体”是一个形式化的、对于共享概念模型明确而详细的规范说明。形式化指本体可通过各种形式化的语言进行描述，这种形式化的语言对于计算机来说都是可读、可操作的。共享指本体体现的是公认的知识，反映的是对相关领域中知识的共同理解。概念模型是将客观世界中一些现象的相关概念抽象出来得到的模型，这些概念及使用这些概念的约束都有明确的定义。本体库可以看成结构化知识库的一个模板，其精练而标准。拥有本体库的知识库层次结构强，且其中的冗余知识比较少。

二、技术架构

知识图谱的技术架构也被称为体系架构，是指其在构建知识图谱时选择的模式结构。知识图谱的构建从最原始的数据出发，采用一系列自动或者半自动的技术手段，从数据库中提取知识，并将其存入知识库的数据层和模式层。

图4展示的是Google知识图谱采用的架构。虚线框的左边是可以输入的3种数据结构：结构化数据（如关系数据库）、半结构化数据

（如XML）和非结构化数据（如图像、文本），数据来源没有限制；虚线框的右边是生成的知识图谱，这个过程循环往复，且随着人的认知能力的提升而不断更新迭代；虚线框内是知识图谱的构建过程，主要包含信息抽取、知识融合、知识加工3个阶段。

图4 Google知识图谱的架构

● 信息抽取：从各种类型的数据源中提取出实体、属性以及实体间的相互关系，在此基础上形成本体化的知识表达。

● 知识融合：在获得新知识之后，需要对其进行整合，以消除矛盾和歧义，比如某些实体可能有多种表达，某个特定称谓也许对应多个不同的实体等。

● 知识加工：经过融合的新知识需要经过质量评估之后（部分需要人工参与甄别），才能将合格的部分加入知识库中，以确保知识库的质量。

知识图谱的构建有自顶向下（Top-Down）与自底向上（BottomUp）两种方式。自顶向下是先为知识图谱定义好本体与数据模式，借助百科类网站等结构化的数据源，从高质量的数据中提取本体和模式信息加入知识库；而自底向上是通过一定的技术手段，从公开的数据中提出资源模式，选择其中置信度较高的模式，经人工审核后加入知识库，之后再构建顶层的本体模式。架构构建过程中涉及的关键技术将在第二章展开介绍。

在知识图谱技术的发展初期，企业和科研机构大多采用自顶向下的方式构建基础知识库，如Google公司的FreeBase以维基百科为主要的数据来源。但随着技术的不断发展和成熟，目前，大多数知识图谱是采用自底向上的方式构建的，其中典型的是Google公司的Knowledge Vault和微软公司的Satori知识库，两者都将公开采集的海

量网页作为数据来源，对现有知识库不断进行丰富和完善。这符合互联网数据内容知识产生的特点。

第四节 知识图谱的特点

知识图谱经历了由人工和群体智慧构建到面向互联网数据利用机器学习和信息抽取技术自动获取的过程。其发展过程中的不断演化使得知识图谱相较于早期的技术有了更多不同的特点。同最早的科学知识图谱相比，现在的知识图谱是动态的，在不同的时间段，各个节点之间的关系是不断更新迭代的。与语义网络、知识库等技术相比，知识图谱有其优势。

一、与早期语义网络的比较

同早期的语义网络相比，知识图谱具有以下特点。

● 关注实体间的关联。早期的语义网络主要应用于对自然语句的表示，而知识图谱强调的是实体自身的属性以及实体之间的相互关联。虽然知识图谱中的概念有层次高低的关系，但这些关系相比实体之间的关系要少得多。

● 自动抽取，快速构建。早期的语义网络主要依靠人工的方式在结构化的数据源中进行构建，而知识图谱可以从百科等半结构化的数据中自动抽取得到。

● 强调知识间的融合。知识图谱绘制中抽取的知识不是独立的某种类型或某个学科。在构建图谱的过程中，强调的是不同来源的知识间的相互融合以及知识的清洗，但这些并不是早期语义网络关注的重点。

二、与早期知识库的比较

同早期的知识库相比，知识图谱具有以下特点。

● 描述更加客观。传统的知识库大多来源于人工编辑的大规模知识库——维基百科，而知识图谱的数据源是确定的、客观的大样本网页数据，在针对实体属性的分析过程中可以消除很多主观因素的影响，绘制的图谱具有客观性。

● 知识发现能力。知识图谱不仅要呈现实体的基本情况，还要揭示各个实体背后隐含的关系、规律和趋势，从而产生新的“事实”，即新的知识。基于知识图谱的交互探索式分析可以在计算机中模拟人的思考过程，进而发现、求证、推理。这使得机器在某种程度上具有了像人一样的分析能力。

● 知识学习能力。知识图谱利用交互式机器学习技术，支持基于推理、纠错、标注等交互动作的学习功能，不断沉淀知识逻辑和模型，提高系统的智能性，降低用户在使用时对经验的依赖。

三、与传统数据技术的比较

同传统的数据技术相比，知识图谱具有以下特点。

● 关系的表达能力更强。传统的数据库通常只能用表格、字段的方式进行读取，知识间的关系层级和表达方式多种多样。知识图谱则可以基于图的数据模型，处理复杂多样的关联分析，满足用户对不同实体关系进行分析和管理的需要。

● 数据的反馈速度更快。相比传统的数据存储方式，采取图式的储存，数据调取速度更快。图库可计算超过百万潜在的实体的属性分

布，可实现秒级返回结果，真正实现人机互动的实时响应，让用户可以做到即时决策。

第五节 知识图谱的应用

不知不觉中，知识图谱的应用已经深刻融入了人们的日常生活。在搜索引擎中，搜索结果给出的联想结果往往来自于知识图谱技术的应用；应用软件依据用户的习惯和爱好进行的个性化推荐也来自于知识图谱技术的应用……越来越多的应用场景依赖知识图谱。

一、知识图谱应用于搜索——查询理解

知识图谱的引入使得传统的基于关键词的搜索引擎得到了补充，搜索出来的结果不再是可能包含答案的网页，而是答案本身，知识图谱技术将传统的链接文本转变为链接数据。

Google、百度等搜索引擎巨头构建知识图谱的重要目标之一是令机器能够更好地理解用户输入的关键词。通常，用户输入的是一个短文本，由一个或几个关键词构成，传统的关键词匹配技术并不能理解关键词背后的含义，因此需要用户自己对搜索结果进行筛选确认，查询效果可能会很差。如搜索“珠穆朗玛峰高度”这样的关键词，传统的搜索引擎只能机械地返回所有含有“珠穆朗玛峰”和“高度”这样词的网页，而现在的百度查询不仅会反馈匹配关键词的网页，也会在页面直接呈现结果——珠穆朗玛峰的高度是8844.43m。另外，采用知识图谱理解用户的查询意图，还可以更好地匹配商业广告信息，提高广告点击率。

二、知识图谱应用于回答——自动问答

多年前，很多学者预测，下一代搜索引擎将能够直接回答人们提出的问题，这种形式被称为自动问答。自动问答系统是具有交互形式的进阶版搜索引擎，而知识图谱的重要应用之一就是为自动问答提供知识库。

例如苹果手机的智能语音助手Siri就依托于Wolfram Alpha公司提供的知识搜索技术。为了使对话系统更加准确地给出用户想要了解的信息，其必须依托强大的知识图谱。

三、知识图谱应用于查阅——文档表示

文档表示是计算机自然语言处理的基础，如文档分类、文档摘要、关键词抽取等。经典的文档表示方案是空间向量模型（Vector Space Model），该模型将文档表示为词汇的向量，而不考虑文档中词汇的顺序信息。这种文档表示方案与基于关键词匹配技术的搜索方案相匹配，由于其表示简单，效率较高，是目前主流搜索引擎采用的技术。

然而，经典的文档表示方案未考虑词汇之间的复杂语义关系，往往难以高效处理稀疏短文本。这些缺陷在一定程度上影响了用户的实际应用。但基于知识图谱的文档表示可以将文档表示为知识图谱的一个子图（Sub-Graph），即用该文档中出现或涉及的实体及其关系构成的图表示该文档。这种知识图谱的子图比词汇向量拥有更丰富的表示空间，为文档分类、文档摘要和关键词抽取等应用提供了更丰富的可供计算和比较的信息。这样的文档表示使得一篇文章不再只是一组代表词汇的字符串，还是一张由文章实体及语义关系构成的图谱。

第六节 知识图谱的重要意义

通过知识图谱技术对海量信息进行智能化处理，可形成大规模的知识库并进而支撑业务应用，使得机器能够更好地理解网络、理解用户、理解资源，最终为用户提供新型智能化服务。大规模构建并应用知识图谱，对于互联网行业、传统行业甚至社会治理具有重要意义。

一、提升互联网服务

作为互联网最重要的入口，搜索引擎正在朝着以知识图谱为基础的智能搜索方向发展。智能搜索将极大地提升现有的互联网搜索效果，更好地对接人、信息、服务，从而提升整个以搜索为核心的产业生态的服务质量和效率。基于知识图谱的智能搜索能够以实体为粒度理解用户意图和展现搜索结果，让人和搜索引擎的交互更加自然；开展基于实体的计算，为人们提供更加直接的答案/服务，使搜索引擎更加智能；提供个性化/场景化的搜索结果，让搜索做到千人千面；提供更全面、更有价值的结果，让优质资源更容易被用户发现；最终，使人们更好地获取知识、应用知识、满足所求，使产业和生态以更高的效率、更好的效果获取相应的回报。

二、升级传统行业

除了互联网上的海量数据外，各个行业也拥有大量的行业数据、专业数据。构建行业知识图谱可为传统行业注入新动能，从而升级传统行业。

以金融为例，金融行业拥有大量机构和个人的存贷款、交易、征信、消费、投资数据等。金融知识图谱可以在智能投顾、反欺诈等领域发挥重要作用。

以客服为例，智能客服利用知识图谱技术，根据不同行业、不同企业的信息和知识构建专用知识库，可以提升客服的效率；通过进一步的需求挖掘、产品改进，推荐高匹配度产品，可以提升订单转化率。

以教育为例，知识图谱一方面可以整合海量的教育资源（包括文档、图书、视频、AR等内容），打造系统化的知识网络；另一方面可以根据用户的特点，直接为用户推荐合适的内容和方法，为用户提供个性化的学习方案，满足用户学习所需。

以医疗为例，知识图谱可以整合大量的专业医学书籍、文献、医疗大数据等，打造医疗知识库。根据收集到的患者信息以及自建医疗知识库里面的海量内容，知识图谱可以为医生诊疗提供临床决策支持。进一步地，知识图谱可以打造“医疗大脑”，为患者提供精准医疗。

三、改善社会治理

政府、企业拥有大量的公众数据和公共服务数据。知识图谱作为基础技术，对这些海量数据进行分析，并将结果应用到社会治理和公共服务的各个领域，可以大幅提升全社会的智能化水平，全面提升人民的生活品质。

在工商方面，政府拥有大量的工商企业信息。通过知识图谱技术构建工商图谱，对接消费者和企业，将有效地强化工商治理，打击假冒伪劣商品和虚假违法信息，有利于营造诚实可信、安全健康的消费

环境，最终提高工商监管的效能，形成精准治理、多方协作的工商治理新模式。

在交通方面，打造便捷、安全、高效的智能交通体系是智慧城市的一个重要部分。作为承载人们日常交通出行的重要平台，手机地图、车载地图积累了海量的数据，通过知识图谱等人工智能技术对交通数据、出行数据、兴趣点（Point of Interest，POI）数据进行分析、推理计算，将极大地改善现有交通管理的难题，降低拥堵指数，提升出行效率。

在舆情方面，知识图谱可以对搜索、即时通信、论坛、新闻评论等众多的用户数据进行综合分析，准确感知、预测、预警社会舆论的重大态势，及时把握群体认知及心理变化，进行主动决策反应，提高社会治理的能力和水平。

在法律法规方面，知识图谱可以充分发挥知识汇聚和推理分析的作用，为大众提供快速和个性化的法律法规咨询。

综上，知识图谱技术不仅具有较为广阔的应用范围、明确的市场前景，而且对于提升社会效率、推动经济发展、打造创新国家具有十分重要的意义。

第七节 代表性的知识图谱

在知识图谱技术的发展过程中，有很多值得关注的代表性知识图谱。表1是目前知名度较高的知识图谱，接下来将重点介绍其中的一些关键图谱。

就知识的覆盖范围而言，知识图谱分为通用知识图谱和行业知识图谱两类。通用知识图谱注重广度，强调融合更多的实体，而行业知识图谱受专属行业的限制，具有特定的行业意义。本节分别介绍经典

的通用知识图谱和行业知识图谱，并着重探讨互联网行业的知识图谱以及近年来兴起的中文开放知识图谱联盟中的重要成员。

表1 现有的代表性开放知识图谱

一、经典的通用知识图谱

事实上，在2012年Google公司发布Knowledge Graph产品以前，知识图谱已经出现了，但并没有明确地提出这一概念。从2005年开始，DBpedia、YAGO等项目纷纷创建，这就是知识图谱的雏形。其中，FreeBase、DBpedia、YAGO、WikiData是具有代表性的高质量大规模开放链接知识图谱。

1.FreeBase

FreeBase是一个开放共享的、协同构建的大规模链接数据库，由硅谷创业公司MetaWeb于2005年启动。后来，Google公司在2010年收购了FreeBase，并将其作为Google知识图谱的数据来源之一。FreeBase主要采用社区成员的协作方式进行人工构建，其数据来源包括维基百科、世界名人数据库NNDB、开放音乐数据库MusicBrainz以及社区用户的贡献等。

FreeBase基于资源描述框架（Resource Description Framework，RDF）三元组模型，底层采用图数据库进行存储。它的特点是不对顶层本体做非常严格的控制，用户可以创建和编辑类和关系的定义。

截至2014年年底，FreeBase拥有6800万个实体、10亿条关系知识、超过24亿条事实三元组知识。2016年，Google公司将FreeBase的数据和API服务迁移至Wikidata，并正式关闭了FreeBase。

2.DBpedia

DBpedia是一个大规模的多语言百科知识图谱，可以看作维基百科的结构化版本，由德国莱比锡大学和柏林自由大学的科研人员在2006年开始创建。

DBpedia的产生原因是维基百科的固有结构限制了某些查询需求的实现，如“18世纪之后的意大利作曲家”或“流过莱茵河的所有河”。DBpedia从维基百科的词条里抽取出结构化的知识，以强化维基百科的搜寻功能，并将其他资料集联结到维基百科。DBpedia使用固定的模式对维基百科中的实体信息进行抽取，包括Abstract、Infobox、Category和Page Link等信息，并提供完整的数据集下载。

DBpedia的第一份公开数据集在2007年发布，通过自由授权的方式允许他人使用。截至2017年7月，DBpedia拥有127种语言的458万个实体和超过30亿个三元组。

3.YAGO

YAGO可以看作一个将维基百科和WordNet整合到一起的大规模链接数据库，由德国马克思·普朗克研究所的研究人员于2007年开始创立。

YAGO集成了Wikipedia、WordNet和GeoNames 3个来源的数据，将WordNet的词汇定义与Wikipedia的分类体系进行了融合，使得YAGO具有更加丰富的类别层次结构。

随着时间的推移，YAGO技术不断升级，开发人员为YAGO中的三元组增加了时间和空间知识，为很多知识条目增加了时间和空间维度的属性描述，完成了YAGO2的构建，又利用相同的方法对不同语言的维基百科进行抽取，构建了YAGO3。

目前，YAGO拥有10种语言、约459万个实体和1.2亿个三元组，支持数据集的完全下载，是IBM Watson的后端知识库之一。

4.Wikidata

Wikidata是由维基媒体基金会在2012年启动的协作式多语言辅助知识库，目标是构建一个免费开放、多语言、可编辑的大规模共享链接数据库。项目早期得到了微软联合创始人Paul Allen、Gordon and Betty Moore基金会以及Google公司的联合资助。

Wikidata是维基百科、维基文库、维基导游中结构化数据的中央存储器，支持以三元组为基础的知识的自由编辑。Wikidata中每个实体可以有多个不同语言的标签、别名或描述。截至2016年，Wikidata支持超过350种语言，拥有超过2470多万个实体以及7000万个对实体的描述。

二、经典的行业知识图谱

接下来介绍早期经典的行业知识图谱。行业知识图谱通常需要依靠特定行业的数据进行构建，具有特定的行业意义，实体的属性与数

据模式往往比较丰富，需要考虑不同的业务场景与使用人员。IMDb（Internet Movie Database）、MusicBrainz、ConceptNet是具有代表性的行业知识图谱。

1.IMDb

IMDb是一个与电影、电影演员、电影制作以及电视节目相关的在线数据库，最早创建于1990年，1998年成为亚马逊旗下的网站。其中的资料按类型进行组织，每个具体的条目都包含了详细的元信息。截至2018年6月21日，IMDb共收录了4734693部作品的资料以及8702001名人物的资料。

2.MusicBrainz

MusicBrainz是一个自由的音乐数据库，致力于成为数字音频和视频的元数据库，而不只包含CD曲目信息，它被称为“开放音乐百科全书”。Music Brainz的创始目的是突破CD数据库（CD Database，CDDb）的限制，但如今的目标已经扩大为一种结构化的“音乐维基百科”。MusicBrainz通过数据库和Web服务两种方式向用户社区提供服务。

3.ConceptNet

ConceptNet是一个大规模多语言的常识知识库，最早源于麻省理工学院（MIT）媒体实验室的Open Mind Common Sense（OMCS）项目。ConceptNet主要依靠互联网众包、专家创建和游戏3种方法进行构建，由三元组形式的关系型知识构成。与链接数据和Google知识图谱相比，ConceptNet比较侧重于词与词之间的关系。目前，ConceptNet拥有304种语言的版本、超过390万个概念和2800个声明（即网状图中的“边”）。

三、基于互联网搜索的知识图谱

自2012年Google公司明确提出知识图谱的概念开始，各大互联网巨头开始构建自己的知识图谱。国外以Google、微软为代表，国内以百度、搜狗为代表。

1.Google Knowledge Graph

如前文所述，Google KG技术的目标主要有3个：首先是为用户提供正确的搜索结果，在多重含义下的信息混淆中，为用户找到最想要的答案；其次是为用户提供结构化的总结，Google KG可以更好地理解用户搜索的目的，用户无须通过点击其他链接搜索相关的信息，即可直接在页面右侧看到整合好的结果；另外，Google KG可以帮助用户发掘更深、更广的信息，例如，用户搜索即将去往的目的地，GoogleKG会随之提供相同名字的餐馆、小说、电影等，帮助用户了解更多的知识。

2.微软“概念图谱”

2016年10月，微软亚洲研究院正式发布了微软“概念图谱（Concept Graph）”，该图谱用于提升计算机的语义计算和人际互动能力。

Concept Graph是一个大型的概念知识图谱系统，建立在由微软构建的Probase知识库的基础之上，核心知识库包括超过540万条实体、近亿条关系/属性。

微软还发布了Concept Tagging模型，将文本词条的实体映射到不同的语义概念，并根据实体文本内容标记相应的概率标签，配合知识图谱系统提升计算机的语义计算能力。例如“微软”这个词可以被自动映射到“软件公司”和“财富500强”等概念，并带有相应的概率标签。

3.百度知心

作为百度下一代搜索引擎的雏形，百度知心于2012年年底上线。百度知心致力于构建宏大的知识网络，以图文并茂的方式全方位地展示知识，其特点是对搜索结果进行细致的甄选和干预，并利用数据挖掘技术，将与关键词相关的知识内容聚合在一起，形成知识集群，满足用户的求知需求，实现搜索即答案的效果。

百度知心的具体表现形式是，将知心搜索结果注入阿拉丁系统，形成标准化页面。用户在通用搜索中发出请求后，除了直接给出搜索答案，还以“为您推荐”和“相关搜索”的方式引导用户进入相关知识的页面，这些页面将向用户呈现更丰富的内容和信息。

百度知心的主要定位是以行业为中心的信息整合搜索系统。目前百度知心已有教育、医疗、游戏等多个行业的专属知识集群，其他行业的知识集群也在逐步发展中。

4.搜狗知立方

搜狗知立方是搜狗搜索打造的战略级衍生产品，于2012年11月22日上线。它可以处理海量的互联网碎片化信息，通过“语义理解”重新优化计算搜索结果，向用户呈现最核心的信息。2013年9月，搜狗移动应用产品搜狗语音助手实现了搜狗知立方数据的接入，标志着搜狗知立方正式进入无线领域。

搜狗知立方的目标是令用户的搜索结果更加精准、更加权威、更加全面。搜狗知立方将知识库中的信息转化为用户可以理解的展现内容；为用户提供更多可以直接消费的富文本信息，增添图片、表格等，结果呈现方式不局限于文字；增加更多的用户交互元素，如点击试听等，提升用户体验。

四、中文开放知识图谱联盟

随着知识图谱技术的逐渐兴起和应用深入，国内从事知识图谱研究与开发的学者和机构在2016年共同发起了一个开放的中文知识图谱联盟——Open KG。

Open KG旨在促进中文知识图谱数据的开放与互联，促进知识图谱和语义技术的普及和广泛应用，促进跨领域的交流，让知识图谱能更多地在垂直行业落地。

联盟搭建了Open KG.CN的技术平台，有35家机构入驻，吸引了国内很多著名知识图谱资源的入驻，如Zhishi.me、CN-DBpedia、PKU Base等，并已经有来自于常识、医疗、金融、城市等15个类目的开放知识图谱。Open KG聚集的知识图谱资源规模见表2。

表2 Open KG聚集的知识图谱资源规模

第二章 通用知识图谱的技术要素

通用知识图谱是对全网大数据的挖掘、抽取、清洗、融合、关联和推理，通过知识库、自然语言理解、机器学习和数据挖掘等多种技术，将无序数据变为知识网络，并包含事物及它们之间的联系，以图文并茂的方式展示知识的方方面面。其关键技术可概括为自底向上的4个过程：知识表示与建模、知识抽取与挖掘、知识存储与融合、知识检索与推理。

如前所述，知识图谱构建主要有自顶向下（Top-Down）和自底向上（Bottom-Up）两种方法。两种方法在具体的构建过程中通常不是从零开始的，前者可以利用现有的结构化的知识库，后者则可以从开放链接数据或在线百科中得到很多实体。以下将以通用知识图谱的技术要素为线，简要介绍在通用知识图谱构建方面的主要技术。

第一节 知识表示与建模

一、知识表示

1.知识表示的概念

知识表示从一般意义上来说是将客观世界符号化、模型化，是认知科学、人工智能两个领域共同面对的问题。在认知科学里，它关系到人类如何存储和处理资料；在人工智能里，其主要目的是将人类知

识表示成机器可理解的数据模式，让程序能够存储、处理和运用知识，进而接近人类的智慧水平。

从其表示特征来看，知识表示可分为过程型表示和说明型表示。在过程型表示中，知识是一些客观存在的方法和规则，实现知识表示时将事实型知识和知识推理融为一体，这种方式的特点是推理过程直接、高效，但是灵活性差，知识不易更新；在说明型表示中，知识是一些已知的客观事实，实现知识表示时将推理和表示分开处理，这种方式的特点是简单灵活，但是推理执行效率低。

早期的知识表示方式包括谓词逻辑、产生式、框架和语义网络。在以上几种知识表示方法之上，产生了一种概念化的表示方法，称为本体。它是对领域实体存在本质的抽象，强调实体间的关联，并通过多种知识表示元素将这些关联表达和反映出来。一个本体形式可以由概念、属性、关系、函数、公理、实例构成。用本体表示知识的目的是统一应用领域的概念，构建本体层级体系表示概念之间的语义关系，可以实现对知识的共享和重用。

为了支持知识图谱数据的通用性、复用性和流动性，基于本体的思想，诞生了知识图谱通用Schema标准。其中，cnSchema是一个面向中文的基于社区维护的开放的Schema标准。其分类、数据类型的词汇集包括上千种概念、属性和关系等常用概念定义。cnSchema复用连接并扩展了Schema.org、Wikidata、Wikipedia等已有的知识图谱Schema标准，为中文领域的开放知识图谱、聊天机器人、搜索引擎优化等提供可供参考和扩展的数据描述和接口定义标准。

2.知识表示的原则

知识表示的原则包括以下几个方面。

（1）具备足够的表示能力

针对特定的应用领域，能正确有效地涵盖该领域的各种知识，而且能够处理知识中的模糊性和不确定性。

（2）适合计算机处理

知识表示的最终目的是通过计算机进行知识的分析、处理，因此适合机器推理的表达方式才能挖掘数据的价值。

（3）清晰自然的模块结构

知识库通常要不断地扩充和完善，具有模块性结构的表示模式有利于新知识的扩充及新旧知识的融合。

3.知识表示的形式

（1）谓词逻辑表示

谓词逻辑是一种基于数理逻辑的说明型知识表示，它使用高度形式化的符号语言，通过引入谓词、函数来描述自然语言的知识。利用逻辑公式，人们能描述对象、性质、状况和关系，进而将其转化为机器内部的代码表示。谓词表示逻辑中典型的例子是一阶谓词表示法。

这种知识表示形式具有以下特点：

- 表达自然，逻辑性强，推理严密，易于实现；
- 推理效率低，推理过程中可能产生“组合爆炸”；
- 不能表示不确定知识。

（2）语义网络

语义网络是一种用图表示知识的结构化方式，它是由奎林（J.R.Quillian）于1968年以人类联想记忆的一个心理学模型提出的，之后被用于自然语言理解。它将概念及其语义关系用节点和节点之间的弧来表示，因此又被称为二元关系有向图，节点表示事物、概念、

事件等，弧表示它们之间的关系，常见的关系包括相近关系、推论关系、因果关系、组成关系和属性关系。

这种知识表示形式具有以下特点：

- 具有匹配推理和属性可继承特性，推理效率高；
- 表达直观，方法灵活；
- 知识之间存在层级关系，不利于新知识的添加和维护；
- 推理规则不明确。

（3）产生式表示

产生式是人工智能常用的过程型知识表示，它是由美国数学家波斯特（E.Post）在1934年提出的，被用于构造波斯特机计算模型。产生式表示由事实和规则构成。事实可看成断言一个语言变量的值或多个语言变量间关系的陈述句。事实又分为确定事实和不确定事实。确定事实一般采用三元组表示，即（对象，属性，值）或者（关系，对象，对象）；不确定事实在此基础上增加可信度，即（对象，属性，值，可信度）或者（关系，对象，对象，可信度）。规则主要用于描述知识和陈述各种过程知识之间的控制及其相互作用的机制。规则的形式一般为IF-THEN，它表示一种条件-结果形式。IF后面的部分描述了规则的先决条件，THEN后面的部分描述了规则的结论。

这种知识表示形式具有以下特点：

- 格式固定，形式简单；
- 表达关系自然，符合思维习惯；
- 无法表示知识结构和层次；
- 推理过程烦琐，效率低。

（4）框架表示

框架（Frame）表示方法是一种层次的、组合式的知识表示方法，它是马文·明斯基于1975年提出的，是把对象、概念的所有信息和知识存储在一起的一种复杂的数据结构。其上层主体是固定的，表

示某个固定的概念、对象或事件；其下层由一些槽（Slot）组成，表示主体每个方面的属性。框架方法采用与语言网络相同的图形表示，是一种层次的数据结构，框架下层的槽可以看成一种子框架，子框架本身还可以进一步划分层次。

这种知识表示形式具有以下特点：

- 能够表达知识的内部结构；
- 框架之间可以继承形成框架网络，减小信息冗余；
- 推理过程不够严密；
- 知识适应性差。

（5）面向对象的知识表示

面向对象的知识表示是按照面向对象的程序设计原则，将对象的属性、行为和处理方法进行封装，组成一种混合知识表示形式。在这种方法中，知识的基本单位就是对象，属性集和关系集的值描述了该对象具有的知识，方法集为该对象作用于知识上的处理方法。面向对象的知识表示一般采用四元组模型，即主题=（对象名，属性，方法，接口）。

这种知识表示形式具有以下特点：

- 具有面向对象的继承特性，知识具备层次化和结构性；
- 易于扩充和维护，推理效率高；
- 具备多态特性，适应性强。

（6）基于本体的知识表示

本体表示方法是由语言网络演化而来的，是一种概念化、结构化的表示方法。它是对领域实体存在本质的抽象，强调实体间的关联，并通过多种知识表示元素将这些关联表达和反映出来。一个本体可以由概念、属性、关系、函数、公理和实例构成。

本体表示方法很多，可分为以下3类。

● 自然语言：以自然语言为基础，用语法、语义定义概念和关联。

● 一阶谓词逻辑：以形式逻辑为基础，应用知识概念的逻辑理论描述知识模型。

● 框架和语义网络：以认知理论和认知模型为基础，使本体符合人类认知规律。

本体表示知识的目的是统一应用领域的概念，并构建本体层级体系表示概念之间的语义关系，实现人类、计算机对知识的共享和重用。

这种知识表示形式具有以下特点：

● 表达实体的固有特性；

● 消除领域知识的分歧；

● 方便人与人、人与组织、系统之间的交流；

● 便于共享和重用。

二、知识建模

知识建模是通过基于本体的知识表示方法组织和表达不同类型的知识，利用形式化的知识表示方法获取知识语义信息的过程。它包括从知识获取到知识完成形式化表示的过程，结果在于形成对知识的有效组织和表示，主要包括知识获取、本体构建、基于本体的知识表示三部分内容。知识建模常用的数据模型主要经历了以下3个阶段的发展。

（1）RDF

资源描述框架（Resource Description Framework，RDF）为描述资源提供的基本元素有国际化资源标识符（IRI）、字面值和空白节

点（Blank Node）。IRI是一个符合特定语法的Uinicode字符串，如http://www.w3.org/1999/02/22-rdf-syntax-ns#HTML，与统一资源定位符（URL）的形式类似。URL是IRI的一种，字面值可以理解为类似时间、人名、数字等常量的表示，由字符串和表示数据类型的IRI构成。例如数字1的字面值可以表示为“1”^^xs:integer，其中xs:integer是表示整型数据类型的IRI。空白节点是指没有IRI的匿名节点，一般是RDF内部使用的一个特殊结构，不可被引用。RDF中对资源的描述称为陈述（Statement），一般用（Subject，Predicate，Object）（SPO）三元组（Triple）表示。其中，Subject的取值可以为IRI、Blank Node；Predicate取值为IRI；Object的取值则是IRI、Blank Node和Predicate。例如，“A person named Eric Miller”在RDF中的基本形式为(xs1:me, xs2:fullName,“Eric Miller”)。一个RDF数据集由一组相关的三元组组成。由于这个三元组集合可以抽象为一张图谱，因此也被称为RDF图谱，并通过边将不同的资源链接起来，形成语义网。

值得注意的是，RDF是一种数据模式，而不是序列化格式，其具体的存储表现形式可以为XML、Turtle、N-Triples和N-Quads。RDF基于XML的表述语法，RDF/XML语法是目前唯一一个符合W3C标准的语法。一般可以定义一个简写的前缀表示形式，如xmlns:cd表示http://www.recshop.fake/cd#。接下来每一个资源对应一个<命名空间：资源名称>标签，其中rdf:about给出了该资源的IRI，也就是三元组中的Subject。<rdf:Description>标签里的其他子标签分别对应着Predicate和Object，XML形式紧凑，从图模型的角度分析，它是以顶点为基本单元进行RDF Graph的描述。流行且更常用的格式是Turtle格式，它是RDF 1.1中的标准语法。Turtle中直接以三元组形式进行表示，三元组中的Subject、Predicate、Object之间用空格隔开，用“。”表示一个三元组的结束。为了对同一个Subject的三元组进行简

化表示，允许Subject的省略，同时三元组的结尾用“；”表示省略的Subject与上一个三元组相同。还有两种表示形式，分别是N-Triples和N-Quads。N-Triples是Turtle的简化版，去掉了Turtle中的高级语法，一行就是一个Triple，没有简写的格式。因此，能够处理Turtle的解析器（Parser）同样能够接受N-Triples的数据格式。N-Quads则在三元组的基础上增加了一个维度，成为四元组。新增加的维度表示图谱名称，即四元组所属的RDF Graph的名称，这就能够进一步区分SPO，有利于进行数据融合和管理。

（2）RDFS

资源描述框架模式（RDF Schema，RDFS）是对RDF的一种扩展。如何将文本数据或者现实世界中的知识表示成RDF数据？这就需要RDF字典，即一般所说的数据的模式层（Schema）。模式层在RDF的基础上提供了术语、概念等定义方式，为RDF模型提供了一个基本的类型系统，在RDF数据层之上增加了模式层，为简单的推理提供了支持。通过RDFS可以表示一些简单的语义，但缺少诸多常用的特征，例如对局部值域的属性定义、不相交类的定义等，不足以支持更加复杂的语义场景。例如，用RDF描述一本书，RDF字典就需要定义这本书包含作者、书名、页数、出版时间、语言类型等。RDF字典定义了数据建模的元数据项，这些元数据项主要包括两种类型：类（Class）和属性（Property）。类是指对象实例的集合，可以理解为面向对象编程中的类。属性可分为两种子类型：一种是表示类的属性（Attribute），另一种是表示多个类之间的关系（Relationship）的属性。有了完整的Schema，用户可以方便地将现实中的知识映射成RDF Graph。复用RDF Schema有利于数据的开放共享，同时避免重复劳动。到目前为止，已经有许多定义好的RDF字典，不过英文的居多，例如朋友的朋友（Friend of a Friend，FOAF）、Schema.org等。Linked Open Vocabularies网站专门汇总了互联网上公开的RDF

字典。国内也开始关注RDF字典的标准化，出现了Cnschema。Cnschema主要对Schema.org进行翻译，同时结合中文特点进行定制和扩充，形成了可复用的符合中文事实的知识图谱的数据字典。复用RDF字典可以大大降低知识图谱构建的成本，同时也有利于数据的标准化。

（3）OWL

网络本体语言（Ontology Web Language，OWL）旨在提供一种可用于描述网络文档和应用之中固有的那些类及其之间关系的语言。它在RDFS的基础上进一步扩充，是W3C组织于2002年7月31日发布的本体语言。OWL已经是获得万维网联盟认可的、用于编纂本体的知识表达语言家族。其功能是为网络文档和应用中固有的类以及其间的逻辑关系提供描述，使得基于此技术的网络应用更加人性化和智能化，节省用户的资源搜索时间，并将这些工作交给计算机系统内部处理。基于不同的语义论特性，此家族语言大致分为两个系统：基于描述逻辑进而丰富表达和精准计算属性的OWL Lite和OWL DL、以资源描述架构提供兼容叙述的OWL Full。网络本体语言已经被认为是语义网技术的基础语言，其3种形式（OWL Lite、OWL DL、OWL Full）前者均为后者的子集。OWL Lite被提供给那些只需要一个分类层次和简单的属性约束的用户。OWL DL包括OWL Lite的所有约束，同时逻辑蕴涵是可判定的。OWL Full允许在预定义的（RDF、OWL）词汇表上增加词汇，导致任何推理软件均不能支持OWL Full的所有特征。OWL Full语言上的逻辑蕴涵通常是不可判定的。

第二节 知识抽取与挖掘

一、知识抽取

知识抽取与挖掘指的是从不同来源、不同结构的数据中，利用实体抽取、关系抽取、属性抽取、事件抽取等技术抽取知识。知识抽取技术是指把蕴含于信息源中的知识经过识别、理解、筛选、归纳等过程抽取出来，存储形成知识元库。目前研究较多的是自然语言文本，已经出现了一些工具或系统，知识抽取已经成为自然语言处理领域一个重要的研究分支，它是知识图谱构建的基础，也是大数据时代自然的产物。在互联网信息呈爆炸式增长的背景下，人们需要这样一种从原始数据中提取高价值信息的方法。知识抽取的应用领域非常广泛，包括恐怖袭击预警、空难事故调查、疾病暴发预测等。

1.知识抽取的主要方法

知识抽取的方法主要有以下几种。

（1）词典标引法

该方法的基本思想是：首先构造一个机内词典（主题词典、关键词词典等），然后设计相应算法与词典匹配，若匹配成功，则将其抽出作为文献的标引词。词典标引法在目前汉语自动标引中占据着主要地位，早期的自动标引试验大部分采取该方法。标引算法基本相同，但具体细节有所不同：有的采取最大匹配法，有的采取最小匹配法，有的采取切分抽词和综合加权确定标引词。

（2）切分标记标引法

该方法的基本思想是：将能够断开句子或表示汉字之间联系的汉字集合组合成切分标记词典输入计算机。切分标记词典有词首字、词尾字和不构成词的单字，也有人用表外字、表内字、非用字、条件用字等组成切分词典。当原文本被切分词典分割成词组或短语后，再按照一定的分解模式将其分成单词或专用词。

（3）单汉字标引法

该方法的基本思想是：在标引时将概念词拆分成单个汉字，以单个汉字作为标引词，采取后组方式，将检索词串分解成单个汉字，以逻辑乘关系进行组配，利用汉字索引文件实现自动标引和逻辑检索。

（4）词频统计标引法

词频统计标引法的理论基础是著名的Zipf定律，它建立在较成熟的语言学统计研究成果基础之上，具有一定的客观性和合理性，而且这种方法简单易行，因此在自动标引中占有较重要的地位。国内外很多公司曾使用这种方法进行标引试验，结果证明此法行之有效。词频统计方法要进一步发挥其功能，就必须融合其他因素，因此这种方法目前多被融合到其他标引方法中使用。在加权统计标引法中，从文献频率加权标引到词区分值，加权标引主要依赖于词的频率特征（标引词在某一特定文献中的出现频率或词的文献频率）和词的区分能力。上述两种方法的主要缺陷是与词的相关性无关。而词相关性加权标引法和价值测度加权标引法不仅考虑了词在某一特定文献或整个文献集合中的频率特征，而且考虑了标引词在相关文献集合和无关文献集合中的频率特征以及检索结果的效益值。理论和实践都证明这两种方法比前两种方法更有效。但这两种方法在实际应用中具有一定的局限性，权值函数中的R等值在标引之前是未知的，只能近似估计。

（5）句法分析标引法

基于深层结构的标引法将文献标题可能反映的主题内容归纳为有限的几种元素基本范畴，并使用简洁的句法规则，减少了句法分析的复杂性。数字化指示符和处理码标识的运用更方便了计算机的识别处理。但是这种方法在主题名称的范畴分析及主题标目的选择等方面需要较多的人工干预，影响了其自动标引的效率。另外，这种方法仅以文献标题为标引对象，虽然主题内容容易突出，但标题句法形式的规范性较差，增加了句法分析的难度，同时过窄的分析范围容易漏标一

些相关主题。句法分析标引法获得的一些有效结果通常来自于一些特殊的小量样本，而在大量样本上的试验往往令人失望，最突出的问题是标引词词义的模糊性，而这一问题又是句法分析标引法本身难以解决的。因此，所有的句法分析必须辅以语义分析，才能保证自动标引的准确性。

（6）基于潜在语义分析的标引法

基于潜在语义分析的标引法通过单值分解，将词、文献和提问根据语义相关程度组织在同一空间结构中。在这一空间中，分散在不同文献和提问中的同义词相近放置，具有不同的词但主题语义接近的文献和提问相邻组织。因此，在文献和提问检索词不匹配的情况下，这种方法仍可以给出合理的检索结果，这一点显然是基于关键词的检索系统无法达到的。因为每个词在潜在语义空间中只有一个位置，所以这种标引法目前不适用于多义词。在简化的奇异值分解（Singular Value Decomposition，SVD）描述中，文献集合中一个含义模糊的词将被置于多个独特含义的矩心，这无疑会对检索产生负面影响。尽管这种方法还存在缺陷，但是许多人对其进行试验后认为，潜在语义分析标引法是一种很有希前景的方法。

语义矢量空间模型在现有的矢量空间模型基础上，融入格式语义结构，通过标引词的语义矢量构造描述文献的语义矩阵，使文献的标引得以在语言的深层结构——语义层上实现。相比句法分析标引法，语义分析标引法无论是使用范围还是实际的使用效果都明显优于前者。语义分析标引与人工智能标引的融合将是今后自动标引技术的研究方向。

（7）人工智能标引法

人工智能应用在标引中的具体技术是专家系统，专家系统的知识表示方法主要有产生式表示法、语义网络表示法和框架表示法。基于产生式表示法的JAKS系统，其规则具有统一的条件-行为表示形式，

各自具有自己的功能，这使得知识容易被定义，也容易被理解。而且规则具有高度模块化的性质，系统对规则的定义、修改、扩充等操作可各自独立进行而不互相干扰。但因为规则之间不存在明显的相互作用，所以难以对规则库进行整体把握，这给规则库的一致性维护带来了困难。另外基于规则的推理缺乏必要的灵活性，难以应付复杂内容标引的变动推理方式的需求。

尽管采用人工智能法进行自动标引比在相同专业领域中运用其他方法复杂，但人工智能标引法是真正从标引员思维的角度模拟标引员的标引过程的，这显然比以被标引文献为出发点的其他自动标引方法更有希望获得理想的标引效果。

2.各类数据的抽取方式

知识图谱的典型数据类型可分为三大类，分别是结构化数据、半结构数据和非结构化数据，各类数据的知识抽取方式各不相同。

（1）结构化数据

结构化数据的抽取通常对应两类知识抽取工作：一种是将关系数据库数据映射为RDF格式数据，可采用的标准化工具有Direct Mapping和R2RML，该工作的难点是复杂表数据的处理，例如嵌套表；另一种是从链接数据（通常是已有的通用知识图谱）中提取出一个子集，形成行业知识图谱，其主要实现方式是图映射，即将通用知识图谱映射到定义好的行业知识图谱Schema上，该工作的主要难点是数据对齐问题。

（2）半结构化数据

半结构化数据通常分为两类，分别是百科类数据和普通网页数据。百科类数据（如Wikipedia）知识结构较为明确，易于抽取。基于

这类数据，已经形成较为成熟的知识图谱，如DBpedia和Zhishi.me，其中DBpedia抽取了Wikipedia的知识，Zhishi.me则抽取融合了百度百科、互动百科和中文版维基百科的知识。

普通网页类数据的通用抽取方法被称为包装器，它是一类能够将数据从HTML网页中抽取出来，并且将其还原为结构化数据的技术。包装器的实现方式主要有3种，分别是手工方法、包装器归纳和自动抽取。

半结构化数据可以通过半监督学习的方式进行信息抽取。基于半监督学习的文本知识抽取技术，把蕴含于信息源中的非结构化知识经过识别、理解、筛选、归纳等过程抽取出来，存储形成知识元库。这个过程主要使用了层次类型约束主题实体识别和关系抽取算法。

（3）非结构化数据

典型的非结构化数据有文本、图片、音频、视频等，它们占据了互联网数据的绝大部分。现阶段，人们更多的是从文本这类非结构化数据中抽取知识。信息抽取于20世纪70年代后期出现在自然语言处理（Natural Language Processing，NLP）领域，目标是自动化地从文本中发现和抽取相关信息，并从多个文本碎片中合并信息。文本信息抽取主要由4个子任务构成，分别是实体抽取、属性抽取、关系抽取、事件抽取。知识图谱以图模型进行表示时，实体抽取产生的便是节点；属性抽取构造节点和关系的属性；关系抽取产生的是节点之间的连接边；事件抽取抽取的是文本中的实际实体和事件关系。

实体抽取指的是抽取文本中的原子信息元素，形成实体节点。实体抽取可作为一个序列标注问题，因此可以使用机器学习中的隐马尔可夫模型（HMM）、条件随机场（CRF）、神经网络等算法进行标注。实体抽取要考虑文本分词的特征，包括词本身的特征（例如词性）、前后缀特征（例如地名中会出现省、市）、字本身的特征（例如是否为数字）。特征模型的选择有隐马尔可夫模型、条件随机场

等，目前流行的做法是将传统方法与深度学习结合，例如利用长短期记忆（LSTM）网络进行特征自动提取，再结合CRF模型，利用模型各自的优势进行实体抽取。

关系抽取指的是从文本中抽取出两个或者多个实体之间的语义关系，常见的关系有二元关系、配偶关系、父子关系、雇佣关系、部分整体关系、会员关系、地理坐标关系。例如：“王XX谈儿子王YY：我期望他稳重一点。”这个句子中的关系为“父子（王XX，王YY）”。其中还涉及一个子问题，即共指消解，上述例子中，“我”指的是“王XX”，“他”指的是“王YY”。根据关系抽取方法的不同，可以将其分为：基于模板的方法（触发词的模板、依存句法分析的模板）、基于监督学习的方法（机器学习方法）、弱监督学习的方法（远程监督、Bootstrapping）。

事件抽取指的是从自然语言中抽取出用户感兴趣的事件信息，并用结构化的形式呈现出来。事件通常具有时间、地点、参与者等属性，属性和属性值的抽取能够将知识图谱中的实体概念维度构建完整，事件的发生可能是因为一个动作的产生或者系统状态的改变。事件抽取任务包括：识别事件触发词及事件类型、抽取事件元素，同时判断其角色、抽出描述事件的词组或句子等。事件抽取问题可转化为多阶段的分类问题，需要的分类器包括用于判断词汇是否事件触发词的分类器、判别词组是否事件元素的分类器以及判定元素角色类别的分类器等。

二、知识挖掘

知识挖掘源于全球范围内数据库中存储的数据量急剧增加，人们的需求已经不只是简单的查询和维护，还希望能够对这些数据进行较

高层次的处理和分析，以得到数据的总体特征和对发展趋势的预测。知识挖掘最新的描述性定义是由Usama M.Fayyyad等人给出的：知识挖掘是从数据集中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的非平凡过程。知识挖掘的基本任务是洞察真相、因果推理和规律探寻，其本质是对目标或事件的来龙去脉、前因后果、特点规律进行建模和表现。比如：目标画像，即对目标人物和组织的真实情况、行为模式、社会关系等进行“全景成像”；事件拼图，即通过证据链拟合，按时间轴将事件发生、发展与演变的真实过程进行反演；因果推理，即揭示事件间的因果关系，包括概率因果推理、基于统计相关的预测型因果推理、从海量文本中自动获取因果规则进行因果推理、事件之间发展脉络因果链生成等；规律探寻，即通过模式识别、可视化分析等揭示潜在规律或行为模式。

1.知识挖掘的流程

知识挖掘的步骤如下。

（1）数据准备

知识挖掘的对象是数据。这些数据一般存储在数据库系统中，是长期积累的结果。但这些数据往往不适合直接进行知识挖掘，首先要清除数据噪声和与挖掘主题明显无关的数据，其次将来自多数据源的相关数据组进行合并，然后将数据转换为易于进行数据挖掘的数据存储形式。这个过程就是数据准备。

（2）知识挖掘

根据知识挖掘的目标，选取相应算法及参数，分析准备好的数据，并产生一个特定的模式或数据集，从而得到可能形成知识的模式模型。

（3）模式评估

在由挖掘算法产生的模式规律中存在无实际意义或无实用价值的情况下，也存在不能准确反映数据真实意义的情况，甚至在某些情况下与事实相反，因此需要对其进行评估，从挖掘结果中筛选出有意义的模式规律。在此过程中，为了取得更为有效的知识，可能会返回前面的某一处理步骤进行反复提取，从而提取出更有效的知识。

2.知识挖掘的主要方法

知识挖掘的常用方法如下。

（1）决策树方法

决策树是一种常用于预测模型的算法，它通过一系列规则将大量数据有目的地分类，从中找到一些有价值的、潜在的信息。它的主要优点是描述简单、分类速度快、易于理解、精度较高，特别适合大规模的数据处理，在知识发现系统中应用较广。它的主要缺点是很难基于多个变量组合发现规则。在数据挖掘中，决策树方法主要用于分类。

（2）神经网络方法

神经网络是模拟人类的形象直觉思维，在生物神经网络研究的基础上，根据生物神经元和神经网络的特点，通过简化、归纳、提炼，总结出来的一类并行处理网络。神经网络利用其非线性映射的思想和并行处理的方法，用其本身结构来表达输入和输出的关联知识。

（3）粗糙集方法

粗糙集理论是一种研究不精确、不确定知识的数学工具。粗糙集处理的对象是类似二维关系表的信息表。目前成熟的关系数据库管理系统和新发展起来的数据仓库管理系统为粗糙集的数据挖掘奠定了坚

实的基础。粗糙集理论能够在缺少先验知识的情况下，对数据进行分类处理。在该方法中知识是以信息系统的形式表示的，先对信息系统进行归约，再从经过归约后的知识库中抽取更有价值、更准确的一系列规则。因此，基于粗糙集的数据挖掘算法实际上就是对大量数据构成的信息系统进行约简并得到属性归约集的过程，最后再进行规则抽取。

（4）遗传算法

遗传算法是一种基于生物自然选择与遗传机理的随机搜索算法。数据挖掘是从大量数据中提取人们感兴趣的知识，这些知识是隐含的、事先未知的、潜在有用的信息。因此，许多数据挖掘问题可以看成搜索问题，数据库或者数据仓库是搜索空间，挖掘算法是搜索策略。应用遗传算法在数据库中进行搜索，对随机产生的一组规则进行进化，直到数据库能被该组规则覆盖，从而挖掘出隐含在数据库中的规则。

第三节 知识存储与融合

一、知识存储

知识存储解决如何管理大量的结构化数据的问题。当经过知识提取得到了结构化的数据，并选择了适当的知识表现语法后，下一步就是如何持久性地存储这些数据。我们可以使用不同的数据库工具解决这个问题。现代的关系数据库适用于大多数需要知识图谱的场合。在一些场合，甚至不需要数据库做这些事情，如果数据只需要按属性键（Key）查找而不是按值查找，也不需要连接（Join），那么文件系统就可以作为数据后端。由于太多的小文件会影响查询效率，常见的

做法是把Key做哈希，并将头几个字符取出来作为分子目录。稍微复杂的处理是把Key放在Redis这样的键值数据库里进行管理，把具体的数据放在文件里，这样可以综合数据库和文件系统的优点。而在某些特殊场合中，我们需要图数据库。因此知识存储主要有3种选择：基于表结构的知识存储关系数据库、图数据库和RDF数据库。

（1）基于表结构的知识存储

基于表结构的知识存储利用二维的数据表对知识图谱中的数据进行存储，典型的有关系型数据库、三元组表、类型表。

● 关系型数据库：表中每一列称为一个属性，也称字段，用来描述实体集的某个特征。元组Tuple以表中每一行表示，由一组相关属性的取值构成，相对完整地描述了一个实体。

● 三元组表：作为一种常用的图谱数据模型，在前文已经详细介绍过，这种存储方式简单直接，扩展性强。

● 类型表：在构建数据表时，考虑了知识图谱的类别体系。每个类型的数据表只记录属于该类型的特有属性，不同类别的公共属性保存在上一级类别对应的数据表中，下级表继承上级表的所有属性。类型表克服了三元组表过大和结构简单的问题，但多表连接操作开销大，并且大量的数据表难以进行管理。

（2）基于图结构的知识存储

基于图结构的知识存储利用图的方式对知识图谱中的数据进行存储。图数据库起源于欧拉图理论，也可称为面向/基于图的数据库。图数据库的基本含义是以“图”这种数据结构存储和查询数据。它的数据模型主要是以节点和关系体现的，也可处理键值对。它的优点是快速解决复杂的关系问题。以下为常用的一些原生图数据库。

● Neo4j：是一个开源的图数据库系统，它将结构化的数据存储在图上而不是表中。Neo4j基于Java实现，是一个具备完全事务特性的高性能的数据库，具有成熟数据库的所有特性。

● OrientDB：是一个开源的文档-图混合数据库，兼具图数据库对数据强大的表示及组织能力和文档数据库的灵活性及可扩展性。

● HyperGraphDB：依托BerkeleyDB数据库的开源存储系统，相较于其他的图数据库具有更强大的表示能力。

（3）基于原生RDF结构的知识存储

Weikum在2008年提出了基于原生数据存储格式的RDF管理系统——RDF3x，根据 RISC架构的设计思想，重新设计RDF管理系统，并开发了多个针对RDF的优化技巧，使得RDF3x成为当时单机性能最好的RDF管理系统。RDF3x沿用了传统数据库的查询优化思路，对用户的查询先通过优化器找到一个合适的执行计划、具体的Join顺序，然后再执行查询，获得结果。另外，RDF3x采用精心设计的多种索引结构减少外存的I/O操作，提升了查询性能。

首先，RDF3x将RDF中的一个Triple视为基础元素，把它作为一行数据进行存储；其次，为了降低存储空间，提高访问效率，将RDF中的字符串统一映射为数字ID，形成字典表；最后，设计了15个压缩的聚集B+Tree索引。在15个索引中，有6个SPO排列组合的索引，支持完整的三元组的查找；6个二维索引，支持部分元组信息和统计信息的快速查找以及3个一维索引。Triple在索引中以字典序进行管理，利用合并连接（Merged Join）可以进一步减少I/O操作。

以下为常用的一些开源的RDF数据库。

● RDF4j：它是处理RDF数据的Java框架，使用简单可用的API实现RDF存储，支持SPARQL查询和两种RDF存储机制，支持所有主流的RDF格式。

● gStore：gStore从图数据库角度存储和检索RDF知识图谱数据，支持W3C定义的SPARQL 1.1标准，支持含有Union、OPTIONAL、FILTER和聚集函数的查询操作，支持有效的增删改操

作。gStore单机可以支持10亿级别三元组规模的RDF知识图谱的数据管理任务。

二、知识融合

知识融合概念的发展经历了“数据融合→信息融合→知识融合”的过程，这3个概念并不是完全独立的，涉及的对象和内容有很大程度的交叉。数据融合这一概念起源于1973年美国国防部资助开发的声呐信号处理系统，主要是利用多个传感器的信息进行分析处理与综合，得到对环境或目标的判断测量，并形成综合的发展趋势预测。到了20世纪90年代，由于信息技术的广泛发展，信息融合的概念逐渐取代了数据融合。信息融合是对来自多源的数据和信息进行组合或综合的处理过程，以期得到比单一信息源更精确、更可靠的估计或推理决策。从这个概念中可以看出，信息融合的对象是来自多个数据源的信息或数据，并不局限于来自传感器的数据。知识融合这一说法最早出现于1983年的文献中，但是对这一概念进行关注与研究则是开始于20世纪90年代后期。

知识融合是通过高层次的知识组织，使来自不同知识源的知识在同一框架规范下通过异构数据整合、消歧、加工、推理验证、更新等步骤，达到数据、信息、方法、经验以及人的思想的融合，形成高质量的知识库。知识融合技术产生的原因，一方面是通过知识抽取与挖掘获取的结果数据中可能包含大量的冗余与错误信息，有必要进行清理和整合；另一方面，知识来源广泛，存在重复、良莠不齐、关联不够明确等问题。知识融合涉及的领域比较广泛，包括工业设计、医疗、商业等。知识融合的内涵和外延随着用户需求、应用领域和外部环境的变化而有所差异。在不同的学科领域中，人们对知识融合的认

识有所差异，在医学、工程科学等领域，将知识融合看作信息融合的高级阶段，融合对象通常是指从物理层（如传感器）获得的数据转化而成的信息；在计算机科学、管理学、图书情报学等学科中，知识融合则处于知识科学视角下，知识融合的对象已经不局限于从传感器获取的信息，而是业已形成的知识库，或者从既有的信息库中抽取而来的知识，甚至扩展到各种方法、专家经验等，应用范围非常广泛。这是对知识融合比较粗略的划分，而在实际应用中，不同学科之间会有很多交叉，因此对知识融合概念的认识还应该从更加微观的层次进行剖析。

知识融合通常由两部分构成，分别是本体匹配和实体对齐。

本体匹配是指建立来自不同本体的实体之间的关系，这些关系可以是实体间的相似值、模糊关系等。本体匹配的研究核心在于如何发现异构本体间的匹配关系，匹配发现是实例共享、查询重写、本体集成等应用的基础。从技术实现上，本体匹配可分为元素层面的匹配方法和结构层面的匹配方法。

实体对齐也被称为实体匹配或实体解析，是判断相同或不同数据集中的两个实体是否指向真实世界同一对象的过程。现在实体对齐普遍采用的是聚类的方法，关键在于定义合适的相似度的阈值，一般从3个维度依次进行考察。第一个维度是从字符相似度考察的，基于的假设是具有相同描述的实体更有可能代表同一实体；第二个维度是从属性的相似度考察的，即具有相同属性以及属性词的实体有可能代表相同的对象；第三个维度是从结构相似度考察的，基于的假设是具有相同邻居的实体更有可能指向同一对象。在数据库领域中，对象共指的消解常被称为记录链接、重复检测或记录匹配。在自然语言处理和信息检索领域，常称之为共指消解，属于指代消解中的一类工作。这些工作在数据清洗、数据集成和数据挖掘等方面起着重要的作用。实体对齐存在许多问题和挑战，尤其是在大数据条件下，较突出的是计

算复杂度、数据质量和先验对齐数据的获取问题，这些都需要根据知识库的实际情况设计有效的算法进行解决。在实体对齐算法的选择上，可以分为只考虑实例及其属性相似程度的成对实体对齐和在成对对齐基础上考虑不同实例之间相互关系用以计算相似度的集体实体对齐两类。

第四节 知识检索与推理

一、知识检索

知识图谱的知识是通过数据库系统进行存储的，而大部分数据库系统通过形式化查询语言为用户提供访问数据的接口。知识图谱数据在逻辑上是一种图结构，因此也可以通过图查询技术完成特定查询图的查找，其核心问题是判断查询图是否为图数据集的子图，也叫子图匹配问题。

图数据库查询SPARQL是由W3C为RDF数据开发的一种查询语言和数据获取协议，是被图数据库广泛支持的查询语言。与SQL类似，SPARQL也是一种结构化的查询语言，用于对数据的获取与管理。

同时，随着移动互联网以及可穿戴设备的飞速发展，人们更需要有效、准确的自然语言形式的知识检索方式，信息服务和交互模式开始向问答系统转变。问答系统被称为下一代搜索引擎的基本形态，目前主流的实现方式便是基于知识图谱的。基于知识库的知识问答从技术上可分为以下3类。

（1）基于模板

基于模板的知识问答实现通常由模板定义、模板生成和模板匹配3个部分构成。模板定义通常没有统一的标准或格式，需要结合KG的

结构以及问句的句式。模板的查询响应速度快、准确率高，但为了尽可能匹配上一个问题的多种不同表述，需要建立庞大的模板库。该过程使用人工定义的方式，往往耗时耗力。

（2）基于语义解析

基于语义解析的知识问答实现通常由短语检测、资源映射、语义组合和查询生成4个部分组成。基于语义解析的知识问答实现的主要挑战是开放域环境。

基于语义检索的方法有以下几种。

● 基于IR：基于IR的检索是单一数据结构和查询算法，针对文本数据进行排序检索，达到优化的目的。它的数据是高度可压缩的、可访问的。其可以处理排序，但不能处理简单的查询（Select）、Join等操作。基于IR的检索工具有Sindice、Falcons。

● 基于DB：如Oracle的RDF扩展及DB2的SOR，其具有各种索引和查询算法，以适应对结构化数据的复杂查询。优点是能够完成复杂的选择、合并等操作。缺点是由于使用B+树，空间的开销大且访问存在局限性，同时来自叶子节点的结果没有集成对检索结果的排序。

● 原生存储（Native Stores）：基于原生存储的检索工具有Dataplore、YARS、RDF-3x。该检索方法的优点是高度可压缩、可访问；类似于DB的Select和Join操作；可在亚秒级时间内在单台机器上完成对TB级数据的查询；支持高动态操作。缺点是没有事务恢复等功能。

语义数据搜索具有以下难点。

● 可扩展性：语义数据搜索对链接数据的有效利用要求基础架构能扩展和应用在大规模和不断增长的内链数据上。

● 异构性：语义查询结果的难点包括数据源的异构性、多源性及对多源数据的合并处理。

● 不确定性：对用户输入的问句需要进行词典匹配、消歧和符号匹配等处理。同时，人工编写规则时工程量依旧很大。

（3）基于深度学习

深度学习与基于知识库的知识问答的结合主要有两个方向：对传统问答方法进行改进和直接基于深度学习的端到端（End to End）模型。基于深度学习的方式目前只能处理简单题和单边关系问题。

二、知识推理

知识推理是指计算机在知识表示的基础上进行问题分析、解答的过程，即根据一个或者一些已知条件得出结论的过程。常见的知识推理方式包括以下几种。

（1）语义推理

语义推理是在相应词项的语义系统框架内，借助特定的意义公设，对分析性词项内涵关系的一种概括或描述。这种语义推理是一种必然性推理，其推理的有效性是以正确分析词项的语义结构为基础，以恰当把握词项间的语义关系为前提的。由于语义推理是脱离特定语境而独立进行的，因此它不同于依赖特定语境的语义推理。知识图谱常用于问答系统中，而语义推理是问答系统的一种实现方式，例如“诸葛亮的主公的二弟是谁”，系统会对问句进行基本的语义分析，提取出潜在的实体和谓词，而后转换为实体的关系搜索或属性搜索。

（2）间接推理

间接推理指的是现有数据或图谱中不包含所有可能的逻辑，需要进行多步计算后产生新的推理逻辑。间接推理包括演绎推理（从一般到个别的推理）、归纳推理（从个别到一般的推理）、生成推理（例如聚合计算，统计后产生新的属性）等。

（3）基于规则引擎的推理

规则引擎也称专家系统，是一种固化条件逻辑推理的实现方式。规则引擎可以体现为一种可以嵌入应用程序中的组件，实现了将业务决策或业务标准从应用程序中分离出来，并使用预定义的语义模块编写业务决策的目的。简单来说，就是接受数据输入，通过引擎进行规则分析，据此做出业务决策。

（4）基于表示学习的推理

对图谱中实体的特征学习已经成为一项非常重要的任务。网络表示学习算法将图谱信息转化为低维稠密的实数向量，并将其用作已有的机器学习算法的输入。比如Trans系列的模型，在这个模型基础上进行语义的推理。TransE模型的思想比较直观，它是将每个词表示成向量，然后向量之间保持一种类比的关系，因此它是无限地接近于伪实体的映射向量（Embedding）。这个模型的特点是比较简单，但是它只能处理实体之间一对一的关系，不能处理多对一与多对多的关系。后来Lin Y提出了TransR模型，TransR实际上解决了前文提到的一对多、多对一、多对多的问题，它首先分别将实体和关系投射到不同的空间里面，一个实体的空间和一个关系的空间，然后在实体空间和关系空间构建实体和关系的嵌入，通过它们在关系空间里面的距离，判断其在实体空间里面是不是具有这样的关系。除了TransE、TransR，还有更多的Trans系列的推理模型，如TransH、TransN、TransG等。

节点的表示可以作为特征，送到类似支持向量机的分类器中。主要算法包括矩阵特征向量计算（谱聚类算法）、简单神经网络（DeepWalk算法）、矩阵分解、深层神经网络、社区发现等。

（5）基于图计算的推理

基于图计算的推理是以图论的思想或者以图为基础建立模型来解决现实中的问题，即基于图之间的关系的特征构建分类器进行预测。基于图提取特征的方法主要有随机游走、广度优先和深度优先遍历，

特征值计算方法有随机游走、路径出现/不出现的二值特征以及路径的出现频次等。基于图的方法的优点是直观、解释性好，但缺点也很明显。图计算技术主要是由点和边组成的，主要特点如下：

- 具有较多迭代次数；
- 图计算模型都是将表视图和图视图分别进行实现的，这意味着图计算模型要针对不同的视图分别进行维护，而且视图间的转换也比较烦琐；
- 图计算很难处理关系稀疏的数据，而且很难处理低连通度的图，对于路径特征提取的效率低且耗时长。

图计算中常用的算法有：特征向量分析（PageRank）、聚集度分析（数三角形）、最大连通图（Kosaraju算法）、最短路径（Dijkstra算法）、社群发现（LPA、Louvain）、中心度分析（GN算法）。

常用的知识推理语言为OWL，在本章第二节中已经介绍过，它是知识图谱语言中较规范、较严谨、表达能力较强的语言。基于RDF语法，OWL促进了统一词汇表的使用，定义了丰富的语义词汇，使表示出来的文档具有语义理解的结构基础。

常见的知识推理策略包括正向推理和反向推理。

- 正向推理又被称为数据驱动策略或者自底向上策略，是由原始数据按照一定的方法，运用知识库中的先验知识推断出结论的方法。正向推理的特征体现为：重复利用已知信息，响应速度快；推理目的性不强。
- 反向推理又被称为目标驱动策略或者自顶向下策略，先假设或者结论，然后验证支持这个假设或者结论成立的条件和证据是否存在。如果条件满足，结论就成立；否则，再提出新假设重复上述过程，直至产生结果。反向推理的特征体现为：推理目的性强、建立目标和条件之间的关联时会造成资源浪费。

第三章 行业知识图谱的应用场景

知识图谱技术与行业应用结合后，知识图谱的价值便得到了更大的发挥。不同行业在知识图谱构建与应用方面也体现出了不同的特点。本章选择了一些行业，介绍知识图谱技术在这些行业的应用场景。

第一节 行业知识图谱的特点

行业知识图谱是研究各具体行业知识的技术工具。相对于采集自通用百科或者互联网网页的通用知识图谱而言，行业知识图谱面向的是具体行业，其对象是该行业的各种实体、属性和关系，同时其专业性也相对较强。本章将对行业知识图谱进行简单的介绍，探讨其与通用知识图谱的不同，并重点研究公安、金融、教育、电信等行业知识图谱的应用场景。

行业知识图谱通常具有如下特点。

（1）面向垂直行业

行业知识图谱是面向某一个特定垂直行业的知识图谱，一般是由行业内企业或机构发起构建的，用于行业内各种复杂的分析和决策支持，使用者一般为行业内不同岗位的从业人员，例如证券行业知识图谱、医疗行业知识图谱。行业知识图谱面向垂直行业，因此又叫垂直行业知识图谱。

（2）知识表示深入而全面

如前所述，行业知识图谱多应用于复杂分析和决策支持，因此这对行业知识图谱的数据刻画深度和全面性提出了更高的要求。例如，

在金融企业知识图谱中，描述企业的字段可能有上百个，这些字段对企业涉及的方方面面进行刻画。不仅如此，行业知识图谱的概念层对知识表示的要求也更加严格，实体类型更加丰富，实体属性更加多样，知识表示的行业意义更加凸显。

（3）数据可靠性、准确性要求高

相对于通用知识图谱，行业知识图谱的应用更加深入，因此对数据可靠性、准确性要求很高。有的行业本身对数据要求很高，比如医疗行业知识图谱；有的行业可能涉及相关方的经济利益，比如证券行业知识图谱。因为对概念层的要求更高，所以行业知识图谱的构建需要专家辅助的程度更高。

第二节 公安行业

知识图谱技术在公安行业天然具有高度的场景适配性，拥有广阔的应用前景。

一、行业应用背景

新形势下的公安工作离不开大数据的支撑，公安信息化更是离不开大数据的挖掘和深度应用。在公安信息化建设中，各类结构化、非结构化数据（例如视频、语音等）和半结构化数据（例如半结构化事件文本）广泛存在，现有的公安系统在数据应用中往往只能将结构化数据做简单应用，多数非结构化数据和半结构化数据并没有发挥应有的作用。

利用知识图谱技术，可以通过以下几个步骤完成行业知识的融合，从而形成行业专属的公安知识图谱，帮助公安行业提升智能化应

用的水平。

步骤1 通过对公安行业数据进行认知、质量分析和安全分析，发现有价值的数据，借助大数据接入平台实现对多源异构数据的接入，对数据内容进行提取、清洗、关联和比对处理，通过对公安内外部数据进一步融合，形成标准库、专题库、主题库等；再整合公安行业现有的人、地、事、物、组织、虚拟身份等基本信息，结合吃、住、行、消费、娱乐等数据，将标准库、主题库、专题库等用知识图谱方式打通，在海量多态数据中进行数据模型的统一，从而形成公安数据的统一知识图谱数据模型。

步骤2 通过整合公安数据、社会数据、碎片化数据，构建公安知识图谱，使计算机读懂数据，完成智能识别；发现数据间联系，改变“数据孤岛”状态，将分散的海量多样数据进行智能分析和关联挖掘；并将全量数据归一为易于被计算机使用、易于被业务人员理解的语言和图形，最大化还原数据的本质。

步骤3 在整合各类数据资源的基础上，深入挖掘目标人员、通信设备、网络身份、交通工具之间的关联关系，根据公安实战经验形成研判战法集（算法模型），从海量数据中抽取社会化的复杂关系，构建深度社会关系网络，实现动态更新。

步骤4 对传统应用中难以实时处理的事务，公安知识图谱通过流式处理技术高效构建包含数亿实体和数十亿关系的全量关系网络，并为关系挖掘、路径推演、全文检索、时空分析等相关应用提供实时处理和计算能力，从而极大地提升公安大数据的智能化应用能力。

二、解决方案

公安知识图谱与具体的公安警种业务结合，能够衍生出一系列服务公安机关实战的解决方案。以下列举3种场景的应用。

1.基于公安知识图谱的超级智能检索

基于知识图谱技术打造的超级智能检索系统可以拥有强大的搜索引擎和语义分析能力，可根据用户搜索内容智能识别用户的搜索意图，秒级完成复杂的检索，匹配出最优的结果，并且可按照业务场景预制多项主题进行关联检索。

基于知识图谱技术和公安领域数据特点，公安知识图谱可以在检索服务的过程中打造一系列为公安领域量身定制的行业检索能力。

● 智能意图识别能力。根据用户输入的内容判断用户的搜索意图，并提供符合用户期望的搜索结果，帮助用户智能化处理简单的分析工作。支持识别、查询语句中的身份证号、手机号、银行卡号、地址、单位、日期、邮箱、护照号、车牌号等各类要素。

● 简洁标签检索。基于公安人、事、地、物、组织数据特有的标签特征进行检索，在辅助用户快捷检索的同时，支撑用户基于要素的快速分析研判。

● 数据感知能力。自动梳理数据关系，整合人物档案、事件视图及相关关系。

● 数据整合能力。整合呈现档案视图、轨迹信息、多媒体信息等。

总之，基于公安知识图谱的超级智能检索将在以下应用方向较传统检索技术呈现明显优势：全量公安数据一键搜索、复杂检索秒级高速返回结果、搜索意图识别与自动化推荐、基于意图的信息优先排序、智能范围搜索、操作优化提示、简洁的公安属性标签检索、精确

检索与模糊检索、智能筛选的二次检索、基于结果的统计分析、基于结果的比对碰撞、多源数据的在线接入与融合能力。

2.基于公安知识图谱的全警实战警务大脑解决方案

以科信部门警务大数据平台为基础，结合禁毒、食药环（食品、药品、环保）、经济犯罪侦查、刑事侦查等行动部门的实战经验积累，基于知识图谱技术融合公安轨迹、标签和关系三大数据体系，运用机器学习、模式识别、数据挖掘等先进技术，构建多种专题业务模型，充分发挥大数据技术“海量处理，全量计算”的优势，变未知为可知，实现犯罪行为的精准识别。

● 创新警务大数据服务模式。构建公安知识图谱，汇总融合人、事、地、物、组织、虚拟身份等要素信息，并基于属性联系、时空联系、语义联系、特征联系等建立信息关联，为情报工作的预知、预警、预测等应用提供更高效的服务保障。

● 创新大数据背景下的公安情报内生能力。基于公安知识图谱，针对违法犯罪团伙或隐性人员，通过对团伙（人员）事件、轨迹活动、团伙人员组成进行多个维度的建模，输出嫌疑（隐性）团伙/人员名单，为人工智能研判提供线索，增强公安机关在大数据环境下内生情报的挖掘能力。

● 创新公安大数据使用交互模式。基于已构建的公安知识图谱，结合人类大脑的发散与定向思维，通过可视化的数据特征分析和交互能力，提供推演可视化、比对可视化、多元轨迹时空可视化等人机交互功能，提高公安机关对数据的使用和掌控能力。

3.基于公安知识图谱的情报信息分析挖掘平台解决方案

针对公安情报信息分析挖掘的场景，可以将公安知识图谱作为底层技术支撑，打造一套集全量数据整合治理、指令上传下达、跨平台功能集成、线索共享、面向专项行动合成作战及大数据智能预警等分析应用于一体的解决方案。

方案采用图数据库技术，将公安业务数据治理为“人事地物组织”的知识图谱关系网络，辅以强大的交互可视化设计，向公安行业提供可视化情报综合研判、电子全息档案、团伙挖掘和行为预警、警情笔录等文本信息挖掘与串并案分析等强大功能的应用。

部分非涉密功能应用介绍如下。

（1）可视化情报综合研判

基于高性能公安知识图谱数据库，可以提供全要素数据关系推演分析工具。在数亿级别的实体、数十亿级别的关系网中，进行关系挖掘、路径推演、全文检索、时空分析等运算，提供强大、灵活的交互体验，确保可快速地为研判人员提供对海量多样数据的智能分析、关联挖掘等图形化操作服务，建设盗抢骗、网络诈骗、治安形势监测等各类专题性综合研判分析系统。

（2）电子全息档案

由于公安知识图谱汇聚了所有人、事、地、物、组织等实体的全维度的相关信息，可以在平台里高度集成的一个页面中，对实体的基础信息、轨迹信息、关系网络、历史研判、文档音视频进行汇总显示，减少警务人员跨多个系统查找、关联数据的不便。当实体本身是关注对象时，显示其积分和预警信息。如果在实体的两度关系网络内

存在关注对象，平台会给出提示性信息，方便研判人员发现关联线索。

（3）团伙挖掘和行为预警

针对不同的团伙和群体类型，建立种类丰富的知识图谱关系模型，基于真实世界的关系网络，实现对涉恐、涉毒等团伙和涉稳群体行为的监控和预警。在某省会城市已经上线的系统挖掘出的数十个团伙中，有十余个团伙被一线民警确认为高度可疑团伙，已经进入后续的管控或抓捕流程。

（4）警情笔录等文本信息挖掘与串并案分析

通过自然语言分析，实现对警情笔录、情报信息等文本数据的自动分类，通过文本特征提取以及搭建模型实现对案件笔录的串并案分析。同时，通过对文本中的实体关系提取，并结合结构化数据，构建知识图谱应用，使得文本数据信息变得更加立体。基于文本分析，发现刑事案件描述、重大事件情报中的案-案、人-案、事-事、人-事关联，自动实现案件的串并，提供破案关键线索。

第三节 金融行业

一、行业应用背景

伴随着科技发展，现代金融已经转换为数据密集型和科技驱动型行业。面对金融科技带来的业务多样化、主体多元化、场景丰富化等行业新态势，银行面临着重大机遇，也同时面临着诸多挑战。在宏观层面，国家以防范金融风险为目标，不断加强监管；在微观层面，面对互联网金融带来的业务冲击，越来越多的银行通过云计算、大数据、人工智能等技术手段寻求创新，提升业务能力。

随着银行IT架构的转型升级，大数据建设的不断完善，利用大数据、人工智能技术挖掘数据价值，实现科技赋能金融、辅助客户管理、精准营销、风险管理、运营优化等场景下的智能业务决策，是未来金融信息化发展的必然方向。

要实现辅助智能决策，首先需要让数据以接近人类认知的方式进行存储，再将业务经验、专家经验、机器学习融合，最终实现银行各条线的业务效率提升。知识图谱抽取数据中的“实体-关系”，将数据以更接近人类认知的形式呈现。在近些年，人工智能技术发展极其迅速，是金融行业颇为关注的技术之一。在金融行业，知识图谱也已经应用于诸多场景。

二、应用场景

1.客户风险监测与预警

基于金融知识图谱平台，利用对公贷款相关客户关系管理系统（CRM）、信贷系统、押品系统、“三查”报告等多源异构行内数据以及企业工商信息、法律诉讼、舆情资讯等外部数据，构建以企业为主要实体，以企业间股权关联、担保关联、资金流转关联等为主要关系的机构/行业知识图谱，支撑对公客户风险视图、贷前风险识别、贷后风险监控预警、风险传导分析等风险管理应用，在贷前、贷中、贷后环节全面提升风险管理能力。

2.内部审计

基于知识图谱平台，构建各业务条线数据、内部分支机构和员工相关数据的审计知识图谱，可以提供与审计计划、审计项目相关的监控预警和追踪查证。审计知识图谱通过强大的分析和挖掘能力、友好并贴合业务的人机交互体验，帮助审计人员高效应对审计业务，为银行建立智能审计科技体系。

3.智能投研

利用投资产品、产品相关主体、投资者以及公告、新闻、行研等外部资讯，构建投资知识图谱，帮助金融机构内部投资顾问和客户经理及时、有效地洞察投资机会，使其能够选择与投资能力、风险偏好、风险承受能力匹配的投资者进行产品推荐。

4.反欺诈

金融知识图谱针对金融行业零售业务的反欺诈场景，可以基于现有黑名单机制，建立反欺诈知识图谱和应用。在业务申请端和交易端，通过图分析的技术手段判断业务相关账号和交易本身与黑名单的关系，提升欺诈识别的覆盖度和有效性。

5.智能营销

集成工商、司法诉讼、舆情等企业相关外部数据，支持与银行对公CRM数据打通，建立对公营销知识图谱，分析隐性的集团企业、利

益共同体、上下游企业，帮助金融机构打通内、外部数据，分析和理解客户特征，识别和预测潜在目标客户，基于行内存量客户情况和外部营销信号，为一线客户经理智能推送营销信息，提升营销转化率和执行效率。

第四节 教育行业

近年来，随着互联网的发展，大数据成为各行各业重点研究和革新的重要手段。教育行业被认为是大数据技术可以大有作为的一个重要应用领域。

一、行业应用背景

教育包括学前教育、中小学教育、高等教育、职业教育等。本节将重点放在知识图谱在中小学教育中的应用。

中小学教育资源丰富，包含学科（语文、数学、英语等）、资料（试卷、视频、课件、教案等）、精品课程、专辑等。如何利用这些资源进行有效的学习，成为教育类公司亟待探索和解决的问题。

教育知识图谱的构建主要解决教育机构的如下几个需求。

（1）知识库建设

知识库是教育行业中最重要的基础资源库，所有的应用开发都要建立在知识库的基础上。知识图谱可以作为知识库的内容组织框架结构，将各类资源链接到相应的知识图谱节点上，从而为应用的进一步开发奠定基础。

（2）语义搜索

● 资源查询：输入要查询的资源标题，展示资源基本属性以及相关联的知识点、教材信息。

● 知识点查询：输入要查询的知识点名，展示相关联的资源节点。

（3）知识导览

根据用户输入的资源或者知识点，将关联数据按照知识图谱的方式进行展示，方便用户按照关联关系学习。

二、解决方案

教育知识图谱构建主要考虑数据层、数据抽取、数据融合3个阶段。本书涉及的数据主要是半结构化的教育行业数据。

知识图谱系统采用的底层存储是MongoDB分片集群和Neo4j，逻辑存储采用的是以RDF、Schema为主的表示方法。

知识图谱查询主要考虑实体查询、关系查询、实体链指等问题。

知识图谱展示主要通过API服务与图谱可视化两种方法实现。

（一）知识图谱构建

1.结构化数据构建

教育资源包含中小学教材、试题试卷、课件、教案、素材、视频、作业、学段、知识点、名校资源、精品课程等，资源的来源多样，不同机构都有自己的资源。有效地整合资源，使其便于学习变得非常重要。表3是资源一般涉及的相关数据字段。

表3 资源涉及的相关数据字段

在学习过程中，学生更多地按照教材、知识点等进行关联学习、串联资源（课件、习题、教案、素材）等。因此对教学资源的组织重点应放在资源、教材、知识点、学段实体以及相互之间关联关系的构建上。

知识点按照逻辑关系组建为树状结构，与资源、学段（年级、科目）、教材建立、知识点关联关系形成教育知识图谱的基础图谱。构建教材、资源（课件、习题、教案、素材）、知识点、学段、学员5种实体。构建关系有：（知识点，父节点，知识点）（知识点，子节点，知识点）（知识点，包含，资源）（资源，包含，知识点）（资源，属于，教材）（资源，属于，学段）（教材，包含，资源）（教材，属于，学段）（学段，包含，教材）（学段，包含，知识点）。

学员对资源、知识点、学段、教材的浏览、学习、评价、下载、点击量等形成知识图谱的动态部分。动态图谱相当于事件信息，需要从文本日志中抽取特定的事实信息。这个过程用到了命名实体识别、共指关系确定、模板元素填充等技术。

2.数据融合

半结构化数据是人工标注的，难免存在问题，因此要对数据进行优化，包括属性融合、日期属性值归一化、值分割。

属性融合包含添加候选关系、删除错误关系。添加候选关系包含人工规则、外部知识库、属性名称相似性等方法。删除错误关系包含人工规则、实体重叠度等方法。

3.数据抽取

资源、知识点为结构化数据，但是标注数据有限，需要从其他数据中补充一些知识点、关联关键词。首先，可以定制一系列规则，按照规则进行信息抽取，将所有可能的知识点、关键词数据抽取出来；然后，通过知识关联性、文本相似度等手段判断抽取结果的优劣；最后，管理员人工审核并入库。

实体抽取主要采用基于规则的方法或基于统计机器学习的方法。基于规则的方法主要根据数据的特点进行分类，并构建正则规则。基于统计机器学习的方法主要解决教育领域的实体识别问题。

（二）知识图谱查询

教育知识图谱可以进行实体查询、实体间关系查询和可视化展示等。

（1）实体查询

在教育知识图谱中，输入要查询的知识点名，相关联的资源节点以及父知识点、子知识点将会被展示出来。知识点按照TF-IDF值计算得分由高到低排序。

（2）实体间关系查询

给定两个实体，通过RDF图遍历、计算两个实体的关联度。这里主要通过定制推理规则，确认边遍历深度及边权重，确认两个实体的关联度。

（3）可视化展示

根据知识图谱返回接口格式，可以定制图谱展示方式。

（三）知识图谱更新

当需要更新时，用户只需要将更新的原始数据按照约定的格式传送到知识图谱系统，就可以实现自动更新。更新支持以下两种方式。

- 主动更新：根据需要，提供一个实时主动触发更新的机制。
- 被动更新：定期（周期可配置）从文件传输协议（FTP）中读取导出的数据库文件，更新数据。

三、应用价值

（一）语义搜索

传统搜索只能针对关键词进行匹配，针对特定字段进行筛选，并不能进行语义层面的处理。知识库是语义搜索进行推理和知识积累的基础，我们考虑将知识图谱引入搜索，建立语义关联，达到根据知识点搜索资源、根据资源搜索知识点的目的，同时可以引入基于图谱的问答，对部分搜索结果产生所问即所答的效果。例如，用户输入“西游记”，传统搜索只能找到包含“西游记”的文本。而利用知识图谱可以知道“西游记”的主要人物有“孙悟空”，从而也可以将与孙悟空相关的资源找到。另外，传统搜索没有知识点、资源的概念，只做文本匹配。利用知识图谱，我们明确知道用户搜索的实体类型，可以根据用户搜索实体类型做语义扩展。例如，用户输入“浮力”时，可以判定用户搜索的是一个知识点；当用户输入“验证阿基米德原理”时，可以通过语义搜索判定用户搜索的是一个实验。最后，当用户输入“关于浮力的资源

有哪些？”时，利用知识图谱问答可以判定用户的意图是搜索浮力知识点对应的资源，从而可直接将结果反馈给用户，提高效率。

（二）智能推荐

智能推荐针对用户当前的学习内容、历史浏览记录，推荐关联资源、知识点。教育知识图谱可以利用知识推理、知识关联度计算，给用户推荐其想要学习的内容，如图5所示。

图5 教育知识图谱示意图1

传统在线教育通过分类和搜索完成资源的查找和学习，总体上学习效率比较低，形式没有新意。推荐展示可以用知识图谱可视化的方式进行图谱导览学习，增加趣味性，如图6所示。

图6 教育知识图谱示意

（三）个性化学习

知识图谱可以针对个人构建模型，实现从群体教育转变为个性化分析的目标。利用大数据技术，可以关注每一个学生的学习记录，包括其点击哪个资源、学习多长时间、习题作答的准确程度、每个习题用时多少、学习路径记录等。利用学员的个人信息、学习目标以及上述信息等数据，将知识图谱、机器学习、深度学习结合起来，分析学员的知识结构，进而给出推荐学习路线、推荐需要加强学习的资源、生成需要加强的习题集等针对性训练。

例如，用户A的学习目标是高考，根据知识图谱可知高考的科目、每个科目大纲对应的知识点、高考的历年试卷以及对应知识点出现的次数等信息。另外，我们知道用户最近一段时间学习的知识点、资源形式以及最近模拟习题的解答情况、答题时间等。根据知识图谱推理，可以确定用户掌握的知识点、没有学习的知识点、没有掌握的知识点、常出错的知识点，最终给出用户的学习计划以及需要加强学习的推荐列表，辅助用户A顺利通关。

第五节 电信行业

知识图谱为各行各业海量、异构、动态的大数据的表达、组织、管理以及利用提供了一种更为有效的方式，使得网络的智能化水平更高，更加接近人类的认知思维。在电信行业内，知识图谱可以和电信业务应用充分融合起来，广泛地应用在智能客服、电信反欺诈等方面。

一、智能客服系统

1.行业需求

各大电信运营商的客服呼叫中心以人工服务为主，客服人员必须随时接听全球用户的电话。他们会根据用户的问题在预先安排好的知识库中搜索相近的问题，依此来回答用户的问题或者受理用户的投诉。但是随着电信业务的不断增加，业务场景越来越复杂，客服人员处理问题的数量和种类都变得越来越多。电信运营商迫切需要一种智

能化的客服系统，因为同其他行业相比，电信行业的客服系统面临着更大的困境。

第一，电信运营商的客户分布在全球，每天服务的人数不少于10亿，这些客户每天产生众多的问题。因此电信的客服人员需要全天候地为客户提供高质量的服务。

第二，随着4G、5G和宽带业务的快速发展，电信运营商的业务线变得更加复杂，人工客服要不断地更新专业知识，才能应对各种各样的业务场景。

第三，电信运营商的客服人员数量比较多，培训成本比较大，同时客服人员加班的成本也非常高昂。

因此，传统的人工客服正遭遇着效率低、成本高、用户满意度低以及营销转化率低的困境。可以说，传统的人工客服是一个低技术、高成本的密集型行业，不适合当前的经济环境。

2.解决方案

随着机器学习以及人工智能的发展，智能客服系统具有传统客服系统无法比拟的优势：①提高服务效率，分流客服人员的压力，节约成本，减少重复劳动力；②通过客服系统的日志，分析用户投诉的数据，对产品进行优化迭代；③智能客服系统可以通过多种方式提供服务，包括微信、短信、电话以及语音等方式。因此，如何构建一个智能客服系统是电信运营商越来越关心的问题。

目前，客服问答系统主要是基于自然语言理解技术和知识图谱技术的。首先分领域建立本体的知识图谱；然后通过引导式对用户的问题进行确认，采用推理技术对问题进行识别；最后在知识图谱中智能检索出答案。知识图谱是关键技术之一。

3.知识图谱的构建

知识图谱的构建分为三大部分：知识获取、知识融合、知识加工。知识获取是指从非结构化、半结构化以及结构化的数据中获取知识；知识融合是指将从不同数据源获取到的知识进行融合，构建数据之间的关联；知识加工是指从融合的知识库拓展出新的知识。

4.知识获取

知识获取面向开放的链接数据。这数据包括：① 百度百科等网页数据；② 电信设备厂商提供的数据；③ 电信领域内人工录入的专业数据；④ 第三方的数据接口提供的数据。使用自动化或者半自动化的技术抽取关键的知识单元。知识单元包括实体、关系以及属性。将知识单元存储在各种数据库中，如Redis、MongoDB、Neo4j等。Redis主要用来存储热数据，MongoDB主要用来存储实体和属性，Neo4j主要用来存储关系。

（1）实体获取

实体获取将文本中具有某个特定含义的实体识别出来。在电信领域内，命名实体识别主要是从一些语料中识别出电信领域的相关实体。

构建电信领域的实体库是一个持续的过程。首先通过专业的电信人员建立一个小规模的电信实体词典；然后从网上爬取大量关于电信领域的文献资料，利用已有词典提取特征，使用统计方法中的各种模型进行训练，得到更多的新实体。通过这些方法，可以不断积累更多的实体，建立更加庞大的实体库。

（2）属性获取

属性获取的任务是为每个实体构造一个属性列表。例如，在电信领域内，实体属性包括：身份证号、身份证上的归属地、出生日期等。电信实体的属性可以从套餐业务表等结构化数据中解析出来。

（3）关系获取

关系获取的目标是解决实体语义连接的问题。在电信领域内，将实体之间的关系分为3种：顺序关系、共现关系、主从（父子）关系。

- 顺序关系：在电信领域中，顺序关系可以用来进行实体之间的相互推荐，比如有人先问剩余流量，再询问加油包。因此，需要抓住用户询问的先后次序，建立实体之间的顺序关系。
- 共现关系：在电信领域中，不同的实体经常一起出现，这些实体之间具有很强的关联性，这种关系被称为共现关系。顺序关系是共现关系中的一种。不同于顺序关系，共现关系不仅可以从用户的提问中获取，还可以从各种文本中获取。在顺序关系中，实体之间存在先后次序；在共现关系中，实体之间不存在先后次序。
- 主从关系：主从关系是指实体之间具有隶属关系。例如在电信领域内，主卡和副卡是主从关系，主卡控制副卡，主卡可以将流量、通话时间以及短信共享给副卡。

专业领域的关系模式较为固定，可以通过统计算法找到关系模式，使用这些关系模式进行关系抽取。目前，基于深度学习的抽取算法非常热门，它能发现人工未能定义到的关系模式。

5.知识融合

通过知识抽取，人们实现了从非结构化和结构化数据中获取实体、关系以及属性的目标。但是由于知识来源广泛，知识存在质量良莠不齐、来自不同数据源的知识重复、层次结构缺失等问题，因此需要进行知识融合。知识融合是高层次的知识组织，使得来自不同数据源的知识在同一框架规范下进行异构整合、消歧、加工以及推理。知识融合技术主要包含实体链接、知识加工。

（1）实体链接

实体链接也称为实体匹配或者实体解析，主要用于消除异构知识中实体冲突、指向不一致的问题。例如在电信领域中，亲情卡和主副卡是一个概念，但在不同套餐中其含义是不一致的。在进行知识库实体链接时，主要面临以下3个问题：算法的复杂度高、数据质量差、先验数据缺乏。

实体链接的主要步骤如下所示。

步骤1 为待对齐的数据建立索引，降低计算的复杂性。

步骤2 利用相似算法查找匹配实体。

步骤3 使用实体对齐算法进行实体融合。

步骤4 将步骤2和步骤3的结果结合起来，形成最终的结果。

（2）知识加工

从原始语料中抽取实体、关系以及属性等知识要素，再经过知识融合，可以消除实体歧义，得到一系列基本的事实表达。然而事实本身并不等于知识，要想获得结构化、网络化的知识体系，还需要经历知识加工的过程。知识加工包含本体构建与知识推理。

本体是对概念进行建模的规范，是描述客观世界的抽象模型，以形式化的方式针对概念及其之间的联系给出明确的定义。本体可以采用人工编辑的方式手动构建，也可以以数据驱动的自动化方式构建。其包含3个阶段：实体并列关系相似度计算、实体上下位关系抽取以及本体生成。

实体并列关系相似度是指两个实体在多大程度上属于同一概念分类的指标测度。实体并列关系相似度的计算方法有两种：模式匹配法和分布相似度。

实体上下位关系抽取是该领域的关键，主要方法是基于语法模式抽取。

本体生成阶段的主要任务是对各层次得到的概念进行聚类，并且对其进行语义类的标定。

6.知识推理

知识推理是指从知识库已有的实体关系数据出发，进行计算推理，建立实体间的新关联，从而丰富知识网路。知识推理是知识图谱构建的重要手段和关键环节，通过知识推理，能够从现有知识中发现新的知识。

知识的推理方法可以分为两大类：基于逻辑的推理和基于图的推理。基于逻辑的推理主要包含一阶逻辑谓词、描述逻辑以及基于规则的推理；基于图的推理方法主要基于神经网络或者Path Ranking算法。

7.知识问答

在知识问答系统中，系统会先在知识图谱的帮助下对用户使用自然语言提出的问题进行语义分析和语法分析，进而将其转化成结构化形式的查询语句，然后在知识图谱中查询答案。对知识图谱通常采用基于图的查询语句（如SPARQL），在查询过程中，通常会基于知识

图谱对查询语句进行多次等价变换，例如，如果用户提问：“月末流量不够用，该使用什么样的套餐？”，该问题有可能被等价变换为“月末流量套餐有哪些优惠？”后再进行推理变换，最终形成等价的三元组查询语句，如（月末流量套餐，优惠，？）。据此进行知识图谱查询，并得到答案。问答系统大致分为2类：基于语义分析的问答系统和基于深度学习的问答系统。

（1）基于语义分析的问答系统

对于给定的问题，首先利用对齐规则将问题中的实体、关系词、疑问词映射成知识库中的实体与关系谓词；然后将相邻的实体、关系谓词进行桥接，产生新的谓词；最后将问题中的所有谓词取交集形成一个精确的查询语句，再直接利用该查询得到答案。

（2）基于深度学习的问答系统

利用卷积神经网络和循环神经网络，把一个问句转换成一个向量的形式，同时通过图谱的表示学习，把知识图谱中所有实体或者关系表示成一个向量形式。现在整个问答的过程就是一个检索的过程。使用问句的向量在这个知识图谱向量中查询，找到距离最近的实体或者关系向量，对应的实体就是当前问句的答案。

8.图谱展示

为了方便计算机的处理和理解，使用三元组表示知识，例如手机号码是一个实体，归属地也是一个实体，两个实体之间的关系称为归属，三者构成三元组：（手机号码，归属，归属地）。根据这个关系，归属地确定下来后，该手机号码能够办理哪些业务套餐也可随之确定。每个套餐都有一些属性，例如：套餐价格、流量余额以及通话时间。综上所述，可以构建一个如图7所示的电信业务的知识图谱。

图7 电信业务的知识图谱

二、电信反欺诈

1.行业需求

电信诈骗是指不法分子通过电话、网络和短信等方式，编造虚假信息，设置骗局，对受害人进行非接触式诱骗，诱骗受害者给不法分子打款的犯罪行为。2016年12月20日，我国发布的《关于办理电信网络诈骗等刑事案件适用法律若干问题的意见》中规定，利用电信网络技术手段实施诈骗，诈骗公私财物价值3000元以上的可判刑，诈骗公私财物价值50万元以上的，最高可判无期徒刑。但是对电信诈骗的打击、预防较为困难。电信诈骗具有非接触性、范围大、诈骗手段花样迭出、团伙作案以及犯罪分子具有一定的反侦察能力等特点。因此，对电信诈骗的研究虽然成为热点，但都缺乏对电信诈骗的全面认识。知识图谱能对研究的问题进行整体的可视化分析。将知识图谱应用于电信反欺诈成为一个重要的研究热点。本节重点介绍知识的原始数据和关键共词分析。

2.原始数据

首先，知识图谱的原始数据来源于中国期刊数据库（CNKI），通过检索“电信诈骗”搜索到560篇相关文献。这些文献集中在2010—2017年。用Refworks格式导出CNKI的文章。然后用CiteSpaceIII工具

对数据进行转换。CiteSpace是美国德雷赛尔大学信息科学与技术学院陈超美博士和大连理工大学WISE实验室联合开发的科学文献分析工具。CiteSpace基于Java平台，能将数据可视化展示，以发现数据的规律，通过这些规律支持用户做出决策。

3.关键共词分析

基于知识图谱的电信诈骗的研究思路是对关键共词进行分析。从关键词出现的频次以及中心性来分析关键词之间的关系。关键词是文献的重要组成部分，通过关键词可以了解文献研究的内容和领域。共词分析是一种文本分析技术，它研究的是同一文本主题中同时出现的单词。这些共词可以用作确认文本代表的学科领域中相关主题的关系。电信诈骗文中的高频词反映了研究热点；中心性反映了该关键词在知识图谱中作用的大小。关键词的中心性越强，该关键词在知识图谱中与其他的关键词共同出现的概率越大。因此频次高和中心性强的关键词反映了这段时间内研究的热点和前沿。

从原始数据中，使用上节介绍的图谱构建技术，构建了如图8所示的电信诈骗的知识图谱。

图8 电信诈骗的知识图谱

从这个简易知识图谱中，可以挖掘出以下几点。

- 当前的诈骗犯罪主要是电信诈骗，由于电信诈骗是非接触的，被识破后罪犯也难以被立即抓获，而且电信诈骗多为跨省，甚至跨国。
- 电信诈骗多为团伙作案，团伙成员冒充公检法机关人员或者扮演客服人员进行电信诈骗。

● 随着移动电子支付的发展，网购和二维码奖品成为电信诈骗的重灾区。

● 电信诈骗的主要原因是公民个人信息泄露，而虚拟运营商是个人信息泄露的主要渠道。

● 在电信诈骗中，受害者大多通过银行汇款给诈骗者。因此，从2016年12月1日开始，客户通过ATM转账时，资金都是在24小时后转出的，在24小时内，客户可以随时取消该笔转账。

第六节 工业

工业生产中会产生海量的数据，这些数据的主要来源有两类：第一类是与生产经营相关的业务数据，如企业资源管理（ERP）数据。第二类是设备物联数据，如工业产业链的各个环节的条形码、二维码、RFID、工业传感器、工业自动控制系统等产生的数据。如何组织和管理工业大数据是“工业4.0”需要解决的核心问题。知识图谱可以作为其数据建模的技术手段。构建工业知识图谱是企业实现自动化、智能化的基础。

一、工业知识图谱构建

工业知识图谱构建主要包括以下两部分。

1.结构化知识抽取

工业生产管理过程中，人、机、料、法、环、仓等环节产生的海量结构化数据经过清洗后会形成标准的基础性和可操作性数据，并统

一载入工业制造业数据仓库。在此基础上，建立不同层级的工业标准数据模型和数据计算。基于规则的和基于深度学习的知识抽取技术，从这些标准的、多层级的工业数据中识别实体及其属性，并在多个实体之间建立关联关系。

2.非结构化知识抽取

在工业制造过程中，会产生海量的非结构化数据，如生产线流动物品检测得到的图像数据、机械运动流的视频数据、针对工作人员的操作规范产生的视频数据、机械声音数据、车间环境动态图像数据、非标准化的文本交互数据等。针对这些非结构化数据，采用基于深度学习的图像、文本处理技术，使之转化为知识图谱的实体-关系。这些非结构化数据与工业生产管理过程中各环节产生的结构化数据有紧密的联系。

二、工业知识图谱应用场景

（1）可视化展示

基于知识图谱，可以提供数据的可视化展示，比如通过一个设备编号，知识图谱会呈现与之相关的状态信息、其他相关的设备和人员信息。通过知识图谱的可视化展示把复杂的信息非常直观地呈现出来，使得人们对工业生产整体关联的情况一目了然。

（2）生产故障诊断

工业生产制造过程在人员生产操作、机械设备运转、企业生产信息管理三者结合的状况下，较难做出精细的防呆机制设计。因此，在生产环节的人与设备、设备与企业生产信息管理、人与企业生产信息

管理之间的交互过程中出现的各种异常状况都可能导致生产流程中断，影响生产效率和进度。工业知识图谱的构建过程，使得工业生产中的人员操作流、机械设备运转流、企业生产管理信息流三者有机地融合在一起，其间形成的大规模、多层级、立体化、标准化的数据能够帮助工业企业最大限度地建立丰富的知识图谱实体库、属性库和实体间关联关系。借助知识图谱，制造企业的生产管理人员可以迅速地从宏观数据层面发现故障位置，并下钻到微观数据层面，定位故障原因。

（3）产品质量追溯

针对工业生产制造的产品生命周期，基于零部件物料采购、仓储、工艺流程设计、生产加工、组装、包装、入库、物流、销售、维保等一系列流程相关的数据集建立起来的知识图谱可以帮助制造企业追溯每一件产品的完整制造过程和生命周期，建立健全产品质量追踪和改进机制，在新产品研发方面也能发挥重大的作用。

（4）机械设备寿命预测

生产线上的各种设备及其零部件在重复的生产中产生了海量的时间序列数据（状态数据、运动流数据、局部和全局图像数据、视频数据）。通过大数据技术和机器学习的相关算法，可以计算出机械设备的故障规律。每台设备及其零部件的早期故障期、偶发故障期、严重故障期规律各不相同，把这些实体关系纳入工业制造知识图谱后，可以帮助企业预测各台机械设备及其零部件的使用寿命，做到提前防呆，减少机械设备故障，降低安全事故。

（5）辅助柔性生产

制造企业想要高效地实现多品种、小批量的柔性生产，需要人、机、料、法、环、仓等多个环节的精密配合。知识图谱的应用可以帮助企业更加便捷地利用这些环节中大量的知识实体及其关联关系，提升生产的灵活性和应变能力。

第四章 知识图谱的发展趋势与挑战

在对通用知识图谱与行业知识图谱的相关内容进行介绍后，本章对知识图谱这一技术工具及其应用领域的发展趋势与挑战进行了简要探讨。知识图谱如何发展？知识图谱面临着哪些困难和挑战？本章将给出作者的一些思考。

第一节 知识图谱的发展趋势

随着关注度越来越高，知识图谱的发展正呈现出诸多趋势。针对基础理论和应用技术，人们展开进一步的研究。同时随着技术的发展和广泛的关注，知识图谱已经从学术研究逐步转移到行业应用中，落实在相关产业发展，应用领域也日趋广泛。目前，知识图谱技术正在呈现如下趋势。

一、与机器学习相互渗透融合

近几年在知识图谱技术的推动下，对机器友好的各类在线知识图谱大量涌现。但是这些蕴含人类大量先验知识的宝库却尚未被深度学习和有效利用。现阶段，越来越多的厂商开始将机器学习技术应用到知识图谱中，大量的机器学习模型可以有效地完成端到端的实体识别、关系抽取和关系补全等任务，进而可以用来构建或丰富知识图谱。

知识图谱与机器学习的结合主要有两种：一是将知识图谱中的语义信息输入机器学习模型中，将离散化知识图谱表达为连续化的向

量，从而使得知识图谱的先验知识能够成为机器学习的输入；二是利用知识作为优化目标的约束，指导机器学习模型的学习，通常做法是将知识图谱中的知识表达为优化目标的后验正则项。

另外，在机器学习的大量应用实践中，人们越来越多地发现机器学习模型的结果往往与人的先验知识或者专家知识冲突。如何让机器学习摆脱对大规模样本的依赖、如何让机器学习模型有效利用大量存在的先验知识、如何让机器学习模型的结果与先验知识一致，已成为当前机器学习领域的重要问题。

因此，融合知识图谱与机器学习已经成为进一步应用知识图谱和提升机器学习技术的重要思路。以知识图谱为代表的符号主义、以机器学习为代表的联结主义日益脱离原先各自独立发展的轨道，走上协同并进的新道路。

二、向更多行业渗透

知识图谱的应用领域日趋广泛，正在从金融、公安、电信等相对成熟的领域向医药、农业、政务、天文气象等领域延伸拓展。下面对知识图谱在行业应用的未来发展趋势进行讨论。

在医药领域，由于研发新药花费较高，医药公司非常关注如何缩短新药研制周期，降低研发成本。欧盟第七框架下的开放药品平台（Open Phacts）项目就是利用来自实验室的理化数据、各种期刊文献中的研究成果以及各种开放数据，包括Clinical Trials.org、美国开放数据中的临床实验数据，加速药物研制中的分子筛选工作，已吸引了辉瑞公司和诺华集团等制药巨头的参与。另外，IBM公司成立了事业部（Watson Group），对各种行业进行认知突破。其中在医疗方面，IBM公司启动了登月计划（Moon Shot），通过整合大量医疗文

献和书籍以及各种电子病历（EMR）形成知识图谱，获取海量高质量的医疗知识，并基于这些知识向医护人员提供辅助临床决策和用药安全等方面的应用。同样在中医诊疗上，可以从医案中抽取临床知识构建知识图谱，帮助用户了解中医特色疗法以及疾病的临床表现、相关疗法、相关养生保健方法等。

在农业领域，大量的农业资料以不同格式分散存储，传统的关系数据库模式不适用于复杂多变的领域，无法定义所有可能的知识点并构建关键数据库模式，而知识图谱这种更加灵活的知识表示模型可以实现对农业数据的管理。利用抽取挖掘技术从各种多源异构数据中获取相应的知识，并用统一图谱进行表示，形成完整的知识库，刻画作物知识、土壤知识、肥料知识、疾病知识和天气知识等。通过图谱关联到图片信息，形成多媒体知识图谱，图片信息相比专业知识更加直观，更方便农民使用。

在政务领域，知识图谱也具有多方面的应用价值：①政务信息服务，知识图谱可以为政府网站提供语义搜索、人机智能问答系统等交互服务；②政务知识库构建，如国家安全生产监督管理总局的“政府垂直行业知识库”、科技部知识库等；③人工智能（AI）+政务层面，知识图谱是AI核心基础能力；④公安部门案情调查、情报分析；⑤司法部门事理图谱、辅助判案；⑥政府部门专题分析、决策研究、舆情监控等。

在天文气象领域，气象文献知识图谱主要基于文献网站以及新闻网站的气象文本数据，如维普、万方以及百度新闻，利用知识图谱技术对气象文本数据进行管理和知识抽取，最终构建的气象文献知识图谱能够实现一些智能应用，如文本数据的路径分析、关联分析、可视化和统计分析等。

三、从学术界转移到产业界

随着技术的发展和大众的广泛关注，知识图谱已经从学术研究逐步转移到行业应用中，落实在相关产业发展中，知识图谱提供了全新的视角和机遇。

知识图谱与人工智能结合之后，产品和服务将具备认知能力，这将对企业产生颠覆性影响，将重塑其所处行业的形态，革新行业的各个关键环节。当前已有越来越多的企业将人工智能提升至企业核心战略的高度，在电商、社交、物流、金融、医疗、司法、制造等众多领域中将涌现出越来越多的人工智能赋能的案例。除了探索发现能力将得到长足进步以外，认知系统接受专业人员的训练，掌握政治、经济、法律、医学、销售和烹调等专业术语后，能够理解和传授复杂的专业技能，将大大缩短社会培养人才所需的时间，甚至取代人类做出部分社会管理层面的决策工作决定。越来越多的知识工作将逐步被机器代替，这将对社会结构产生深远的影响。

第二节 知识图谱面临的挑战

目前，人们对知识图谱的研究已有一定的进展，也陆续形成了一些开放知识图谱和相应的应用工具。但是，成熟、大规模的知识图谱应用仍然非常有限。除了搜索、问答、推荐等少数场景外，知识图谱在不同行业中的应用仍然处于非常初级的阶段，有非常广阔的研究和扩展空间。对于客户而言，按照目前学术界提出的方法构建的知识图谱未必能够在实际中直接投入使用，更多时候需要融合不同的行业经验或已积累的大量规则。作者认为，知识图谱仍然面临着诸多挑战。这里简要列举如下。

一、知识获取效率较低

知识获取是构建知识图谱、理解深层语义的主要任务，最主要的任务就是从互联网的网页文本中抽取实体关系。已有的知识元素抽取技术虽有一定的成效，但由于方法可扩展性不强，在很多方面尤其在大规模开放领域的知识抽取仍面临着准确率低、覆盖率低、效率低的问题。知识抽取如何在自动化的基础上实现实用化，是亟待解决的难题。

已有的实体抽取、关系抽取、属性抽取工具都面临着效率较低的问题。受限于数据源，这些工具的通用性不强，需要针对数据源进行相应调整。而调整的方法和过程需要大量的人工投入，这样效率低下也成为制约知识获取的瓶颈。

二、知识融合的难点难以突破

知识融合主要是指在知识图谱构建的过程中，对多来源数据知识进行融合的过程，这对知识图谱构建过程中的准确率与执行效率均具有重要意义。目前知识融合的难点主要有以下4点：

- 不同来源、不同形态数据的融合；
- 海量数据的高效融合；
- 新增知识的实时融合；
- 多语言的融合（特别是中英文的融合）。

具体地，这些难点的主要原因是从不同数据源抽取的知识没有统一的发布规范，数据质量参差不齐，从中挖掘出的知识也会有大量噪声以及冗余。如实体通常会有多个名称，从海量的数据中找到这些名称并且将它们规约到同一个实体下非常重要。针对这个问题，目前的

研究主要是从开发并行与分布式的对齐算法、众包算法一级跨语言知识库对齐的角度进行探讨的。但是要构建高质量的知识图谱，目前的知识质量评估仍然过多地依赖人工，图谱的自动化更新以及确保动态更新的有效性也是面临的重大挑战。

三、知识推理应用进展缓慢

知识推理是目前学术界的研究热点之一，但是已有的学术成果在实际领域的适用性较弱。首先，通过知识推理可以推导出新的关系，这种关系的精度难以得到保证。尤其是在大规模的知识图谱中，预测准确率低、效率低的问题有待进一步研究。其次，目前的知识推理学习和推理方法大多基于通用知识图谱，在实际应用过程中，利用旧关系推导出新关系只能在很小范围内、明确规则下进行尝试，这也意味着专用领域知识图谱的构建才刚开始。最后，目前通用的知识图谱大多是英文的，如何将现有的基于英文的推理方法应用于中文知识图谱的构建，是我们需要努力的方向。

四、缺乏高质量知识库

此外，缺乏高质量的知识库是制约知识图谱技术发展的又一重大问题。对于在业务中将知识图谱作为核心技术的公司来说，获得高质量的训练数据极为关键。虽然很多算法和软件工具是开源和共享的，但好的数据集通常是专有的，且很难创建。可以说，没有大量且高质量的数据集，知识图谱就仍然停留在纸上谈兵的阶段。这也是目前制约许多知识图谱初创企业发展的重大障碍。而从数据集到知识库的构建也有较高的技术门槛。

五、行业知识图谱构建困难

在技术实践中，对于金融、法律、制造、人事等行业，相关的词典或其他NLP方面的资源较少，再加上目前很多开源的工具不具备商业实用性，这给企业构建知识图谱平台带来了极大的挑战。过于专业的知识也给一般的工程技术人员造成了较大困难。如何实现专业人员与技术人员的协调，使得行业知识图谱的质量得以提升，并真正服务于行业的实际需求，仍然是行业知识图谱面临的挑战。

六、商业模式面临阻碍

目前，知识图谱的应用场景仍然非常受限，有些场景存在“伪需求”的可能性。相对于学术界的热烈讨论（各种新算法不断提出），真正的企业应用仍然相对滞后。缺乏解决方案与最佳实践，使得知识图谱技术的“知名度”仍待提升。而知识图谱的商业模式仍然存在多种不确定性。

目前，知识图谱企业的商业模式主要包括3类：第一类，以“产品+定制化”解决方案的形式进行客户服务，优点是能够与客户深度绑定，积累行业经验，缺点是该模式通常耗时耗力；第二类，通过集成商销售通用性较高的模块化功能，其优点是节省人力，缺点是收益的性价比较低；第三类，以第三方技术提供商的角度专注于特定技术环节，通过与不同客户合作，以产品分成或项目方式获得营收，其优点是应用领域相对宽泛灵活，缺点是对技术要求较高。总之，随着技术的发展，这些商业模式的缺陷暴露得越来越深刻。如何构建成熟的商业模式，始终是值得知识图谱企业深入探索的问题。

第五章 知识图谱实战案例

知识图谱的实际应用体现了知识图谱技术的价值。近年来，我国的知识图谱企业发展迅速，在通用知识图谱和行业知识图谱等各领域均涌现了一大批优秀企业。本章介绍了国内一些知识图谱企业在医疗、社交、金融、公安方面的优秀案例。每个案例都从痛点难点、实践路径与应用效果3个角度入手，深入剖析应用背景，简要叙述技术路线，并着重探讨应用效果。希望读者可以从本章的案例中感受到知识图谱技术给业界带来的变化。

第一节 基于知识图谱的医疗决策辅助系统

一、痛点难点

随着人们对健康问题的重视程度不断加深，医疗保健需求的增长与优质医疗资源的不足之间的矛盾亟待解决。现阶段，培养一名合格的医生至少需要8年的时间，而一名资深的诊疗师则需要25年以上的临床经验。优质医生资源的严重匮乏致使“好医难求”的问题突出。因此，急需一个能够辅助决策以提高诊断效率的平台。

二、实现路径

“精准医疗”基于大样本研究疾病预防与处置方法，主要依据各种医学知识实现疾病的精准预防、精准诊断和精准治疗，因人而异地确

定治疗方案和药物用法用量，从而达到提高医疗的有效性、减少治疗方案副作用的目标。

知识图谱拥有强大的多级语义推理能力和网络可扩展能力，以此为基础搭建数据平台，收集、整合并处理从医学词典、论文著作以及行业标准等来源获取的非结构化数据，生成结构化的大规模医疗知识图谱，使得医生和患者可以直接使用基于医疗知识图谱构建的诊断系统分析或辅助分析病情，达到精准医疗和实时预防检测疾病的目的。下面以渊亭医疗决策辅助系统平台为例，具体说明基于知识图谱的医疗决策辅助系统的实现路径。

（1）医疗数据集成

渊亭医疗决策辅助系统支持多种方式的医疗知识来源，如临床医学知识库，SNOMED-CT、ICD-10以及智能硬件，医学影像与电子病历单。这些非结构化的医疗数据具有规模大、冗余性高、类型多样、增长快速、价值巨大等特点，主要存储在关系型数据库、NoSQL、文件、HDFS中。图谱数据集成总线是所有多源异构数据源的处理入口，总线会自动识别数据源的格式和数据规模、决定使用什么数据提取向导进行处理、是否使用分布式计算或并行计算资源。针对专家知识情报的自动分类技术，通过集成学习（也称为多重学习或分类器组合），借助决策优化或覆盖优化两种方法将若干弱分类器的能力进行综合，以优化分类系统的总体性能。决策优化对于不同的分类器均采用完整的样本集进行训练，测试时，通过对所有分类器的决策进行投票或评价，确定整个系统输出的类别，系统利用概率方法综合不同分类器的输出，确定最后的决策。通过获取一条或多条包含实体名称及对应实体属性信息的结构化数据，提取所述结构化数据中包含的实体名称及其属性信息的映射关系，生成对应的数据结构对，将生成的数据结构对作为知识图谱数据项进行存储，完成结构化数据的映射匹配。

（2）医疗数据知识表示语言

医疗知识本体表示语言的设计理论源于本体语义学，具备以下优点。

第一，它强调对意义的处理不需要通过句法分析，至少不是主要通过句法分析。在它看来，机器对意义的接受、表征、加工、生成和输出，或者使机器的句法加工具有语义性或意向性，主要靠的不是原先的关键词匹配、句法转换，而是对人类智能的全方位模拟。

第二，它认识到了人类心理状态具有意向性、自然语言的语义对于复杂因素的依赖性，并在这种认识的基础上形成了一种具有研究意义的综合性方案。在具体的工程学实践中，医疗知识本体表示语言关注了意义处理中的多方面因素，即不仅仅注意到了知识性因素，而且还重视潜藏在人类智能中的非知识因素，并通过特定的方式将这些非知识因素“内化”到他们构建的人工智能系统之中。

第三，重视本体论的图式化在人类心理状态意向性、自然语言语义性中的作用，并将这一认识成果向工程技术领域转化，进而让自然语言加工系统在这一语义生成的重要枢纽、机制方面做了大胆探索，取得了富有启发意义的初步成果。

第四，最重要的应用价值是它能产生文本意义表征。因为它的语义处理系统可以借助静态知识资源对输入文本做出分析，借助加工器的动态能力将存储的知识动态地提取出来，并用于知识表征，然后借助这些知识资源产生文本意义表征，并由特定输出设备完成在意义交流层的人机对话。

（3）医疗知识抽取

医疗知识图谱的构建主要是将分散在知识库中的非结构化数据进行人工的或自动化的处理，从非结构化数据中提取实体、关系、属性的三元组。人工进行知识抽取的代价太大，知识的自动提取是目前重点的研究方向，也是将来构建知识图谱的重要过程。对于医疗知识抽

取，渊亭医疗决策辅助系统主要采用机器学习、人工智能、数据挖掘等信息抽取技术，提供了以下几种方式的抽取：NLP全自动化提取、正则表达式、规则引擎、CSS选择器、XPath等。基于半监督学习的文本知识抽取技术把蕴含于信息源中的非结构化知识经过识别、理解、筛选、归纳等过程抽取出来，存储形成知识元库。其主要使用了层次类型约束主题实体识别和关系抽取算法。对于知识抽取任务，有两个重要的部分：其一是数据源实体识别，即将非结构化的文本数据中的主题实体识别出来，并与知识图谱做实体链接；其二是谓词映射。对于主题实体识别任务，主流的做法为依靠字符串相似度，再辅以人工抽取的特征和规则来完成的。但是这样的做法并没有将问题的语义与实体类型、实体关系这样的实体信息考虑进来。实体类型和实体关系在很大程度上是与问题的上下文语义相关的。当只考虑实体关系时，会遇到零样本（Zero-Shot）的问题，即测试集中某实体的关系是在训练集中没有遇到过的，这样的实体关系无法准确地用向量表达。因此，渊亭医疗决策辅助系统使用的技术首先利用实体类型（Entity Type）的层次结构（主要为实体类型之间的父子关系）解决Zero-Shot的问题。如同利用Wordnet计算Word相似度的做法，将父类型的“语义”视为所有子类型的“语义”之和。一个实体总是能够与粗颗粒的父类型相关，例如一个实体至少能够与最粗颗粒的Person、Location等类型相连。这样，利用实体所述的类型，在考虑实体上下文时，就可以在一定程度上弥补实体关系的Zero-Shot问题。此外，使用深度学习技术建立了一个神经网络模型——层次型约束主题实习监测（Hierarchical Type Constrained Topic Entity Detection，HTTED）模型，利用问题上下文、实体类型、实体关系的语义，计算候选实体与问题上下文的相似度，选取最相似的实体，解决命名实体识别（NER）问题。

（4）医疗知识融合

渊亭医疗决策辅助系统使用了多种来源的医疗知识库，为了解决知识复用的问题，并增强知识库内部的逻辑性和表达能力，需要在同一框架规范下进行知识的整合、消歧、加工、推理验证以及更新。针对知识图谱中不同粒度的知识对象，知识融合可以分为实体对齐、冲突解决、实体链接，在医疗信息系统中主要集中于实体链接。渊亭医疗决策辅助系统使用的实体链接方法由3个模块构成：候选实体生成模块、实体相关图构造模块以及集成实体链接模块。候选实体生成模块的主要功能是对于给定的输入语料，识别出其中的所有实体指称项，据此分别查找本地知识库，得到与该实体指称项同名的候选实体集合，作为后续构造实体相关图的顶点集合。实体相关图构造模块的主要功能是针对从同一文本中抽取得到的所有实体指称项和相应的候选实体集合，构造出一张该文本的实体参考关系图，作为集成实体链接的依据。集成实体链接模块的主要功能是借助实体相关图实现对输入语料中歧义实体的消歧，将其正确地链接到本地知识库中的正确的实体对象上。

（5）医疗知识利用与推理

推理是从已有的结构化知识中挖掘出有价值的隐含信息，医疗知识推理相对而言，更注重知识与方法的选择与运用，尽量增强推理过程的可解释性，添补缺失事实，降低疾病诊断过程中的误诊率。即使对于相同症状的疾病，渊亭医疗决策辅助系统仍然会根据病人的真实体态特征与历史治疗记录，做出特殊的诊断。知识推理帮助医生完成了从病患数据收集到疾病诊断与治疗的全过程，医生只需要定批地监控和校验医疗决策辅助系统的工作状态与诊断结果即可。传统事件演化预测技术是基于自然语言分析构造文本相似模型进行的，当文本相似度高、概率分布均匀时，对事件的预测相对准确；但是如果文本之间语义差距较大，事件预测的准确度会直线下降。为了解决该问题，采用属性图特征模型，综合知识图谱的属性图模型（PGM）表征、形

式概念分析（FCA）等多项技术。形式概念分析强调以认知为中心，提供了一种与传统的、统计的数据分析和知识表示完全不同的方法，用*N*层结构模型对主题中的事件进行属性提取，并依据特征频率因子进行属性选择，利用多个事件之间的属性关系建立形式背景，以此为基础形成概念格，用基于概念格的相似度分析发现事件之间的潜在联系。它不依赖文本语义的情况，即使事件之前的文本语义差距比较大也能分析出关联程度，而且形式概念分析建立在数学基础之上，对组成本体的概念、属性以及关系等用形式化的语境表述出来，具备一定的解释性。

（6）系统总体架构

基于知识图谱的渊亭医疗决策辅助系统的总体架构如图9所示。

图9 医疗决策辅助系统的总体架构

医疗数据采集层实现对多源异构数据的采集。医疗数据的来源可以是多种多样的，包括网站数据、病历数据、各种医学词典、医学影像、各类医学知识库、各种医疗智能硬件等。而这些不同来源的数据结构本身也是多样的，包括结构化数据（比如医疗系统关系型数据库中的数据）、半结构化数据（比如网页上的数据）、非结构化数据（比如文本信息、图像信息等）。

数据池层通过知识提取、知识融合、知识存储等方法，构建出各类知识图谱。

知识加工层实现对知识、人员和案例等的管理。

推理引擎层利用机器学习、深度学习、图计算等方式，为医疗知识检索、分析等提供知识推理引擎。

医疗人工智能平台层利用知识推理引擎，为个人用户、医疗机构以及医疗站等用户群体提供各类医疗智能应用功能，包括分诊导流、健康管理、健康方案、智能问答、辅助诊疗、风险评估、健康专家咨

询等。同时，该系统还为其他不同的软硬件提供多种形式的接口，包括对智能移动设备、医疗机器人等的接口。

三、应用效果

渊亭医疗决策辅助系统加载待诊断病人的电子病历单与医疗检测数据，通过系统提供的NLP引擎对非结构化的信息进行抽取，以获取该病人的真实体态特征和疾病史信息。

渊亭医疗决策辅助系统的疾病与症状知识图谱示意图如图10所示。

如图10所示，原发性高血压可能引起耳鸣、头痛等症状，那么原发性高血压与各种症状之间就建立了关系。类似地，渊亭医疗决策辅助系统建立起各种疾病与症状之间的关系。

图10 疾病与症状知识图谱示意图

渊亭医疗决策辅助系统通过提取病人病历数据，建立病人、疾病、症状等各种实体之间的各种类型的关系图谱，再利用算法模型等分析引擎，就可以在诊疗中为医生和病人提供可能疾病和医疗方式的智能诊断建议，还可以针对医生的诊疗方案进行分析，查漏补缺，减少甚至避免误诊。

渊亭医疗决策辅助系统被应用于多家三甲医院，促进了医学领域大数据与精准医疗图谱分析的发展，实现了科学研究和临床应用的有机融合，最终促进了医学大数据项目和学校的协同健康发展。渊亭医疗决策辅助系统为客户解决了以下问题。

● 实现了对年轻医生的辅助培养：基于平台已有诊疗过程中的问题和处置办法以及病人术前术后的各种综合数据，年轻医生可自主学

习。

● 可通过平台做研究：通过数据的梳理和整合，为医院的重点学科提供数据参考。

● 交叉学科学习，提高临床诊治水平：各学科可以查找相关学科的数据库，从相关的数据中剖析病理。

第二节 利用知识图谱构建“虚拟生命”

一、痛点难点

虚拟生命是利用人工智能技术对生命的延伸，具备生命的主要特征，不仅要有感知能力，更要有认知能力，还要具备自我进化的能力。在现有条件下，具有里程碑意义的技术集大成者是聊天机器人，即一种通过自然语言模拟人类进行对话的程序。近年来，在大数据基础上，自然语言处理、深度学习和知识图谱技术的结合造就了聊天机器人的崛起。它将每一个单独的能力连接起来，形成功能更强大的智能引擎，可以灵活运用多种表达方式与用户进行交互。典型的聊天机器人有苹果公司的Siri以及IBM的Waston等。

但目前的聊天机器人还停留在初步的“感知”层面，离“认知”还有非常大的差距，与人类的交互并没有做到真正的理解，只能看作大数据和规则的体现，缺乏自然的多轮交互和上下文的一致性，无法实现真正的情感表达，并且与用户的交互形式也较为单一。虚拟生命涵盖了聊天机器人的基本范畴，并延伸了聊天机器人的定义。从感知能力角度来看，虚拟生命需要能够听得到、看得见、可交互。这一切都依赖于语音识别、计算机视觉、语音合成等技术的发展，这些技术使得虚拟生命和人类能够进行特定领域下的、基于规则的简单交互。从认知

能力角度来看，虚拟生命需要能够与人以及周围环境进行“真实自然”的交流，包括规划、推理、联想、情感和学习能力，需要具有非常强的可用性和可交互性。

二、实现路径

作为聊天机器人的下一代范式，虚拟生命能够进行多模态交互、融合多源知识，具有联想推理和个性化认知能力。技术的进步是永无止境的，虚拟生命的实现并不能被动地等待技术的成熟，而是要利用现有技术实现落地，并且不断完善。因此，在实现虚拟生命从问答到感知以及自我进化的过程中，必须借助各个领域的技术发展，尤其是大数据、机器学习和知识图谱的发展。

图11所示为基于GAVE引擎搭建的虚拟生命“琥珀·虚颜”的整体技术架构，从数据到开放平台共分为4个层次。

图11 “琥珀·虚颜”的整体技术架构

最底层的数据层是资源和运营的结合，强调知识图谱的构建以及人工智能（AI）和人类智能（HI）的结合。由于技术的局限性，知识图谱的自动构建、数据的自动标注等还未达到令人满意的效果。在自动化方法的基础上，借助众包数据，加上专家的介入和修正，可以产生一个良好的闭环（Human in the Loop），从而不断增强知识的质量。

第二层是基础技术层和服务支撑层，在底层的基础上实现了虚拟生命的多模态交互以及知识驱动的自然语言理解，包括细粒度实体识别与链接、远程监督的语义标注等功能，体现了感知智能和认知智能的一体化以及算法的引擎化。作为中间层，该层建立从语义到语用的

桥梁，实现从独立感官到融合感官、从应用到认知的跨越。通过基础技术的引擎化和内部平台化，该层可以将底层知识和上层服务关联起来，支撑快速产品开发。

第三层是服务层。现阶段的虚拟生命技术落地一定是分场景切入的，还无法达到通用性的强人工智能。因此，如何进行场景建模、快速定制、快速配置，是服务层关注的重点。服务层将智能家居控制、多模态问答等聊天机器人的核心展示形式作为底层支撑，结合深度问答技术，实现理解、记忆、推理、表达等认知智能。

最上层是开放平台接入点，可以通过开放API为用户提供更多的服务。

三、应用效果

作为聊天机器人的下一代范式，虚拟生命已经有了产品化的体现。深圳狗尾草智能科技有限公司开发了世界上第一款结合了AR+VR以及GAVE引擎（Gowild AI Virtual Engine）的虚拟生命产品——琥珀·虚颜。琥珀·虚颜搭载HoloEra硬件平台及360°全息投影，创造了一个基于大规模知识图谱和AI技术的，有情感、可养成、可进化的虚拟存在，但这种存在又可以与周边世界进行多模态的真实互动，并针对用户行为习惯形成不同的性格体系。

GAVE引擎在投入使用后，有效地加快了智能机器人产品研发的进度，降低了智能机器人行业的准入门槛，并对智能机器人行业的产业升级起到了促进作用，尤其是对提升智能机器人的类人程度具有显著意义。聊天机器人作为现有技术的集大成者，可以看作自然语言处理上的明珠。再进一步，虚拟生命是聊天机器人的最终目标和下一代范式。在现阶段，虽然虚拟生命还没有发挥出最大的价值，但已经开

始真正地与用户进行自然交互，通过声音、视频、图像，让用户体验在特定场景下真实的聊天感受和情感陪伴。

第三节 股份制银行知识图谱案例

一、痛点难点

目前，金融科技的发展主要存在如下痛点、难点。

（1）数据维度

在互联网金融时代，全国性的股份制银行需要建立一个覆盖全行业、全业务条线的数据分析平台，以支撑跨地区业务的协调运行。然而，电子支付、去中心化等功能需求带来了海量低价值密度数据，这对平台的数据治理能力、数据存储能力和业务理解能力提出了更高的要求。如何在数据中挖掘价值，建立匹配业务的人工智能学习型算法是数据维度面临的挑战。

（2）技术维度

建立知识图谱数据库，需要对现有全维度数据进行“关系-实体”数据建模与治理，实现复杂业务条线下的数据架构整合。如何运用自然语言处理技术解析半结构化文本数据，解决消歧、对齐和融合等技术难点，拓展数据维度，提升数据处理能力、知识图谱构建能力和存储能力是技术维度面临的挑战。

（3）业务维度

面对内部的员工操作风险、审计不当和外部的非法金融活动、网络黑产等问题，银行的风险防范需求不断提升。如何将数据与业务重构，建立百亿级的实体关系网络，使业务人员能够通过资金流向分

析、关联客户分析等模型有效地识别内外部运营风险，是业务维度面临的挑战。

二、实现路径

知识图谱项目一般可以分为知识构建、知识计算、知识存储、知识应用4个主要部分。

（1）知识构建——从海量文本到行业图谱

行业内数据资源的获取和整合不仅依赖数据爬取、多源异构数据治理、分布式数据存储等技术，也依赖强大的外部数据资源协作能力和内部推动能力。知识图谱构建技术从多源异构的数据中抽取实体信息，链接和融合实体，推理补全属性，识别语义并建立关系，最终将知识存储于知识图谱数据库中。当金融知识图谱构建完成后，金融机构可以获得包含基本实体、属性以及从数据中可以构建的显性关系的基础知识图谱。这其中包含的“实体-关系”抽取技术是知识图谱构建的核心技术之一。

对于银行客户，构建全行级知识图谱需要将数十个维度的行内客户类数据、机构类数据、企业类数据、业务流程类数据等进行治理和整合，构建完成后在这个阶段获得以企业客户、账户、产品为核心的，包括企业客户之间、企业与产品之间基本关系、担保关系、资金往来关系的知识图谱。

（2）知识计算——行业知识的数学表达

知识计算阶段的核心任务是计算隐性关系和扩展属性，这是知识图谱体现智能的重要特点。在银行知识图谱平台的构建中，需要结合营销或风险控制的业务分析和设计企业客户之间形成的集团、一致行动、实际控制等潜在的隐性关系，形成相应的计算规则和模型，构建

资金流转网络、担保关系网络等，并将这些关系和网络存储在基础知识图谱数据库中。

知识计算引擎具有实体查询、事件查询、在线隐性关系计算和挖掘、基于事件的动态关系推演等基础查询功能，同时支持对不同数据类型的复杂检索（多属性组合、多关系组合、多事件组合查询）。金融知识图谱平台内部通过智能的查询分析和复杂的逻辑计算优化过程，结合分布式计算和NQL等用户友好的查询语言，将一个查询分解成不同类型的子查询进行分布式处理。

知识计算应用的技术除了自然语言处理外，还包括规则引擎、机器学习和图挖掘等数据挖掘技术，需要银行专家、工程师、数据科学家协同参与。同时，为了检验构建的显性和隐性知识的完备性、相关性和一致性，需要结合专家知识和特定的知识计算方法进行校验，处理其中的缺失、冲突、冗余知识。

完成知识计算阶段后，将包含经过验证的显性和隐性知识的完整知识图谱作为知识应用的数据模型基础。

（3）知识存储——知识应用的重要工程保障

知识存储阶段承担的使命不仅仅是存储知识，更重要的是为知识应用提供稳定、准确、高效的运转能力，同时还要支持增量数据和业务变化带来的海量知识更新。

从数据库技术选型的角度来看，传统的关系型数据库、KeyValue数据库及时下流行的各种图数据库都可以作为知识存储的基础，可以结合数据规模、应用规模、性能要求和整体IT架构规划综合做出选择。

混合型数据存储技术可支持海量数据图谱的高效存储和查询。在知识图谱数据库中，核心图谱数据将存储对象抽象为“实体-关系-事件-文档”，根据不同的数据对象类型，使用最合适的存储方式以及对应的查询方式，包括图存储、列式存储、索引存储、文件存储。知识图谱

数据库可以通过Java API或者NQL查询语言（类似SQL）实现对存储数据的快速访问。此外，知识图谱数据库还提供了批量数据导入工具和内部状态及性能检测工具。

（4）知识应用——搜索、业务应用和问答

知识应用是最直接体现知识图谱智能化的部分，也是使用者能直观感受到其价值的部分。从Google公司提出知识图谱的概念到微软、百度、搜狗的快速跟进，搜索一直都是知识应用最典型的场景。在完整的知识图谱上，搜索需求可以被解构，搜索体验完成了从匹配文本内容到“问题-推理-答案”的重大升级。

三、应用效果

全行级知识图谱平台的建立可以让银行采用复杂网络、图计算等大数据算法和人工智能技术搭建远程监控体系下复杂计算及非结构化模型建设的框架，实现海量数据和非结构化数据的分析和探索，解决传统技术无法解决的问题，加强远程监控的工作广度与深度，从而提升远程监控水平和能力，并且从多维度带来业务效率的提升。

● 数据维度：整合全行多业务条线数据，构建全行大数据关联关系分析平台，重构银行数据架构，建立银行内部的知识图谱数据库。

● 技术维度：运用机器学习和知识图谱技术，实现复杂业务条线下的数据架构整合。运用自然语言处理技术，解析半结构化文本数据，拓展数据维度，提升数据处理能力。

● 业务维度：通过平台的构建，链接银行全场景50余个维度的实体、关系、事件等百亿级实体关系网络，全面提升银行风险内控水平，辅助业务人员提高效率。

第四节 基于公安知识图谱的禁毒大数据分析平台

一、痛点难点

毒品犯罪是新时期刑事犯罪最突出的表现形式。毒品犯罪并不是一种单一的犯罪行为，因吸毒引发的盗、扒、抢、伤害、杀人等各类治安案件和刑事案件层出不穷。毒品问题的发展蔓延给政治、经济和社会生活带来了很大的危害，毒品问题已是当前影响社会稳定和经济发展的重大问题之一。

随着大数据、物联网时代的到来，丰富的“人、车、电、网、像”轨迹数据大量被采集，这为公安机关获得案件侦破线索提供了更多的可能性，但是传统的业务系统在承载大数据方面又存在诸多不足，这在一定程度上增加了业务员发现线索的难度。涉毒违法犯罪分子反侦查意识的日渐加强也为公安打击工作带来了新的挑战。毒品犯罪的特点如下所示。

（1）毒品交易隐蔽，取证难

毒品犯罪与通常的犯罪不同，不仅没有直接的被害人，而且毒品贩卖极为秘密，往往有组织、有规则，局外人一般很难介入其中，用通常的方法进行侦查极为困难。

（2）流通渠道多样化，管控难

现有物流渠道多样，社会综合整治力度不强，这为涉毒犯罪提供了可乘之机，比如高速公路运输查处力度不够，贩毒人员常常将大量毒品放到私家车内从外省运回省内出售。邮政、快递、物流等行业对毒品疏于防范，不可避免地沦为运毒渠道。

（3）联络方式多元化，打击难

在毒品犯罪中，犯罪分子大多利用现代化通信工具及互联网社交软件进行联系，通话内容大多使用行话、黑话，用语隐蔽。犯罪分子隐瞒真实身份，在微信、陌陌等社交软件上注册多个账号，使用化名联系涉毒的“犯罪业务”。相对于熟人介绍等传统方式，通过手机社交软件推介联系毒品交易等方式更容易避开审查，犯罪行为具有更强的隐蔽性。

（4）毒品犯罪群体化，发现难

从毒品货源到吸食购买会经过多个层次之间的周转，在长期进行的毒品犯罪活动中，多数犯罪分子都是以贩养吸，形成了一个相对稳定的团伙和固定贩毒、吸毒人员构成的消费网络，其内部人员相对固定，分工明确，有的负责联系毒品货源，有的负责分装，有的负责送货等，群体构成复杂，发现难。

二、实现路径

基于涉毒类案件的特征和大数据技术的成熟度，紧紧围绕警务大数据顶层设计和建设，构建公安知识图谱，建立大数据背景下的新型工作流程，实现更高层次的涉毒违法犯罪预测功能。下面以明略禁毒大数据分析平台为例，说明知识图谱技术在公安系统的具体应用。

明略禁毒大数据分析平台建设包括以下4个方面。

（1）构筑涉毒违法犯罪预测的数据基础

近年来，各地公安机关汇聚了大量数据资源，但数据如何融合使用还是个难题，传统上公安机关基于关系数据库的技术思维，围绕“人、地、事、物、组织、线索”等要素建立的数据关联库、要素库、专题库只能提供信息查询、检索方面的实战支撑，数据价值远未得到充分发挥。明略禁毒大数据分析平台利用关系网络及标签技术对

数据体系进行重构，建立新型的数据研判体系，为涉毒违法犯罪预测、分析研判提供可靠、高效的数据支撑。

（2）构建公安知识图谱，建立涉毒违法犯罪人员及团伙的挖掘和预测机制

涉毒违法犯罪分子采用多种隐蔽、伪装方式以逃避法律制裁，但此类犯罪客观还是存在较为明显的特征，即：吸毒人员复吸率较高，多次入所，涉毒人员关系固定且相互交叉，从吸毒、以贩养吸到大宗贩毒逐层递进等。明略禁毒大数据分析平台从这些特征入手，以社交通联轨迹为切入点，结合特征识别和机器学习等手段，建立涉毒违法犯罪人员及团伙的识别、挖掘和预测模型，充分发挥大数据海量处理、全量计算的优势，为打击涉毒违法犯罪提供高质量的情报线索。

（3）强化涉毒犯罪情报的深度研判能力

在大数据背景下，传统依靠“人脑串烧”的分析研判模式已经越来越力不从心了。明略禁毒大数据分析平台结合关系网络、图计算、自然语言识别等技术，为业务人员提供可视化的分析研判工具，帮助业务人员搭建物理世界与数字世界的沟通桥梁，便于他们掌控大数据内部蕴含的关联关系，提高侦查、研判、情报线索串并的能力，减轻日常侦案过程中核查工作的强度，提高工作效率。

（4）探索新型的涉毒情报产品服务机制

客观上，通过数据挖掘获得的线索存在准确率的问题，而在公安机关现行的机制下，由于警力、精力有限，基层一线部门相对更重视依靠传统方式获得的情报和线索，对通过大数据手段和模型运算获得的线索还存在疑虑。基于明略禁毒大数据分析平台，未来将探索涉毒情报产品服务的新机制，即建立起线索归口推送、接收反馈、核查反馈的闭环管理机制，不断推动模型迭代优化和改进，从而使模型越来越准确，实战性越来越强。

三、应用效果

2017年9月，东部沿海某市公安局利用基于涉毒挖掘模型实现的明略禁毒大数据分析平台挖掘出了一个高危涉毒嫌疑人，通过大数据分析实现了嫌疑人的身份判定、活动区域判定、鲜活度判定，再由点到面，深度研判，拓展出该嫌疑人关联关系网络图，进而挖掘出一个分布在多省的制贩毒网络团伙。2017年10月，专案组在多地同时收网，成功缴获上百公斤的毒品，抓获涉案人员20余人，彻底摧毁了这个分布在多省的制贩毒网络团伙。此案的侦破是传统情报模式向大数据分析主动式情报模式转变的一个创新的尝试。

第六章 知识图谱构建工具

本章将介绍目前较为常用的8种知识图谱构建过程中的软件工具，其中，国内常用的工具有Pajek、CiteSpace，国外常用的工具有UCINET、Gephi、VOSviewer、VantagePoint、Sci^2 和SciMAT。下面将对各工具的主要功能和特点进行介绍。

第一节 Pajek

一、Pajek软件概述

Pajek是由来自斯洛文尼亚的Vladimir Batagelj和Andrej Mrvar 共同编写的大型复杂网络分析工具，主要用于研究目前存在的各种复杂非线性网络。Pajek是基于图论、网络分析和可视化技术发展而来的，允许对海量抽象的数据进行分析和可视化，为科学发现、工程开发和业务决策提供依据。它在Windows环境下运行，免费提供给非商业用途，提供一整套快速分析复杂网络的算法和可视化的界面，让用户可以从视觉角度直观地了解复杂网络的结构特性。

对于知识图谱构建软件来说，Pajek缺少数据预处理和数据标准化处理的功能，但是它可以将大型复杂网络分解为几个小的子网络并可视化呈现，很好地解决了复杂网络可视化的难题。Pajek支持构建一些特殊的网络，例如：多关系网络、2-mode网络（二分图—网络由两类异质结点构成）、暂时性网络（动态图—网络随时间演化）等。同时，Pajek还支持多种数据格式的输入，也可以辅助其他知识图谱的软

件工具生成可视化图谱，例如，Ucinet可以将数据和数据处理结果输出到Pajek后进行可视化呈现。

二、Pajek的主要特点

1.计算快速

Pajek最主要的特点是提供一整套针对大型网络的快速有效的算法。一个算法的复杂度主要表现在时间复杂度和空间复杂度两方面。随着存储技术的快速发展，空间复杂度的重要程度日益减弱。当复杂网络的节点数目逐渐变得庞大时，计算机运算速度对于解决问题的时间来说已经无足轻重，算法的时间复杂度开始起到至关重要的作用。

表4展示的是5种算法的时间复杂度及其在不同节点数量下的计算时间。可以得出，随着节点数量级的上升，在不同时间复杂度下，各个算法计算时间的差距越来越大，当复杂网络的节点数达到100万的时候，时间复杂度为$O(n)$的算法耗时仅为2.22s,而时间复杂度为$O(n^3)$的算法耗时却需要3.17年。

表4 几种算法的时间复杂度比较

Pajek中所有算法的时间复杂度都低于$O(n^2)$，达到$O(n)$、或者$O(n \log n)$，因而可以实现对大型复杂网络的快速处理。

2.可视化

Pajek为用户提供了一个可视化的平台。随着计算机软件的飞速发展，从早期的Turbo C到现在的Visual C++，可视化是软件发展的必然趋势。Pajek提供了一个非常人性化的可视化平台，在Pajek的主菜单里有Draw的菜单命令，用户只要执行Draw→Draw的菜单命令，就可以绘制出网络图谱；用户还可以使用Draw菜单下一系列的其他命令，自动或手动地调整网络图谱的布局和视觉效果，从视觉的角度直观地分析整个复杂网络的结构。

3.抽象化

Pajek还为复杂网络的全局结构分析提供了一整套抽象的方法。

图12（a）是一个由若干节点和边组成的图，不同的圈代表了不同的“类”。我们可以很清楚地看到各个节点之间的联系，却很难直接看出这个网络的整体结构。在Pajek中，可以将每一类看作一个整体，发掘整体的结构关系（如图12（b）所示），再将每个“整体”作为新的节点得到新的网络图（如图12（c）所示），这样就可以很轻松地看出原网络的整体关系。Pajek还可以令图12（a）中间的类保持不变，将其他的各类看作各个整体，这样就可以方便地得出中间的类的各个节点在整体结构中的作用。

图12 Pajek的抽象化分析示例

Pajek提供的这样一整套算法可以方便地计算复杂网络结构的各个特性，使用户可以具体地分析复杂网络中的各个节点和各条边的特点，从具体和抽象两方面综合分析复杂网络。

三、Pajek的数据结构

在数据结构上，Pajek支持多种数据格式的输入，同时能识别其他软件处理的数据，如Ucinet的DL格式等。

● 网络（Network）：它是Pajek最基本、最重要的数据类型，包含整个复杂网络的最基本信息，如节点数、各节点名称、边及权值等。默认扩展名为.net。输入文件中的网络表现格式可以是边、弧线、序列、矩阵及Ucinet等其他软件处理的数据格式。

● 分类（Partition）：它指明了每个节点分别归属的类别。Pajek可以自动地按照用户指定的分类标准（用户也可以手动设定），根据各个节点的不同特性对节点进行分类，分类结果以Partition文件输出。默认扩展名为.clu。

● 排序（Permutation）：它展示了各节点的重新排序。用户可以如Partition一样选择手动或Pajek自动根据某种算法进行重新排序。默认扩展名为.per。

● 类（Cluster）：它表示某种具有相同特性节点的集合。用户可以利用它对一类节点进行操作，避免多次处理单个节点的麻烦。默认扩展名为.cls。

● 层次（Hierarchy）：它展示了复杂网络中各个节点的层次关系，其结构类似于数据结构中的树，常用于家谱图一类的分析。默认扩展名为.hie。

● 向量（Vector）：它以向量的形式为某些操作提供各节点所需的相关数据，也指明了每个节点具有的数字属性（实数）。默认扩展名为.ver。

第二节 CiteSpace

一、CiteSpace软件概述

CiteSpace是美国雷德赛尔大学信息科学与技术学院的陈超美博士于2004年使用Java语言开发的信息可视化软件，可在所有系统环境下运行。CiteSpace的主要着力点在于科学分析蕴含的潜在知识，通过可视化的手段呈现现有科学知识的结构、规律和分布情况，并显示科学发展的新趋势和新动态，因此，更多被应用在科学研究界。

CiteSpace的前身是StarWalker，在三维虚拟现实中演示科学引文逐年增长的过程。软件注重功能的升级创新，不断发布新的版本，免费提供给研究者使用，是国内研究者使用最多的一款软件。

二、CiteSpace的主要特点

1.动态追踪

作为主要应用在科学研究界的文献图谱构建工具，CiteSpace主要对特定领域文献的相关数据进行计量，通过对各学科领域或研究方向的文献数据进行动态追踪，探寻出学科领域演化的关键路径及知识转折点。

CiteSpace的灵感主要来源于库恩（Thomas Kuhn）的科学发展模式理论。库恩认为科学研究的重点随着时间变化，有时候速度缓慢，有时候又比较剧烈，而科学发展的足迹是可以通过已发表的文献中提取出来的。CiteSpace可以直观捕捉相关研究领域的热点话题、重要学者和研究机构，还能展示出特定时间跨度内新研究话题的突然激增情况。

2.可视化与序列化兼具

CiteSpace展示的数据信息包含作者、期刊、研究机构、关键词、被引文献、国别等，通过分析、建立耦合、共作者、共引等关系形成可视化的图谱。不仅如此，通过以上关系的建立，CiteSpace展示的数据是顺序化的、结构化的，可以清晰地展示学科发展的脉络。

CiteSpace展示的既是可视化的知识图形，又是序列化的知识谱系，显示了知识单元或知识群之间的网络、结构、互动、交叉、演化或衍生等诸多复杂的关系。利用CiteSpace可以帮助刚进入某一特定领域研究的学者对该领域建立全面的认识，识别学科的研究热点，并预测学科的未来发展趋势。

3.知识图谱构建功能完整

在整个知识图谱的构建过程中，CiteSpace在各个流程的处理都能满足不同研究者的需要。

- 从支持的数据格式上看，CiteSpace不仅支持TXT格式文档，也支持用软件转化了的CSSCI格式，因此广受国内研究者的欢迎。
- 从数据的预处理上看，CiteSpace提供时间切片、数据和网络的缩减功能。
- 从数据建模上看，CiteSpace提供多种方法进行关系矩阵的构建，不仅包含耦合、工作者、共引的常见关系，也可以构建Co-Grant矩阵，充分满足学者研究的需求。
- 从数据的标准化处理上看，CiteSpace具备Salton余弦、DICE强度和Jaccard强度等方法。

● 从数据分析上看，CiteSpace也支持突发检测、构建网络、时序分析等基本方法，且在地理空间分析方面具有很大优势。

三、CiteSpace的结果呈现

CiteSpace可以呈现4类可视化图谱，第一类是作者、研究机构、国别；第二类是参考文献（Cited Reference）之间以及被引作者（Cited Author）之间的共引关系；第三类是关键词和术语；第四类是研究基金。

第三节 UCINET

一、UCINET软件概述

UCINET是由加州大学欧文分校社会网研究的权威学者Linton Freeman编写的功能强大的社会网络分析软件。目前，该软件主要由新一代学者肯塔基大学的Stephen Borgatti和曼彻斯特大学社会科学学院的Martin Everett维护更新。UCINET是一个商业软件，需要向用户收费，但是它也会提供免费的使用版本。

作为一个综合性功能的软件，UCINET提供了大量数据管理和转化工具，例如，在CNKI上搜索的数据，经过格式的转换，也能被UCINET识别；它也可以将图论程序转化为矩阵代数程序。同时，UCINET也拥有大量的知识图谱分析方法，包含中心性分析、子群分析、角色分析和基于置换的统计分析等。在数据的标准化处理上，UCINET也提供了Jaccard指数、cohen’s kappa、Identity系数、Correlation、Hamming-Sim等方法。

二、UCINET的主要特点

1.矩阵格式的数据处理

UCINET最大的优点在于它能够将原始数据转化为矩阵格式，从而构建各种关系矩阵，比如：作者共现矩阵、关键词共现矩阵、期刊共被引矩阵等。

在数据的输入方式上，可以用UCINET本身的数据矩阵表（MatrixSpreadsheet）进行输入，也可以读取Excel或常见的文本文件（如KrackPlot、Pajek、Negopy、VNA等格式）。同时，UCINET的矩阵格式数据输入支持全矩阵、半矩阵及多个矩阵的同时输入，用户可以自由对各行、各列添加说明标签，也可以直接导入关联列表格式的数据，方便用户指定数据间的关系。

UCINET在矩阵式处理上的强大功能使得它成为目前较流行、较适合初学者、容易快速上手的知识图谱分析软件。

2.多样化的分析方法

在分析数据、构建网络图谱时，UCINET提供了大量的网络分析方法，如中心度、二方关系凝聚力测度、位置分析算法、拍戏分析、随机二方关系模型、P1以及对网络假设进行检验的程序。UCINET还包含众多的基于过程的分析程序，如角色和地位分析、拟合中心-边缘模型。此外，UCINET也有很多常见的多元统计分析工具，如多维量表（MDS）、对应分析、因子分析（Factor Analysis）、聚类分析

（Cluster Analysis）、针对矩阵的多元回归（Multiple Regression）等。

3.集成可视化软件包

虽然在分析功能上，UCINET足够强大，但它在构建知识图谱的整个流程上缺少数据预处理功能（这或许会对数据分析结果质量的好坏有一定影响）和结果可视化模块。

在可视化功能上，UCINET提供了集成软件包的形式，包括一维与二维数据分析的NetDraw、正在发展应用的三维展示分析软件Mage以及Pajek用于大型网络分析的Free应用软件程序。

三、UCINET的主要分析方法

在UCINET提供的构建知识图谱的方法中，最主要的是网络密度分析、网络中心性分析以及凝聚子群分析3种。

● 网络密度指各成员之间联系的紧密度，可以通过比较图谱网络中实际存在的关系数与理论上可能存在的关系数得到，成员之间联系越多，图谱的网络密度越大。整体图谱的网络密度越大，则各节点的行为对图谱产生影响的可能性越大。

● 网络中心性是度量整个图谱中心化程度的重要指标。例如，在城市群网络中，处于中心位置的城市更容易获得资源和信息，对其他城市有更强的影响力。中心性也可以分为点度中心度、接近中心度和中间中心度3个指标。

● 凝聚子群是指满足特定条件的节点的子集合，即在这个子集合中的节点之间具有较强、直接、紧密、经常的或者积极的关系。例

如，城市网络凝聚子群是用于刻画城市群体内部的子结构状态。找到城市网络凝聚子群的个数以及每个凝聚子群包含的成员，分析凝聚子群之间的关系和连接方式，可以激发用户从新的视角和维度对城市网络群体的发展情况进行考察。

第四节 Gephi

一、Gephi软件概述

Gephi是一款跨平台基于JVM的复杂网络分析软件，主要用于各种网络和复杂系统，可以实现动态和分层图的交互可视化与探测开源工具。同Pajek一样，Gephi可以处理的数据规模巨大，支持100000个节点和1000000条边，适合搭建大型的知识图谱。

Gephi被誉为“数据可视化领域的Photoshop”，它的界面优美，构建在NetBeans平台上，使用Java语言，并且以OpenGL为可视化引擎。它是免费的开源软件，允许开发者扩展和重复使用，依赖于Gephi的API，开发者可以编写自己感兴趣的插件，创建新功能。

作为图谱分析与构建的工具，Gephi有自己的数据实验室，可提供类似Excel的界面来操作数据列、搜索和转换数据。同时Gephi提供了中间中心性、紧密性、直径、聚类系数、页秩、社区检测（模块化）、随机发生器以及最短路径等多种分析方法用于图谱网络的构建。

二、Gephi的主要特点

1.可扩展性

通过可扩展的功能为用户提供个性化和创新性的空间，是Gephi的最大特点。在数据标准化处理上，用户可以自定义插件。同时，Gephi内置的插件中心自动从Gephi插件门户获取可用插件列表，并负责所有软件更新。其他用户可以自由在Gephi的官方网站插件一栏中下载任意需要的插件，增强了用户的自主性。目前，有几十个用户社区构建的插件扩展了Gephi的功能。

2.实时可视化

Gephi由其特别的OpenGL引擎提供支持，致力于研究如何进行交互式和高效的网络探索。Gephi是动态图形分析创新的前沿，支持使用GEXF文件格式导入时间图等方式，也提供了丰富的图形处理工具，用户可以通过操纵嵌入的时间线来展示可视化图谱是如何随着时间而发展的。这可以使用户从图形可视化引擎中最快获益，以加速对大型图谱的理解和模式发现。

3.探索性

Gephi的设计对象是数据分析师和热衷于探索、理解数据图表的科学家。其目的是帮助数据分析师在数据源中进行假设，直观地发现模式，隔离结构奇点或故障。它是传统统计学的一个补充工具，通过

视觉思维与交互界面促进推理。这是一个用于探索性数据分析的软件，一个出现在视觉分析研究领域的范例。

4.动态过滤

Gephi根据网络结构或数据选择节点和边，搭建动态的交互式网络。通过创建不带脚本的复杂筛选器建立以筛选结果为依据的新网络，保存用户的查询习惯，最终实现实时过滤。

第五节 VOSviewer

一、VOSviewer软件概述

VOSviewer是由荷兰莱顿大学的Nees Janvan Eck和Ludo Waltman共同开发的，构建和可视化网络图谱的计量分析软件。虽然VOSviewer软件的开发原理基于文献的共引和共被引原理，但其可以应用在各个领域的数据网络图谱构建中。

VOSviewer免费且专业，具有可视化能力强、适合于大规模样本数据的特点，支持用户通过VOS映射技术和VOS聚类技术创建知识图谱、利用网络数据。在导入数据时，提供文本集（Text Corpus）和网络形式（Network）两种方式；在导出结果时，提供标签视图、密度视图、聚类视图和散点视图4种图谱浏览方式以及缩放、滚动等功能，帮助用户轻松绘制、详细观察知识图谱。此外，VOSviewer还具有文本挖掘功能。

二、VOSviewer的主要特点

1.基于关联强度的数据处理

VOSviewer的最主要特点是在呈现图谱时使用的数据处理技术是关联强度的相似性测量技术，这样在图谱中可以突出最重要（出现频率最高）的标签。同时，用户还可以在视图中放大某个具体的区域，发现隐藏在重要的关键词后面的一些词条，在显示数据集重要信息的同时又可以避免一些重要节点和标签相互覆盖的情况出现。

2.傻瓜化操作

操作简单是VOSviewer受欢迎的重要原因之一。用户只需点击VOSviewer界面左侧Action条目下的Create按钮，按照分析需求依次选择导入形式，导入Web of Science文本文档后，再选择分析范围。用户根据实际情况进行阈值设置后，系统会提示选择分析数据的数量，选定后系统弹出所选数据的详细信息，用户确认后点击完成，即可生成图谱。

3.智能可视化

VOSviewer使用类似于谷歌地图的缩放和滚动功能，可以详细探索图谱，并利用智能标记算法防止标签相互重叠。VOSviewer还应用高级布局和群集技术以及自然语言处理技术，提供图谱关键部分的快

速概述和图谱随时间变化的演变轨迹。同时，VOSviewer可以以高分辨率创建屏幕截图，并支持多种图形文件保存格式，如位图和矢量格式。

三、VOSviewer的结果呈现

● 标签视图（Label View）：每个节点用一个圆圈和标签表示，圆圈的大小代表节点的重要程度，如果节点被划分为不同的聚类，其圆圈则呈现不同颜色。为了避免标签重叠，这一视图下一般只显示标签的子集，可以通过“放大”操作详细查看图谱上每个节点的标签。

● 密度视图（Density View）：图谱上的每个节点都根据其密度进行颜色填充，缺省颜色是红色和蓝色。一个节点越大，表示其权重越大，颜色越接近于红色。相反，如果其权重越小，则颜色越接近于蓝色。密度视图中各节点的标识与其在标签视图下相同，这一模式的作用是可以快速查看图谱中的关键区域。

● 聚类密度视图（Cluster Density View）：只有节点被分成了不同聚类族才会使用这种视图。在这种模式下，每个节点的颜色和其特定的聚类族一致，视图效果虽与密度视图接近，但这种视图对于快速发现每个聚类族更有效果。

● 散点视图（Scatter View）：散点视图是一个简单视图，图中每个节点都表现为一个小的圆环。如果节点被聚为不同的类，则每个圆环用同样的颜色表示。散点视图不展现节点的标签，有利于快速查看图谱的结构。

第六节 VantagePoint

一、VantagePoint软件概述

VantagePoint是开发商Search Technology开发的一种数据挖掘产品，是一种将信息转化为知识的桌面软件，目前，主要应用于深层次挖掘专利信息。VantagePointt拥有业界领先的信息摄取、提炼和分析工具以及灵活的报告界面。它采用多种算法，通过模型匹配、基础规则和自然语言加工技术等进行文本挖掘。

VantagePoint是一个基于Windows操作环境的商业软件，只提供免费的试用版本，但软件操作简单、价格合理。作为数据挖掘工具，VantagePoint不提供数据库服务，系统使用的数据由用户直接向数据供应商购买。

在数据的预处理上，VantagePoint支持去重处理、时间切片和数据缩减。在关系矩阵的构建上，VantagePoint可以直接用具体的字段构建一些矩阵，如异质网络，通过在行和列中使用不同的字段，可以抽取节点不同时间点的变化矩阵。在数据的标准化处理上，Vantage Point有Pearson's、Salton余弦、最大均衡等方法。在图谱搭建方法上，VantagePoint的地理空间分析效果很好。在处理结果上，VantagePoint可以提供一维表格和二维交叉列表，也可提供多维分析功能，使用户更方便地进行聚类分析等处理。

VantagePoint的设计目标是令数据分析师能够利用他们的专业知识产生卓越的结果，旨在提高分析效率和增加吞吐量，节省人力和资源。在分析时，VantagePoint被固定在信息的统计分析视角中，这样的方法可以提高用户的检索速度。

二、VantagePoint的主要特点

1.优秀的数据预处理和数据清理功能

VantagePoint有多达180个过滤器，通过专门的过滤器可以允许用户输入很多格式的数据，包括：*.txt, *.dat, *.csv, *.tag, *.trn,*.xls, *.xlsx, *.mdb, *.accdb, *.ris, *.xml等。

VantagePoint提供数据清洗工具（Data Cleaning Tool），使用Clean Up功能，对数据进行清理，它应用模糊匹配技术识别和整理数据，以减少重复和不规范的数据量。例如该工具可以处理拼写错误、连字符号、大小写以及不同人名拼写习惯等，从而提高数据质量。

2.有效的数据归档工具

VantagePoint 允许用户创建用户管理辞典（User-Managed Thesauruses），提炼特定数据。利用辞典，用户可以方便地综合某一数据变量的多种形式。例如，美国有United States、US、U.S、USA等描述形式，通过用户辞典，系统会将这些有关美国的不同描述形式视为同义词，自动进行归一化处理。

此外，用户可以进行其他类型的整理，如综合数据要素到更宽泛的目录中。例如将“Aluminum Alloys”“Magnesium Alloys”“Carbon Fiber Reinforced Plastics”和“Copper Alloys”等词条归类到材料类（Materials）中，或将美国、加拿大、墨西哥归类到“北美”类目中。

3.安全的应用环境

VantagePoint是一个灵活的桌面应用程序，它提供的是一个本地分析环境。用户通过数据库供应商提供的搜索引擎进行数据检索，将原始数据下载到用户计算机上后将数据导入VantagePoint，系统为每个数据库或数据供应商提供唯一的数据库文件结构。通过模型匹配、基础规则等进行数据挖掘。在这样的过程中，所有敏感信息和过程都在防火墙的一侧。

第七节 Sci^2

一、Sci^2 软件概述

Sci^2 （Science of Science）是美国印第安纳大学Katy Borner及其团队开发的一款知识图谱分析软件，它是专门为科学研究设计的模块化工具集。它支持时间、地理空间、主题和网络分析以及微观（个体）、中微观（局部）和宏观（全局）层面的数据集可视化。

目前，美国国家科学基金会（National Science Foundation，NSF）、美国国立卫生研究院（National Institutes of Health，NIH）、美国农业部（US Department of Agriculture，USDA）以及美国国家海洋和大气管理局（National Oceanic and Atmospheric Administration，NOAA）等都在使用Sci^2 进行数据可视化分析。

Sci^2 构建在CyberInFrastructure Shell（CIshell）上，CIshell是一个功能强大的开放源代码的Eclipse插件框架，它以赛百平台（Cyber in Frastructure）为理论基础，主要用于轻松集成和利用数据集、算法、工具和计算资源。Sci^2 支持的示例性参考系统包括时间条形图、共面图、UCSD科学图、双峰网络可视化。

与其他知识图谱工具相比，Sci^2 在插件、算法等方面具有优势，但也存在一定的局限性。首先，运行Sci^2 时，需要占用大量的内存，对计算机系统的要求比较高，这主要是由Java虚拟机的限制造成的。另外，由于Sci^2 是国外学者主持开发的，目前还没有出现中文版的Sci^2 ，对于国内用户来说，使用Sci^2 前还需要一定的文本转换。

二、Sci^2 的主要特点

1.丰富的插件

Sci^2 的一大优势是拥有丰富多样的插件可供使用，这就为用户使用Sci^2 绘制各类知识图谱提供了强力支持。其中，OSGi、CIShell等插件运行在核心框架上；另外一些算法插件因其自身功能的不同，分布在不同菜单栏中，为数据准备、预处理、分析、建模、可视化等操作提供服务。这样，用户不仅可以使用该软件预先打包好的各种插件，而且可以根据自己的不同需求，创建、下载、共享并导入插件，不断丰富Sci^2 的现有功能。例如，用户可以从Sci^2 的官网上下载有关数据库、气球图（Balloon Graph）、国会地理编码（Congressional District Geocoder）、Cytoscape不同插件，将这些文件复制到Sci^2 的目录文件夹下的插件文件夹中，即可使用这些插件。

2.强大的数据处理能力

Sci^2 集成了各种数据处理功能，它具有强大的数据处理能力，在数据预处理方面的优势明显，提供去重处理、时间切片、数据和网络

的缩减等功能，能够根据具体的需要对数据进行对应的预处理。

当数据量很大时，用户可以根据自己的需要，选择对数据进行相关处理，去除一些孤立节点，抽取前*N* 个节点和边，选择先进行数据处理，再分析网络情况。用户也可以先分析网络情况，事先了解网络中孤立节点的数量以及边的权重（最大值、最小值、均值），再根据需要选择提取前*N* 个节点和边，或者进行其他处理。

3.多种可视化模式

在进行数据可视化时，Sci^2 支持绘制多种形式的可视化模式（如GUESS、Cytoscape）等。一方面，Sci^2 可以很容易地整合各种数据集、方程、工具和计算机资源；另一方面，许多可视化插件也可以根据需要，帮助用户完成对特定数据集的交互式探索和分析。

目前，比较常用的可视化插件是GUESS。当网络数据很大时（如进行引文分析），可以用DrL算法先将网络进行一定的缩减。

Cytoscape则是一种网络分析和可视化的通用平台，含有多种布局算法（如Cyclic、Tree、Force-Directed、Edge-Weight等）。在最新发布的软件版本中，R语言和Gephi可视化工具还可以以插件的形式与Sci^2 结合，使得Sci^2 的可视化功能更加强大。

第八节 SciMAT

一、SciMAT软件概述

SciMAT是由西班牙格拉纳达大学于2012年开发的一种较新的知识图谱构建工具，具有强大的前处理模块和使用向导配置分析的功能。它提供了知识图谱构建流程中从预处理到结果可视化的所有步骤的方法、算法和相似性度量。

SciMAT是一个开源的工具，用于在纵向框架下执行知识图谱分析。它使用Java语言开发，代码规范，可读性好；接口设置灵活，扩展性好。由于SciMAT的设计初衷是基于文献进行科学知识的图谱绘制，因此SciMAT还提供了内置的数据库，用于文献元数据的管理、数据本地储存，有利于用户进行个人数据资料管理。

SciMAT可以构建多种知识网络，并采用计量学指标对结果的影响力进行定量分析。在可视化模块中，通过战略图、集群网络、重叠图和进化区域4种方式进行展示，帮助用户更好地理解分析结果。它的用户界面友好，操作简单，分析过程通过配置向导的方式引导用户进行分析，每个过程提供相应的选项供用户选择。

二、SciMAT的主要特点

1.全流程分析

SciMAT集成了知识图谱构建的所有步骤所需的全部模块，这些步骤可以进行即时配置。SciMAT提供了不同的模块，有专门负责管理知识库和实体的模块，有负责进行图谱网络分析的模块，也有生成可视化结果和映射的模块。这些模块为用户提供了贯穿从数据采集和预处理到结果的可视化和解释的知识图谱构建的全过程。但是SciMAT的可视化效果一般，比不上Pajek、VOSviewer等专业可视化工具。

2.时序分析

SciMAT是知识图谱分析工具中，进行纵向时序分析最好的软件。它允许数据分析师在纵向框架中进行知识图谱的构建，以便在连续的时间段内分析和跟踪研究领域的概念、智力或社会演变。它的时间序列呈现方式简捷，方便用户轻松判定分析对象的发展过程，并判断出关键的节点或边。

3.数据管理和预处理能力

SciMAT的数据管理面板可以实现记录的删除和更新。元数据采用分组管理，分析过程以分组为基本单位，可以有效地解决实体名称的规范问题、相同概念的甄别问题。

SciMAT具有强大的预处理能力，提供了广泛的数据预处理工具，如检测重复和拼写错误的项目、时间切片、数据简化和网络预处理等，用户可以根据不同的分析问题构建不同的项目，以方便地实现项目数据和分析过程的导入导出。在预处理模块中，SciMAT还内置了数据批量管理和替换功能以及一些相似概念的检测算法，帮助用户快速完成数据的分类和预处理。

参考文献

[1] 徐增林，盛泳潘，贺丽荣，等.知识图谱技术综述[J].电子科技大学学报，2016, 45(4): 589-606.

[2] 曹倩，赵一鸣.知识图谱的技术实现流程及相关应用[J].情报理论与实践，2015, 12(38): 127-132.

[3] 刘峤，李杨，杨段宏，等.知识图谱构建技术综述[J].计算机研究与发展，2016, 53(3): 582-600.

[4] 庄严，李国良，冯建华.知识库实体对齐技术综述[J].计算机研究与发展，2016, 53(1): 165-192.

[5] BIZER C, LEHMANN J, KOBILAROV G, et al.DBpedia—a crystallization point for the Web of data[J].Web Semantics Science Services & Agents on the World Wide Web, 2009, 7(3):154-165.

[6] SUCHANEK F M, KASNECI G, WEIKUM G.YAGO: a large ontology from wikipedia and wordnet[J].Web Semantics Science Services & Agents on the World Wide Web, 2007, 6(3):203-217.

[7] MILLER G A.WordNet: a lexical database for English[J].Communications of the ACM, 1995, 38(11): 39-41.

[8] 孙镇，王惠临.命名实体识别研究进展综述[J].现代图书情报技术,2010(6): 42-47.

[9] 刘知远，孙茂松，林衍凯，等.知识表示学习研究进展[J].计算机研究与发展，2016, 53(2): 1-16.

[10] 欧艳鹏.知识图谱技术研究综述[J].电子世界，2018, 547(13): 56, 58.

[11] 段宏.知识图谱构建技术综述[J].计算机研究与发展，2016, 5(3):582-600.

[12] 李涓子，侯磊.知识图谱研究综述[J].山西大学学报（自然科学版）,2017(3): 454-459.

[13] 梁秀娟.科学知识图谱研究综述[J].图书馆杂志，2009(6): 58-62.

[14] 胡泽文，孙建军，武夷山.国内知识图谱应用研究综述[J].图书情报工作，2013, 57(3): 131-137.

[15] 张香玲，陈跃国，马登豪，等.实体搜索综述[J].软件学报，2017(6):1584-1605.

[16] HAN J W, KAMBE M.Data mining: concepts and techniques[M].San Francisco: Morgan Kaufmann, 2006.

[17] VAPNIK V.The nature of statistical learning theory[M].Berlin:Springer, 2000.

[18] KANTARDZIC M.Data mining[M].Hoboken: John Wiley &Sons, 2011.

[19] COCHINWALA M, KURIEN V, LALK G, et al.Efficient data reconciliation[J].Information Sciences, 2011, 137(14): 1-15.

[20] CHRISTEN P.Automatic training example selection for scalable unsupervised record linkage[C]// The 12th PacificAsia Conference on Advances in Knowledge Discovery and Data Mining, May 20-23, 2008, Osaka, Japan.Berlin: Springer, 2008.

[21] TAN C H, AGICHTEIN E, IPEIROTIS P, et al.Trust, butverify:predicting contribution quality for knowledge baseconstruction and curation[C]//The 7th ACM International Conference on Web Search and Data Mining, February 24-28, 2014, New York,USA.New York: ACM Press, 2014: 553-562.

[22] 耿霞，张继军，李蔚妍.知识图谱构建技术综述[J].计算机科学，2014,41(7): 148-152.

[23] LACOSTE-JULIEN S, PALLA K, DAVIES A, et al.SIGMA:simple greedy matching for aligning large knowledgebases[C]//The 2013 ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 11-14, 2013, Chicago, USA.New York: ACM Press, 2013: 572-580.

[24] BHATTACHARYA I, GETOOR L.Alatent dirichletallocation model for unsupervised entity resolution[C]//The 6th SIAM Int Conf on Data Mining, April 20–22, 2006, Bethesda, USA.Philadelphia: SIAM, 2006: 47-58.

[25] DOMINGOS P.Multi-relational record linkage[C]//The 3rd International Workshop on Muti-Relational Data Mining,August 22, 2004, Seattle, USA.New York: ACM Press, 2004.

[26] 史树明.自动和半自动知识提取[J].中国计算机学会通讯，2013, 9(8):65-73.

[27] HARRIS Z S.Distributional structure[J].Word, 1954, 10(23):146-162.

[28] 雷二庆.h指数知识图谱分析[J].科研管理，2010(s1): 20-23.

[29] 雷会珠，姚立会.知识地图与科学知识图谱辨析[J].中国科技信息,2012(10): 59.

[30] 李涛，王次臣，李华康.知识图谱的发展与构建[J].南京理工大学学报（自然科学版），2017, 41(1): 22-34.

[31] ZENG Y, WANG D S, ZHANG T L, et al.CASIA-KB: a multisource Chinese semantic knowledgebase built from structured and unstructured Web data[C]//The 3rd Joint International

Conference, November 28-30, 2013, Seoul, Korea.Berlin:Springer, 2014: 75-88.

[32] BLANCO R, CAMBAZOGLU B B, MIKE P, et al.Entity recommendation in web search[C]//The 12th International Semantic Web Conference(ISWC), October 21-25, 2013,Sydney, Australia.Berlin: Springer-Verlag, 2013: 33-48.

[33] 漆桂林, 高桓, 吴天星.知识图谱研究进展[J].情报工程, 2017, 3(1):4-25.

[34] PRICE D J.Networks of scientific papers[J].Science, 1965,149(3683): 510-515.

[35] HODGE G.Next generation knowledge organization systems:integration challenges and strategies[C]//The 5th ACM/IEEECS Joint Conference on Digital Libraries, June 7-11, 2005,Denver, USA.Piscataway: IEEE Press, 2005.

[36] ZHANG X, DU C, LI P, et al.Knowledge graph completion via local semantic contexts[C]//The International Conference on Database Systems for Advanced Applications, April 16-19, 2016, Dallas, Texas.Berlin: Springer International Publishing,2016.

[37] 朱木易洁，鲍秉坤，徐常胜.知识图谱发展与构建的研究进展[J].南京信息工程大学学报（自然科学版），2017，9(6): 575-582.

[38] 焦晓静，王兰成.知识图谱的概念辨析与学科定位研究[J].图书情报工作，2015(15): 5-11.

[39] 王新才，丁家友.大数据知识图谱: 概念、特征、应用与影响[J].情报科学，2013(9): 10-14.

[40] 秦长江，侯汉清.知识图谱——信息管理与知识管理的新领域[J].大学图书馆学报，2009, 27(1): 30-37, 96.

[41] FANG L J, SARMA A D, YU C, et al.REX: explaining relationships between entity pairs[J].Proceedings of the VLDB Endowment, 2011, 5(3): 241-252.

[42] SUCHANEK F M, KASNECI G, WEIKUM G.Yago: a core of semantic knowledge[C]//The 16th International Conference on World Wide Web, May 8-12, 2007, Banff, Canada.New York:ACM Press, 2007: 697-706.

[43] LAO N, MITCHELL T M, COHEN W W.Random walk inference and learning in a large scale knowledge base[C]//The 2011 Conference on Empirical Methods in Natural Language Processing, July 27-31, 2011, Edinburgh, UK.Stroudsburg:Association for Computational Linguistics, 2011: 529-539.

[44] 冯新翎，何胜，熊太纯，等.“科学知识图谱”与“Google知识图谱”比较分析——基于知识管理理论视角[J].情报杂志，2017, 36(1):149-153.

[45] 陈悦，刘则渊.悄然兴起的科学知识图谱[J].科学学研究，2005,23(2): 149-154.

[46] 杨思洛，韩瑞珍.国外知识图谱的应用研究现状分析[J].情报资料工作，2013(6): 15-20.

[47] JIA Y, WANG Y, CHENG X, et al.OpenKN: an open knowledge computational engine for network big data[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), August 17-20,2014, Beijing, China.Piscataway: IEEE Press, 2014.

[48] HOFFART J, SUCHANEK F M, BERBERICH K, et al.YAGO2:a spatially and temporally enhanced knowledge base from Wikipedia[J].Artificial Intelligence, 2013, 194: 28-61.

[49] SUCHANEK F M, KASNECI G, WEIKUM A G.Yago - a large ontology from wikipedia and wordnet[J].Web Semantics Science Services & Agents on the World Wide Web, 2008, 6(3):203-217.

[50] AUER S, BIZER C, KOBILAROV G, et al.DBpedia: a nucleus for a web of open data[J].Semantic Web, 2007, 4825:11-15.

[51] LENAT D B.CYC: a large-scale investment in knowledge infrastructure[J].Communications of the ACM, 1995, 38(11):33-38.

[52] BOLLACKER K, COOK R, TUFTS P.Freebase: a shared database of structured general human knowledge[C]//The 22nd National Conference on Artificial Intelligence, July 22-26,2007, Vancouver, Canada.Palo Alto: AAAI Press, 2007.

[53] MA Y, QI G.An analysis of data quality in DBpedia and Zhishi.me[C]// China Semantic Web Symposium and Web Science Conference, August 12-18, 2013, Shanghai, China.Berlin:Springer, 2013.

[54] XING N, SUN X, WANG H, et al.Zhishi.me: weaving Chinese linking open data[C]// the 10th International Conference on the Semantic Web, October 23-27, 2011, Bonn, Germany.Berlin:Springer-Verlag, 2011.

[55] 胡芳槐.基于多种数据源的中文知识图谱构建方法研究[D].上海：华东理工大学，2015.

[56] 杨思洛，韩瑞珍.知识图谱研究现状及趋势的可视化分析[J].情报资料工作，2012(4): 22-28.

[57] SOWA J F.Principles of semantic networks: exploration in the representation of knowledge[J].Frame Problem in Artificial Intelligence, 1991(2-3): 135-157.

[58] 袁国铭，李洪奇，樊波.关于知识工程的发展综述[J].计算技术与自动化，2011, 30(1): 138-143.

[59] 陈和.机构知识库发展趋势探析[J].图书情报工作，2012, 56(21):62-66.

[60] 张晓林.机构知识库的发展趋势与挑战[J].数据分析与知识发现,2014, 30(2): 1-7.

[61] ETZIONI O, CAFARELLA M, DOWNEY D, et al.Web-scale information extraction in knowitall:(preliminary results)[C]//The 13th International Conference on World Wide Web, May 17-20,2004, New York, USA.New York: ACM Press, 2004: 100-110.

[62] YATES A, CAFARELLA M, BANKO M, et al.TextRunner:open information extraction on the web[C]// Human Language Technologies: the Conference of the North American Chapter of the Association for Computational Linguistics:Demonstrations, April 23-25, 2007, Rochester, USA.Stroudsburg: Association for Computational Linguistics, 2007.

[63] NICKEL M, TRESP V, KRIEGEL H P.A three-way model for collective learning on multi-relational data[C]//The 28th International Conference on Machine Learning, June 28-July 2,2011, Bellevue, USA.Athens: OmniPress, 2011: 809-816.

[64] CARLSON A, BETTERIDGE J, KISIEL B, et al.Toward an architecture for never-ending language learning[C]//The 24th AAAI Conference on Artificial Intelligence, July 11-15, 2010,Atlanta, Georgia.Palo Alto: AAAI Press, 2010.

[65] 赵军，刘康，周光有，等.开放式文本信息抽取[J].中文信息学报，2011, 25(6): 98-111.

[66] 杨博，蔡东风，杨华.开放式信息抽取研究进展[J].中文信息学报，2014, 28(4): 1-11.

[67] 王元卓，贾岩涛，刘大伟，等.基于开放网络知识的信息检索与数据挖掘[J].计算机研究与发展，2015, 52(2): 456-474.

[68] 王宇，谭松波，廖祥文，等.基于扩展领域模型的有名属性抽取[J].计算机研究与发展，2010, 47(9): 1567-1573.

[69] ZHANG Y, AI Q, CHEN X, et al.Learning over knowledgebase embeddings for recommendation[J].Computer Science,2018, arXiv:1803.06540.

[70] CRAVEN M, DAN D P, FREITAG D, et al.Learning to construct knowledge bases from the World Wide Web[J].Artificial Intelligence, 2000, 118(1): 69-113.

[71] KUMAR R, RAGHAVAN P, RAJAGOPALAN S, et al.Extracting large-scale knowledge bases from the Web[C]//The 25th International Conference on Very Large Data Bases,September 7-10, 1999, Edinburgh, Scotland.San Francisco:Morgan Kaufmann Publishers Inc., 1999: 639-650.

[72] SOCHER R, CHEN D, MANNING C D, et al.Reasoning with neural tensor networks for knowledge base completion[C]//The 26th International Conference on Neural Information Processing Systems, December 5-10, 2013, Lake Tahoe, USA.New York:Curran Associates Inc., 2013.

[73] CRAVEN M, KUMLIEN J.Constructing biological knowledge bases by extracting information from text sources[C]//The 7th International Conference on Intelligent Systems for Molecular Biology, August 6-10, 1999, Heidelberg, Germany.Palo Alto:AAAI Press, 1999: 77-86.

[74] 杨莉，胡守仁.知识库推理和维护系统 (KBIMS)[J].国防科技大学学报，1991(2): 127-133.

[75] 鄢珞青.知识库的知识表达方式探讨[J].情报杂志，2003, 22(4):63-64.

[76] 卢道设，杨世瀚，吴尽昭，等.基于描述逻辑的组合知识库推理[J].计算机应用研究，2012, 29(12)： 4503-4506.

[77] 蒋勋，徐绪堪.面向知识服务的知识库逻辑结构模型[J].图书与情报，2013, 2013(6)： 23-31.

[78] 祝忠明.机构知识库开源软件DSpace的扩展开发与应用[J].数据分析与知识发现，2009, 25(7-8)： 11-17.

[79] 王志，夏士雄，牛强.本体知识库的自然语言查询重写研究[J].微电子学与计算机，2009, 26(8)： 137-139.

[80] WANG Z, XIA S X, NIU Q.Research on natual language query rewriting for ontology-based knowledge base[J].Microellectronics & Computer, 2009, 26(8)： 137-139.

[81] LIN Y, LIU Z, XUAN Z, et al.Learning entity and relation embeddings for knowledge graph completion[C]//The 29th AAAI Conference on Artificial Intelligence, January 25-30,2015, Austin, USA.Palo Alto: AAAI Press, 2015： 2181-2187.

[82] SHI B, WENINGER T.ProjE: embedding projection for knowledge graph completion[J].Computer Science, 2016,arXiv:1611.05425.

[83] MENG W.Predicting rich drug-drug interactions via biomedical knowledge graphs and text jointly embedding[J].Computer Science, 2018, arXiv:1712.08875.

[84] FAN M, ZHOU Q, ZHENG T F, et al.Distributed representation learning for knowledge graphs with entity

descriptions[J].Pattern Recognition Letters, 2016: S0167865516302380.

[85] 金嘉晖.面向大规模知识图谱的分布式查询技术研究[D].南京：东南大学，2015.

[86] 韩先培.基于语义知识挖掘与融合的实体消歧技术研究[D].北京：中国科学院研究生院，2010.

[87] 李宁，李秉严.知识挖掘技术及应用[J].情报杂志，2003, 22(6): 34-36.

[88] 王知津.论知识组织的十大原则[J].国家图书馆学刊，2012, 21(4):3-11.

[89] LIN Y, HAN X, XIE R, et al.Knowledge representation learning: a quantitative review[J].Computer Science, 2018,arXiv:1812.10901.

[90] WU W, LI H, WANG H, et al.Probase: a probabilistic taxonomy for text understanding[C]//ACM International Conference on Management of Data, May 20-24, 2012,Scottsdale, AZ, USA.New York: ACM Press, 2012.

[91] ZETTLEMOYER L S, COLLINS M.Learning to map sentences to logical form: structured classification with probabilistic categorial grammars[C]//The 21st Conference on Uncertainty in Artificial Intelligence, July 26-29, 2005, Edinburgh, Scotland.Arlington: AUAI Press, 2012: 658-666.

[92] 陈祖香.面向科学计量分析的知识图谱构建与应用研究[D].南京: 南京理工大学，2010.

[93] 苏永浩，张驰，程文亮，等.CLEQS——基于知识图谱构建的跨语言实体查询系统[J].计算机应用，2016, 36(s1): 204-206.

[94] 陈悦，陈超美，刘则渊，等.CiteSpace知识图谱的方法论功能[J].科学学研究, 2015, 33(2): 242-253.

[95] HERZOG T N, SCHEUREN F J, WINKLER W E.Data quality and record linkage techniques[M].New York: Springer New York, 2007.

[96] THAMARAISELVI G, KALIAMMAL A.Data mining: concepts and techniques[J].Data Mining Concepts Models Methods &Algorithms Second Edition, 2006, 5(4): 1-18.

[97] WITTEN I.Data mining[M].New York: Elsevier Science Inc.,2012.

[98] 唐晓波，魏巍.知识融合: 大数据时代知识服务的增长点[J].图书馆学研究，2015(5): 9-14.

[99] 郭强，关欣，曹昕莹，等.知识融合理论研究发展与展望[J].中国电子科学研究院学报，2012, 7(3): 252-257.

[100] 缑锦.知识融合中若干关键技术研究[D].杭州：浙江大学，2005.

[101] 林海伦，王元卓，贾岩涛，等.面向网络大数据的知识融合方法综述[J].计算机学报，2017(1): 1-27.

[102] XIN L D, GABRILOVICH E, HEITZ G, et al.From data fusion to knowledge fusion[J].Proceedings of the VLDB Endowment,2015, 7(10): 881-892.

[103] GRAY P M D, PREECE A, FIDDIAN N J, et al.KRAFT:knowledge fusion from distributed databases and knowledge[C]//The 8th International Conference on Database and Expert Systems Applications, September 1-2, 1997, Toulouse,France.Piscataway: IEEE Press, 1997.

[104] ROEMER M J, KACPRZYNSKI G J, ORSAGH R F.Assessment of data and knowledge fusion strategies for prognostics and health management[C]// 2001 IEEE Aerospace Conference Proceedings, March 10-17, 2001, Big Sky, USA.Piscataway:IEEE Press, 2001.

[105] 张旭.知识图谱技术落地金融行业的关键四步[J].金融电子化，2017(11): 92.

[106] 邵领.基于知识图谱的搜索引擎技术研究与应用[D].成都：电子科技大学, 2016.

[107] 赵鑫.刍议搜索引擎中知识图谱技术[J].辽宁行政学院学报，2014,16(10): 150-151.

[108] 张观林，欧阳纯萍，邹银凤，等.知识图谱及其在医疗领域的应用[J].湖南科技学院学报，2016, 37(10): 73-75.

[109] FADER A, ZETTLEMOYER L, ETZIONI O.Open question answering over curated and extracted knowledge bases[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2014, New York,USA.New York: ACM Press, 2014.

[110] BORDES A, CHOPRA S, WESTON J.Question answering with subgraph embeddings[J].Computer Science, 2014,arXiv:1406.3676.

[111] BORDES A, WESTON J, USUNIER N.Open question answering with weakly supervised embedding models[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases, September 15-19, 2014, Nancy, France. [S.l.:s.n.],2014.

[112] 鲁轶奇.知识图谱的数据清理和应用探索[D].上海：复旦大学，2013.

[113] 雷丰羽.知识图谱在金融信贷领域的应用[J].现代商业，2018,491(10): 91-92.

[114] 刘柳.知识图谱的行业应用与未来发展[J].互联网经济，2018, 38(4):18-23.

[115] BLANCO R, CAMBAZOGLU B B, MIKE P, et al.Entity recom mendation in web search[C]//The 12th International Semantic Web Conference(ISWC), October 21-25, 2013, Zurich,Switzerland.Berlin: Springer-Verlag, 2013: 33-48.

[116] BRACHMAN R J.What IS-A is and isn’t: an analysis of taxonomic links in semantic networks[J].Computer, 1983,10(1): 5-13.