知识图谱标准化白皮书

（2019版）

主编单位：中国电子技术标准化研究院

二零一九年八月

牵头编写单位：

中国电子技术标准化研究院

联合编写单位：

第一章 背景介绍

北京京航计算通讯研究所

南华大学

中电科大数据研究院有限公司

第二章 知识图谱需求

联想(北京)有限公司

成都数联铭品科技有限公司

阿里巴巴网络技术有限公司

上海思贤信息技术股份有限公司

创新工场人工智能工程院

南华大学

南京柯基数据科技有限公司

中车株洲电力机车研究所有限公司

上海交通大学

第三章 知识图谱推动条件和驱动因素

中电科大数据研究院有限公司

阿里巴巴网络技术有限公司

上海思贤信息技术股份有限公司

创新工场人工智能工程院

中国电子科技集团公司第二十八研究所

成都数联铭品科技有限公司

北京京航计算通讯研究所

湖北畅云时讯软件技术有限公司

中车株洲电力机车研究所有限公司

南华大学

上海交通大学

阿里云计算有限公司

第四章 知识图谱主要技术

南华大学

中电科大数据研究院有限公司

东软集团股份有限公司

联想(北京)有限公司

星环信息科技（上海）有限公司

成都数联铭品科技有限公司

阿里巴巴网络技术有限公司

上海思贤信息技术股份有限公司

创新工场人工智能工程院

南京柯基数据科技有限公司

拓尔思信息技术股份有限公司

天津大学

上海交通大学

第五章 知识图谱应用

中电科大数据研究院有限公司

东软集团股份有限公司

联想(北京)有限公司

星环信息科技（上海）有限公司

南华大学

成都数联铭品科技有限公司

阿里巴巴网络技术有限公司

上海思贤信息技术股份有限公司

创新工场人工智能工程院
南京柯基数据科技有限公司
拓尔思信息技术股份有限公司
北京京航计算通讯研究所
湖北畅云时讯软件技术有限公司
中车株洲电力机车研究所有限公司
中国电子科技集团公司第二十八研究所
京东集团
北京欧拉认知智能科技有限公司
上海交通大学
北京三快在线科技有限公司（美团点评集团）

第六章 知识图谱存在的挑战

中电科大数据研究院有限公司
东软集团股份有限公司
南京柯基数据科技有限公司
联想(北京)有限公司
星环信息科技（上海）有限公司
创新工场人工智能工程院
成都数联铭品科技有限公司

第七章 知识图谱标准化现状与需求

中电科大数据研究院有限公司
成都数联铭品科技有限公司

第八章 结论与展望

中电科大数据研究院有限公司
南京柯基数据科技有限公司

编写组成员

中国电子技术标准化研究院

郭　楠　　韦　莎　　李瑞琪　　李　佳　　张　欣

夏娣娜　　马原野　　何宏宏　　程雨航　　纪婷钰

中电科大数据研究院有限公司

曹　扬　　程　序　　王　进　　张婧慧　　蔡惠民

邹丽华　　阚丹会

东软集团股份有限公司

赵耕弘　　汤一凡　　蔡　巍　　崔朝辉

联想(北京)有限公司

杨　帆　　范雪琴　　邓人菠　　陶宏芝

南华大学

刘永彬　　欧阳纯萍　罗凌云　　阳小华　　万亚平

星环信息科技（上海）有限公司

杨一帆　　宋群豪　　李　逸　　安　磊　　朱哲伟

上海思贤信息技术股份有限公司

王　莹　　邱梦娟　　何志亮　　安自珍

成都数联铭品科技有限公司

任　渝　　查　琳　　岳　冰　　黄　蓉

阿里巴巴网络技术有限公司

胡　宁　　王　昊　　戴　健　　杨勇华

南京柯基数据科技有限公司

漆桂林　　吴　刚　　王虎斌

创新工场人工智能工程院

王博宇　　程正涛

上海交通大学

金耀辉

拓尔思信息技术股份有限公司

王亚强　　岳　松

北京京航计算通讯研究所

王凤敏　　白　洋　　顾楠希

中车株洲电力机车研究所有限公司

张慧源　　李　晨　　熊敏君

中国电子科技集团公司第二十八研究所

张　超　　周　颖

湖北畅云时讯软件技术有限公司

袁　野　　陈　榴

京东集团

赵　楠

阿里云计算有限公司

李龑翔

北京欧拉认知智能科技有限公司

王　涛　　王绪刚

天津大学

刘安安

北京三快在线科技有限公司（美团点评集团）

万峻辰

编辑组成员

李瑞琪　　王　进　　张　欣

目 录

第一章 背景介绍 …… 1

一、知识图谱的起源与发展 …… 1

二、知识图谱的定义 …… 5

三、白皮书范围 …… 6

四、白皮书概述 …… 7

第二章 知识图谱需求 …… 9

一、公共安全 …… 9

二、数字经济 …… 10

三、金融科技 …… 12

四、资源优化 …… 15

五、科学研究 …… 16

六、制造业转型 …… 18

七、公众健康 …… 19

八、人文发展 …… 21

九、政府管理 …… 23

第三章 知识图谱推动条件和驱动因素 …… 27

一、 推动条件 …… 27

二、 驱动因素 …… 36

第四章　知识图谱的主要技术 …… 47
一、知识获取 …… 47
二、知识表示 …… 52
三、知识存储 …… 57
四、知识融合 …… 63
五、知识建模 …… 68
六、知识计算 …… 71
七、知识运维 …… 80

第五章　知识图谱应用 …… 87
一、智慧金融 …… 87
二、智慧医疗 …… 94
三、智能制造 …… 99
四、智慧教育 …… 104
五、智慧政务 …… 108
六、智慧司法 …… 112
七、智慧交通 …… 117
八、智能电网 …… 121
九、智能公安 …… 125
十、智慧农业 …… 130
十一、社交网络 …… 133
十二、新零售 …… 136
十三、智慧外交 …… 141

十四、智慧城建……143
十五、其他……147

第六章　知识图谱存在的挑战……149
一、数据相关的挑战……149
二、算法相关的挑战……150
三、基础知识库相关的挑战……154
四、开发工具相关的挑战……155
五、隐私、安全和伦理相关的挑战……157
六、测试认证相关的挑战……160
七、商业模式相关的挑战……162
八、人才相关的挑战……164

第七章　知识图谱的标准化现状与需求……167
一、知识图谱标准化现状……167
二、知识图谱标准体系……170
三、知识图谱标准化需求……175

第八章　结论与展望……179
一、趋势展望……179
二、支持与监管政策相关建议……180
三、技术开发与应用相关建议……182

图目录

图1.1 知识图谱发展历史 ……1
图1.2 人工智能发展过程中数据关联的必要性 ……4
图1.3《知识图谱标准化白皮书》整体结构 ……7
图2.1 我国数字经济发展情况 …… 11
图2.2 金融生态的金融科技要素分布 …… 13
图3.1 全球数据总量预测2016–2025年 …… 29
图3.2 全球数据圈规模 …… 33
图3.3 全球计算机视觉行业投资趋势 …… 44
图4.1 知识图谱技术架构图 …… 47
图4.2 知识获取示意图 …… 48
图4.3 知识图谱的向量表示 …… 53
图4.4 TransE模型 …… 55
图4.5 带条件损失度量的KR–EAR方法 …… 55
图4.6 知识存储方式 …… 58
图4.7 属性图结构 …… 59
图4.8 普通图与超图 …… 60
图4.9关系数据库的使用热度趋势 …… 61
图4.10 基于图模型数据库的使用热度趋势 …… 62
图4.11 知识融合概念分解 …… 64
图4.12 自顶向下的构建方法 …… 68

图4.13 自底向上的构建方法 …… 68
图4.14 手工建模方式 …… 69
图4.15 半自动建模方式 …… 70
图4.16 知识计算概念 …… 72
图4.17 图信息检索示例 …… 73
图4.18 图特征统计示例 …… 74
图4.19 关联分析计算示例 …… 75
图4.20 异常检测示例 …… 76
图4.21 知识推理示例 …… 78
图4.22 知识运维 …… 81
图4.23 数据从消息队列导入图谱 …… 81
图4.24 利用工作流引擎定时更新图谱 …… 82
图4.25 知识容灾备份 …… 84
图5.1 智慧金融全景图 …… 87
图5.2 智慧金融知识图谱构建 …… 88
图5.3 知识图谱在智慧金融中的应用 …… 89
图5.4 基于知识图谱的个人信用反欺诈应用 …… 90
图5.5 基于知识图谱的产业链分析 …… 91
图5.6 营销流程图 …… 92
图5.7 风控流程图 …… 93
图5.8 基于知识图谱的医疗应用系统 …… 94
图5.9 医疗知识图谱支撑智慧医疗应用 …… 95
图5.10 炼油化工产业链示意图 ……100

图5.11 炼化过程影响因子分析（局部） ……100
图5.12 钢铁产业的流程图 ……102
图5.13 设备制造商全数据链整合示例 ……103
图5.14 知识图谱在智慧教育中的应用示意图 ……104
图5.15 基于知识图谱的学习成效分析示意图 ……107
图5.16 智能校园管理系统架构图 ……108
图5.17 基于知识图谱的智慧政务全景图 ……109
图5.18 基于知识图谱的政策公文智能应用 ……110
图5.19 医保审计知识图谱 ……111
图5.20 政务服务审批系统运行流程 ……112
图5.21 知识图谱在智慧司法中的应用 ……113
图5.22 司法知识图谱的构建 ……114
图5.23 文书生成流程 ……116
图5.24 智慧调节流程图 ……116
图5.25 智慧交通中知识图谱应用框架 ……117
图5.26“智慧交通系统”应用分类……118
图5.27 利用知识图谱构建人脸天网的流程图 ……118
图5.28 人、车识别与标签实例 ……119
图5.29 路径规划实例 ……119
图5.30 城市交通知识图谱 ……120
图5.31 知识图谱在智能电网中的应用分类 ……122
图5.32 智能客服问题系统的知识检索架构 ……123
图5.33 知识图谱智能搜索视图 ……124

图5.34 设备故障检测知识图谱示意图 ……125
图5.35 智能公安全景图 ……126
图5.36 基于公安知识图谱的应用分类图 ……127
图5.37 基于知识图谱产品进行犯罪分析的案例图 ……128
图5.38 基于知识图谱的农业信息检索 ……131
图5.39 农业病虫害知识问答系统架构图 ……132
图5.40 苹果产业实体关联结构图 ……133
图5.41 知识图谱在餐饮娱乐中的应用示例 ……135
图5.42 基于知识图谱的智能检索结果示例 ……136
图5.43 电商知识图谱示意图 ……137
图5.44 阿里巴巴电商认知图谱概览 ……138
表5.1 认知图谱关系举例（部分） ……139
图5.45 手机淘宝搜索框下搜索词提示 ……139
图5.46 手机淘宝首页猜你喜欢主题卡片推荐 ……140
图5.47 国家间关系分析的一般流程 ……141
图5.48 Palantir系统锁定嫌疑人分析功能界面 ……142
图5.49 智慧城建结构概念图 ……144
图5.50 人工承载力知识图谱示意图 ……145
图5.51 城市市政公用设施知识图谱 ……146
图6.1 美国伦斯勒理工学院知识图谱测评系统架构图 ……160
图7.1 知识表示关键技术标准 ……168
图7.2 知识图谱标准体系结构图……171
图7.3 知识图谱标准体系框架 ……173

第一章 背景介绍

一、知识图谱的起源与发展

（一）知识图谱的发展历史

知识图谱始于20世纪50年代，至今大致分为三个发展阶段：第一阶段（1955年—1977年）是知识图谱的起源阶段，在这一阶段中引文网络分析开始成为一种研究当代科学发展脉络的常用方法；第二阶段（1977年-2012年）是知识图谱的发展阶段，语义网得到快速发展，“知识本体”的研究开始成为计算机科学的一个重要领域，知识图谱吸收了语义网、本体在知识组织和表达方面的理念，使得知识更易于在计算机之间和计算机与人之间交换、流通和加工；第三阶段（2012年—至今）是知识图谱繁荣阶段，2012年谷歌提出Google Knowledge Graph，知识图谱正式得名，谷歌通过知识图谱技术改善了搜索引擎性能。在人工智能的蓬勃发展下，知识图谱涉及到的知识抽取、表示、融合、推理、问答等关键问题得到一定程度的解决和突破，知识图谱成为知识服务领域的一个新热点，受到国内外学者和工业界广泛关注。知识图谱具体的发展历程如图 1.1 所示。

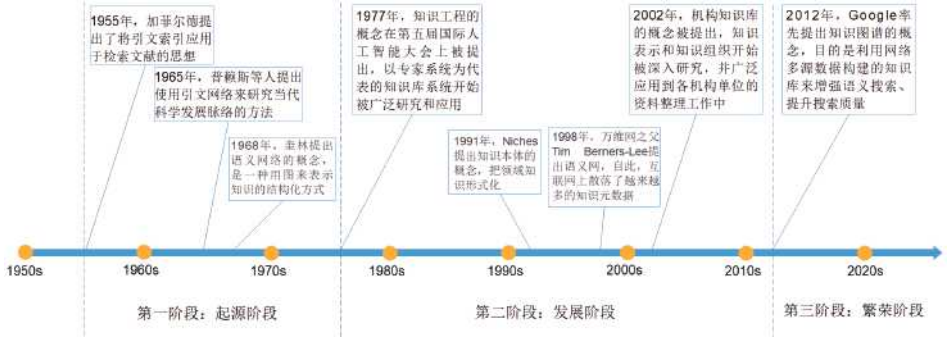

图1.1 知识图谱发展历史

起源阶段（1955年—1977年）：1955年，加菲尔德提出了将引文索引应用于检索文献的思想。1965年，普赖斯在《Networks of Scientific Papers》一文中指出，引证网络--科学文献之间的引证关系，类似于当代科学发展的“地形图”，从此分析引文网络开始成为一种研究当代科学发展脉络的常用方法，进而形成了知识图谱的概念。奎林(J. R. Quillian)于1968年提出语义网络，最初作为人类联想记忆的一个明显公理模型提出，随后在AI中用于自然语言理解，表示命题信息，语义网络是一种以网络格式表达人类知识构造的形式，是人工智能程序运用的表示方式之一。

发展阶段（1977年-2012年）：1977年，在第五届国际人工智能会议上，美国计算机科学家B.A. Feigenbaum首次提出知识工程的概念，知识工程是通过存储现存的知识来实现对用户的提问进行求解的系统，其中最典型和成功的知识工程的应用是基于规则的专家系统，此后，以专家系统为代表的知识库系统开始被广泛研究和应用。1991年，美国计算机专家尼彻斯（R. Niches）等人在完成美国国防部高级研究计划局（Defense Advanced Research Projects Agency, 简称DARPA）关于知识共享的科研项目中，提出了一种构建智能系统的新思想，该智能系统由两个部分组成，一个部分是“知识本体”（ontologies），另一部分是“问题求解方法”（Problem Solving Methods，简称PSMs），知识本体是知识库的核心，涉及特定领域共有的知识结构，是静态的知识；后者（PSMs）涉及在相应领域的推理知识，是动态的知识，PSMs使用知识本体中的静态知识进行动态推理。自1998年万维网之父Tim Berners-Lee提出语义网，同时随着链接开放数据(Linked Open Data)的规模激增，互联网上散落了越来越多的知识元数据。2002年，机构知识库的概念被提出，知识表示和知识组织开始被深入研究，并广泛应用到各机构单位的资料整理工作中。

繁荣阶段（2012年—至今）：21世纪，随着互联网的蓬勃发展，信息量呈爆炸式增长以及搜索引擎的出现，人们开始渴望更加快速、准确地

获取所需的信息。知识图谱强调语义检索能力,关键技术包括从互联网的网页中抽取实体、属性及关系，旨在解决自动问答、个性化推荐和智能信息检索等方面的问题。目前，知识图谱技术正逐渐改变现有的信息检索方式，如谷歌、百度等主流搜索引擎都在采用知识图谱技术提供信息检索，一方面通过推理实现概念检索（相对于现有的字符串模糊匹配方式而言）；另一方面以图形化方式向用户展示经过分类整理的结构化知识，从而使人们从人工过滤网页寻找答案的模式中解脱出来。

（二）知识图谱的重要性

哲学家柏拉图把知识定义为“Justified True Belief”，即知识需要满足三个核心要素：合理性（Justified）、真实性（True）、被相信（Believed）。简单而言，知识是人类通过观察、学习和思考有关客观世界的各种现象而获得和总结出的所有事实（Facts）、概念（Concepts）、规则或原则（Rules & Principles）的集合。人类发明了各种手段来描述、表示和传承知识，如自然语言、绘画、音乐、数学语言、物理模型、化学公式等，可见对于客观世界规律的知识化描述对于人类社会发展的重要性。具有获取、表示和处理知识的能力是人类心智区别于其它物种心智的重要特征，知识图谱已成为推动机器基于人类知识获取认知能力的重要途径，并将逐渐成为未来智能社会的重要生产资料。

1、知识图谱是人工智能的重要基石

人工智能分为两个层次：感知层与认知层。首先感知层，即计算机的视觉、听觉、触觉等感知能力，目前人类在语音识别、图像识别等感知领域已取得重要突破，机器在感知智能方面已越来越接近于人类；第二个层次是认知层，是指机器能够理解世界和具有思考的能力。认知世界是通过大量的知识积累实现的，要使机器具有认知能力，就需要建立一个丰富完善的知识库，因此从这个角度说，知识图谱是人工智能的一个重要分支，

也是机器具有认知能力的基石，在人工智能领域具有非常重要的地位。

2、知识图谱推动智能应用

知识图谱将人与知识智能地连接起来，能够对各类应用进行智能化升级，为用户带来更智能的应用体验。知识图谱是一个宏大的数据模型，可以构建庞大的“知识”网络，包含客观世界存在的大量实体、属性以及关系，为人们提供一种快速便捷进行知识检索与推理的方式。近些年蓬勃发展的人工智能本质上是一次知识革命，其核心在于通过数据观察与感知世界，实现分类预测、自动化等智能化服务。知识图谱作为人类知识描述的重要载体，推动着信息检索、智能问答等众多智能应用。

3、知识图谱是强人工智能发展的核心驱动力之一

尽管人工智能依靠机器学习和深度学习取得了快速进展，但严重依赖于人类的监督以及大量的标注数据，属于弱人工智能智能范畴，离强人工智能仍然具有较大差距，而强人工智能的实现需要机器掌握大量的常识性知识，同时以人的思维模式和知识结构来进行语言理解、视觉场景解析和决策分析。如图1.2所示，知识图谱技术将信息中的知识或者数据加以关联，实现人类知识的描述及推理计算，并最终实现像人类一样对事物进行理解与解释。知识图谱技术是由弱人工智能发展到强人工智能过程中的必然趋势，对于实现强人工智能有着重要的意义。

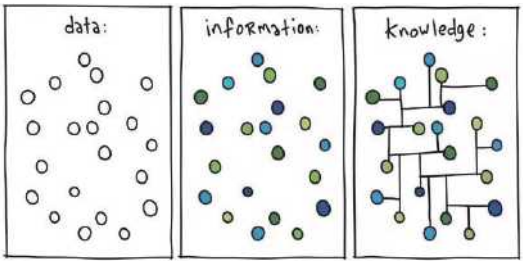

图1.2 人工智能发展过程中数据关联的必要性[1]

[1] Gapingvoid Culture Design Group https://www.gapingvoid.com/

（三）知识图谱与Ontology、语义网络之间的区别

知识图谱与Ontology、语义网络等概念之间具有密切的相互联系。语义网络（Semantic Networks）是由Quillian于上世纪60年代提出的知识表达模式，主要用于自然语言理解领域，其用相互连接的节点和边来表示知识，节点表示对象、概念，边表示节点之间的关系。语义网络具有容易理解和展示、相关概念容易聚类的优点，同时也有以下几个方面的缺点：一是节点和边的值没有标准，完全由用户自己定义；二是多元数据融合比较困难，没有标准；三是无法区分概念节点和对象节点；四是无法对节点和边的标签进行定义。语义网络虽然可以让我们比较容易理解语义间的关系，但由于缺少标准，比较难以应用于实践。

1980年，本体论（Ontology）哲学概念“本体”被引入到人工智能领域用来刻画知识。本体是共享概念模型的明确的形式化规范说明，该定义体现了本体的四层含义：概念模型、明确、形式化、共享。本体是实体存在形式的描述，往往表述为一组概念定义和概念之间的层级关系，本体框架形成树状结构，通常被用来为知识图谱定义Schema。

二、知识图谱的定义

知识图谱（Knowledge Graph）以结构化的形式描述客观世界中概念、实体及其关系，将互联网的信息表达成更接近人类认知世界的形式，提供了一种更好地组织、管理和理解互联网海量信息的能力[1]。知识图谱给互联网语义搜索带来了活力，同时也在智能问答中显示出强大威力，已经成为互联网知识驱动的智能应用的基础设施。知识图谱与大数据和深度学习一起，成为推动互联网和人工智能发展的核心驱动力之一[2]。

[1] 中国中文信息学会语言与知识计算专委会，《知识图谱发展报告(2018)》.
[2] 中国中文信息学会语言与知识计算专委会，《知识图谱发展报告(2018)》.

知识图谱不是一种新的知识表示方法，而是知识表示在工业界的大规模知识应用，它将互联网上可以识别的客观对象进行关联，以形成客观世界实体和实体关系的知识库，其本质上是一种语义网络，其中的节点代表实体(entity)或者概念(concept)，边代表实体/概念之间的各种语义关系。知识图谱的架构，包括知识图谱自身的逻辑结构以及构建知识图谱所采用的技术（体系）架构。知识图谱的逻辑结构可分为模式层与数据层，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，通常采用本体库来管理知识图谱的模式层，借助本体库对公理、规则和约束条件的支持能力来规范实体、关系以及实体的类型和属性等对象之间的联系。数据层主要是由一系列的事实组成，而知识将以事实为单位进行存储。在知识图谱的数据层，知识以事实（fact）为单位存储在图数据库。如果以“实体 – 关系 – 实体”或者“实体 – 属性 – 性值”三元组作为事实的基本表达方式，则存储在图数据库中的所有数据将构成庞大的实体关系网络，形成“知识图谱”。

三、白皮书范围

根据当前知识图谱技术发展情况及在多个领域的成功实践，本白皮书从哲学层面、政策层面、产业层面、行业层面、技术层面、工具层面、支撑技术等多个层面对知识图谱的实际需求、关键技术、面临的问题与挑战、标准化需求、展望与建议等进行了梳理，以期对未来知识图谱在更多行业的推广应用及标准研制提供支撑，白皮书整体结构如图1.3所示。

图1.3《知识图谱标准化白皮书》整体结构

四、白皮书概述

本白皮书通过梳理知识图谱技术、应用和产业演进情况，分析知识图谱的技术热点、行业动态和未来趋势，从支撑知识图谱产业整体发展的角度出发，研究制定了能够适应和引导知识图谱产业发展的标准体系，进而提出近期急需研制的基础和关键标准项目。

本白皮书力求以较为浅显易懂的语言和方式进行阐述，针对目前知识图谱涵盖的技术热点、应用领域及产业情况进行分析，研究提出知识图谱标准体系。知识图谱标准化工作尚处于起步阶段，本白皮书只作为知识图谱领域技术、产业和标准化之间初始的连接纽带，并将不断根据技术、产

业和标准化的发展需求进行修订。

本白皮书的意义在于与业界分享知识图谱领域的研究成果和实践经验，呼吁社会各界共同加强知识图谱领域的技术研究、产业投入、标准建设与服务应用， 共同推动人工智能及其产业发展。此外，知识图谱作为人工智能产业中较为前沿的技术，受到越来越多的关注，这也说明了知识图谱技术在一定层面上可以解决一些当前技术无法解决的问题。对于技术厂商及第三方服务机构，有待着力推动知识图谱技术研究与应用市场健康发展，避免其成为技术噱头；对于需求方，有待不断加深对知识图谱在构建和持续维护时存在困难与挑战的理解。各方在构建知识图谱相关应用时能够各担其责，客观面对存在的技术难点与问题，构建出真正对实际业务场景有意义的知识图谱及相关产品。

由于时间仓促，难免有疏漏，甚至错误的地方，仅供有志于知识图谱研究和开发及标准研制的同仁参考，激发更广泛的思考和讨论，期待共同努力推动知识图谱及其标准化的发展。

第二章 知识图谱需求

一、公共安全

公共安全领域信息化以现代通信、网络、数据库技术为基础，将所研究对象各要素汇总至数据库，并针对各个业务领域进行定制化开发，以满足公共安全实战需求。近年来我国公安信息化建设不断加强，发展至今，形成了涵盖1300余项标准的公安信息化标准体系，组织建设了9个国家工程实验室和27个公安部重点实验室，依托一站式工作平台“警综平台”，有效降低了基层民警工作量。虽然我国公安信息化发展较快，但受传统观念、资金投入、工作机制等影响，公安信息化建设仍有较大发展空间，而且由于公共安全相关信息中实体关系重要性较高，对知识图谱的需求凸显。

1、违法犯罪活动呈现复杂化、动态化、智能化的特征

公安机关为提升整体效能，在打击违法犯罪活动中占据信息主导权，能够充分利用现代信息技术和信息资源，对现有的警务组织、机制、模式、装备等警务实践进行信息化改造。当前社会组成要素的人流、物流、信息流激增，在社会生产力日益提升的背景下逐步使犯罪活动呈现出复杂化、动态化、智能化的特征，针对这些特征，公安机关能够基于已有的结构化及非结构化数据，通过知识图谱技术从少量已知信息中获取和挖掘出图谱的潜在网络拓扑信息及行为信息，并构建一系列基于知识图谱的应用工具提升公安面对犯罪类型多变等挑战的能力，有效应对新型犯罪活动及多类型、多渠道的犯罪模式。

2、公安信息庞大而且实体关系复杂

公安数据信息规模庞大且实体关系复杂。以人为例，身份证号、籍

贯、姓名、指纹、出行记录、住宿记录等都属于公安范畴的数据信息，以及嫌疑人在案件中的各种信息、涉及物品信息、涉及案发地、案发时间等信息也属于公安数据，这些数量巨大的公安信息能够通过知识图谱海量的概念实体、丰富的属性信息、多样化的关联关系进行直观准确的描述，同时能够从时间序列维度、类别维度对图谱知识进行统计分析，能够根据公安业务需求实现定制化的智能应用，为公共安全部门提供准确、高效的技术工具和辅助决策支撑。

3、海量数据快速检索与理解需求难以满足

随着公共安全数据规模、案件关系复杂程度的增加，传统数据检索功能已经无法满足当前的公安业务需求，同时海量公安数据难以实现语义化理解和知识化转化。知识图谱技术为公安大数据环境下的人员分析方法提供了一种有效技术手段，知识图谱通过从海量公安数据中获取与融合丰富的语义知识，形成庞大、相互关联的公安知识网络，实现潜在隐藏信息的自动挖掘与推理，满足海量公安数据的语义理解与快速准确检索。

二、数字经济

数字经济是指以使用数字化的知识和信息作为关键生产要素、以现代信息网络作为重要载体、以信息通信技术的有效使用作为重要推动力的一系列经济活动。数字经济是继农业经济、工业经济之后的一种新的经济社会发展形态，更容易实现规模经济和范围经济，日益成为全球经济发展的新动能。数字经济占我国经济发展规模的比重接近三分之一，对经济增长具有显著拉动效应。数字经济不仅是我国经济增长的主要动力源泉和转型升级的重要驱动力，同时也是全球新一轮产业竞争的制高点。我国数字经济发展情况如图2.1所示。

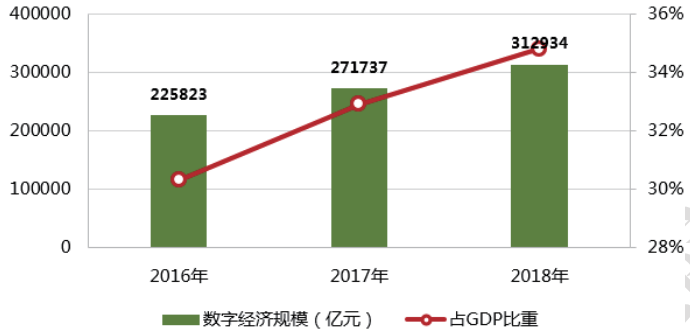

图2.1 我国数字经济发展情况

数字经济可被认作为一个经济系统，在这个系统中，数字技术被用户广泛使用，使得整个经济环境和经济活动都发生了根本变化。同时数字经济也是一个全新的社会政治与经济系统，其中所有的信息和商务活动是以数字化的方式进行，企业、消费者和政府之间进行的网络交易量迅速增长。当前信息流蓬勃增长的浪潮中，我国发展数字经济具有良好机遇的同时，也面临着诸多挑战，知识图谱可以为数字经济背景下所面临的一些具体问题提供有效的解决方案。

1、数字经济的发展离不开大规模有效数据的关联

在数字经济时代，数据是最重要的生产要素，大数据时代的数据资源虽然规模庞大，但在数据的关联使用上仍有不足。在知识驱动的数字经济中，需掌握数据的“萃取”技术，方能“提炼”知识。未来的数据资源会越来越开放，知识谱图绘制和深度学习的能力极有可能超越数据本身，并成为核心竞争力。大数据之“大”并非只强调上数据量之大，也指数据的汇聚、关联和使用之宽广。数据本身需要通过理解，分析才能够有效利用，从而服务于人类，而构建知识图谱即是把数据提炼为有效知识的重要途径之一。

2、将“数据中心”向“知识中心”过渡是数字经济升级发展的重要需求

在互联网背景下，知识图谱技术推动数字经济向数据知识化方向升级发展。互联网与数字经济一同在不断地蓬勃发展，深刻且全面地改变了世界和人类本身。从“流量变现”到“数据变现”使互联网发展的转变为以技术和数据为核心，而数字经济知识化的发展需求将推动知识图谱作为数据知识化的重要工具深入发展，数字经济中的数据知识化汇集领域知识库、专家系统、机器学习等核心技术动能，将大大提升生产效率和自动化水平，因此将“数据中心”向“知识中心”过渡是数字经济升级发展的重要需求。

3、知识图谱成为推动数字经济创新场景下智能应用的重要力量

随着数字经济与大数据的不断发展，传统经济模式和应用场景持续变革，以新零售场景为例，由于顾客需求的变迁，新型零售方式不断出现，线上线下渠道融合趋势日益明显，零售渠道从单一结构演变为多渠道与全渠道结构，在这样的应用背景下，数据的搜集与挖掘便是一个很大的难题，需要把线下和线上的数据汇总起来，形成以商品为中心的知识图谱，把不同类型的数据，按照既定模型组合在一起，最终形成商品与商品之间、商品与门店之间、商品与消费者之间的关系网络，最终实现新零售场景下的智能化推荐和资源优化配置，因此知识图谱技术将成为推动数字经济创新场景下智能应用的重要力量。

三、金融科技

全球范围内伴随着云计算、大数据、人工智能等新兴技术的落地，互联网的数据量每两年翻一番，基于数据的生产变革和业务模式创新正驱动着全球范围内经济社会各个领域的数字化与智能化转型，实现了数字经济在数字化程度高的行业中赋能，加快了其数字化转型进程。金融行业首当

其冲，成为人工智能最先应用的行业之一。金融生态的数字化、智能化转型过程中，业务应用场景化、数据垄断化特征成为金融科技不断优化金融生态环境的必要条件。如图2.2所示，金融科技作为更为开放的信息技术知识领域渗透金融行业的技术载体，深刻地影响着金融的业务场景、机构与金融监管构成的金融生态系统。

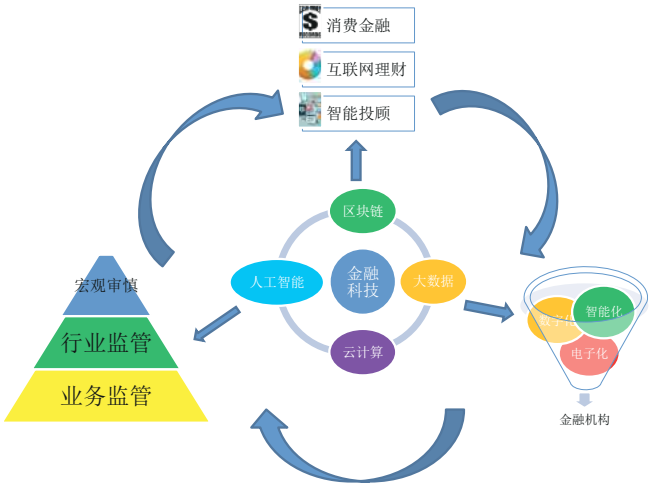

图2.2 金融生态的金融科技要素分布

知识图谱在金融科技领域是语义理解和知识搜索基础，可以为金融行业的风险评估、预测、反欺诈、精准营销、智能搜索等提供技术支撑。越来越多的金融机构及企业在探索构建金融领域知识图谱，为金融领域应用决策提供更精准可靠的依据。

1、知识图谱能促进金融科技信息技术与开放知识领域的融合

国家对人工智能发展的重点规划（国办发2017〔35〕号）中，智能金融作为高端智能经济发展的重点领域。智能金融存在的数据碎片化、业务数据质量问题约束了智能推送功能，大数据时代的数据多源异构特征加剧了智能金融的建设过程中的数据质量问题。金融业务知识图谱的建设为解

决上述问题创造了条件，如金融知识图谱在政府数据、商业数据（社交、电商）等海量、多元维度数据的基础上，建立金融业务的元数据，并以图谱节点与关联关系，实现了大数据与人工智能知识图谱的融合，为金融机构的数据治理提供了设计思路。

2、知识图谱为金融科技服务于实体经济提供了深度技术

基于知识图谱的应用能够提高金融行业、金融机构与金融业务的微观、中观与宏观全产业链数据治理能力，以数据为核心构建的金融知识图谱，梳理不同业务模块的业务知识、融合海量外部领域数据，形成以金融科技为中心的全维知识网络，加强各知识领域多维度数据的关联性。金融领域知识图谱在一定程度上能够通过信息精准识别，解决长尾信贷中小企业客户的信贷供给不足问题，优化与引导金融资源配置服务于中小企业实体，增大为实体经济中的活跃主体“输血”规模，为实体经济作贡献。2019年由国家信息中心牵头，工商银行在重庆两江新区践行的“信易贷”将大数据与人工智能技术融合，通过中小企业信用的业务知识图谱底层建设，通过一站式授信平台实现了中小微企业融资服务，实现了传统金融产品与服务的边界拓展。

3、知识图谱践行金融科技的风险管理创新机制

金融生态的业务、技术、监管、资源配置在金融科技的支撑下，金融风险的复杂性与多样化使风险监管的知识图谱建设迫在眉睫。如宏观审慎管理中系统性风险的实体对象，不在局限于“大而不倒”的系统性金融机构，互联网金融所拓展的金融产品与业务边界，因网络效应而更具传染性风险的“网而不倒”产品与服务，也应当纳入风险管理的实体对象之中；另外，基于网络的相关利益者关联关系，更大范围的金融产品与服务的消费者、金融科技平台以及各类非正规金融机构，拓展了风险管理的主体对象，2018年证监会关于企业知识图谱的建设，广泛应用于中央监管信息平台的上市公司画像、拟上市公司画像和中央监管私募机构画像，则是通过

关联关系（KYC：充分了解你的客户、AML：反洗钱）识别面向企业层面的风险防范，如非法集资、财务管理风险等，使金融风险管理适应于金融科技产品、服务、行业的发展，成为了可能。

四、资源优化

面对新的经济形势和市场竞争格局，许多大型集团都提出了“主动拥抱产业变革调整带来的新模式、新机遇，积极应对市场竞争呈现的新变化、新挑战”的数字化需求，其核心在于利用“大数据、云计算和人工智能”等相关新型数字化技术，从市场需求出发，优化整个资源配置，快速适应市场变化，促进降本增效，提升服务水平。此外，当前农业资源、电力资源等重要社会资源持续紧张。以农业为例，我国农业人口比例大，耕地面积不足，而且近年来随着农业用地面积的逐渐减少，耕地面积更加紧张，据数据显示，1996年中国耕地面积约1.3亿hm2，而2016年统计为1.2亿hm2，下降了0.1亿hm2。而且，耕地面积分布不均衡，而且随着生态环境逐渐恶化，土壤退化严重，耕地难以得到休整，化肥和农药的施用使水污染加剧，我国农业现代化发展面临着资源和环境的双重压力。同时，我国农业劳动力文化水平相对较低，如何在多重限制条件下提升农业资源的利用效率一直是推进我国农业现代化建设的一大难点。

1、企业资源优化

大型集团企业倒三角支撑，“将问题甩单，让一线更简单”，用移动互联网手段推进全公司倒三角支撑。重点通过一线问题甩单，打造跨流程、跨部门、跨层级的扁平化透明化快速响应协调机制，实现全业务、全流程、全渠道覆盖，高效解决一线日常生产问题。一线人员需要的业务知识遍布在许多垂直类的系统中，这类系统之间彼此相对独立，形成了一个个“烟囱”，知识信息很难流通，形成 “企业知识孤岛”。利用知识图谱技术，透过纷繁复杂的海量文本和结构化数据，可以实现有效的语义化

数据治理和知识获取。同时，知识图谱可以作为企业中各个业务系统的数据源，为业务系统提供知识和可计算服务。

2、社会资源优化

农业资源、电力资源作为紧缺且重要的社会资源，具有重要的代表性，如何快速、准确获取知识，并能够提供应用指导对于农业从业者、电力企业员工等资源领域的从业者有着重要的意义。在农业资源优化方面，以往受限于人员分布广泛、高质量知识获取难度大、数据积累多以非结构化为主等因素，从业者通常以经验作为重要的操作考虑因素，存在部分情况下误操作、漏操作等风险。如由于农业从业者对病虫害缺乏了解，往往凭借经验诊断已发作的农作物病虫害进行盲目用药，造成防治用药难以有的放矢。通过建设病虫害知识图谱，搭建知识获取及应用、推广平台可以提供更加科学的用药和栽培辅导，进而提升整体农业土地资源的利用效率。电力设备运维、电网安全措施推荐、客户服务等方面也存在类似需求。

五、科学研究

随着数字时代向数据时代演进，科学研究进入“第四范式”，无论是研究个体还是科学共同体都在渴求能够获得更深入全面、更高效便捷的数据服务。科学研究需要主动推进知识服务升级，帮助用户扩展资源发现，揭示资源中蕴含的知识，构筑知识间的关联[1]。如今学术搜索引擎使得科研人员能够高效便捷地查阅文献和相关资源，它已经成为当前科研工作者不可或缺的一项技能。科研社交网络有效地提升了科技成果的传播速度，降低了科研人员之间交流的时间成本，扩大了研究人员之间的合作机会。尽管搜索引擎和社交网络等平台给科学研究带来了极大的方便，但是由于科学研究数据量的急速增长，相关平台并没有很好的组织和挖掘这些信

[1] 张颖.美国研究型图书馆研究数据服务的实践进展及趋势[J].图书情报工作,2017,(9):33–41

息，这也给科研人员带来了很大的不便。如何有效的组织和管理这些科学研究相关的信息是科研工作者迫切的需求。

1、各学科资源整合困难，学科知识服务有待提高

传统基于文本形式综述分析的学科研究方法存在难以揭示学科间相似性、学科发展继承性等问题,致使科研人员难以客观、高效理解学科静态结构、动态演化及发展趋势。在这种形势下,伴随文本挖掘和知识图谱技术发展,学科领域知识图谱的构建为该问题解决提供了可能性,其用可视图的方式高效立体揭示学科发展特征[1]。学科知识图谱基于引文分析和可视化,将知识作为研究对象,直观地显示出知识的发展进程及其之间结构关系的图谱，并作为知识可视化的重要载体,描述了学科知识的内部结构及知识节点间的联系。

2、学术网络信息分散，专业群体建设不足

随着科学研究的飞速发展，越来越多的学者在各个学术领域间进行交流合作。学术合作可以交换学者之间的学术知识、提高工作效率，已经成为学术领域的发展趋势。学者们之间通过合作进行课题研究，共同发表文献，从而形成了复杂的科研合作关系网络，这种科研合作关系网络，包含了学者和学者之间、学者和科学研究领域之间、各个科学研究领域之间等多种类型的复杂关系。如何在这种复杂的学术关系网络中展现知识演化、发现隐含的关联知识是目前众多学者关注的问题。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,构成一张巨大的语义网络图，因此，利用知识图谱可以较好地描述学术关系网络中的各种复杂关系，并利用知识图谱强大的语义处理能力去实现学者的影响力分析、领域研究热点之间的关联关系发现、未来研究热点的趋势预测等多类应用。

[1] 廖盼. 基于多方法融合的学科知识图谱构建研究[D].湖北工业大学,2018.

3、识别前沿热点问题，预测技术发展趋势

如何快速识别前沿问题并把握技术发展的方向，是当前很多科研机构和学者迫切的需求。通过知识图谱技术，在上亿篇文献信息中对技术来源、热度、发展趋势进行研究进行深入挖掘，进而实现技术发展趋势的准确预测。同时分析技术的出现、变迁和消亡的全过程，可以帮助研究人员理解领域的研究历史和现状，快速识别研究的前沿热点问题，把握技术的发展趋势。使用各领域科学知识图谱，通过对各领域研究的主要学者、重点文章、最新成果等信息加以分析，发现了研究的核心论文及知识谱系，建立专业方向和技术发展的知识图谱，并在最新的研究现状梳理的基础上对各个领域的前沿热点问题和技术研究的发展趋势进行了预测。

六、制造业转型

近几十年来，随着产业的不断发展和技术水平的进步，在人工红利、政策引导和市场需求的大力推动下，我国的制造业充分发挥成本、规模以及制度上的优势迅速发展，在体量和覆盖面上均位居世界前列。然而，相比较发达国家而言，我国制造业竞争力还有较大差距，如缺乏发达的产业分工体系、未形成强大的产业链、自主创新能力弱、产业结构不合理等尤其是高端装备制造业和生产性服务业发展滞后，部分企业还是建立在高能耗、高污染、高投入、低效益的基础之上，发展理念亟待转变。随着资源、环境和成本压力的加大，一方面，部分中、低端制造业已出现外流的趋势，另一方面在ICT技术的快速发展的驱动下，制造技术和信息技术融合带来新的工业革命，全球主要的制造大国均提出了制造业转型升级的战略方向，我国制造业从资源密集型和劳动密集型向技术密集型、数字密集型转型升级的任务已非常紧迫。

1、数据粗放采集管理，智能化应用的水平无法满足工业需求

现阶段大部分企业普遍信息化、自动化基础都较为薄弱，受限于成本

压力、技术风险及人才缺口等因素，目前对于数据积累和知识理解都处于初级阶端。而现阶段提供智能制造转型服务的多为软件或硬件服务商，对制造业行业知识和数据信息理解有限，而知识图谱技术可能为精益的数字化智能化升级提供解决方案。

2、人力成本增加、人员流动率大

随着竞争加剧，知识密集型行业的人力成本逐渐提高。同时，人才的培养需要管理、技术等多方面人才的密切配合，人员流动率的增大也会导致企业流失掉员工的经验，培训成本也随之增大。因此，企业也希望用知识图谱技术去沉淀经验，培训和赋能员工。

3、技术瓶颈、人才断层，导致创新能力不足

传统的制造业领域经历了几十年的发展，基础理论、材料、工艺等已经到了发展的瓶颈期，另一方面在金融、IT等行业的冲击下，制造业人才流失和断层十分严重，再者受制于技术壁垒、市场竞争、成本压力，这些都是导致制造业整体创新能力不足的因素。因此，加强知识图谱建设，增强技术流通，将有助于减缓人力和研发成本压力，提升行业创新能力，从而赋能制造业转型升级。

七、公众健康

人民健康是民族昌盛和国家富强的重要标志。习近平总书记作出重要指示，实施健康中国战略，提高保障和改善民生水平，加强和创新社会治理。人口老龄化、慢性病高发、医疗资源分配不均是公众健康领域中国乃至世界面临的主要问题。迫切需要通过人工智能技术高效的打造出可预测、可预防、精准医疗的健康解决方案，对公众健康领域的诸多痛点提供革命性的辅助与支持。

1、医疗资源不足、分布不均衡以及效率低下

目前，国内有280万注册执业医生，每天有2000万门诊量，产业供需

严重不平衡。特别是进入老龄化社会之后，对医生的需求量有增无减。《2016年中国卫生和计划生育事业发展统计公报》显示，医院卫生技术人员数同比增长5.57%，低于诊疗人次6.17%的增速，医疗资源供给将会越来越紧张。面对医疗健康产业供需严重不平衡的大背景，以医疗知识图谱为支撑的知识系统，可以一定程度上将医生从繁重的重复性知识工作中解放出来，专注于治病救人和疑难病症突破，进而提升医疗服务的效率、扩大医疗服务的供给、提供基于知识服务的患者院内院外高效服务，所以医疗健康知识图谱技术可以在医学辅助诊疗、知识推荐、医院管理、医疗控费等环节体现出强大的降本增效能力。

2、分级诊疗政策难以有效落实

在分级诊疗问题上，可以通过优化医疗资源的配置，让不同层级的医院形成有效的分工协作，从而实现缓解目前医疗服务供给不足的矛盾。然而基层的医疗服务水平的薄弱，导致患者对于基层医生的广泛不信任，造成分级诊疗难以落实。医疗知识图谱可以将顶尖医学专家的知识和诊治经验进行快速复制，以医疗知识图谱为基础搭建的各类应用系统可以为基层医生提供实效、实时的决策支持，提高广大经验不足的医疗工作者的工作能力。借助医疗知识图谱海量、结构化的医学知识体，有望实现计算机辅助下的医疗辅助诊疗赋能，提高基层医院医生的诊疗能力，实现智能化的分诊导诊，配合并推动实现国家的分级诊疗制度和愿景规划，提升患者就医效率。

3、医疗行业诊疗规范分散

2006年，国家卫生部、国家中医药管理局、总后卫生部三家联合委托中华医学会，由其各医学分会制定了相关的学科《临床诊疗指南》。迄今为止，《临床诊疗指南》已经出版发行47分册，涉及临床各主要学科。但在《临床诊疗指南》之外，由中华医学各专业分会制定了大量的指南、专家意见、专家共识、指导原则等，这些资源对于临床实践具有重要意义。

但是由于资源分散，缺少有效的管理，难以发挥效力，我国平均误诊率仍然高达30%左右。而基于医疗知识图谱构建的各类应用系统可以有效的加速诊疗规范的形成与验证，并且通过及时的诊疗知识推送和推广，更好的普及诊疗规范，同时可以实时发现诊疗违规，全面规范诊疗过程，降低误诊率和漏诊率，提升患者就医体验、降低就医成本。

4、医疗资源浪费现象严重，医疗服务效率有待提高

以大型医院急诊科为例，急诊科总是人满为患，但是实际上，非急诊患者至少占1/3，需要立刻抢救的大概只占急诊总量的5%～10%。在美国，每年在医疗上的开支接近3万亿美元，占GDP比例超过18%，其中7500亿美元是可避免的开支。经常会出现病人在急诊室门口等待却无人医治，或者医护人员和病床资源空闲时，却无人可医的情况。通过构建医疗知识图谱，以知识图谱为支撑的智能问诊系统，可以实现院内的智能分诊问诊，减少不必要的事务性资源占用，优化医疗服务流程，进而实现优化国家医疗支出，使患者享受优质的医疗服务的目的。

5、医疗保险监督管理问题

当前，我国医保基金一方面面临收支平衡的严峻压力，另一方面，各地违规使用医保基金，诈骗套取医保基金的案例时有发生。通过临床海量病例结合医疗知识进行医疗保险监督知识图谱构建，借此辅助医保信息系统智能化，建立的一种更加科学的医疗保险监管体系，涵盖事前、事中、事后的诊疗全过程，管理对象包括参保人、医师、医院和药店，管理的处方内容包括药品、检验检查及医用材料，识别医保费用报销数据中的各种异常疑似情况，辅助医保监控，一方面减少了医保欺诈对医疗资源的浪费与冲击；另一方面也提高了医保监控的工作效率，降低国家医疗支出。

八、人文发展

十九大报告中指出中国社会主要矛盾已转化为人民日益增长的美好生

活需要和不平衡不充分的发展之间的矛盾。随着我国居民生活水平的稳步提高和市场供给端的长足进步，我国消费市场取得较快发展，大众餐饮、文化娱乐、休闲旅游、教育培训等服务消费已成为新的消费热点。2018年中国教育市场规模达到人民币2.68万亿元，至2020年民办教育的总体规模将达到3.36万亿元，复合增速12%；在文化娱乐消费方面，相关市场建设不断完善和市场供给能力日益增强等因素带动文化消费持续增长；2018年社会消费品零售总额中餐饮收入超过4万亿元，是1952年的3029倍，年均增长12.9%，比社会消费品零售总额年均增速高1.3个百分点；我国餐饮市场保持较快增长，餐饮收入占社会消费品零售总额比重由1952年的5.1%提升至2018年的11.2%。然而随着文化、教育娱乐等人文领域市场规模的快速增长，社会面临一系列重要挑战，对于知识图谱等技术需求日益明显，主要表现在如下几个方面：

1、知识图谱有助于提升传统文化数据知识化治理水平，拓展传播渠道

在国家大剧院，一场演出最多能容纳2000多名观众观看，同样的演出进入网络直播间，观众会立刻变成几十万甚至上百万。网络直播技术的出现，大大拓展了文化艺术的传播边界，在它的“加持”下，京剧、昆曲、民乐、书法、泥塑等传统文化艺术，可以从各自的小天地飞向社会的大舞台。不仅如此，网络直播中，受众的年轻化以及审美的多元化，也反过来推动了传统文化表达方式的创新转型。以诗词为代表的传统文化领域作为数字人文研究者关注的重要领域之一，存在大量的数据资源。由于唐诗在用词、句式、语法、对仗等方面要求较高，有着特殊的文法和习惯，传统的文本挖掘技术难以处理这样的古文文本，唐诗知识图谱的构建能够统一表示和组织唐诗领域数据并进一步实现智能化唐诗知识服务。

2、知识图谱有助于提升休闲娱乐消费体验和发展水平

近年来，休闲娱乐消费市场总体规模持续扩大，第三产业增加值占国

内生产总值的比重从1980年的20%左右增加到2018年的52.2%，居民对休闲娱乐方面的需求不断提升，而且需求不断朝个性化、差异化方向发展。深入挖掘客户需求、发现店铺特色及能力需求不断提升，知识图谱作为重要的客户画像及辅助决策工具正在得到注意，其中美团点评作为重要的在线本地生活服务平台之一，覆盖了餐饮娱乐领域的众多生活场景，连接了数亿用户和数千万商户，积累了宝贵的业务数据，蕴含着丰富的日常生活相关知识。通过这些海量的休闲娱乐业务数据，构建具有数十亿实体关系的用户休闲娱乐知识图谱（如美团大脑知识图谱），能够更好的为用户提供休闲娱乐服务推荐，提升休闲娱乐体验与消费质量。

3、知识图谱助力破解教育资源不平衡

十九大报告提出，努力让每个孩子都能享有公平而有质量的教育，这既是我们努力的目标也是广大人民群众的强烈愿望。建设教育强国是中华民族伟大复兴的基础工程，必须把教育事业放在优先发展位置，加快教育现代化，加强教育信息化，办好人民满意的教育。教育信息化一方面是把提升信息素养纳入培养目标；另一方面是把信息技术手段有效应用于教学与科研，注重教育信息资源的开发和利用。人工智能在第二方面显示出了极大的优势，可以远程教育教学，完全可以解决教育师资不均衡的问题。知识图谱通过实现课程资源、教研内容、教学管理方案之间的链接，进而在课前、课中、课后推动以学生为中心的智慧课堂、个性化学习资源推荐、学习进度评估等，提升教学质量，降低学生学习压力。

九、政府管理

政府的“数字化”建设阶段包括信息数字化、业务数字化、组织数字化三个阶段，由此从“电子政府”走向“数字政府”并最终发展成为“智慧政府”。我国数字政府第一个阶段是上世纪90年代至2006年，这一时期数字政府建设的主要任务是通过普及电脑和网络使政府办公与管理实现网

络链接和互通，以提升政府的行政业务效率。2006年至2012年国家开展的“两网一站四库十二金”建设是数字政府第二阶段建设的标志性工程。2012年至今，随着大数据、人工智能技术的发展，我国数字政府建设进入第三阶段。近年来，在线政务服务建设成为热点，并朝着多元化方向发展。2018年，我国“互联网+政务服务”深化发展，各级政府依托网上政务服务平台，推动线上线下集成融合，实时汇聚网上申报、排队预约、审批审查结果等信息，实现“数据多跑路，群众少跑腿”。

通常认为，数字政府是以大数据、人工智能等新一代信息技术为支撑，以数据治理为核心，通过数据挖掘、分析和运营等方式驱动政府管理、业务和组织架构的革新，是一种“用数据说话、用数据决策、用数据管理、用数据创新”[1]的新型现代化政府治理模式。数字政府既是“互联网+政务服务”深度发展的结果，也是大数据时代政府自身优化的必然。在数字政府环境下，政务服务数据的流通与共享更加高效，能够有效促进政务服务的业务重组与流程再造，提升政府治理能力。

在国家及各级地方政府的大力推动下，数字政府建设尽管取得了一定成效，但在某些方面依然存在问题，知识图谱技术在数字政府领域的应用需求表现在如下几个方面：

1、政府数据碎片化导致“信息孤岛”现象严重

我国进入大数据时代以来，通过统筹国家政务信息化工程建设，从政务服务入手实施了一系列惠民工程，但政务信息化建设中的条块分割和信息孤岛问题依旧突出。一方面，由于政府各部门在对自身利益的利益倾向不同而造成各部门在协同办公过程中未能及时互动和资源共享，从而没有形成整体性的政府治理能力，从而出现政府数据碎片以及信息孤岛的产生，制约部门间的交流与合作。另一方面，由于现行的科层体

[1]《促进大数据发展行动纲要》（国发〔2015〕50号）

制下的政府服务模式呈现典型的“金字塔”结构，权利和决策集中在顶端，随着信息的逐层递减，信息的精确性在某种程度上出现偏差以及明显分化。数字政府建设的目的在于打破部门间壁垒和信息垄断的格局，实现信息的平等共享。

2、政府数据共享开放缺乏统一标准

政府数据共享开放是数字政府建设的一项基础性工作，长期以来条块分割、各自为政造成的“信息孤岛”、“数据烟囱”在跨部门尤其是横向部门之间存在巨大的数据“壁垒”，而由于缺乏统一标准体系的支撑，各部门各系统采集的数据格式不统一、标准不一致，数据处理技术各异，数据库接口不互通。要突破标准不统一的“数据烟囱”，技术壁垒成为数据共享和开放首先要跨越的高墙。在共享与开放工作中，政府部门普遍对于共享和开放的范围、数据分级分类、数据脱敏以及数据审查等问题存在疑问，但由于缺乏操作性强的指导性文件、统一的标准规范和技术规范，各部门无据可依，不会共享、开放。

3、政府决策缺乏有效数据支撑

当前数字政府建设中，普遍存在政务数据“拥而难用、汇而不慧”[1]的现象，由于跨部门的政务数据之间未建立逻辑关联，多源异构政务数据缺乏深度融合而导致政务数据价值难以释放，政府决策尚缺乏有效数据支撑，距离“用数据说话、用书决策、用数据管理、用数据创新”的要求还有显著差距。随着政务数据共享开放工作不断深入，跨部门数据流通渠道基本建立，但对海量政务数据缺乏有效整合分析，数据挖掘分析多停留在简单的相关性分析层面，碎片化政务数据难以转化为可供决策使用的知识和智慧。

[1] 国脉研究院，《政务数据资源体系建设白皮书》

4、政府数据应用模式单一

当前政府数据的应用多是简单的检索与分析，缺乏基于深度推理的智能化应用，对海量数据之间存在的隐式关系挖掘不足，限制了大数据价值的挖掘分析。政府数据应用模式单一导致对社会治理的应用支撑能力不足，无法满足政府在宏观调控、社会管理、公共服务、市场监管方面的应用需要。在宏观调控方面，政府数据应用缺乏对金融交易、商贸活动、企业运行等的精准掌控。在社会管理方面，目前的应用无法满足对人口、交通、市政、安全、网络等重点社会管理领域的精准管理需求。在公共服务方面，个性化、精准化服务供给缺乏相关政府数据应用支撑。在市场监管方面，对政府数据的应用无法匹配事前事中事后全流程的监管要求。另外，这种相对单一的应用模式甚至会对政府数据的应用产生负面效应，存在国家秘密被泄露、个人隐私被暴露的消极影响。

第三章 知识图谱推动条件和驱动因素

一、 推动条件

（一）计算设备及硬件的发展

知识图谱的发展离不开计算硬件的支撑，特别是知识图谱构建、推理、应用过程中的机器学习算法的训练和预测等过程，对计算硬件的依赖显得尤为明显。随着不越来越多不同类型的硬件设备连接到互联网，生成了海量有用业务数据，同时基于这些业务数据在一定程度上改善该行业领域的用户体验。现阶段知识图谱对算力的需求体现在两方面，一是知识图谱算法包括大量的卷积、残差网络、全连接等计算需求，在摩尔定律接近物理极限、工艺性能提升对计算能力升级性价比日益降低的前提下，仅基于工艺节点的演进已经无法满足算力快速增长的需求；二是知识图谱需要对海量数据样本进行处理，强调芯片的高并行计算能力，同时大量数据搬运操作意味着对内存存取带宽的高要求，而对内存进行读写操作尤其是对片外内存进行读写访问的消耗的功耗要远大于计算的功耗，因而高能效的内存读写架构设计对芯片至关重要。目前市场上知识图谱技术使用的主流硬件加速器有三类：GPU、FPGA、ASIC。

1、GPU对知识图谱发展的推动

CPU与GPU两者都是由控制器、逻辑单元和寄存器组成，在CPU中控制器和寄存器占很大比重，而在GPU中逻辑单元的规模远远高于CPU。这种不同的构架决定了GPU在数据的算术、逻辑运算等方面的并行处理能力显著增强，而基于神经网络的知识图谱技术，对并行处理能力的要求远远高于计算精度。除了计算核心的增加，GPU在每个流处理器集群末端

设有共享内存，在共享内存中就可以完成线程之间的数据通讯，通讯速度显著提高，而CPU每次都返回内存，进行数据调用。另外，GPU采用的是GDDR5显存颗粒，具备很高的工作频率和显存带宽，对大规模深度神经网络的训练更有优势。越来越多的深度学习标准库支持基于GPU加速，如OpenCL、CUDA等。英伟达的GPU云平台NGC，提供Caffe、Caffe2、MXNet、CNTK、Theano、TensorFlow、Torch等框架、深度学习SDK等，大大促进了知识图谱技术的发展与应用效果。

2、FPGA对知识图谱发展的推动

GPU价格昂贵且功耗大，相对于GPU，价格便宜、能耗较低的可编辑门阵列（FPGA）引起了大家的注意。目前，FPGA可以采用OpenCL、C等更高效的编程语言，降低了硬件编程的难度，还可以集成重要的控制功能，整合系统模块，提高了应用的灵活性，FPGA以门电路直接运算，速度快，而用户可以自由定义这些门电路和存储器之间的布线，改变执行方案，以期得到最佳效果。FPGA是承载知识图谱技术的重要手段之一，以其价格低、功耗低的特性，使得知识图谱在智能终端上推广成为了可能。

3、ASIC对知识图谱发展的推动

随着机器学习技术的蓬勃发展，越来越多的机器学习算法被应用到知识图谱的构建、演化、推理等各个环节中，因此机器学习模型的运算速度决定了知识图谱更新周期。虽然GPU并行处理能力高，但不是针对机器学习而设计的，而FPGA要求用户自主编程，对用户的要求过高。芯片要同时具备并行化、低功耗、高性能等特性，还需要实现本地即时计算，这时ASIC（专用集成电路）的优势就显现出来了。但是，ASIC的研发周期长，可能无法跟上市场的变化。所以，SoC+IP模式较为流行。SoC（System-on-Chip，片上系统）可以在芯片上集成许多不同模块的芯片，每个模块可自行设计或根据需求进行第三方定制。与ASIC相比，该模式具有成本低、上市快、灵活地适配用户需求等特点。

（二）可用数据规模的提升

以互联网、物联网、感知网络及社交网络等为代表的新型信息技术的快速发展,推动数据获取的规模化和低成本化，引发了数据规模以爆炸式态势增长。根据智研咨询集团的预测，如图3.1所示，2020年，预计全球数据规模超过50亿ZB，到2025年，其规模超过163ZB。工信部副部长陈肇雄表示，我国海量数据快速增长，数据量年均增速超过50%，预计到2020年，数据总量全球占比将达到20%，将成为数据量最大、数据类型最丰富的国家之一。数据总量高速增长，线上线下产业新形态不断出现，数据资源日益成为贯通经济发展的大动脉的关键要素。

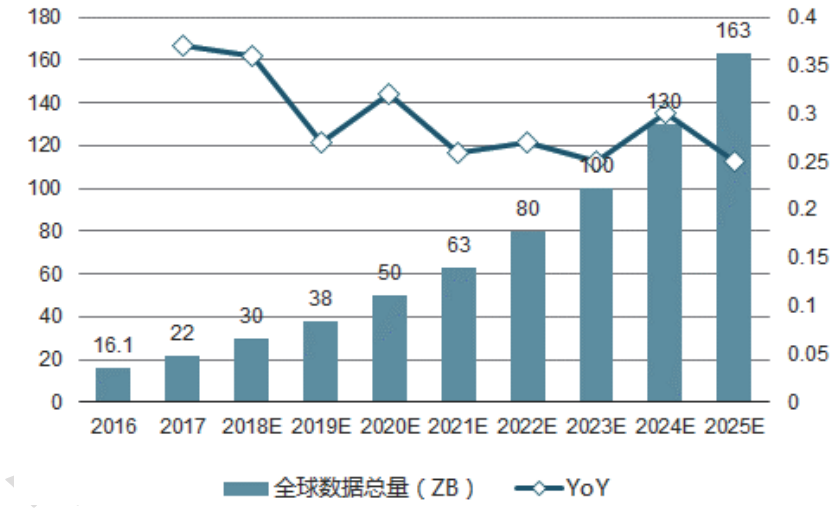

图3.1 全球数据总量预测2016–2025年

资料来源：智研咨询，《2018–2024年中国数据中心建设市场深度调研及投资战略分析报告》

1、丰富的数据资源储备奠定知识图谱工程化的知识基础

1970—1990年的专家系统未能进一步的完善和发展的重要原因之一在于知识局限性。专家知识的稀缺性和获取的高成本化导致专家系统高度领域化，缺乏良好拓展性，规模化难度大，技术功能价值不明显。随着可用

数据规模不断攀升，数据资源储备得以极大丰富，数据短缺困境得以快速破解。同时，各种技术工具的进步完善，提升了知识获取效率、降低了成本，使得大量“专家知识”可从数据中自动化获取和推理拓展，进一步巩固了知识图谱的知识基础。

2、数据规模攀升推动知识图谱技术的演进

可用数据规模的提升对传统知识工程技术提出了新的挑战和要求，进而催生新兴技术，推动知识图谱技术发展。所谓的知识图谱应用，主要是通过构建实体与关系的语义网络将大规模数据/知识进行整合、交叉关联、分析比对，对数据进行深度挖掘，支撑知识的智能化理解表示、推理、检索和服务，向用户提供自助的即席、迭代分析能力。相比较传统结构化数据处理工具，知识图谱在非结构化和半结构化数据的特征提取、内容检索、表示理解方面更具优势。对于非结构化、半结构化数据整合关联、知识提取表示等的需求和应用，无论在技术上还是功能上，传统数据库和数据分析挖掘技术工具都几乎束手无策。因此，可用数据规模攀升的推动下，数据服务提供商逐步转向选择和研发能支撑知识图谱应用业务需求的开源数据库技术和新兴数据处理、加工进而分析技术，进而推动了知识图谱技术的快速发展。

3、推动知识图谱规模化和商业化

专家系统没落的另外一个原因在于难以充分发挥规模效应导致商业价值释放困难。对于知识图谱总服务引擎而言，通用知识域和业务知识域是相辅相成的，可用数据规模和种类的丰富以及持续融合更新，底层通用知识域得以快速迭代、拓展和完善，通过通用知识服务引擎支撑业务知识提供领域化、精准化、精细化服务，同时业务领域通过服务产生新知识反馈至通用知识域进一步完善更新。基于该模式，随着数据规模的增大和数据内容的丰富，图谱底层通用知识域就越完善，知识重用和技术迁移能力就越高，新型智能化应用的开发边际成本越低，进而规模化程度提高，规模

效益显现，商业价值得以释放。最终，企业在规模报酬的驱动下，知识图谱应用研发和技术投入不断增加，推动知识图谱商业化。市场逐利驱动下的知识图谱的技术演变将持续化、常态化。

（三）算法演进

知识图谱是人工智能重要分支、是知识工程在大数据时代的重要成果，算法在人工智能各个阶段的发展推动了知识工程从最初基于语义网络的知识表示到现阶段知识图谱自动化构建和智能化知识服务。

1、算法在符号主义和连结主义时代对于知识图谱的推动

在人工智能的“推理”时代和“知识工程”时代，认知学家Allan M. Collins在提出了语义网络的知识表示方法，概念之间的语义关系通过网络的形式得以表达。 Newell和Simon基于符号表示和规则推理提出了“逻辑推理家”程序，实现了机器定理证明。这一时期主要是基于人解决问题的能力建立智能系统，机器只具备了逻辑推理的能力难以在实际中发挥作用。1977年Feigenbaum提出了知识工程的概念，开启了基于知识系统的“知识库”+“推理机”的机器智能，这一时期涌现出了大批量限定领域的专家系统，如斯坦福大学研制出的第一个专家系统DENDRAL、医疗专家系统MYCIN、基于模式匹配的自然语言理解系统STUDENT等。专家系统实现了某一特定领域的问题解答，但需要大量的人力工作，因此存在成本高、扩展性差等问题。

2、算法在大数据时代对于知识图谱的推动

在人工智能的“数据挖掘”时代，2001年Tim Berners-Lee提出了语义网的概念，利用本体描述互联网内容的语义结构。万维网使得知识由封闭走向开放，这个过程出现了群体智能，如维基百科。互联网时代数据的聚集，机器学习算法的发展推动了知识图谱的构建过程的自动化，增强了海量信息提取和分析能力。早期图谱构建方式主要基于规则，例如通过人工

编写规则和启发式算法从文本中自动提取公司名称；Galarraga利用AMIE算法实现不完备知识图谱中的关联规则挖掘。随着统计学习的发展，基于统计学习的方法图谱构建越来越多，基于统计学习的方法弥补了基于规则方法扩展性差、人力成本高的缺陷。Sundheim基于最大熵分类模型、支持向量机、条件随机场模型等统计学习方法实现实体识别和链接；Culotta基于依存核函数实现关系抽取；Zhu将本体学习用贝叶斯描述逻辑网络表示出来通过推理获得新公理。随着深度学习方法的迅速发展，利用深度学习方法构建知识图谱称为研究热点。与统计学习相比，深度学习方法端到端的方式减少了人工构建特征的工作，基于深度学习[Bengio, et al., 2009][1]为代表的向量知识表示克服了原有网络形式的知识表示存在的计算效率低和数据稀疏的问题。

（四）数据/知识检索需求攀升

随着信息技术和互联网的普及，全球逐步信息化和互联网、物联网时代的到来，随之而来的是政府数据、企业数据和互联网数据的几何级增长，如图3.2所示。2004年，全球数据总量是30EB。随后，2005年达到了50EB，2006年达到了161EB。这个规律，被称为新摩尔定律，即：人类有史以来的数据总量，每过18个月就会翻一番。2018年，希捷科技与IDC共同发布了一份名为《数据时代2025》的数据发展趋势报告，IDC在报告中预测，全球数据圈将从2018年的33ZB（泽字节）增至2025年的175ZB。

[1] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends in Machine Learning. 2009, 2(1): 1–127.

图3.2 全球数据圈规模

1、精准搜索正在成为搜索引擎的目标和挑战

信息的爆炸式增长，在带来信息获取便利的同时也带来了精准信息获取的难度和复杂度。据统计，在海量信息时代，人们在日常工作和信息处理中，查找信息的时间往往占据了近80%，而真正处理数据和利用数据，往往占据不到20%的时间。不管是哪一类搜索引擎，精准搜索成为当下搜索引擎的共同目标和挑战，也成为用户搜索体验的关键所在。精准搜索至少包含两层涵义，即首先搜索引擎精准理解用户的搜索意图，其次是在此基础上，通过一定的算法，搜索出与用户意图精准匹配的搜索结果，并呈现给用户。

2、个性化信息的精准智能化推荐成为全新的趋势

搜索是人们主动获取信息主要方式，而个性化信息的智能化推荐，则是人们被动获取信息的重要方式。个性化信息的智能化推荐是建立在海量数据挖掘与分析基础上的一种高级智能系统，向用户提供个性化的信息服务和决策支持。智能化推荐在日常生活中已经很常见，例如在电子商务网站（如京东、淘宝等）会根据用户的浏览记录、使用习惯等，为用户智能推荐用户可能喜欢的商品；资讯类APP中（如今日头条、网易新闻客户端等）会根据用户的阅读记录和内容风格，为用户推荐个性化的新闻资讯，以满足用户的阅读习惯，提升用户体验。这种智能化的个性化信息精准推

荐，不单单在互联网领域，在企业信息获取、政府信息获取、科研信息获取等各方面都已经成为了全新的趋势。

（五）开源知识库建设

由于支持力度大的社区不断引进和创新，使得开源软件始终处于技术发展前列，并且具有很强的、长期的生命力，可以一直满足企业用户的需求。同时，开源知识库的建设也不断促进者知识图谱的发展，知识库的开源能够吸引更多有才能的人加入知识库的建设中，为知识图谱的应用提供多种解决方案，共同促进知识图谱的不断创新与长期发展。开源知识库可分为：开放链接知识库和行业知识库。开放链接知识库的典型代表有：Freebase、Wikidata、DBpedia；垂直行业知识库的典型代表有：IMDB、MusicBrainz等。此外，国内外也建立了相关的开放链接数据网站，其中Linked Open Data是链接数据（Linked Data）作为语义网的一种实现的设想。开源知识库建设对推动知识图谱发展体现在以下两个方面：

1、开源知识库辅助知识图谱的知识抽取和本体构建。

现有的大规模知识图谱在知识获取上大多依赖于开放社区的协作和开源数据库，开源知识库的建设能为知识图谱的构建提供丰富的数据来源。开源知识库中拥有丰富的概念及实体知识，且这些知识都是以一定的形式组织在一起，支持概念及实体的快速查询。因此，开源知识库的知识可以有效支持知识图谱的构建，为知识图谱提供所需要的概念及实体知识，从而提升知识图谱的构建效率。开源知识库中的大量实体，丰富了知识图谱的内容。而开源知识库中的概念知识及分类体系，可有效支撑知识图谱的本体构建，从而降低图谱本体的构建难度。

2、开源知识库推动知识图谱融合，扩大知识图谱规模。

知识图谱之间的融合通常需要解决概念、实体的冲突问题。概念、实体的冲突问题源于缺乏一个统一的概念、实体应用规范与标准，导致相同

的概念、实体可能存在不同的名字，而相同的名字可能对应不同的概念和实体。而开源知识库可以为部分概念和实体的统一提供一个参考标准。开源知识库中的概念、实体知识可以起到桥梁作用，用于辅助知识图谱中概念、实体的共指消歧与实体消解。因此，开源知识库将有利于知识图谱间的融合，从而扩展知识图谱的规模丰富图谱信息，最大限度的发挥知识图谱的强大功能。

（六）专业人才培养

伴随着AI这块新的投资风口，新兴企业对AI人才的需求激增。据LinkedIn研究报告显示：过去三年间，通过领英平台发布的AI职位数量从2014年的5万飙升至2016年的44万，增长近8倍。截至2017年一季度，基于领英平台的全球AI领域技术人才数量超过190万，其中美国相关人才总数超过85万，高居榜首，而中国的相关人总数也才5万人，缺口极大。人才是引领互联网科学和技术实现内在价值的关键所在，互联网人才具有重要性和紧缺性的双重特征，谁抢占了人才高地，谁就能够掌握互联网领域发展的前沿方向。知识图谱的构建同样需要专业的计算机AI人才，基于目前互联网人才资源不足的现状，相关AI专业人才的培养已经成为了推动知识图谱从发展到成型的关键因素。国务院在2017年7月印发的《新一代人工智能发展规划》指出将“加快培养聚集人工智能高端人才”列为重点任务，并强调“把高端人才队伍建设作为人工智能发展的重中之重，完善人工智能教育体系，加强人才储备和梯队建设，形成我国人工智能高地”。除政府出台的AI人才培养政策以外，知识图谱相关的人才培养计划还需要做到以下几点：

1、构建新型校园AI人才教学体系

高校作为一个最大的人才输出平台，对于社会人才的建设十分重要，专业人才的培养需要从校内做起。在校园内通过优化专业课程体系，开展

有效的专业核心课程教学改革，深化实践教学改革，构建多维融合的AI才培养平台和建设应用校本特色的专业教学资源库，可以提高学生的综合能力和培养质量，以满足区域经济发展所急需的高素质AI人才的需求。

2、构建校企合作培养AI人才

知识图谱属于应用型学科，产生于企业应用和服务于社会。所以，培养适合经济社会发展需求的AI人才，必须走校企合作道路，构建新型校企合作机制，可以使学生离开校园进入企业学习，将理论结合实际在企业中进行运用，通过实践不断学习完善理论基础，技术基础，通过企业经验积累，可以提高学生的实践能力和判断力，满足社会对AI相关人才的需求。

3、优化人才政策、释放人才效能

要创新教育培训体系，充分结合互联网人才特点，以培养批判思维、独立思考为育才政策重点，推动新时代的AI人才的脱颖而出。要解放思想，慧眼识才，爱才惜才，不拘一格降人才，全球范围聚人才，精准化、针对性制定适用特殊人才的特殊政策。要放眼全球，下大力气引进优秀高端人才，重视青年人才的作用，营造优化AI人才发展环境，在国际视野下优化互联网人力资源配置。要顺势而为，构建有利于互联网人才创新聚集的聚才环境，确立同行评议、国际认可的互联网人才认证标准，突出专业性实用性，增强人才的成就感获得感。要完善薪酬激励、产权优化、财政税收等方面政策，破除互联网人才在国家区域机构间的流动障碍，推动互联网人才的有序顺畅流动，不断释放互联网人才的效能。

二、 驱动因素

（一）云计算

云计算的市场规模增长趋于稳定，目前主要以IaaS、PaaS、SaaS为代表的全球公有云为主，根据相关机构预测，市场规模已经达到2000亿美

元，平均增长幅度达到29%。知识图谱最初是由谷歌提出用来优化搜索引擎的技术，目前已经应用到了很多热门领域中，例如语音助手、聊天机器人、智能问答等。其可以组织现实世界中的知识，描述客观概念、实体、关系。这种基于符号语义的计算模型，一方面可以促成人和机器的有效沟通，另一方面可以为深度学习模型提供先验知识，将机器学习结果转化为可复用的符号知识累积起来。

云计算作为知识图谱的基础，能够更好的帮助知识图谱完成图谱的构建，数据的采集、存储、计算等。例如边缘计算可以使得知识图谱能够采集到更加完备的数据，云上的GPU可以帮助知识图谱更快的进行分析计算，云中的分布式存储使得数据存储上限近乎没有上限。同时，Web数据源的数量激增，大量RDF数据被发布。互联网正从仅包含网页和网页之间超链接的文档万维网(Document Web)转变成包含大量描述各种实体和实体之间丰富关系的数据万维网(Data Web)，云计算中的各种应用对知识图谱的依赖也越来越深，尤其以改进搜索质量、在复杂关系中寻找特殊关联关系等应用中尤为突出。由此可以看出云计算和知识图谱相辅相成，互相促进，知识图谱从云计算的各个基础技术中获得了极大的便利性，包括但不限于分布式存储，大规模并行计算，边缘计算等。同时知识图谱也给云计算中的应用带来了极大的帮助，例如安全方面，传统的方式不利于发现某些特殊的安全漏洞，利用知识图谱擅长的关系计算分析可以更为容易的发现安全隐患。

（二）大数据

大数据时代的到来，为人工智能的飞速发展带来了前所未有的数据红利，其中以知识图谱为代表的知识工程以及以深度学习为代表的机器学习等相关领域均取得了长足的进步。然而，随着大数据红利逐渐消失殆尽，深度学习的局限性日益凸显。相反，基于大数据种类和来源多样化、数据增长速度快、处理速度快、数据真实等特点，知识工程摆脱了传统人工构

建、数据规模局限的束缚，迎来了发展新机遇，特别在知识图谱技术的推动下，对机器友好的知识图谱重要性不断攀升。得益于海量数据规模和强大数据计算能力，大数据为知识获取获取提供了数据支撑，辅助了知识存储、知识表示、知识理解地不断优化，有利于大规模高质量知识库的构建。反之，作为实现机器认知智能的基础，知识图谱的关联分析、可解释性推理为大数据智能搜索、推荐、分析预测提供了可靠地先验支撑。

1、大数据为知识图谱提供数据支撑

传统知识工程与知识图谱的差异体现在数据容量规模上。传统知识工程在上世纪七八十年代后逐渐销声匿迹其的根本原因由于传统知识库主要依靠人工构建、成本高昂、规模有限。而在大数据时代，依赖海量数据规模以及大数据强大计算能力，利用自动化构建、众包等形式可实现大规模知识库的构建。而且，大数据已经有了很多成熟的分析技术，各个行业的数据量规模也达到了十亿级，甚至千亿级，知识图谱可以利用这些技术和数据，其分析计算能力将远远超越专家系统。以知识规模量变带动知识效用的质变，大数据为良好的知识表示提供基础。

2、知识图谱为大数据提供高质量结构化数据

人工智能的出现极大提高了数据可利用的广度和深度。随着互联网时代社交媒体的兴起，非结构化数据体量激增，如何对其进行有效处理也一直是大数据价值挖掘的重点和难点。知识图谱根据数据不同的结构化形式采用不用方法将数据结构转化为机器能够有效识别的形式，这是使机器具有理解和解释能力的前提。友好的数据结构有助于提升大数据的存储、分析、数据处理能力辅助大数据应用。此外，知识图谱可充分利用大数据多源特性对已有知识库进行交叉验证以保证知识库的质量。高质量数据也极大减少了大数据处理成本，提升了数据使用价值，有利于数据的深入挖掘。

3、知识图谱为大数据提供可靠的先验知识

语义关系足够丰富可实现对知识的精细分析和精准推理，以满足大

数据对智能服务的实际需求。大数据价值的分析、挖掘与创造的目的就是发现知识关联的模式，并对关联化的知识进行有效组织管理，将其运用到实践中以产生更多的价值。知识图谱通过知识表示、知识融合将不同种类的信息连接在一起形成巨大关系网，并以关系的视角推理出数据间的隐含知识。知识图谱用关系建立起数据间的联系，实现了具有可解释性的关联分析，这为大数据价值挖掘提供了可靠的先验知识，节约了大数据分析成本，符号化的知识表示有助充分利用知识图谱中的已有知识，也提高了大数据分析效率。

(三) 物联网

从物联网概念兴起发展至今，受基础设施建设、基础性行业转型和消费升级三大周期性发展动能的驱动，处于不同发展水平的领域和行业成波次地动态推进物联网的发展。当前，基础性、规模化行业需求凸显，一方面，全球制造业正面临严峻发展形势，主要国家纷纷量身定制国家制造业新战略，以物联网为代表的新一代信息技术成为重建工业基础性行业竞争优势的主要推动力量，物联网持续创新并与工业融合，推动传统产品、设备、流程、服务向数字化、网络化、智能化发展，加速重构产业发展新体系。另一方面，市场化的内在增长机制推动物联网行业逐步向规模化消费市场聚焦。受规模联网设备数量、高附加值、商业模式清晰等因素推动，车联网、社会公共事业、智能家居等成为当前物联网发展的热点行业。

1、物联网为知识图谱构建提供数据支撑

物联网通过前面所述的各种技术，汇集了无论是规模上还是质量上都前所未有的数据集，而知识图谱的第一环就是数据信息的抽取，数据的规模和质量直接影响到知识图谱的构建，而经过知识图谱的计算分析，能为物联网提供更好的决策，使得物联网中的数据采集部分更加高效。

2、物联网为知识图谱的应用提供落地渠道

物联网的传感器源源不断产生的大量数据，同时也为决策数据的下达提供了高速通路，当基于知识图谱应用生成决策数据时，能够确保快速准确下达至执行设备，从而保证了知识图谱与应用场景间的融合交互。例如，在智能家居领域中，由于设备的多样性和规则的个性化容易导致设备间控制冲突，知识图谱能够为动作器功能的自动分类与作用效果判断提供支撑，进而避免相关问题的出现。

3、物联网为知识图谱在各行业中相关平台建设提供底层支持

针对知识图谱在医疗、金融、工业等领域的平台建设而言，物联网是平台数据层的重要组成部分，通过可穿戴设备、手机APP、工业生产线等不同设备采集数据，为平台提供除静态数据录入与应用以外的动态数据支持，进而支撑平台的上层应用。同时，为已建设知识图谱的验证、更新与维护提供数据依据,提升平台的自我完善能力与适应能力。

（四）信息安全

近年来，网络空间安全形势快速变化，国家级博弈更为凸出、攻防对抗更为激烈、数字经济安全保障要求不断提升，网络安全形势演变对网络安全产业发展产生深刻影响。众所周知，知识图谱的核心就是知识提取、数据整合、知识表示，这些都涉及到了数据信息，如前所述，现今信息的安全受到了越来越广泛的关注，特别是个人和公司的隐私数据，一旦遭到泄露，将会导致严重的后果。在安全数据越来越多的大环境下，根据这些数据构建知识图谱，用来发现更多的安全隐患和攻击链条势在必行。

网络中的安全事件与攻击行为正从零散的、炫耀式的，变为有组织、有规模的行动，一般以窃取企业核心数据、破坏机构系统、个人隐私信息、金融欺诈、骗取勒索个人财产为目的。这些事件与攻击行为有几个特点：攻击链长，数据稀疏且流量庞大，病毒木马家族化，隐蔽性强，具有

产业链。在上述特点中，相同的是事件、攻击节点之间的关联性，如描述APT安全事件的三元组（URL、MD5、IP地址）。数据关联正是知识图谱擅长解决的问题。通过人工定义或自动发现关联关系，可以构建出基于安全行为和流量分析的知识图谱，从流量和用户行为中识别出隐藏的信息，如IP地址、文件哈希值、URL、访问行为、流量特征、设备日志等信息之间的复杂关系，基于统计模型或深度学习模型推理攻击行为，从而达到事中拦截攻击，事后溯源攻击的效果。

（五）机器学习

机器学习被广泛应用在知识获取、知识融合、知识建模、知识计算等知识图谱构建与应用环节。其中，知识计算目前主要分为单步推理和多步推理。在单步推理和多步推理中主要包括4个研究方向：基于传统的规则推理、基于分布式表示推理、基于神经网络的推理、基于上述方法的混合推理。常见的推理算法有单层神经网络SLM、矩阵分解模型RESCAL、TransE等。在知识表示方面，表示学习旨在将研究对象的语义信息表示为稠密低维实值向量。在该低维向量空间中，2个对象距离越近则说明其语义相似度越高。顾名思义，知识表示学习是面向知识图谱中的实体和关系进行表示学习。该方向最近取得了重要进展，可以在低维空间中高效计算实体和关系的语义联系，有效解决数据稀疏问题，使知识获取、融合和推理的性能得到显著提升。在基于知识图谱的应用方面，在视觉问答中大多数方法将CNN和RNN结合起来，通过CNN提取给定图像中的概念特征，然后使用RNN/LSTM与知识库中表示类似概念的节点相关联，最终得到直接从输入图像和问题到答案的学习映射[1]。近年来，以深度学习为代表的表示学习技

[1] Wang P , Wu Q , Shen C , et al. Explicit Knowledge–based Reasoning for Visual Question Answering[J]. Computer Science, 2015.
Wang P , Wu Q , Shen C , et al. FVQA: Fact–based Visual Question Answering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017:1–1.

术异军突起，在语音识别、图像分析和自然语言处理领域获得广泛关注。

在知识融合方面，ERNIE模型在BERT基础上，将外部知识引入大规模预训练语言模型中，有望增强预训练模型的鲁棒性，提高在知识驱动任务上的性能。在知识建模方面，基于先验知识的模型有助于解决知识图谱在实体和实体关系构建的两大瓶颈，自动补充知识获取阶段未成功提取的知识，智能补全未识别的实体间关系。在知识计算方面，缺失连接的推理、知识图谱的自生长自适应、推理的可解释性等问题的解决都依赖于机器学习相关算法模型的进一步发展。

（六）自然语言处理

自然语言处理是指利用计算机对人类日常使用的自然语言进行处理，通过对字、词、句、篇章的输入、识别、分析、理解、生成等操作，实现人机之间的信息交流，是一门涵盖了计算科学、人工智能和计算机语言学等多领域的交叉学科[自然语言处理研究报告，2018]。自动化、半自动化的知识图谱构建以及知识融合和挖掘过程中均离不开自然语言处理技术的支撑，同时自然语言处理的各个任务也离不开基础语料库和知识库的建设，如机器翻译、自动问答、会话系统、实体抽取等。随着知识图谱在自然语言处理的各个领域中的大获成功，现在自然语言处理有明显和知识图谱结合的趋势。特别是在特定领域的知识模型中，这种趋势就更明显，因为这些模型往往要关联很多领域的知识，而这种知识的整合和表示，很适合用知识图谱来解决。

1、自然语言处理为知识图谱提供技术支撑

与传统的语义网络相比，数据爆炸背景下的知识谱图构建需要从结构化、非结构化和半结构化数据源中获取实体、属性以及关系，这就需要利用词性标注、命名实体识别、关系抽取等技术从各类数据源中提取特定类型的信息，通过信息归并、冗余消除和冲突消解等手段将非结构化文本转

换为结构化信息，再通过信息集成技术中实体链接和共指消解实现知识融合。在知识图谱应用中，知识图谱的典型应用智能问答系统中需要利用自然语言处理的句法语义分析技术以及信息检索、文本生成技术，正确理解用户提出的问题，将从中抽取关键信息与知识图谱进行检索匹配，最终将获取的答案反馈给用户。

2、知识图谱为自然语言处理提供知识背景

自然语言存在歧义性、多样性，语义理解模糊且依赖上下文。机器理解自然语言困难的根本原因在于缺乏人的认知形成的背景知识。相对于传统知识表示形式，知识图谱具有覆盖率高、语义关系多样、结构友好以及质量较高等优势，机器在理解自然语言的同时需要这样的知识背景。因此，利用知识图谱，将图谱中的语义信息作为输入，使得知识图谱的已有的知识成为模型的输入，进而降低模型对大样本的依赖；利用知识图谱结合深度模型，基于从图谱中提取的相关已有知识，在训练样本减少的同时保证了模型预测的准确度，另一方面知识作为优化约束，指导模型训练可以更好的提升自然语言处理技术的模型效果。

3、自然语言处理推动了知识图谱的自动化构建

知识图谱的构建已由早期的专家手工和依靠群体智能建设，发展到利用自然语言处理技术自动化的构建的时代。随着大数据时代的到来，传统手工的方式对于知识图谱的构建可以说基本上是无法完成的任务，必须依赖自然语言处理的技术在海量数据中自动化的抽取知识，有监督的构建知识体系以及可持续的补充完善知识图谱。目前,通用知识图谱的规模越来越大,很多都是包含亿级的实体以及百千亿级的各类关系。还有一些针对不同的领域的知识图谱,如医疗知识图谱、金融知识图谱、司法知识图谱和教育知识图谱等，节点与关系的规模也都是在亿级以上。这些知识图谱的构建都受益于自然语言技术的发展。

4、自然语言处理推动了知识图谱的智能化应用

知识图谱是一个网状的知识库，如今有着非常广泛的应用，这不仅依赖于现阶段知识图谱质量的提升和人们对知识的重视，自然语言处理技术的发展也对知识图谱智能化的应用起到了推动作用。自然语言处理主要解决自然语言和计算机的交互问题，包括分析、理解、变化、检索、生成等方面，与图谱应用的很多环节都息息相关。搜索引擎是知识图谱常见的应用之一，基于自然语言处理中的词法分析、句法分析、相似度匹配、信息检索等技术，将提取用户检索的关键字和知识图谱中的数据进行转化比对，按照相关度为用户呈现检索结果。在智能问答系统中，通过语音识别技术、语义理解、智能交互等技术理解用户需求，基于知识图谱为用户提供分析决策。

（七）计算机视觉

随着计算机视觉技术日益成熟，在政策、资本、技术等方面都受到良好的待遇，应用场景不断拓展，自动驾驶、安防、医疗、政务等领域开始全面应用计算机视觉技术。随着消费者在安全和效率需求也不断提升，计算机视觉技术在各行业应用能有效满足人们需求，市场发展空间巨大。根据Forrester统计，全球在计算机视觉领域的投资持续增长，如图3.3所示，截止到2018年11月末投资较2017年增长113%，在过去的五年中复合增长率高达135%，未来五年内全球计算机视觉软件及服务市场规模将超 过200亿美金。

知识图谱对于计算机视觉作用主要是弥补深度学习模型对现实世界的大量背景知识的忽略。一方面，知识图谱可以作为深度学习的输入并优化模型。通常首先完成知识图谱中的语义信息表示学习，然后将离散化知识图谱组成元素（节点与边）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深度学习的输入，从而将符号推理引入以改进计算机视觉

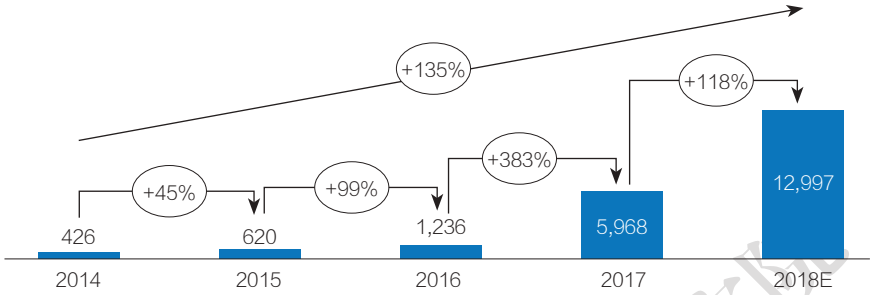

图3.3 全球计算机视觉行业投资趋势

资料来源：Forrester Research and Venture Scanner

中深度学习模型的性能[1]。此类问题的关键在于如何获取各类知识的高质量连续化表示，避免知识图谱的表示学习产生语义损失。

另一方面将知识作为计算机视觉算法中目标优化的约束，可以实现视觉算法模型与背景知识更好的一致性。例如将知识图谱中知识表达为优化目标的后验正则项[2]，通过量化成对概念之间语义一致性与目标检测模型集成实现知识感知检测[3]，使用链接预测方法集成语义模型提高视觉关系检测的结果[4]。这方面后续将重点探索或构建专门针对视觉任务的知识图并融合在深度学习模型中，降低模型对于大规模标注样本的依赖，实现超越对封闭数据集的有监督的学习，并具备不断扩展知识的能力。

1 Marino K , Salakhutdinov R , Gupta A . The More You Know: Using Knowledge Graphs for Image Classification.[C] IEEE Conference on Computer Vision & Pattern Recognition 2017.
Junyu Gao, Tianzhu Zhang, Changsheng Xu. I Know the Relationships: Zero–Shot Action Recognition via Two–Stream Graph Convolutional Networks and Knowledge Graphs[C] . Association for the Advancement of Artificial Intelligence, 2019.

2 Bilen H , Pedersoli M , Tuytelaars T . Weakly Supervised Object Detection with Posterior Regularization[C]. The British Machine Vision Conference. 2014.

3 Fang Y , Kuan K , Lin J , et al. Object Detection Meets Knowledge Graphs[C]. Twenty–Sixth International Joint Conference on Artificial Intelligence. 2017.

4 Baier S , Ma Y , Tresp V . Improving Visual Relationship Detection using Semantic Modeling of Scene Descriptions[J]. 2018.

第四章 知识图谱的主要技术

知识图谱主要技术包括知识获取、知识表示、知识存储、知识建模、知识融合、知识理解、知识运维等七个方面，通过面向结构化、半结构化和非结构化数据构建知识图谱为不同领域的应用提供支持，具体的技术架构图如图4.1所示。

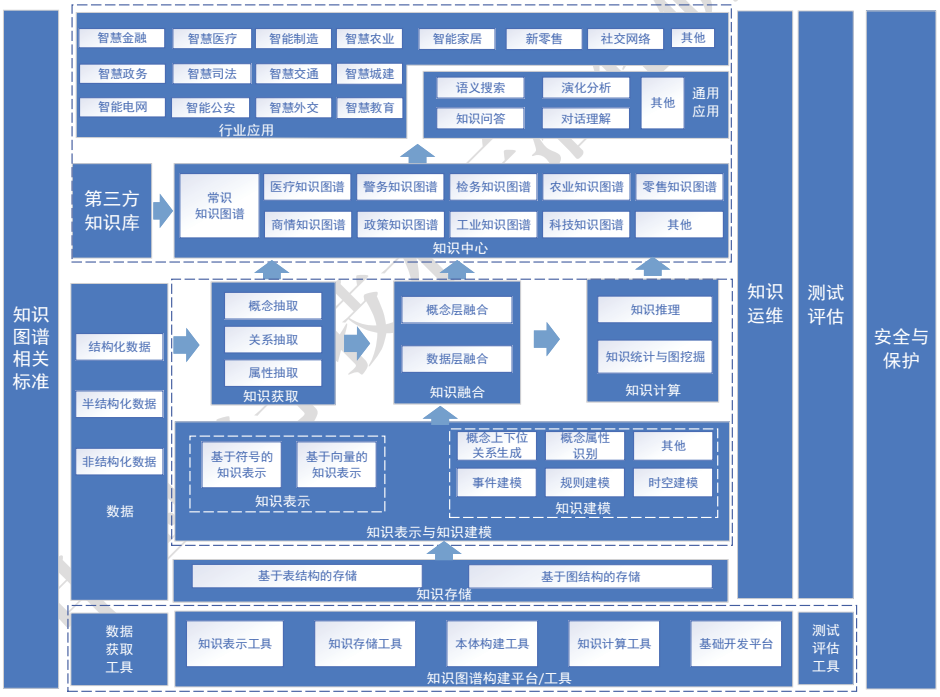

图4.1 知识图谱技术架构图

一、知识获取

（一）知识获取概述

知识图谱中的知识来源于结构化、半结构化和非结构化的信息资源，

如图4.2所示。通过知识抽取技术从这些不同结构和类型的数据中提取出计算机可理解和计算的结构化数据，以供进一步的分析和利用。知识获取即是从不同来源、不同结构的数据中进行知识提取，形成结构化的知识并存入到知识图谱中。当前，知识获取主要针对文本数据进行，需要解决的抽取问题包括：实体抽取、关系抽取、属性抽取和事件抽取。

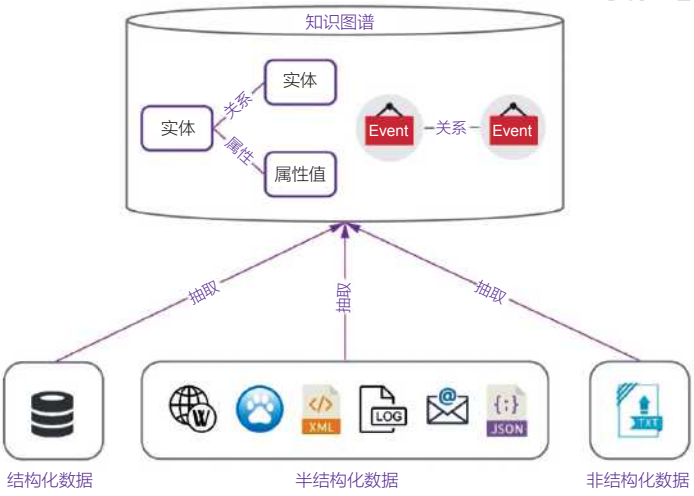

图4.2 知识获取示意图

知识获取作为构建知识图谱的第一步，通常有以下四种方式：众包法、爬虫、机器学习、专家法[1]。

众包法：允许任何人创建、修改、查询的知识库，就是常说的众包模式，百度百科，维基百科就是典型的例子。此类场景下知识库存储的不是大量的杂乱的文本，而是机器可读，具有一定结构的数据格式（以百度某词条为例）。现代通过众包法建立的知识图谱如谷歌和百度的知识图谱都已经包含超过千亿级别的三元组，阿里巴巴于2017年8月份发布的仅包含核心商品数据的知识图谱也已经达到百亿级别。

[1] 杨玉基,许斌,胡家威,仝美涵,张鹏,郑莉.一种准确而高效的领域知识图谱构建方法.软件学报,2018,29(10): 2931−2947.

爬虫：网页开发者将网页中出现的实体、实体属性、关系按照某种规则做上标记，Google、百度等搜索引擎通过爬虫就能获取到这些数据，从而达到知识图谱数据积累。当前不同语言的爬虫框架有不少，例如python的Scrapy，java的WebMagic等，通过简单的配置即可完成爬虫的规则定义、爬取、清洗、去重、入库等操作，从而获取知识。

机器学习：通过机器学习将数据变成了可理解的知识，例如通过文本分类、主题模型等机器学习模型，可以获取文本的特征，而这些特征就可以理解为知识。

专家法：专家法通常用于垂直领域的工程实践，通过专家的经验，归纳总结后形成知识，例如在知识图谱中的事件图谱通常是由专家的经验形成的。

（二）研究现状

目前，面向互联网海量文本数据的知识抽取是研究的主流。已有很多信息抽取的方法被提出用来解决该问题，按照抽取对象的不同，可分为实体抽取、关系抽取、属性抽取和事件抽取。

1、实体抽取

实体抽取也称为命名实体识别（named entity recognition, NER），是指从文本语料库中自动识别出专有名词（如机构名、地名、人名、时间等）[1]或有意义的名词性短语，实体抽取的准确性直接影响知识获取的质量和效率。因此，实体抽取是知识图谱构建和知识获取的基础和关键。为了解决早期的实体抽取方法存在的问题，规则和监督学习相结合的方法、半监督方法、远程监督方法以及海量数据的自学习方法等被相继提出。

2、关系抽取

关系抽取是利用多种技术自动从文本中发现命名实体之间的语义关

[1] Chinchor N, Marsh E. Muc–7 information extraction task definition [C],1998:359–357.

系，将文本中的关系映射到实体关系三元组上。较于实体抽取，关系抽取更加复杂，研究的难点主要体现在并非所有的关系都很明显，即关系表达的隐含性；实体关系不仅有二元，还有多元，即关系的复杂性；一种关系可能会有多种表述形式，如A位于B或B的省会是A，即语言的多样性。

3、属性抽取

属性主要是针对实体而言的，以实现对实体的完整描述，由于可以把实体的属性看作实体与属性值之间的一种名词性关系，所以属性抽取任务就可以转化为关系抽取任务，例如，[Guo et al., 2012][1]采用SVM方法将人物属性抽取问题转化为人物的关系抽取，[Suchanek et al.,2007][2]提出的基于规则与启发式的抽取方法能够从Wikipedia和WordNet半结构化网页中自动抽取出属性和属性值，其抽取的准确率可达95%，并因此得到了著名的本体知识库YOGO，还有直接从非结构化文本中挖掘出实体属性名称和属性值之间的位置关系模式的属性抽取的方法[刘峤等, 2016][3]。

4、事件抽取

事件是发生在某个特定时间点或时间段、某个特定地域范围内，由一个或者多个角色参与的一个或者多个动作组成的事情或者状态的改变[4]。目前已存在的知识资源（如维基百科等）所描述实体及实体间的关联关系大多是静态的，事件能描述粒度更大的、动态的、结构化的知识，是现有知识资源的重要补充[5]。

---

1. Jian–Yi G , Zhen L , Zheng–Tao Y , et al. Extraction and relation prediction of domain ontology concept instance, attribute and attribute value[J]. Journal of Nanjing University(Natural Sciences), 2012.
2. Suchanek F M , Kasneci G , Weikum G . YAGO: a core of semantic knowledge[C]// Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8–12, 2007. OAI, 2007.
3. 刘峤,李杨,段宏,刘瑶,秦志光. 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. Doddington G R, Mitchell A, Przybocki M A, et al. The automatic content extraction (ace) program–tasks, data, and evaluation. [J]. In Proceedings of the International Conference on Language Resources and Evaluation, 2004, 2: 1.
4. 中国中文信息学会语言与知识计算专委会，《知识图谱发展报告(2018)》.

（三）技术发展趋势

现有对知识抽取的研究虽在特定领域数据集上取得了较好的效果，但远远满足不了实际任务的要求。因此，资源缺乏、面向开放域、跨语言及跨媒体等方向的知识抽取成为未来的研究方向。

1、资源缺乏下的知识抽取

目前，大多数知识抽取的研究主要采用的是有监督的方法，需要大量的标注训练集去训练模型参数，然而构建标注数据集的成本比较大，且在更换语料类型后，算法的性能往往不尽如人意，即算法的性能严重依赖大规模的训练数据。因此，如何构建资源缺乏下的知识抽取系统成为研究的热点。

2、面向开放域的知识抽取

知识抽取是知识图谱构建的基础，虽然当前技术对于知识单元（实体、关系、属性）的抽取在特定领域取得了较好的效果，但是由于数据主题或规模等条件的制约，方法的可移植性与可扩展性不强，不能很好的满足大规模开放领域环境下的知识抽取要求。因此，知识抽取发展方向之一是研究大规模面向开放领域的知识抽取技术。具体包括：（1）数据规模上的可扩展性，能够高效完成海量数据的抽取任务；（2）数据主题上的鲁棒性，能够在面对不同主题的数据时具有鲁棒性。

3、跨语言的知识抽取

随着英文知识图谱技术的发展，多语种知识库的构建任务也在快速开展，因此，跨语言的知识抽取任务也成为当下研究的热点。跨语言的知识抽取为研究语言间的互补性和冗余性提供了机会，具体的研究包括：（1）自然语言表达的多样性，不同的语种在表示方式上均具有多样性，需要将实体关系知识映射到三元组上；（2）不同语种在知识表达方式上的差异性，通过比较不同语种对同一知识的表述，可以达到删除或更新错误知识的目的。目前针对跨语言的知识抽取，已有学者进行了研究并取得

了一些成果，例如，清华大学李涓子教授团队融合中英文维基百科、法语维基以及百度百科构建成了的跨语言知识库XLORE，并在此基础上实现了实体链接系统XLink。

4、跨媒体的知识抽取

随着深度学习创新理论的出现及大数据和算力的强力支撑，文本已经可以和图像/视频、音频等跨媒体数据采用相同的深度学习框架进行分析和建模，使得不同模态数据在人类语义层面耦合。因此，跨媒体的知识抽取任务逐渐成为研究热点之一。跨媒体的知识抽取可以利用视觉、听觉等多模态已标注信息来辅助文本标注缺乏下的知识抽取，又可以作为类似跨语言知识抽取的另一维度为实体间未知关系的挖掘及已标注关系的消歧提供互补信息。此外，跨媒体知识抽取在上述开放应用域的基础上，从多模态数据域维度进一步扩展，对模型鲁棒性等方面提出更大挑战。具体研究内容包括：

（1）视觉实体和关系的抽取：相对于文本中的实体和关系相对确定性，视觉实体和关系呈现出尺度、表型、空间关系等多样性，需要通过鲁棒语义模型的构建实现视觉实体和关系的抽取，从而将视觉局部区域映射到三元组上；（2）视觉事件的自然语言描述：针对图像/视频，基于人工智能理论自动生成一段语法和逻辑合理的视觉内容自然语言描述，从而实现语义丰富的视觉信息到抽象的语义事件描述的映射；（3）跨媒体信息融合：跨媒体信息在知识载体上存在差异，通过多模态信息在相同粒度和语义上的对齐，进一步实现特征和语义层面的融合，可以综合利用多模态信息，来辅助后续知识表示、建模、计算等关键技术，并形成面向跨媒体知识图谱构建的创新理论体系和关键技术。

二、知识表示

（一）知识表示概述

知识是人类在认识和改造客观世界的过程中总结出的客观事实、概念、定理和公理的集合。知识具有不同的分类方式，例如按照知识的作用范围可分为常识性知识与领域性知识。知识表示是将现实世界中存在的知识转换成计算机可识别和处理的内容，是一种描述知识的数据结构，用于对知识的一种描述或约定。知识表示在人工智能的构建中具有关键作用，通过适当的方式表示知识，形成尽可能全面的知识表达，使机器通过学习这些知识，表现出类似于人类的行为。知识表示是知识工程中一个重要的研究课题，也是知识图谱研究中知识获取、融合、建模、计算与应用的基础，如图4.3所示。

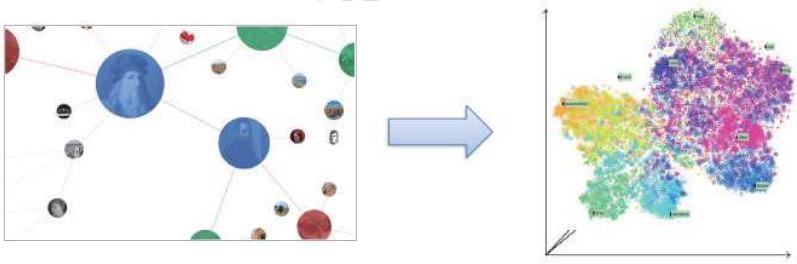

图4.3 知识图谱的向量表示[1]

（二）研究现状

知识表示方法主要分为基于符号的知识表示方法与基于表示学习的知识表示方法。

1、基于符号的知识表示方法

基于符号的知识表示方法分为早期知识表示方法与语义网知识表示方

[1] 谷歌知识图谱项目 Richard Socher, et al. Zero-Shot Learning Through Cross-Modal Transfer. ICLR 2013.

法。其中，早期的知识表示方法包括一阶谓词逻辑表示法、产生式规则表示法、框架表示法与语义网络表示法。

（1）一阶谓词逻辑表示法

基于谓词逻辑的知识表示方法，通过命题、逻辑联结词、个体、谓词与量词等要素组成的谓词公式描述事物的对象、性质、状况和关系。一阶谓词逻辑表示法以数理逻辑为基础，表示结果较为精确，表达较为自然，形式上接近人类自然语言。但是也存在表示能力较差，只能表达确定性知识，对于过程性和非确定性知识表达有限的问题。

（2）产生式规则表示法

上世纪40年代，逻辑学家Post提出了产生式规则表示。根据知识之间具有因果关联关系的逻辑，形成了“IF-THEN”的知识表示形式，该形式是早期专家系统常用的知识表示方法之一。这种表示方法与人类的因果判断方式大致相同，直观，自然，便于推理。除此之外，产生式规则表示法知识的表达范畴较广，包括确定性知识，设置置信度的不确定性知识，启发式知识与过程性知识。但是产生式规则表示法由于具有统一的表示格式，当知识规模较大时，知识推理效率较低，容易出现组合爆炸问题。

（3）框架表示法

20世纪70年代初，美国人工智能专家M.Minsky提出了一种用于表示知识的“框架理论”。来源于人们对客观世界中各种事物的认识都是以一种类似框架的架构存储在记忆中的思想，形成了框架表示法。框架是一种通用数据结构，用于存储人们过去积累的信息和经验。在框架结构中，能够借助过去经验中的概念分析和解释新的信息情况。在表达知识时，框架能够表示事物的类别、个体、属性和关系等内容。框架结构一般由“框架名-槽名-侧面-值”四部分组成，即一个框架由若干各个槽组成，其中槽用于描述所论事物某一方面的属性；一个槽由若干个侧面组成，用于描述相应属性的一个方面，每个侧面拥有若干值。框架具有继承性、结构化、自

然性等优点，但复杂的框架构建成本较高，对知识库的质量要求较高，同时表达不够灵活，很难与其他的数据集相互关联使用。

（4）语义网络表示法

1960年，认知科学家Allan M.Collins提出了语义网络（Semantic Network）的知识表示方法。语义网络是一种通过实体以及实体间语义关系表达知识的有向图。在图中，节点表示事物、属性、概念、状态、事件、情况、动作等含义，节点之间的弧表示它所连接的两个节点之间的语义关系，根据表示的知识情况需要定义弧上的标识，一般该标识是谓词逻辑中的谓词，常用的标识包括实例关系、分类关系、成员关系、属性关系、包含关系、时间关系、位置关系等。语义网络由语义基元构成，语义基元可通过三元组（节点1，弧，节点2）描述，语义网络由若干个语义基元及其之间的语义关联关系组成。语义网络表示法具有广泛的表示范围和强大的表示能力，表示形式简单直接、容易理解、符合自然。然而语义网络存在节点与边的值没有标准，完全由用户自己定义，不便于知识的共享问题、无法区分知识描述与知识实例等问题。

2、基于表示学习的知识表示方法

早期知识表示方法与语义网知识表示法通过符号显式地表示概念及其关系。事实上，许多知识具有不易符号化、隐含性等特点，因此仅通过显式表示的知识无法获得全面的知识特征。此外，语义计算是知识表示的重要目标，基于符号的知识表示方法无法有效计算实体间的语义关系，如图4.4和4.5所示。

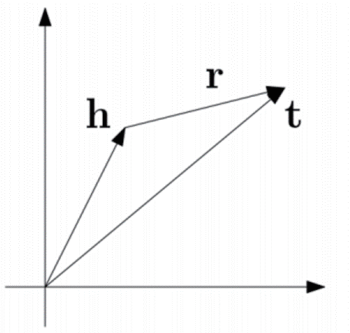

图4.4 TransE模型[1]

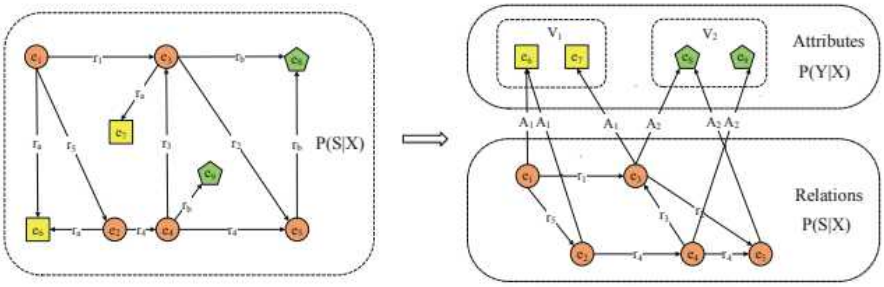

图4.5 带条件损失度量的KR-EAR方法[2]

（三）技术发展趋势

知识表示作为知识抽取、融合、建模、计算、应用的基础，侧重于表达实体、概念之间的语义关联，针对知识图谱的语义增强在未来依旧是知识表示重要任务。知识表示的研究趋势和动态包括以下四个方面。

1、符号与表示学习的融合统一

基于符号的知识表示方法由于考虑了人类的自然语言理解方式，具有严密性、自然性、通用性、知识易表达等优点，但是也存在计算效率低、

[1] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C]//Advances in neural information processing systems. 2013: 2787-2795.
[2] Lin Y, Liu Z, Sun M. Knowledge representation learning with entities, attributes and relations[J]. ethnicity, 2016, 1: 41.52.

无法捕捉隐含语义知识等不足。而基于表示学习的知识表示方法计算效率高却存在可靠性低，推理效果不佳等问题。因此研究基于符号逻辑与表示学习融合统一的知识表示方法有助于知识表达性能的进一步提升，也是未来的发展方向。

2、面向事理逻辑的知识表示

事理逻辑是指事件之间的演化规律和模式。已有的以实体、实体属性、实体与实体或属性之间关系为核心的知识图谱缺乏针对事件之间的演化规律与模式的知识挖掘。事实上，事理逻辑是一种非常有价值的常识知识，挖掘这种知识对认识和分析人类行为与社会发展变化规律意义重大。哈尔滨工业大学社会计算与信息检索研究中心首先提出了“事理图谱”的概念，认为事理图谱是以“事件”为核心的新一代动态知识图谱，并于2018年9月正式对外发布了基于大规模财经新闻文本的金融事理图谱。面向事件实体、事理逻辑关系，如顺承、因果、条件、上下位、组成等的事件知识表示方法是表达和丰富事理图谱的重要基础。

3、融合时空间维度的知识表示

现实世界中，许多知识具有时间和空间属性，例如“王菲的丈夫是李亚鹏”这条知识具有潜在的时间信息；“早餐是豆浆和油条”这条知识潜在的空间信息是中国的某些地区，从时空维度拓展知识表示对许多特定领域具有较强的现实意义。德国马普研究所研制的YAGO知识库为许多知识条目增加了时间和空间维度的属性描述，丰富了知识库内容。人们关心当前事实的同时，也会关注过去和未来的知识情况以及不同空间的知识表达含义，形成融合时间或空间维度的知识表示是增强知识表达的有效方式。

4、融合跨媒体元素的知识表示

当前的知识图谱主要以文本为主，事实上，跨媒体元素包括声音、图片、视频、音频等数据对于丰富和增强知识图谱的知识语义具有重要作用。2017年，国务院发布的《新一代人工智能发展规划》中，明确提出要

加强跨媒体分析推理技术的研究。不同的跨媒体元素能够表达相同的语义信息，能比单一模态反映更加全面正确知识内容。建立基于跨媒体元素的统一知识表示方法对于分析挖掘跨媒体要素的语义信息，构建跨媒体知识图谱具有重要意义。

三、知识存储

（一）知识存储概述

知识存储是针对知识图谱的知识表示形式设计底层存储方式，完成各类知识的存储，以支持对大规模图数据的有效管理和计算。知识存储的对象包括基本属性知识、关联知识、事件知识、时序知识和资源类知识等。知识存储方式的质量直接影响到知识图谱中知识查询、知识计算及知识更新的效率。

（二）知识存储方式

从存储结构划分，知识存储分为基于表结构的存储和基于图结构的存储[1]，如图4.6所示。

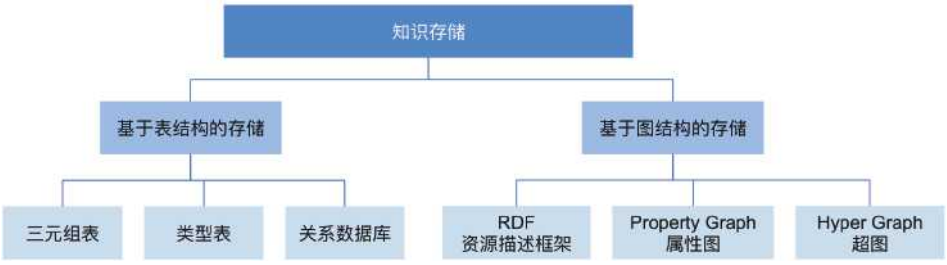

图4.6 知识存储方式

[1] 赵军.知识图谱[M].高等教育出版社:北京,2018:250.

1、基于表结构的存储

基于表结构的存储，是指运用二维的数据表对知识图谱中的数据进行存储。根据不同的设计原则，可以具有不同的表结构，如：三元组表、类型表和关系数据库。三元组表如jena等，优点是简单直接，易于理解。缺点是整个知识图谱都存储在一张表中，导致单表的规模太大。相应的插入，删除，查询，修改的操作开销也大。对实用性大打折扣。复杂查询在这种存储结构上的开销巨大。复杂查询拆分成若干个简单查询的操作，降低了查询的效率。

2、基于图结构的存储

基于图结构的存储即使用图模型描述和存储图谱数据。这种方式能直接反应图谱的内部结构，有利于知识的查询，结合图计算算法，进行知识的深度挖掘与推理。目前业界公认的图模型有3种，分别是属性图（Property Graph）、资源描述框架（RDF）和三元组超图（HyperGraph），其中属性图和资源描述框架已广泛运用到多个图数据库产品中。

（1）属性图（Property Graph）

属性图或带标签的属性图（Labeled-Property Graph），由顶点（圆圈）、边（箭头）、属性（key:value）和标签组成，顶点和边可以有标签。属性图的表达很贴近现实生活中的场景，也可以很好地描述业务中所包含的逻辑。常见的属性图结构如图4.7所示。其中，节点的标签是User，边的标签是FOLLOWS。

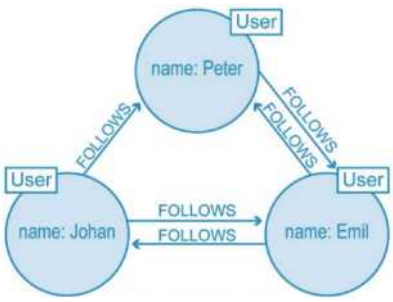

图4.7 属性图结构

（2）资源描述框架（RDF）

鉴于传统关系数据库拥有较高的通用性、可靠性、稳定性及成熟的技术，基于RDF的知识形式也广泛使用关系数据库作为其存储方式。目前主要有以下三种存储方案：基于三元组的三列表存储、水平存储、基于类型的属性表存储和基于谓词存储等[1]。对于基于RDF知识的三列表存储，该存储方式将关系数据库表的3列分别存储为RDF知识三元组的主语、谓语和宾语，即对应（实体，关系，实体）或者（实体，属性，属性值）。该三列表存储方式与传统的结构化数据存储方式相兼容，通用性好。但面向大规模的知识图谱，其本身包含大量的三元组，从而会造成关系数据库低效的查询性能。

（3）超图（Hyper Graph）

超图[2]概念的提出，是为了解决简单图中的共指消解和分割等问题。对于我们熟悉的图而言，简单图的一个边（edge）只能和两个顶点连接；而对于超图来讲，人们定义它的边（超边hyperedge）可以和任意个数的顶点连接。超图可以完美刻画标签网络中一条边包含多节点的问题。图和超图的示意图如图4.8所示：

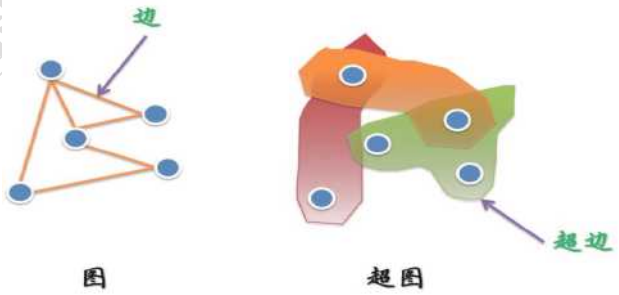

图4.8 普通图与超图

[1] 邹磊. 海量RDF数据管理[J]. 中国计算机学会通讯，2012，8（11）:32–43.
[2] Yue Gao ; Rongrong Ji ; Peng Cui ; Qionghai Dai ; Gang Hua.Hyperspectral Image Classification Through Bilayer Graph–Based Learning[EB/OL].https://ieeexplore.ieee.org/abstract/document/6804696,2014–7.

（三）知识存储基础工具

知识图谱的存储并不依赖特定的底层结构，一般的做法是按数据和应用的需求采用不同的底层存储，甚至可以基于现有关系数据库或NoSQL数据库进行构建。关系型数据库是典型的基于表结构的存储，图数据库是典型的基于图结构的存储。

1、关系型数据库

关系数据库通过属性对现实世界中的事物进行描述，采用关系模型来组织数据的数据库，其以行和列的形式存储数据。一行一个记录，一列表示一个属性。用户通过查询来检索数据库中的数据，而查询是一个用于限定数据库中某些区域的执行代码。根据DB-Engines的统计数据，从2013年到2019年，Oracle、MySQL和Microsoft SQL Server稳居关系型数据库的前三甲，如图4.9所示。

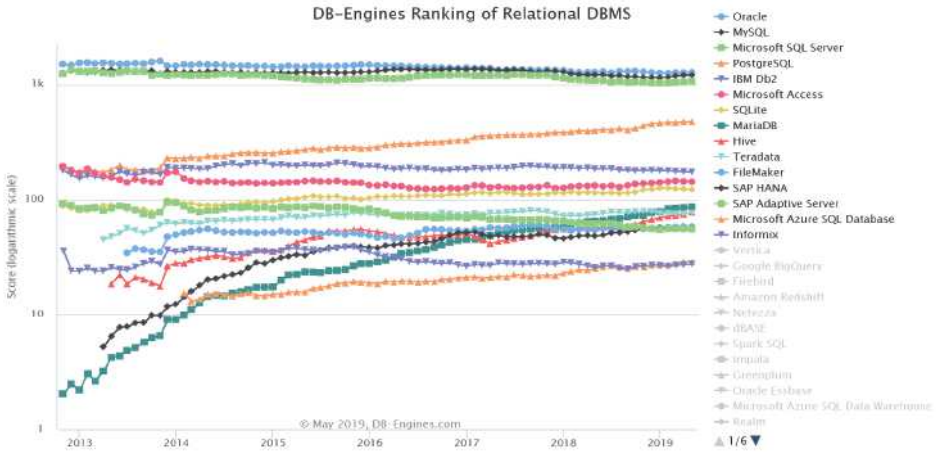

图4.9关系数据库的使用热度趋势

（数据来源：https://db-engines.com）

2、图数据库

图数据库源起欧拉和图理论(graph theory)，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(边)来体现，也可处理键值对，优点是快速解决复杂的关系问题。图数据库是一种非关系型数据库，支持对图结构进行查询、增加、删除、更新等操作。相对传统的关系型数据库，查询速度快、操作简单、能提供更为丰富的关系展现方式。根据DB-Engines的统计数据，列举了一些近几年来常用图数据库的使用热度趋势曲线，如图4.10所示。

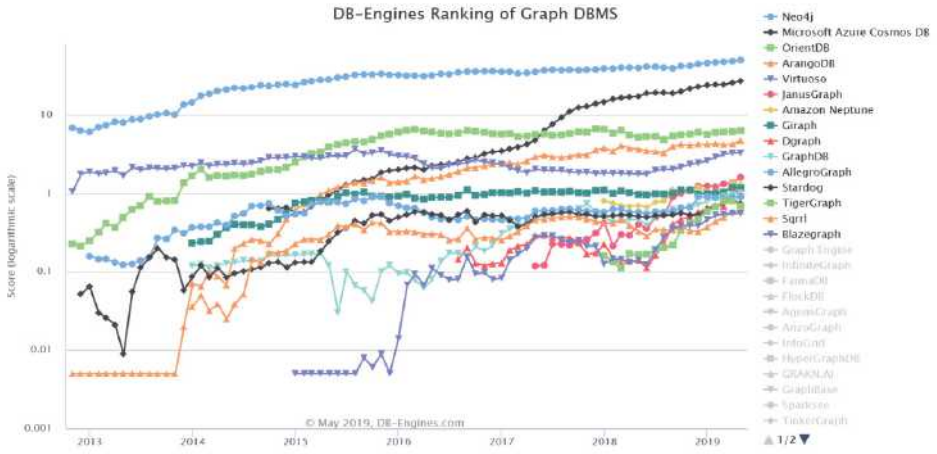

图4.10 基于图模型数据库的使用热度趋势
（数据来源：https://db-engines.com）

（四）技术发展趋势

知识存储过程完成对各类知识图谱组成元素的存储，以支持对大规模图数据的有效管理，满足快速查询、读取、写入的操作，为知识计算等应用提供服务支撑。针对知识图谱的分布式存储、知识存储的伸缩性和灵活性，以及基于LOD的知识存储，是技术发展的重点方向，同时，超图

（Hyper Graph）也是未来知识存储的研究热点。知识存储相关技术发展趋势包括以下几个方面：

1、基于RDF知识表示的分布式存储

随着互联网及各领域数据的持续积累，各个领域方向的知识图谱数量与规模日益增长，单机系统的数据存储能力难以应对知识图谱数据的大规模增长。同时，基于资源描述框架（RDF）存储的知识图谱更便于知识推理和计算，符合知识图谱应用的未来需求。但RDF存储模式所含有大量三元组的数据，使其索引效率与更新维护成本大于其它图存储模式。因此，未来知识存储的一种研究趋势为如何利用分布式数据库系统来解决RDF数据的大规模增长问题。

2、设计高适应性的知识存储

随着知识图谱的规模越来越庞大、知识的表示方式越来越复杂，这对目前的知识存储方式提出了挑战。如何设计出可支持对复杂节点的定制、具有良好可伸缩性和灵活性的知识存储模式，满足复杂的查询、读取、计算和应用需求成为面向知识图谱的知识存储的迫切要求。

3、基于LOD（Linked Open Data）的知识存储

由于知识表示RDF模型的通用性和灵活性，知识图谱供应方越来越倾向将自身的知识图谱数据表示成RDF格式并发布到互联网上。通过URI相互链接起来，这些发布在互联网上的RDF数据共同构成了一个覆盖整个互联网的庞大知识图谱。为了让这个庞大知识图谱网络更加丰富和完善，W3C积极推进LOD项目。LOD已成功将数百个RDF数据集相互链接在一起以增强数据的可用性。

4、超图的进一步研究和应用

超图所拥有的简单图无可比拟的复杂关系表示方式，能更加全面详尽地描述业务、还原场景。但目前对超图的可视化表示方法还没有理想方案，对于超图的划分方式、深度学习及应用，大部分仍处于实验室研究阶

段。推广到各领域进行工程化运用，无论在计算效率和成本上都存在较大问题。但随着知识图谱的普及，未来对于复杂关系的表示的需求，将逐步增多，超图技术的研究和应用探索将是知识图谱的下一个方向。

四、知识融合

（一）知识融合概述

知识融合的概念最早出现在1983年发表的文献[HOLSAPPLE C，et al,1983][1]中，并在20世纪九十年代得到研究者的广泛关注。而另一种知识融合的定义是指对来自多源的不同概念、上下文和不同表达等信息进行融合的过程[维基百科]。[A.Smirnov，et al,2002][2]认为知识融合的目标是产生新的知识，是对松耦合来源中的知识进行集成，构成一个合成的资源，用来补充不完全的知识和获取新知识。[唐晓波，魏巍，et al,2015][3]在总结众多知识融合概念的基础上认为知识融合是知识组织与信息融合的交叉学科，它面向需求和创新，通过对众多分散、异构资源上知识的获取、匹配、集成、挖掘等处理，获取隐含的或有价值的新知识，同时优化知识的结构和内涵，提供知识服务。

（二）知识融合过程

知识融合是一个不断发展变化的概念，尽管以往研究人员的具体表述不同、所站角度不同、强调的侧重点不同，但这些研究成果中还是存在很多共

1 HOLSAPPLE C，WHINSTON， A Software tools for knowledge fusion［J］. Computer world，1983，17(15):11－18.
2 SMI R NOV A，LEVASHOVA T，SHILOV N. Context-based knowledge fusion patterns in decision support system for emergency response［C］//Proceedings of the 10th international ISC R AM Conference. Karlsruhe: Karlsruhe Institute of Technology，2013:597－606.
3 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.

性，这些共性反应了知识融合的固有特征，可以将知识融合与其他类似或相近的概念区分开来。知识融合是面向知识服务和决策问题，以多源异构数据为基础，在本体库和规则库的支持下，通过知识抽取和转换获得隐藏在数据资源中的知识因子及其关联关系，进而在语义层次上组合、推理、创造出新知识的过程，并且这个过程需要根据数据源的变化和用户反馈进行实时动态调整。从流程角度对知识融合概念进行分解，如图4.11所示：

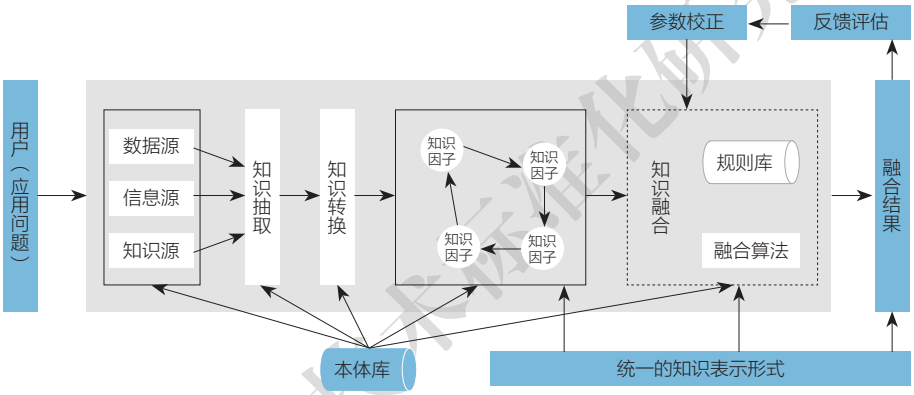

图4.11 知识融合概念分解

（三）研究现状

知识融合从融合层面划分可以分为数据层知识融合与概念层知识融合，数据层知识融合主要研究实体链接、实体消解，是面向知识图谱实例层的知识融合；概念层知识融合主要研究本体对齐、跨语言融合等技术。

1、数据层知识融合技术

实体链接问题是数据层知识融合研究的主要任务，其核心是构建多类型多模态上下文及知识的统一表示，并建模不同信息、不同证据之间的相互交互，主要的实体链接方法有：基于实体知识的链接方法、基于篇章主题的链接方法和融合实体知识与篇章主题的实体链接方法。

2、概念层知识融合技术

概念层知识融合是对多个知识库或者信息源在概念层进行模式对齐的过程。本体对齐或者本体匹配是概念层知识融合主要研究任务，是指确定本体概念之间映射关系的过程。本体匹配可以分为单语言本体匹配和跨语言本体匹配，单语言本体匹配是指同一自然语言中本体的对齐映射，跨语言本体匹配是指从两个或多个独立的语言本体中建立本体之间映射关系的过程[1]。本体匹配的研究核心就在于如何通过本体概念之间的相似性度量，发现异构本体间的匹配关系，本体匹配基本方法包括基于结构的方法、基于实例的方法、基于语言学的匹配算法、基于文本的匹配算法和基于已知本体实体联结的匹配算法。

3、跨语言知识融合技术

在大数据时代背景下，如何将跨语言的知识图谱进行对齐与融合，实现知识的全球共享，为跨语言知识服务提供便利，是知识图谱进一步研究的过程中需要解决的问题[2]。跨语言知识图谱研究的目的是构建一个包含当前重要知识库的大规模跨语言知识库，提高不同语言之间链接数据的国际化以及知识共享全球化，便于跨语言信息检索、机器翻译和跨语言知识问答等跨语言处理任务的研究与应用。[Hailong Jin，et al，2019][3]构建了一个有42万中英跨语言实体链接的双语言知识图谱（XLORE2），自动化融合了来自维基百科、百度百科和互动百科的信息。

4、现有知识融合工具

现有的知识融合工具包括：Falcon-AO、YAM++、Dedupe等。以Falcon-AO为例，其是由南京大学计算机软件新技术国家重点实验室开发

1 Bo F, Brennan R, O’Sullivan D. A configurable translation-based cross-lingual ontology mapping system to adjust mapping outcomes[J]. Social Science Electronic Publishing, 2012, 15(6):15-36.
2 费宇星. 跨语言知识图谱的对齐与融合研究[D].哈尔滨工业大学,2018.
3 Jin H, Li C, Zhang J, et al. XLORE2: Large-scale Cross-lingual Knowledge Graph Construction and Application[J]. Data Intelligence, 2019, 1(1): 77-98.

的一个基于Java的自动本体匹配系统，已经成为RDF(S)和OWL所表达的Web本体相匹配的一种实用和流行的选择。Falcon-AO系统采用了相似度组合策略，首先使用PMO进行分而治之，然后使用语言学算法(V-Doc、I-Sub)进行处理，然后使用结构学算法（GMO）接收前两者结果再做处理，最后连通前面两者的输出使用贪心算法进行选取。

（四）技术发展趋势

尽管知识融合已经在学术和工业应用中取得了非常显著的成效，然而随着网络社会数据特征、跨语言融合、知识规模增加等带来挑战越发紧迫，针对短文本及资源缺乏环境下的实体链接方法、融合先验知识的深度学习端到端实体链接方法、大规模本体的高效匹配方法将成为未来研究的重要趋势。

1、短文本及资源缺乏环境下的实体链接方法

传统的实体链接任务主要是针对长文档，长文档拥有在写的上下文信息能辅助实体的歧义消解并完成链接。而由于日常生活中人们在社交网络中常常会产生大量短文本数据，相比之下，短文本的实体链接存在口语化严重、短文本上下文语境不丰富等巨大挑战，因而面向短文本的实体链接方法研究将会成为未来的研究热点。另外目前绝大部分的实体链接模型依赖于有监督模型，需要大量标签数据集训练来达到实用目的。因此短文本及资源缺乏环境下，基于无监督/半监督和迁移学习的实体链接模型是解决问题的关键。

2、融合先验知识的端到端深度学习实体链接方法

今年来，基于深度学习模型（如BiLSTM-CRF）在实体链接任务上取得了较大的进展，同时展现出了巨大的应用潜力，然而基于深度学习的算法训练需要大量标注数据集，缺少面向特定领域特点和任务的针对性设计。另一方面当前实体链接方法易受到实体识别等前序过程的误差影响，

因此结合先验知识训练端到端深度学习实体链接模型成为未来的一大研究趋势。针对这个问题，一方面，当前许多算法尝试已经证明结合先验知识的思路在实体链接任务中的有效性，如在深度学习模型中增加句法结构、语言学知识、特定领域任务约束、现有知识库知识和特征结构等，如何更好的结合有效利用这些先验知识是提升实体链接算法性能的有效手段。同时设计基于端到端的深度学习模型将有助于降低实体链接过程中的误差传播效应，提高实体链接准确度。

3、大规模本体的高效匹配方法

随着当前各类型知识库的出现和知识规模的快速增长，而由于通常本体匹配的计算复杂度与本体规模成正比，因此大规模跨语言本体匹配成为知识库融合的重大挑战，主要面临的挑战有：大规模本体匹配的快速并行计算问题和人机协同匹配问题。针对这个问题主要的思路有：①研究基于分布式处理技术的大规模本体匹配分布式处理算法，如研究利用MapReduce、GPU等技术的并行匹配算法，提高匹配效率；②研究利用现有本体匹配结果实现潜在本体匹配的方法，同时利用启发式相似度计算方法提高计算效率；③通过对实体匹配进行预剪枝，预先过滤不匹配的实体对，避免本体之间一对一的相似度计算。

五、知识建模

（一）知识建模概述

知识建模是指建立知识图谱的数据模型，即采用什么样的方式来表达知识，构建一个本体模型对知识进行描述。在本体模型中需要构建本体的概念，属性以及概念之间的关系。知识建模的过程是知识图谱构建的基础，高质量的数据模型能避免许多不必要、重复性的知识获取工作，有效提高知识图谱构建的效率，降低领域数据融合的成本。不同领域的知识具

有不同的数据特点，可分别构建不同的本体模型。

知识建模一般有自顶向下和自底向上两种途径：1. 自顶向下的方法（如图4.12所示）是指在构建知识图谱时首先定义数据模式即本体，一般通过领域专家人工编制。从最顶层的概念开始定义，然后逐步细化，形成结构良好的分类层次结构。2. 自底向上的方法则相反（如图4.13所示），首先对现有实体进行归纳组织，形成底层的概念，再逐步往上抽象形成上层的概念。自底向上的方法则多用于开放域知识图普的本体构建，因为开放的世界太过复杂，用自顶向下的方法无法考虑周全，且随着世界变化，对应的概念还在增长，自底向上的方法则可满足概念不断增长的需要。

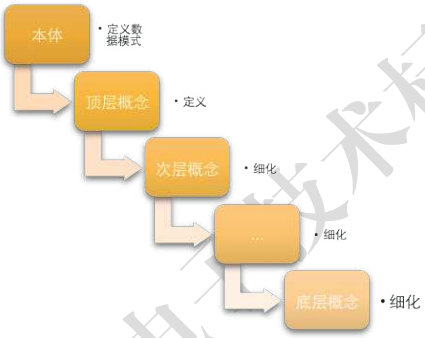

图4.12 自顶向下的构建方法

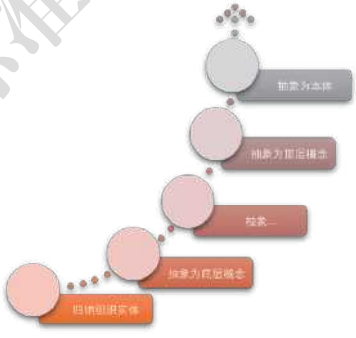

图4.13 自底向上的构建方法

（二）知识建模方法

知识建模目前的实际操作过程，可分为手工建模方式和半自动建模方式[1]。手工建模方式适用于对知识建模容量小、质量高的要求，但是无法满足大规模的知构建，是一个耗时、昂贵、需要专业知识的任务；混合方式将自然语言处理与手工方式结合，适于规模大且语义复杂的图谱。

1、手工建模方式

手工建模方式过程主要可以分为以下的六个步骤：明确领域本体及任

[1] 胡兆芹.本体与知识组织[M].中国文史出版社:北京,2014:63.

务、模型复用、列出本体涉及领域中的元素、明确分类体系、定义属性及关系、定义约束条件。在人工建模的过程中，以上的六个步骤并不是一一顺序执行的，可以根据知识建模的具体需求，组合其中的步骤达到知识建模的目的。下面分别对这些步骤作详细的介绍，如图4.14所示。

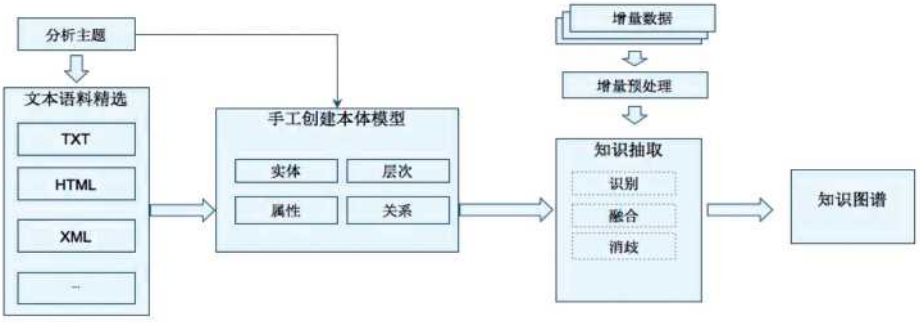

图4.14 手工建模方式

2、半自动建模方式

半自动建模方式先通过自动方式获取知识图谱，然后进行大量的人工干预过程，如图4.15所示。运用自然语言处理技术先自动建模的方法可以分为三大类[1]：基于结构化数据的知识建模方法，基于半结构化数据的知识建模方法和基于非结构化数据的知识建模方法。近年来，对于非结构化数据的知识建模方法研究较多，涌现出一批优秀的基于非结构化数据的知识建模方法的高水平研究成果。

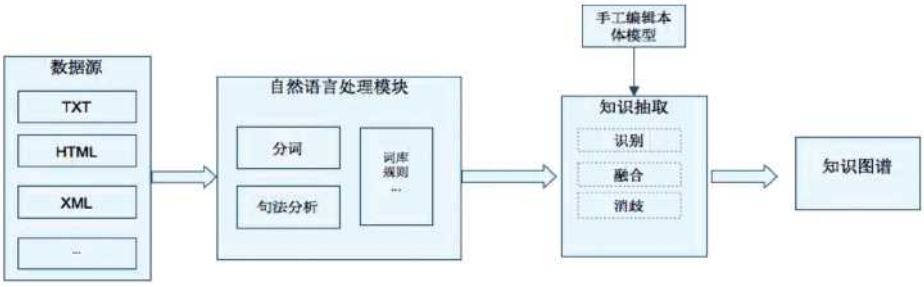

图4.15 半自动建模方式

[1] Al-Arfaj A, Al-Salman A. Ontology construction from text: challenges and trends[J]. International Journal of Artificial Intelligence and Expert Systems (IJAE), 2015, 6(2): 15-26.

（三）知识建模评价

对知识建模质量评价也是知识建模的重要组成部分，通常与实体对齐任务一起进行的。质量评价的作用在于可以对知识模型的可信度进行量化，通过舍弃置信度较低的知识来保障知识库的质量。一个合理的本体模型宜满足以下标准：

- 明确性和客观性：用自然语言对所定义术语给出明确的、客观的语义定义。
- 完全性：定义是完整的，完全能表达所描述领域内术语的含义。
- 一致性：正确一致地展示数据、对象和信息，由术语得出的推论与术语本身含义不会产生矛盾。
- 最大单调可扩展性：添加通用或专用的术语时，不需要修改已有的内容，便于知识图谱扩展。
- 最小承诺：尽可能少的约束，指本体约定应该最小，对建模对象尽可能少的约束。
- 易用性：有效地支撑业务的分析和决策需求。

（四）技术发展趋势

知识建模核心解决了采用什么样的形式高效组织和表达知识的问题，偏向于知识建模的方法论，在未来的发展趋势中，将会解决知识建模的规范化和标准化。同时随着大数据时代的到来，知识建模将会朝着对大规模数据的进行建模的方向发展，届时多人在线编辑，并且实时更新知识建模将成为可能。针对传统人工知识建模耗时、耗力、效率低下等弊端，知识建模可与自动语义处理算法进行结合，实现全自动建模方式，避免人工干预和操作；另外，快速集成现有的结构化知识模型，支撑起事件、时序等复杂知识形式的表达模式，建立功能更加完善、表达更加强大的知识模型。

六、知识计算

（一）知识计算概述

随着知识图谱技术及应用的不断发展，图谱质量和知识完备性成为影响知识图谱应用的两大重要难题，以图谱质量提升、潜在关系挖掘与补全、知识统计与知识推理作为主要研究内容的知识计算成为知识图谱应用的重要研究方向。知识计算是基于已构建的知识图谱进行能力输出的过程，是知识图谱能力输出的主要方式。知识计算概念内涵如图4.16所示，主要包括知识统计与图挖掘、知识推理两大部分内容，知识统计与图挖掘重点研究的是知识查询、指标统计和图挖掘；知识推理重点研究的是基于图谱的逻辑推理算法，主要包括基于符号的推理和基于统计的推理。

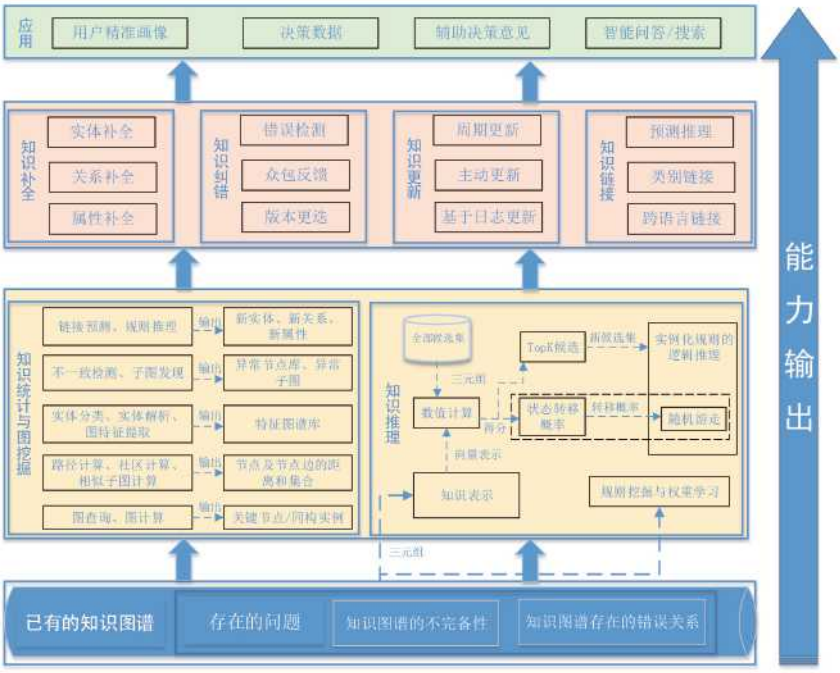

图4.16 知识计算概念

知识计算的概念中明确了以下几个层面的问题：

（1）知识计算是针对已构建的知识图谱所存在的问题：不完备性和存在错误信息，在此基础上通过将知识统计与图挖掘、知识推理等方法与传统应用相结合进行能力输出，为传统应用形态进行赋能，进而提高知识的完备性和扩大知识的覆盖面。

（2）知识计算中两种具有代表性的能力：知识统计与图挖掘、知识推理。知识统计和图挖掘的方法是基于图特征的算法来进行社区计算、相似子图计算、链接预测、不一致检测等；知识推理的目标在于从给定知识图谱中推导出新的实体、关系和属性。通过这两种能力实现对已有图谱的知识补全、知识纠错、知识更新、知识链接等功能。在此基础上，知识计算的能力输出可应用于用户精准画像、决策数据、辅助决策意见、智能问答/搜索等方面。

（二）研究现状

知识图谱的应用大多基于对复杂网络的大规模计算，计算的结果或以在线服务，或以离线结果的形式提供给应用者。知识计算的能力输出方法包括：知识统计与图挖掘、知识推理。

1、知识统计与图挖掘

知识统计与图挖掘是指基于图论的相关算法，实现对知识图谱的基础性查询、统计分析和图挖掘计算，主要包括：图查询检索、图特征统计、关联分析、时序分析、节点分类、异常检测、预测推理等。

（1）图查询检索

图查询和检索是最常见的计算，常用于查询目标节点的n度关联方、或者查询某子图结构，主要是以深度优先或广度优先等方式遍历网络，输出关联节点或同构实例。图4.17为图信息检索示例。

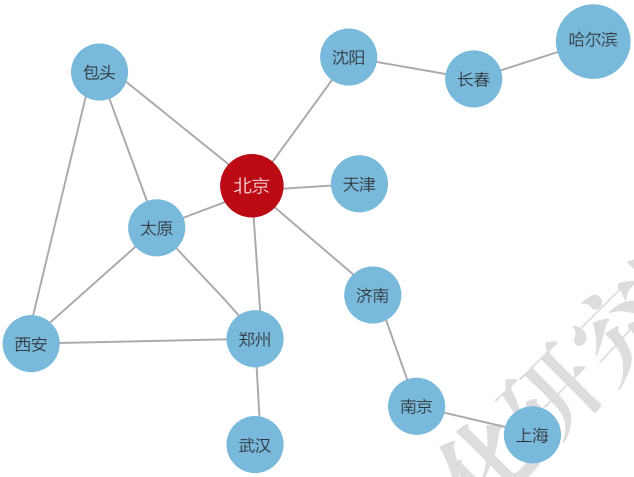

图4.17 图信息检索示例

（2）图特征统计

图特征统计是指对图谱中单一节点、或多个节点的图特征以及属性特征进行统计计算的过程，如图4.18所示。其中，单主体图特征包括出度、入度、介度和中心度，出度表征某节点发出的边的多少，定义为统计节点发出的关系总条数；入度表征某节点接收到的边的多少，定义为统计指向该节点的关系总条数；介度表征某节点桥接作用的重要性；中心度表征节点在当前子网中的重要性。

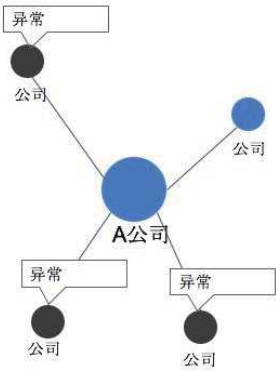

图4.18 图特征统计示例

（3）关联分析

关联分析是指分析图谱中两个节点间或多个节点间的关联关系、紧密程度，进而可以实现社群发现和分割，例如两个公司间的多度投资关系、个人与某公司的最短路径、两人之间的加权距离，多个账户之间的转账关系等等。常用的方法有路径查询、距离计算，输出结果为节点及节点间边的距离和边的集合（路径）。图4.19为关系分析计算示例。

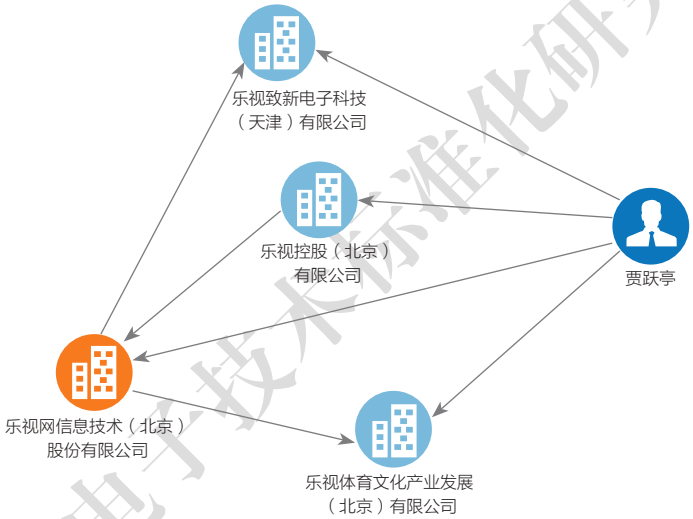

图4.19 关联分析计算示例

（4）节点分类

节点分类是指对节点根据图特征或者关联属性特征进行分类、例如信用违约公司具有典型的风险路径，且一度关联方中的违约公司数量等均可作为分类特征指标、洗钱账户的转账关系子图特征可作为洗钱标签的分类依据。常用方法为标注目标节点，图特征提取，分类算法等，输出结果为特征图谱库等。

（5）异常检测

异常检测是指在全网内发现异常节点、异常子图子图模式，例如出入

度数值离群的节点、闭环的投资关系、未知业务含义但是罕见的频繁子图等。常见方法是聚类、子图发现算法等。输出结果异常节点库、异常子图结构模式库为主。图4.20为异常检测示例。

模体	扰动 10 万次	扰动 20 万次	扰动 30 万次	10E	扰动以前的数量
	2740605	2748075	2754519	2832558	2732439
	2388862	2407153	2414451	2497608	2389916
	1474789233	1474795510	1474800814	1474865345	1474782466
	1062	878	747	0	1291
	93386	85916	79472	1432	101552

图4.20 异常检测示例

（6）预测推理

预测推理是指通过规则或者机器学习等方法，从已有知识图谱中预测推理新的关系和信息，适用于弱关系的推理、链接预测、概率推理等。常见方法有规则推理、机器学习等。输出结果为新节点、新关系、新属性等信息。

（7）时序分析

时序分析是指对单一关系、事件做时序分析，或者对网络拓扑结构的变化做时序分析，例如频繁工商变更的行为、风险在网络中的传播等。常见方法有时序分析、风险传播模型等。输出结果为时序异常、风险评分等。

2、知识推理

知识推理可以定义为按照某种策略，根据已有知识推出新知识的过

程。由于人类认知世界的过程正是一个知识推理的过程，因而如何让机器学会这种方式来理解知识显得尤为重要，一个具备知识推理能力的知识图谱，将挖掘出数据更深层次的内在价值。知识推理可以分为基于符号推理和基于统计的推理。在知识理解的基础上构建应用，知识图谱的应用大多基于对复杂网络的大规模计算，计算的结果，或以在线服务，或以离线结果的形式提供给应用侧。从目标节点的检索、到关联方查询、到探寻节点间的关联关系、统计整图及单一节点指标，结合图算法模型和时间信息，挖掘深层次的潜在信息，提供分析和决策支持，为数据赋能。基于知识推理的典型应用主要包括智能搜索、智能推荐、智能问答等。

（1）智能搜索

传统的搜索引擎，以关键词检索为核心技术，通过关键词比对，从海量网页信息中找出与关键词相匹配的信息并返回，用户可以输入一条关键词或者几条关键词进行查询。然而这种查询方式得到的结果无关信息较多，且查询方式较为机械，并不是按照人的思维方式查询并给出结果。而引入知识图谱技术后的智能搜索引擎，使得上述问题得到了更加智能的解决，现在主流的搜索引擎，比如谷歌，百度等，均采用了知识图谱作为底层技术。

（2）智能推荐

商家对于客户的商品推销活动，从古至今就没有停止过，传统的推广往往靠人海战术，商家会发布铺天盖地的广告来吸引客户，这样的方式有一定的好处，但投入产出不高。为此商家开始提出精准营销的概念，互联网电子商务的兴起，让精准营销变得越来越智能，通过大数据智能算法，根据用户行为，勾勒出用户画像，再基于用户画像，为用户提供更多的精准营销服务。然而这种精准营销服务，是基于已知数据做的，而实际情况往往获取不到那么多的数据来进行分析，因此人们需要借助“推理”的模式来“猜测”用户的喜好，这就好比人们了解一个事务一样，当无法了解

到事务的全貌时，人们往往会结合已有的线索，做推理假设，这也是人区别于其它动物的特性，如图4.21所示。

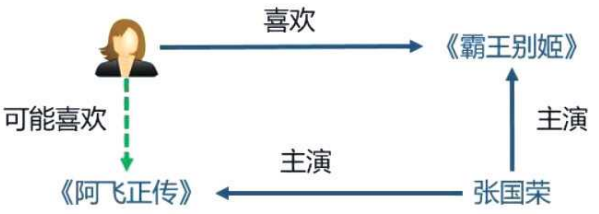

图4.21 知识推理示例

（3）智能问答

传统的问答系统，简单来讲多用配置问答对的方式来实现各种问答场景，举一个常见的例子：问：“姚明的女儿的母亲的身高是多少？”这个问句中的多个词之间带有明显的关联关系，需要通过多次关联查询才能得到最终结果，如果答案存储采用传统方式，比如存储在关系型数据库中，上述查询语句将相对复杂。这个例子也相对简单，设想一下，如果面对的数据量是海量的，数据之间存在更加复杂的多跳关系，传统的问答查询机制将变得很困难。

（三）技术发展趋势

如何解决小样本量场景的知识计算、一致性动态变化下的知识计算以及面向多元关系和多源信息等方面的知识计算将成为未来重要且亟待突破的方向。

1、基于小样本学习的知识计算

现阶段的知识计算模型往往依赖于大规模高质量的学习样本，但在实际应用中，大规模高质量样本的获取会耗费巨大的人力和时间，并成为限制知识计算模型训练及应用的重要挑战。研究表明，融合已有先验知识和

极少量的样本进行模型训练，能够很好的解决小样本量场景下的知识计算模型训练所面临的问题，基于开发世界假设，在进行知识图谱的逻辑推理和计算时，对未发现、不存在的、新产生的知识看作是未知的，将已经建立的知识图谱看作先验知识，通过少量高质量样本进行快速学习计算和整合。该研究方向将成为知识计算研究的重要趋势之一。

2、面向一致性动态变化的知识计算

知识的一致性维护是知识计算需要解决的重要问题，但现有的知识计算方法主要针对的是静态知识图谱，然而知识图谱并不是一成不变的，随着时间和空间的推移，知识图谱也在动态的变化，即在不同的时间轴下相同问题的答案也可能存在不同，如针对“某人现阶段的年龄是多少？”这一问题，在不同的年份所得到的答案是不一样的。因此，如何在保证一致性条件下，引入时间要素进行有效地动态计算是未来知识计算发展的重要趋势。现阶段针对动态知识图谱已有一些学者进行了研究，如基于新提出的演化网络的推理研究、基于时间要素的研究等，Tay等人提出了puTransE模型通过分治策略实现了图谱的分块学习和集成推理，有效处理了动态图谱的增删改。

3、面向多元关系和多源信息的知识计算

现有的知识计算大都是集中在二元和三元关系上，但多元关系相对二元和三元关系来说结构更多样、上下文语义关系更复杂，所涉及到的常识也更多样，针对复杂多元关系的常用处理方法是将复杂的多元关系简化为二元关系或三元关系，但这个过程中大量的语义信息和上下文关系会被损失，面向多元关系的知识计算准确度不高，另外，信息的多源性使得知识计算过程中需要考虑更多上下文问题和额外的常识等信息，来降低知识图谱稀疏性和不连通性，从而进行高效的知识计算，使知识计算的结果更加趋向于完备，因此，如何更好的解决多元关系和多源信息融合场景下的知识计算问题将成为重要的研究方向。

七、知识运维

（一）知识运维概述

由于构建全量的行业知识图谱成本很高，在真实的场景落地过程中，一般遵循小步快走、快速迭代的原则进行知识图谱的构建和逐步演化。知识运维是指在知识图谱初次构建完成之后，根据用户的使用反馈、不断出现的同类型知识以及增加的新的知识来源进行全量行业知识图谱的演化和完善的过程，运维过程中需要保证知识图谱的质量可控及逐步的丰富衍化。知识图谱的运维过程是个工程化的体系，覆盖了知识图谱的从知识获取至知识计算等的整个生命周期。

知识图谱的运维包括两个方面的关注点：一个是从数据源方面的基于增量数据的知识图谱的构建过程监控，另一个是通过知识图谱的应用层发现的知识错误和新的业务需求：例如错误的实体属性值、缺失的实体间关系、未识别的实体、重复实体等问题。这些运维暴露的问题会在知识图谱构建的流程、算法组合、算法调整、可新增业务知识优先级排列等方面进行修正，提升知识质量和丰富知识的内容。知识图谱运维需要基于用户反馈和专家人工的问题发现及修正、自动的运行监控、算法调整后的更新相结合，因此是一个人机协同，专家和算法相互配合的一个过程。知识运维的全过程如图4.22所示：

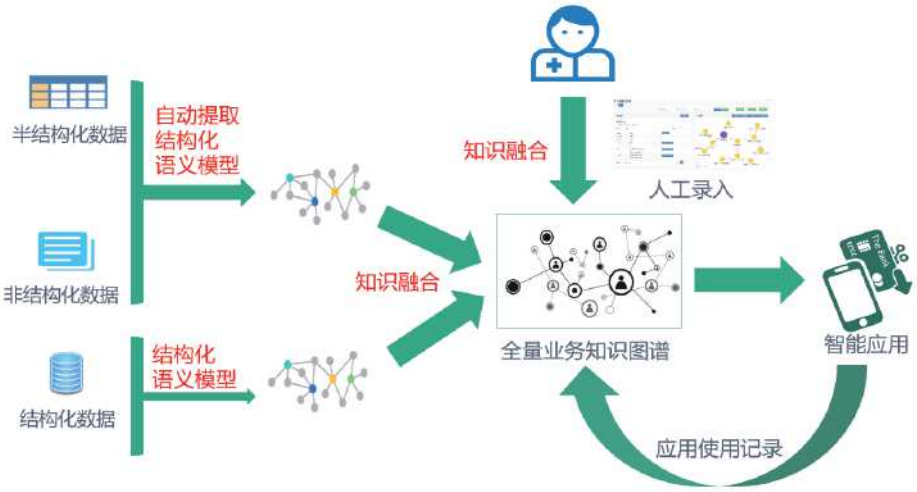

图4.22 知识运维

（二）研究现状

1、基于增量数据的知识运维

构建知识图谱是一个持续和增量的过程，随着数据的不断更新（爬虫数据不断积累、业务数据持续更新等），如何持续地对图谱进行更新成为了一个重要的问题。普通的知识图谱增量更新包括新元素的加入（节点、边或对应的属性）、旧元素属性的更改。在更复杂的场景下可能会涉及到已有元素的删除操作。工程上高效、自动的增量更新策略对于维护一个动态更新、准确性高的知识图谱意义重大。根据不同的使用场景和不同的数据来源，主要存在以下两种增量方式: 数据从消息队列导入图谱（如图4.23所示）、利用工作流引擎定时更新图谱（如图4.24所示）。

图4.23 数据从消息队列导入图谱

图4.24 利用工作流引擎定时更新图谱

2、图谱内容统计监控

如果知识运维人员要高效的进行知识图谱的运维工作，需要对知识图谱中的实体、本体、属性、关系进行统计，掌握目前的知识图谱的规模和状况，也可以对图谱中的知识进行上传或者下载操作，方便进行图谱内知识的管理。同时系统需具备对图谱运行中间产生的各种异常情况进行集中的展示、问题提醒等功能，报告知识图谱中出现的问题，方便运维人员及时进行修正。

3、知识审核与修正

按照业务的正确率的要求程度、数据的量级等角度考虑，需要对知识图谱有明确的新增知识入库的标准和流程。对于准确率要求高的知识图谱支撑应用，对新识别的实体、变更的实体属性、实体或关系冲突等，需要通过明确的列表的方式呈现并由有相关知识背景的专家来进行审核确认后方能入库，审核入库过程要有记录。对已经构建好的知识图谱需要有可以直接增、删、改的途径。此外由于知识图谱中非事实型的行业知识往往具有模糊性，在构建和运维图谱的时候需要有套冲突检测以及多人协同编辑的功能，如果系统自动检测到冲突点或者不同的知识运维人员运维同一知识点产生认知的不一致，那么需要系统提供多人协同工作讨论确定对知识的统一认知和编辑加入知识图谱。

4、知识版本管理

在知识图谱的管理中，可以引入版本概念，按照知识的更迭进行管

理，可以设置当前对外服务的知识版本，可以对历史的知识版本进行作废或者回滚处理。基于版本的知识图谱运维可以实现知识图谱的升级切换，方便线上应用业务的平滑升级，也可以在新版本知识图谱上线出现问题情况下快速切换回原有版本，降低对业务的冲击，同时避免误操作后的知识丢失。

5、知识安全管理

知识图谱在构建的过程中往往倾向于将各种不同来源的数据进行融合构建成为一个完整的知识体系，这样的好处是打破数据壁垒造成的知识缺失。融合的知识对于决策与分析价值更大，但是也减低了原始数据源中的数据访问权限的控制，带来了数据的安全风险。因此对于不同部门或者层级的人员可见或者使用的知识范围要有明确的限定，对知识的上层应用要控制开放的知识范围，降低因为知识融合产生的知识泄密风险。针对此挑战，需要引入权限管理，对维护、使用知识的人员、系统进行账号分配、权限分配，对于人员权限可能需要对接组织已经建设的统一登录和单点登录系统，将知识图谱管理和使用的权限和人员在组织内的角色有机结合，降低人员变动后的数据安全风险。在整个知识图谱的运行过程中，要有日志监控、操作记录、变更内容的记录等，便于追踪异常，堵住漏洞。

6、知识容灾备份

一个知识图谱可能含有上亿个节点以及上百亿的边，单台机器很明显无法存储和处理如此海量的数据，保证分布式图谱服务在某个或者某些节点失效时还能稳定可用就是知识图谱高可用的定义。一个完善的知识图谱通常拥有重大的价值，高可用只保证了服务阶段的可用性，保证在意外发生的时候图谱数据不至于完全丢失是图谱灾备需要解决的重要问题，如图4.25所示。

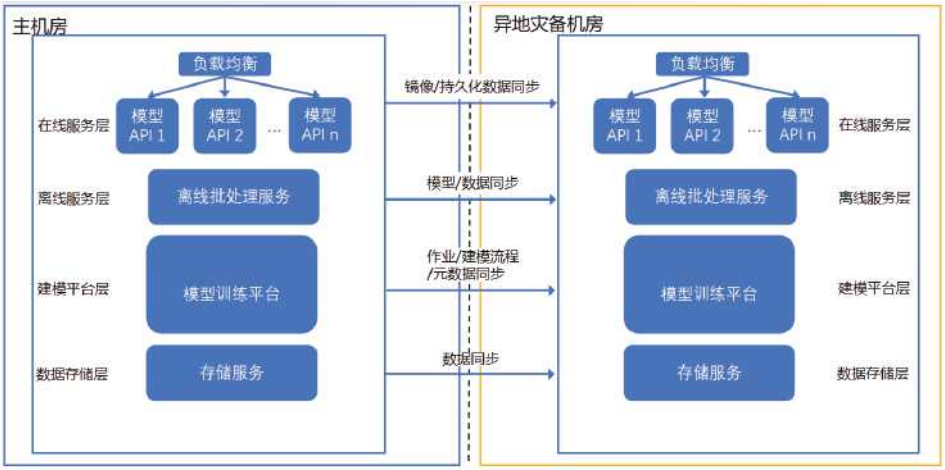

图4.25 知识容灾备份

（三）技术发展趋势

知识运维是知识图谱可以持续健康发展并且支持业务的关键，但是运维工作往往技术难度大，涉及步骤多，所有知识构建及使用人员都是参与者，各部分人员的技能有一定差异，沟通成本比较高，因此需要将知识运维纳入到整个知识图谱的构建体系中来看待，为了降低知识图谱构建的难度并提高运维的可控性，便于知识高效运维，出现了知识图谱平台化的发展趋势。通过建设知识图谱平台的方式将知识图谱构建过程中涉及的所有步骤、技术进行整合，实现知识图谱构建过程的全生命周期管理，加快知识图谱的构建速度，降低知识运维难度，同时对知识图谱的质量和扩展可控可管理。

1、知识图谱的全生命周期质量保障

具备知识运维能力的知识图谱平台主要功能宜包括：本体的构建，针对多种数据来源的结构化、半结构化、非结构化的数据类型在不同的技术下的知识获取，实体识别、关系识别、实体链接、实体属性抽取的实现，基于本体概念和实体知识图谱间的验证，构建流程与运维过程的监控，对

知识图谱构建过程中的各种异常情况的记录和反馈，对入库知识图谱的人工审核。此外，通过在知识图谱平台的知识库以版本的形式进行管理，避免知识运维中因为新知识的错误发布对现有业务的影响，提供给运维人员上线发布前的质量检测方法，并将经过严格测试验证的知识图谱版本正式生效上线，最终保证知识图谱全生命周期各环节的数据质量。

2、多知识图谱的运维管控

此外，面向按照不同领域和范围下多个知识图谱的构建和运维，有待开发一套完备的平台对多个不同知识应用提供支撑。该平台本身需具备完整的安全管控及权限管理，并可满足动态本体的多人协同构建、冲突检测及讨论确定统一的版本的机制及功能，最终可对外开放给上层应用，提高应用的智能化。同时，通过应用的使用记录及问题反馈带动知识图谱的运维优化，形成闭环全周期的多知识图谱间的运维管控。

第五章 知识图谱应用

一、智慧金融

（一）背景介绍

智慧金融作为一个有机整体，知识图谱提供了金融领域知识提取、融合、分析、推断、决策等功能[1]，如图5.1所示。在场景方面，智慧金融涵盖智慧支付、智慧财富管理、智慧银行、智慧证券、智慧保险、智慧风控等诸多方面。在应用功能方面，从KYC、舆情分析、个人/企业信用分析、风险传导、营销推荐、智能问答、知识库等都是典型的知识图谱应用。金融知识图谱构建过程如图5.2所示，包括数据抽取、信息提取、知识融合、知识加工等。

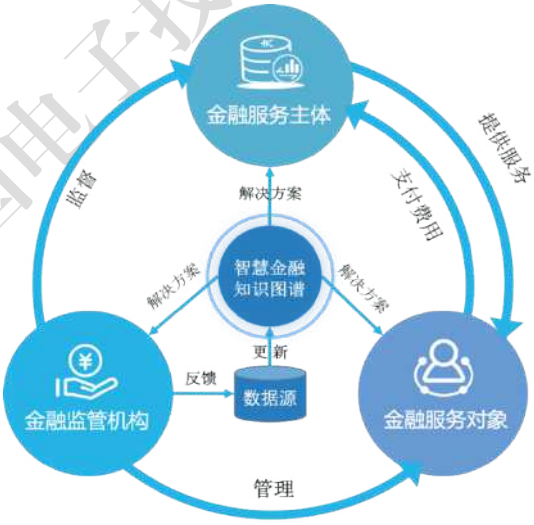

图5.1 智慧金融全景图

[1] 鲍捷. 知识图谱如何助力实现智能金融[J]. 金卡工程, 2016 (7): 45–49.

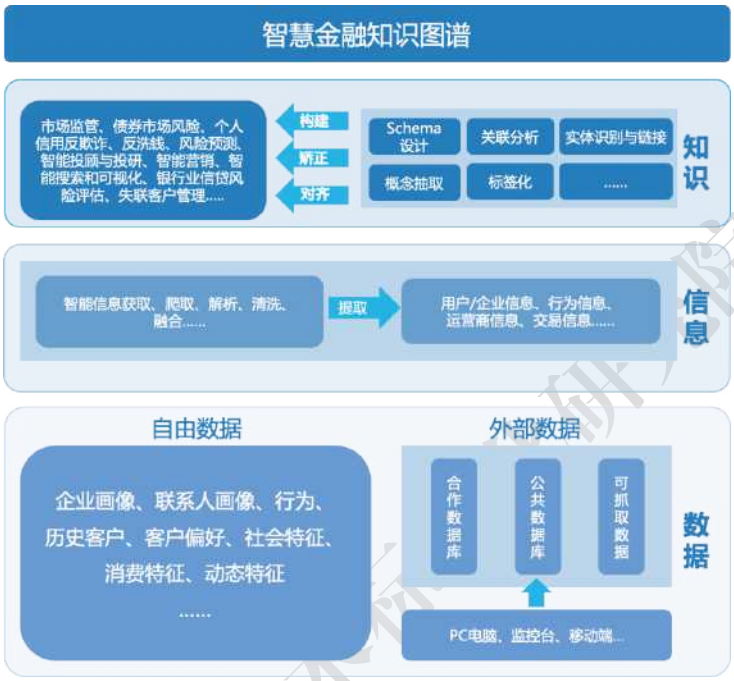

图5.2 智慧金融知识图谱构建

（二）典型应用

知识图谱在智慧金融中的应用可分为金融监管、金融机构应用和金融服务。金融监管是国家金融监管机构金融市场及相关机构与个人的监督管理，金融机构应用是指金融参与者利用知识图谱技术实现的风险预测、智能营销等应用，金融服务是指金融机构面向企业或公众提供的智能化金融服务，如图5.3所示。

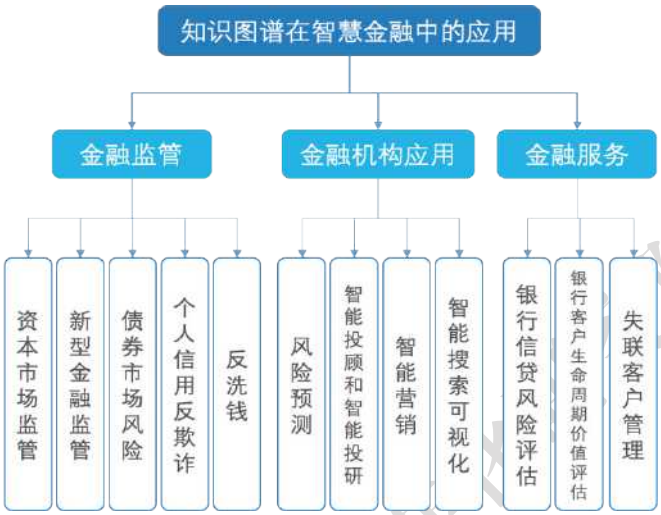

图5.3 知识图谱在智慧金融中的应用

1、金融监管

知识图谱在金融监管领域的应用包括资本市场监管、新型金融监管、债券市场风险、个人信用反欺诈、反洗钱。

（1）资本市场监管

基于知识图谱的资本市场监管从企业关系分析出发，探索企业及其关联方在资本市场的行为表现，结合舆情事件的传递效应，构筑资本市场中知识图谱，全方面识别企业行为风险，实现资本市场的风险监管与预警。

（2）新型金融监管

新型金融智能监管系统紧扣新金融行业特点，运用知识图谱技术构建新型金融企业的实体风险画像，通过对全国所有企业信息的大数据挖掘分析，识别出新金融业态的企业，根据新金融企业的行业分类、风险特征、数据维度构建分行业的不同风险类型的特征风险模型，并按照风险指数等级进行分级管理。

（3）债券市场风险

基于知识图谱的债券市场风险监管通过构建包括债券发行人的产业

链上下游关系、投融资关系、债券发行人的信用状况、债券发行人日常经营状况、投融资关系等信息在内容的知识图谱，实现提前判断企业经营效益，推理挖掘隐含的关联方资金占用倾向、洗钱骗税倾向等问题对可能的违规行为进行提前预警，从而实现对信用评级的及时调整。

（4）个人信用反欺诈

如图5.4所示，通过构建已知的主要欺诈要素（包括手机号码、账号和密码、地理位置等）的关系图谱，全方位了解借款人风险数据的统计分析，基于知识图谱挖掘疑似欺诈用户，并对疑似欺诈用户进行规则判定、图谱验证、欺诈判定等过程，对潜在的欺诈行为做出及时而迅速的反应。

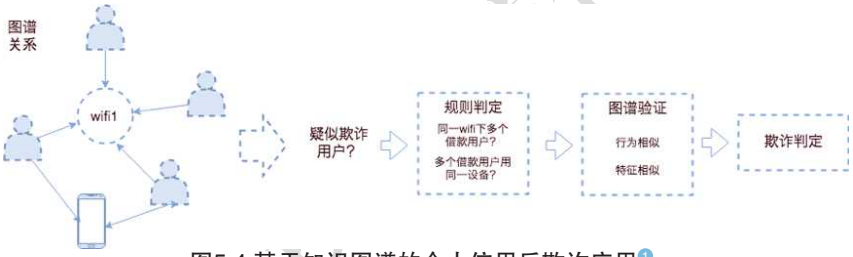

图5.4 基于知识图谱的个人信用反欺诈应用[1]

（5）反洗钱

基于知识图谱的反洗钱系统充分运用大数据技术中分析和图挖掘技术，基于客户标签、画像开展客户立体化识别，并结合互联网大数据、第三方场景的数据等进行图层构建，对企业的关联网络特征进行图编码，并基于图层数据搭建目标企业关联概率网络，将图层进行叠加得到知识图谱，对企业关联结构进行深度解析，实现对隐性风险结构及关联主体的深度挖掘，以全息式多维度实时监控企业洗钱风险。

2、金融机构应用

知识图谱在金融应用领域包括风险预测、智能投顾与智能投研、智能

[1] https://tech.meituan.com/2018/11/22/meituan-brain-nlp-01.html

营销、智能搜索和可视化。

（1）风险预测

风险预测基于多维度的数据建立客户、企业和行业间的知识图谱，从行业关联的角度预测行业或企业在未来可能面临的风险。风险预测包括两部分内容，其一是对潜在风险行业预测，其二是对潜在风险客户预测。在潜在风险行业预测方面，在基于多维度数据的基础上，对行业进行细分，依托行业信息、贷款信息等数据建立起的行业之间的知识图谱，可以发现不同行业间的关联程度。

（2）智能投顾与智能投研

智能投顾是指根据投资者不同的理财需求，通过算法和产品搭建数据模型，实现传统上由人工提供的理财顾问服务。智能投顾可分为机器导向、人机结合以及以人为主三种模式，且人机结合将是未来投顾发展趋势。智能投研指利用大数据和机器学习，将数据、信息、决策进行智能整合，并实现数据之间的智能化关联，从而提高投资者工作效率和投资能力。智能投顾是近年证券公司应用大数据技术匹配客户多样化需求的新尝试之一，目前已经成为财富管理新蓝海。

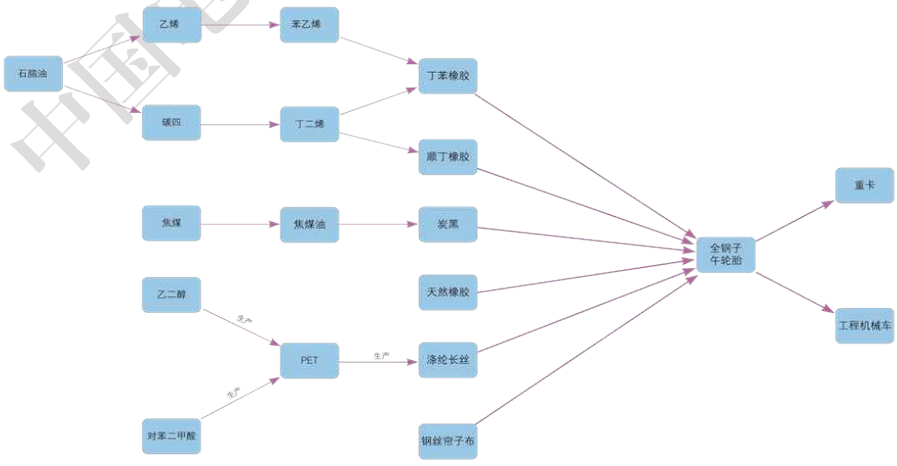

图5.5 基于知识图谱的产业链分析

（3）智能营销

知识图谱能够整合更丰富、更全面的用户信息，根据精准营销的不同角度设定不同类别的场景标签，通过知识图谱技术提供的分类标准，进行用户的标签化分类工作，建立合理的客户类型初分体系。同时结合黑白名单技术，对客户进行判断，对客户质量进行筛选与把控，并最终实现互联网金融产品推荐、客户准入、客户跟踪管理等高级营销策略。。

图5.6 营销流程图

（4）智能搜索和可视化

利用知识图谱在信息检索与可视化方面的优点，实现用户信息、借贷信息、失信信息、网络行为信息的快速精准搜索，能够将复杂的关系信息进行直观明了的可视化处理，从而使得用户对潜在的关系信息了解得更为透彻。

3、金融服务

知识图谱在金融服务领域的应用包括银行业信贷风险评估、银行客户全生命周期价值评估、失联客户管理。

（1）银行业信贷风险评估

基于知识图谱的银行业信贷风险评估能够整合内外部数据资源，根据

行业信息建立关系挖掘模型，展示每个行业的风险状况与关联程度，能够全面准确评估信贷风险，如果某一行业发生了行业风险或高风险事件，可以及时预测未来有潜在风险的关联行业，信贷评估可对相关行业的风险做出预判，尽早地发现并规避风险。

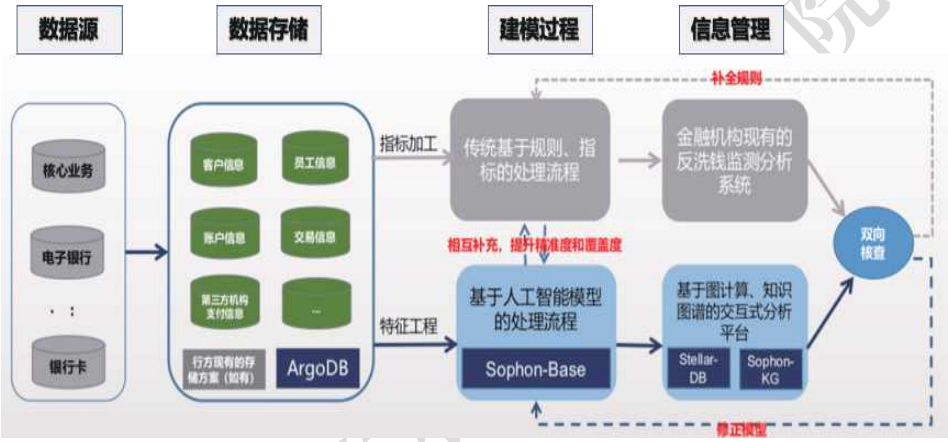

图5.7 风控流程图

（2）银行客户全生命周期价值评估

运用知识图谱技术对银行公司客户进行全生命周期价值评估，能够对企业生命周期中任何阶段的任何一家企业做多维度的全面的价值评估，为银行挖掘潜力客户和进行全量客户关系管理提供支持。整合行内贡献价值、企业价值、关联系价值三个维度的数据，构建多维度企业信息图谱，通过企业的行内价值表现、企业经营情况、资本人才实力和企业关联方规模、结构、质量来评估企业行内价值表现。

（3）失联客户管理

失联客户管理是指知识图谱能够在贷后失联客户管理层面，通过关系图挖掘更多新的潜在联系人来提高催收成功率，减少贷款损失。现实中，很多情况下银行联系不上借款本人，无法进行催收管理，银行一般会联系该借款人曾经提供的其他联系人，但如果银行通过这层关系仍旧无法联系

上该借款人，那么这些借款人就处于“失联”状态。而知识图谱可以帮助银行在借款人出现失联的情况下，去挖掘跟借款人有关系的新的联系人，可以大大提高催收成功率。

二、智慧医疗

（一）背景介绍

智慧医疗是利用先进的物联网与移动通信技术、大数据及人工智能等新一代IT技术，实现医疗信息系统与医疗过程的智能化辅助与自动化处理，实现医疗业务流程的数字化运作，实现患者与医务人员、医疗机构、医疗设备之间的互动。短期来看，在医疗过程辅助、患者服务、医学科研以及临床医学教育等方面，知识驱动的智能化服务，能够辅助医院提供更优质的医疗服务，实现更好的医院管理；辅助患者的整个诊疗与健康管理过程；辅助医生开展知识和数据驱动的前沿医学研究；辅助医学生、低年资医生和基层医生开展临床思维训练与实践演练。知识图谱则是只是驱动的智慧医疗服务与应用的关键基石，如图5.8所示。

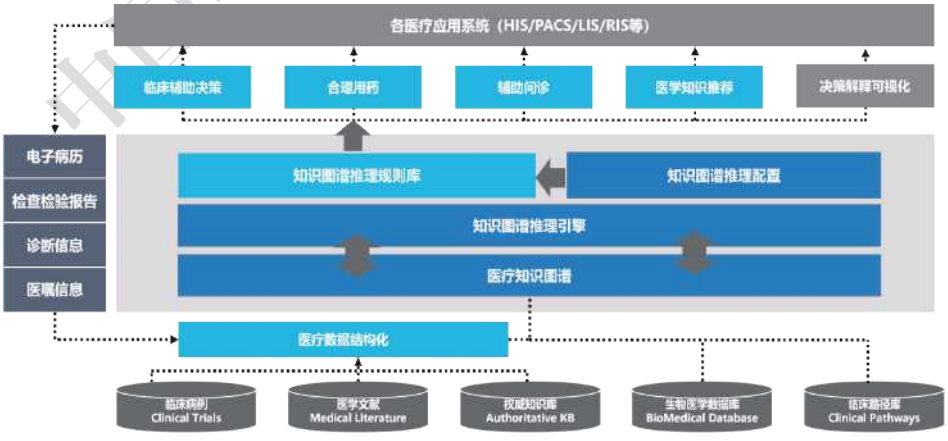

图5.8 基于知识图谱的医疗应用系统

（二）典型应用

如下图所示，围绕医疗核心知识库，智慧医疗应用包括医疗过程智能辅助、患者智慧服务、医学科研、医学教学。医疗过程智能辅助实现基于医疗知识库的临床决策辅助、合理用药、临床数据洞察和病历质检等；患者智慧服务能够根据患者历史就医及医疗知识，为患者提供健康知识推送、健康评估、智能分诊等智能应用服务；医学科研基于医学基础知识库，辅助医疗科研工作者实现疾病风险预测、药物研发及医学证据挖掘与综合分析；智能医学教学针对医学生、低年资医生存在的经验不足、缺乏实战经验等问题提供虚拟问诊、病案学习、场景化实训等智能服务。

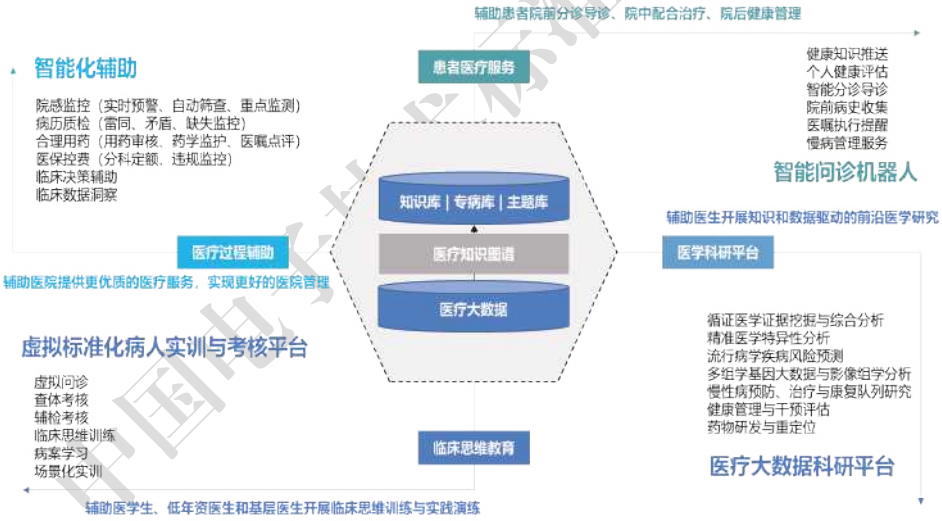

图5.9 医疗知识图谱支撑智慧医疗应用

1、医疗过程智能辅助

在医疗的实施过程中，通过智能化信息收集、信息分析以及辅助决策给与医护人员知识推荐、实时提醒与决策参考。早期的医疗过程辅助，更多的是基于知识库的查询和提醒，以及基于规则的判断和决策，基本用

于解决临床医生在遇到不熟悉的临床问题时进行知识检索的碎片化场景。“人工智能+辅助诊疗”是指让计算机认知和学习医疗知识，模拟医生的思维和诊断推理，按照规范的管理流程，给出可靠的医疗过程辅助方案。医疗过程辅助是医疗领域智能化最重要的场景，知识图谱支撑的医疗过程智能辅助典型应用有：

（1）安全合理用药辅助

系统化的药学专业知识关系十分复杂，涉及药物与不同年龄人群、不同疾病人群的适应或禁忌关系、药物与不同病因导致的疾病的治疗关系、药物与药物副作用、药物之间的相互作用、注射类药物之间的配伍禁忌等多方面知识。应用知识图谱技术可以针对药事服务场景建立大规模、跨语言、高覆盖、高准确性的药学知识采集系统，完成多元异构结构非结构药学文档数据到统一知识表达，形成药物使用指导手册，辅助医生安全合理用药。

（2）临床辅助决策

临床决策支持系统是一个基于人机交互的医疗信息技术应用系统，旨在为医生和其他卫生从业人员提供临床决策支持，通过数据、模型等辅助完成临床决策。基于知识图谱的临床辅助决策，从临床数据、医学文献和医学经验等数据中获取知识、更新知识，支持基于细粒度特征的推理查询，具有推理结果可解释性强，知识体系兼容性好等突出特点。

（3）安全用药

国家食药监总局的一项统计数据显示，我国每年大约有250万人因为错服药损害健康，而导致死亡的就有20万，是全国交通事故致死人数的2倍，所以，医院中临床药师的重要性越来越突出。药品知识图谱用节点（药物、疾病、症状等）和关系（禁忌、适用症等）所组成的图谱，可以为真实的医疗场景直观地建模。形成知识图谱的过程本质是建立对医疗领域的认知和理解。传统用药方案推荐是由医生首先开具处方，药师对处方

进行审核并开具使用方案及注意事项。而每位药师和医生存在知识广度、深度和知识结构的局限性，无法很好的满足日渐增长的安全用药需求。而由于知识图谱运用了“图”这种基础性、通用性的“语言”，“高保真”地表达安全用药领域中的各种关系，并且非常直观、自然和高效，不需要其他方式所需的中间过程的转换和处理。

2、患者智慧服务

患者对于医疗服务，尤其是医疗知识服务有着非常广泛的需求，美国有超过1.6亿的人在网络上搜索健康信息，而中国每天大概有超过6000万次的医疗相关的信息检索。知识图谱支撑的患者智慧服务典型应用有：

（1）医疗知识服务与智能助理

提供医疗知识服务的数字助理可以理解患者的查询意图，推理并输出各类医疗知识和信息，辅助患者或医生。应用场景包括：个人问诊、用药咨询、慢病管理等。基于知识图谱的问答方案能够更好的理解语言中的抽象部分，通过智能对话来完成患者病状问询。在推荐上能够锁定较少、较为精确的推荐结果。

（2）智能分诊导诊

医疗分诊导诊类服务机器人主要是通过患者的输入进行语义分析，然后给出医院的分诊和导诊建议，节约人力，方便患者。依赖知识图谱的推理，还能够达到比普通护士更专业更深入的分诊，并实现基于分诊导诊平台的预问诊，提前收集患者的病史信息，为医生和患者节约时间，并提供更加细致的诊断信息。

（3）辅助问诊

医疗知识图谱可以通过海量的病历、医疗文献、临床手册等文本数据将各种琐碎、零散的医疗信息知识相互连接，以支持综合型知识检索问答、辅助决策和智能医疗诊断。将精准医学知识与大数据相结合，利用庞大的全人类对疾病的理解和医生的经验形成知识库，让医生能够通过智能

诊断系统直接根据病人的个体实际情况来对他们进行针对性的诊断和治疗，辅助医生的诊疗过程，使得普通医生也能够像最好的资深医生一样为病人提供高质量的诊疗服务。

3、医学科研

基于知识图谱技术的医学科研，使得科研工作中一些耗时耗力的基础知识工作能够实现自动化，形成智能化的科研辅助应用，基于知识图谱的医学科研典型应用有：

（1）医疗文献辅助阅读与洞察

医疗文献辅助阅读应用利用人工智能技术从海量医疗文献中针对性抽取有用的信息，结合自动摘要、语义检索、关联分析等技术手段，形成基于医学文献的知识图谱，帮助医生快速阅读和理解医学文献，提高学习效率。并提供精准的检索和推理能力，基于已有文献的知识，为科研选题预测、文献综述分析等提供自动化智能化的辅助。

（2）医疗信息搜索引擎

传统的医疗搜索引擎需要对浩如烟海的医疗文献等资源进行检索、存储、处理，但难以支持和理解用户的语义搜索。基于医学知识图谱的搜索，提供用户不同类型实体间丰富的语义关系的搜索。知识图谱对于传统医学信息搜索的优化主要体现在查询扩展，即从知识图谱中抽取与查询相关的若干实体及实体关系和属性进行扩展查询，以便能更好地理解用户的查询需求。目前，国外典型的医疗专用搜索引擎有WebMD、OmniMedicalSearch、Healthline等，国内主流医疗搜索引擎有搜狗明医、360良医等。

（3）辅助药物研发

深度学习和其它智能算法都有望对药物研发产生重大影响。利用现有数据，构建药物相关的知识图谱表示体系，并基于知识图谱融合多元海量信息，建立智能模型，通过图谱模型推断辅助探索药物研发，降低研发成

本，提高研发成功率。

三、智能制造

（一）背景介绍

随着云计算、大数据、人工智能技术的快速发展，越来越多的新技术正在应用于传统工业领域，并在帮助企业实现产业转型、技术升级及效益提升方面起到了关键作用。目前在提升良品率方面，知识图谱通过深度计算所有的关联参数，可精准分析出与生产质量强相关的关键参数并基于分析结果搭建出参数的曲线模型，结合生产过程中实时监测和调控变量，最终将最优参数在大规模生产中精准落地。此外，知识图谱在提升质检效率、提升测试效率、优化能耗、降低设备维护成本、优化生产工艺等方面同样较多应用场景。

（二）典型应用

1、石油化工领域的应用

石油化工业具有易燃易爆、流程工艺复杂、控制要求精细、信息高度集成等鲜明特点，在加工过程中从原料到中间馏分与产品的物性分析数据纷繁多样，产业链示意图如图5.10所示。由于炼化的复杂性，现实中仍存在大量无法通过机理模型或模拟软件所不能解释的现象是。针对上述问题，可借助知识图谱在较短的时间内从众多影响因子的因果变化关系中找出满足优化目标的操作参数，从而为解决许多生产问题提供帮助。炼化生产过程中众多影响因子间的因果变化关系是一个复杂的关系网络，可以用知识图谱来分析，图5.11展示了与某因子相关联的其它影响因子。

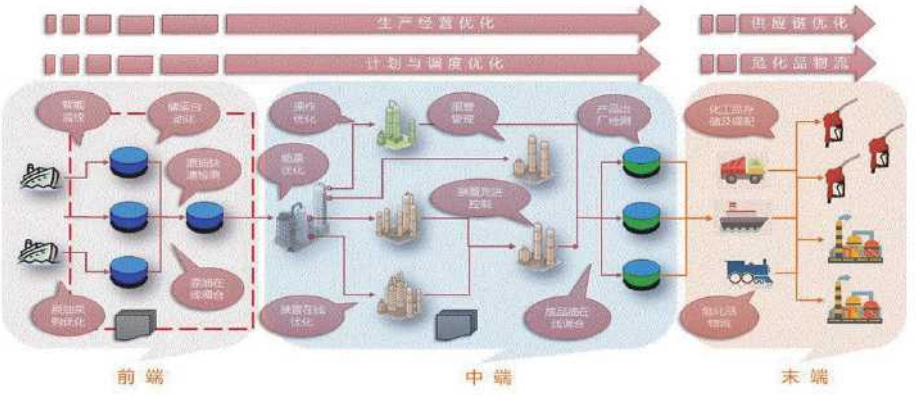

图5.10 炼油化工产业链示意图

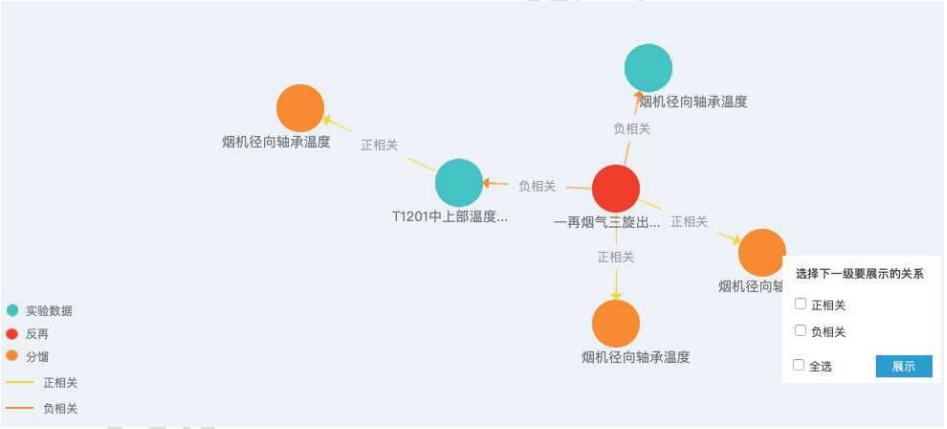

图5.11 炼化过程影响因子分析（局部）

通过炼化知识图谱，可以完成以下功能：

（1）辅助生产操作控制：当生产线工人准备改变某个可操作变量时，可以通过知识图谱直观地看到该操作变量改变时会直接引起其它因子变化的趋势；当试图改变某个非操作变量时，可以通过知识图谱观察到哪些可操作变量值的改变会对该值产生直接影响。

（2）生产预测：在石化生产中，各种影响因子之间关联度高且关系高度非线性，而且大型生产设备各个位置的浓度、温度、速度及化学反

应、物理变化过程相互影响。仅通过机理模型（如集总动力方程、人工经验）和经验模型（如神经网络、统计回归等）都很难建立准确有效的模型。在机理模型与经验模型融合的基础上，结合炼化生产知识图谱可通过图迭代计算，计算出当某些因子变化时，整个关系网络达到稳定后各个产物结点的状态值，进而实现更准确的生产预测。

（3）生产异常追溯：当生产线某个指标发生异常时，通过找出在炼化生产知识图谱中对应的结点，分析以该结点为中心的子图，可发现该异常的影响因素；通过知识图谱推理，还可寻找该异常发生的原因。前文虽然仅以石化行业举例，在其它复杂的生产过程中，也可以有类似的知识图谱应用。

2、钢铁行业生产知识整合

由于炼钢连铸的生产工艺较复杂，工序繁杂，如何在合理时间内掌握各个生产环节的关系并进行生产过程的智能调整或者智能调度是极其困难的事情。借助知识图谱可进行关联性分析，发现一些有价值的新现象和新规律，而传统数据分析方法限于某个局部，造成了数据之间的断裂，数据中的规律无法充分挖掘。此外，知识图谱解决方案可以不改变企业原有数据体系，并以统一模型和实体为上层应用提供数据共享服务。知识图谱在钢铁工业智能制造中的主要应用是实现机器的感知、传递和自我诊断，通过分析工业生产中收集的数据，优化设备的产出和减少资源的浪费。

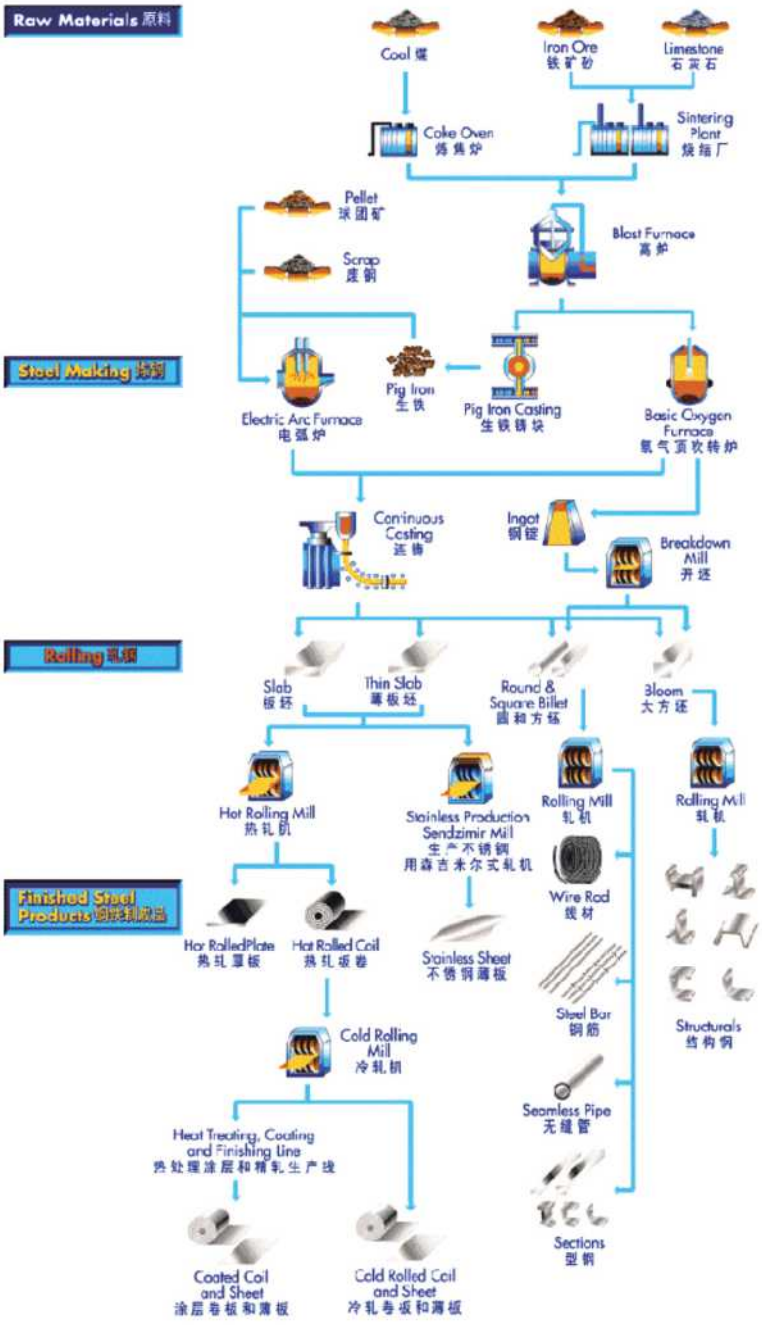

图5.12 钢铁产业的流程图

3、设备制造商全数据链整合

生产企业在不同的应用方向往往存在多个独立系统，如库存管理系统、生产执行系统、物流追溯系统、订单管理系统、维保管理系统、人力资源管理系统、市场监控系统、运营管理系统等。在企业内，不同的应用领域，往往会有相应的专家知识库，例如：市场预测方法、故障维修方法、生产排期方法、销售违规判定方法等，这些孤立的专家知识库仅利用了局部知识，无法形成完整的知识脉络。通过构建全面的企业知识图谱，能够实现企业内各业务数据及专家知识的全面整合，有助于消除数据孤岛并帮助生产企业进行决策管理优化。图5.13展示了一个设备生产企业将供应链、生产、维保、销售、市场监控等数据整合后构建的知识图谱案例，可基于此实现综合决策。

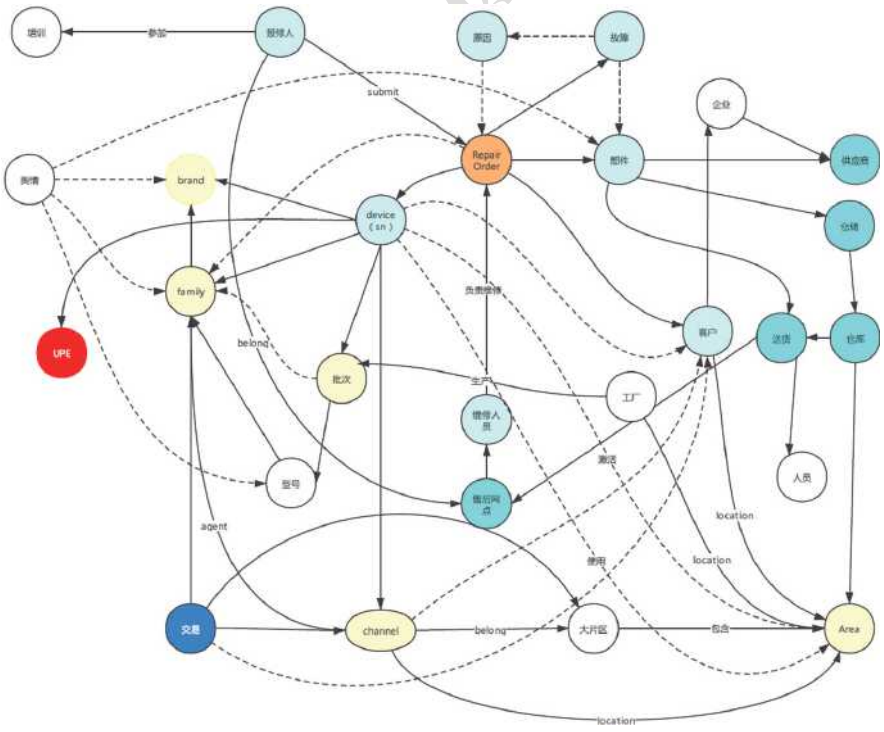

图5.13 设备制造商全数据链整合示例

四、智慧教育

（一）背景介绍

根据《中国互联网教育平台专题分析2018》,2017年中国互联网教育市场规模达到2502亿元人民币,同比增长56.3%,但相比线下整体教育行业而言,互联网教育占比仍然较低,尚有巨大的发展空间。在消费升级、知识付费浪潮的大背景下,居民在教育领域的投入有望进一步增加。我国目前已发布多项政策，加速推动教育信息化进程,逐步落实智慧教育，其中包括《国家中长期教育改革和发展规划纲要(2010-2020年)》、《国家教育事业发展“十三五”规划》、《教育部2018工作要点》等。智慧教育市场的巨大发展潜力吸引了诸多企业涌入,并尝试从不同角度切入市场，其中涉及百度、阿里巴巴、腾讯、网易、新东方和沪江等企业。

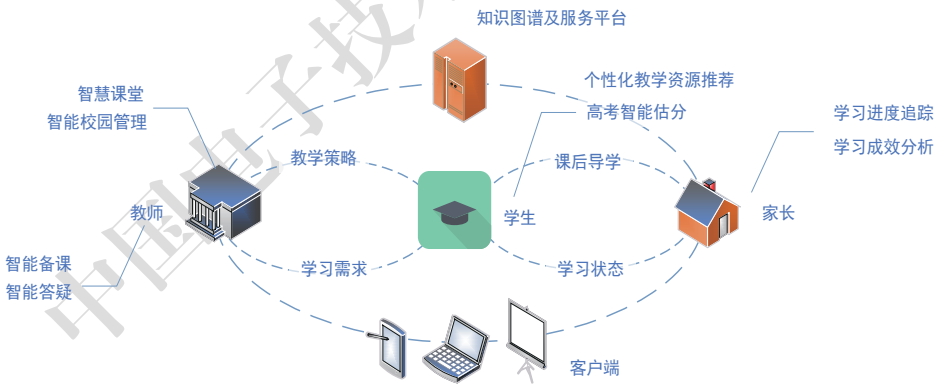

图5.14 知识图谱在智慧教育中的应用示意图

通过构建知识图谱，可建立起教师、学生、家长间的良性互动关系，通过网络阅卷、在线考试、题库诊断、教学资源推荐、智慧课堂等多种渠道为学生提供针对性、个性化学习，帮助学生提高学习成绩，建立“以学生发展为中心”的个性化学习方式。而且，家校互动、互动课堂、学生成

长档案等系统间数据的自动对接与交换还可减轻教师的工作负担。

（二）典型应用

目前，知识图谱技术已运用到智慧课堂、教学资源智能推荐、高考智能估分、学习成效分析等多元场景，同时为教师、学生、家长、校园管理人员等多方提供服务，并借助智能终端的覆盖和更新不断提升用户的使用体验。

1、智慧课堂

智慧课堂是基于新一代信息技术打造的智能、高效的课堂[13]，是信息化课堂发展的新形态。目前对教育信息化的研究不断向课堂、向教与学聚焦，智慧课堂成为实施智慧教育的核心载体，也是当前学校信息化教学改革和企业教育信息化研发推广所关注的焦点。通过打造智能、高效的课堂，实现课前、课中、课后全过程应用，促进学生的智慧发展。与传统课堂教学流程结构相比，智慧课堂实用教学流程在课前阶段以学情分析为核心，在课中阶段以师生互动为关键，在课后阶段以个性化辅导为重点。以百度智慧课堂为例，其将新课标下多版本新教材的多种资源进行匹配，目前包含新课标14学科方向的知识图谱。

2、智能备课

教学体系的变更，教学大纲的更新、迭代，给老师们备课环节提出了更高的要求。通过感知教师教学需求，关联教材版本，并通过对互联网上碎片化的备课内容进行系统化梳理，实现在资源有序组织的前提下实现备课资源与教师诉求的精准匹配，降低老师寻找、筛选资源的成本，满足个性化备课需求。同时，智能备课还可追踪老师的教学进程，持续性推送符合教学需求的备课资源，并搭配搜索技术同步老师下一个章节所需要的备课内容，系统化满足教师在一学期甚至一学年内的连续备课服务需求。

3、智能答疑

目前广泛应用的答疑系统一般通过邮件、聊天工具等方式在线答疑，促进了师生之间的互动，学生的疑问得到反馈，教师的工作压力也得以减轻。但缺乏对以往问题答案的自动归纳总结和课程知识的智能表示，在自动答疑等智能化方面还有所欠缺。 智能答疑系统以知识图谱的形式将学生提问问题以课程知识点分类，并与教学案例课件匹配，将学生问题关键字按照知识点树中的知识点进行归类，逐展现与该课程某个知识点相匹配的全部提问。而且，学生通过点击关联出的与问题相关知识点演示课件页面，提升对知识的查漏补缺和消化吸收的效率。

4、高考智能估分

百度教育连续四年在高考期间推出“高考智能估分”系统，为数百万学子提供了估分便利。2017年，百度不仅在既有的大数据分析技术、OCR识别技术、智能检索技术基础上完成了进一步的升级，同时，基于百度教育知识图谱构建的高考学习链路 逻辑，还能够将题目解析与对应高考知识点进行更为精准、快速的匹配，完成对高考考点的抓取与智能定位，令估分更快、更准确。2017年百度教育闪电智能估分服务实现了考后 2 小时上线真题答案解析。

5、个性化教学资源推荐

个性化教学资源推荐主要服务于学生自主学习、课后练 习以及教师备课等相关环节，可通过用户画像、资源画像及构建知识图谱，实现学习资源的个性化推荐。目前，百度推出的教育知识图谱将知识体系化精细梳理，覆盖高校12个一级学科、92个二级学科、504个专业、7590门课程。个性化课程知识匹配全网内容，满足学生在学习中的使用需求，还能识别学习的意图，根据学习诉求展现不同内容。知识捕手技术能帮助用户快速查找知识点，结合教育知识图谱，将知识点和学习资料快速建立联系，提升使用者的学习效率。

6、学习成效分析

学习结果数据是智慧课堂教学成效的基本体现。通过连续多次考试排名建立对学生成绩上升/下降、学习成绩分档模型。对学生考试成绩偏科情况探索，从整体角度分析偏科人数及偏优和偏弱学科，从个人角度分析学生偏科行为。通过对学生历史考试成绩排名数据以及近期在作业平台上的行为数据进行未来成绩趋势预测。通过学生在智课平台的学习行为来自动还原知识图谱网络，并与人工构建知识图谱进行对比分析，描述学生的知识结构情况。

图5.15 基于知识图谱的学习成效分析示意图[1]

7、智能校园管理

通过对校园特定管理区域内产生的学习行为、教学行为、资源访问行为、学习效果、学习教学评价等多个方面的数据进行关联分析和挖掘，实现教育数据的可视化，为教学过程的不断优化提供参考，也为校园管理者的宏观督导提供依据和方向。比如在中小学校园中，校园管理者可以看到教育资源的消费情况，从而调整相关资源的采购投入比例。同时，可依据每位学生的学习情况，产生系统化的学习报告，向管理者、教学者、家

[1] http://www.foundertech.com/bz/list_164.html

长及时反馈，便于其准确地进行学习诊断和评价，按需机动地提供学习辅助、引导。下图为方正推出的教育数据平台，通过统一数据建模、统一数据管控、统一数据展示、统一数据服务以及统一标准规范，带动全省/市/区/县教育信息化的全面发展。

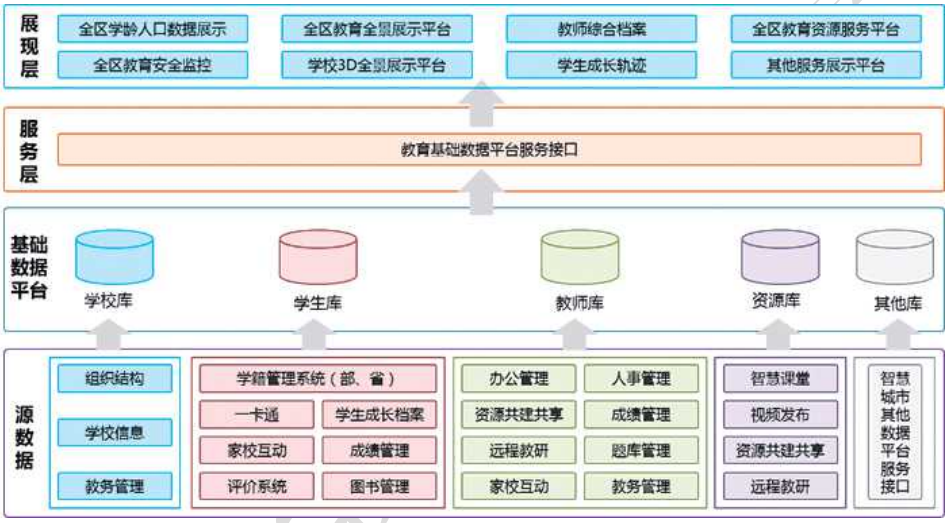

图5.16 智能校园管理系统架构图[1]

五、智慧政务

（一）背景介绍

智慧政务即通过“互联网+政务服务”构建智慧型政府，利用云计算、移动物联网、人工智能、数据挖掘、知识管理等技术，提高政府在办公、监管、服务、决策中的智能水平，形成高效、敏捷、公开、便民的新型政府，实现由“电子政务”向“智慧政务”的转变[2]。智慧政务生态中

[1] http://www.foundertech.com/bz/list_164.html
[2] 金江军. 智慧政府: 电子政务发展的新阶段[J]. 信息化建设, 2011 (11): 16–17.

包含政府部门、企业、非营利组织、企业、公民、监管方、协调方、独立第三方和平台运营公司等角色，其中政府部门、公共部门、非营利组织、企业和公民既是数据的提供者，又是数据和知识服务的需求者；政府大数据主管部门通过授权平台公司运营政务数据共享开放平台、契约式开放平台等数据中台，为各类数据需求者提供数据服务和知识服务，同时大数据主管部门对平台公司具有指导、监督职能；为确保平台公司高效、合规运营，需要协同方对政务大数局参与者进行协调，同时授权独立第三方机构或权威组织对各政务部门进行绩效考核。

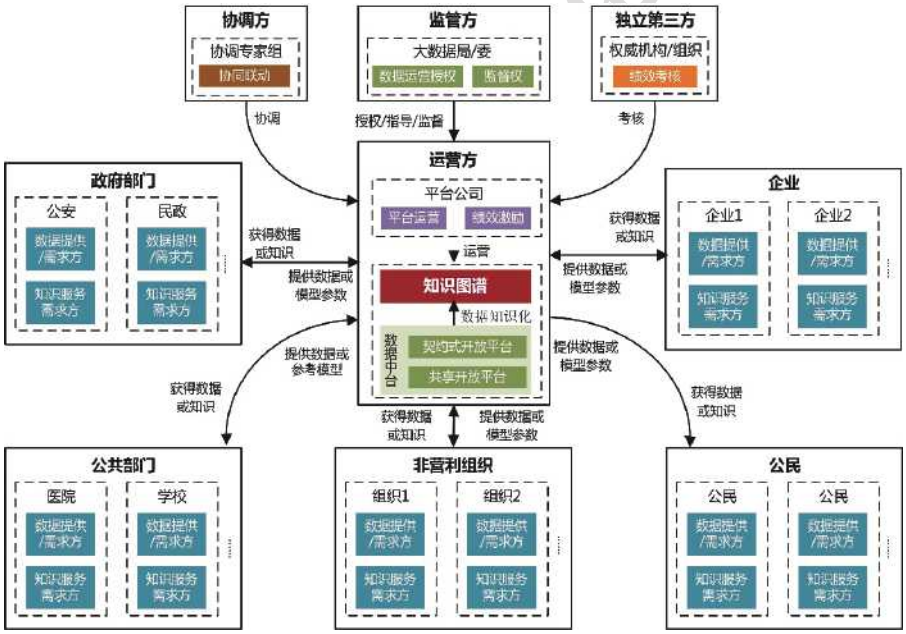

图5.17 基于知识图谱的智慧政务全景图

在政务生态链中，知识图谱的意义在于将大规模、碎片化的多源异构政务数据进行关联，以实体为基本单位对政务数据进行挖掘分析，揭示各实体间的复杂关系，实现知识层面的数据融合与集成，更大程度释放政务数据价值，为政府部门、企业、非营利组织、企业、公民提供知识服务。

（二）典型应用

知识图谱有助于实现事前事中事后全流程综合市场监管，提高监管效率和效能。在政府市场监管由单一行业监管向新技术引领下综合市场监管转变的大趋势下，基于国家到地方市场监管部门的市场监管事项建立监管目录，通过政府数据共享交换平台完成数据采集汇聚，以企业为对象构建综合市场监管知识图谱，实现食药品安全监管、生态环境监管、企业信用监管等监管事项的集成和优化。

1、场景一：知识图谱在政策公文领域应用

基于知识图谱的政策公文智能应用基于全国海量政策数据建设的政策知识图谱构建了机构、政策、公文、法律法规、解读等实体相互关联的复杂网络,实现了知识层面的数据融合与集成,并以知识图谱为核心搜索引擎完成了政策大数据知识服务平台的建设。以知识图谱为核心搜索引擎,打破了原有基于关键词的政策获取单一模式,转型升级到多维度立体知识检索发现服务模式，并为政府机构、企业、政策研究院所、普通民众等聚焦国家党政政策的客户提供精准知识服务，支撑政府用户、企业用户从政策的研究视角,深度探索政策的关系等。

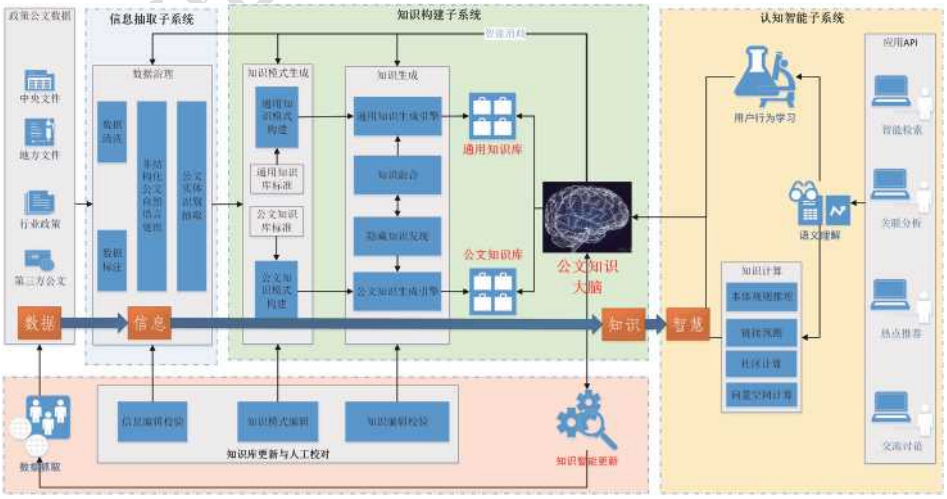

图5.18 基于知识图谱的政策公文智能应用

2、场景二：知识图谱在医保审计的应用

由于我国庞大的机构与人口构成，审计部门由于人力限制，只能牺牲时效性与覆盖广度，采取事后抽查的形式进行审计。为解决这些问题，依托知识图谱技术将医保数据、民生数据等结合关联，通过建立人、医院、药品等实体的关联关系，构建医保审计知识图谱，改变以往抽样审计形式，利用全数据进行审计。通过知识图谱，数据的相互关联充分解决了医保数据分散，精度低的问题，同时通过利用关联各种来源数据的方法可以进一步挖掘数据价值，极大拓展了可审计内容的范围。

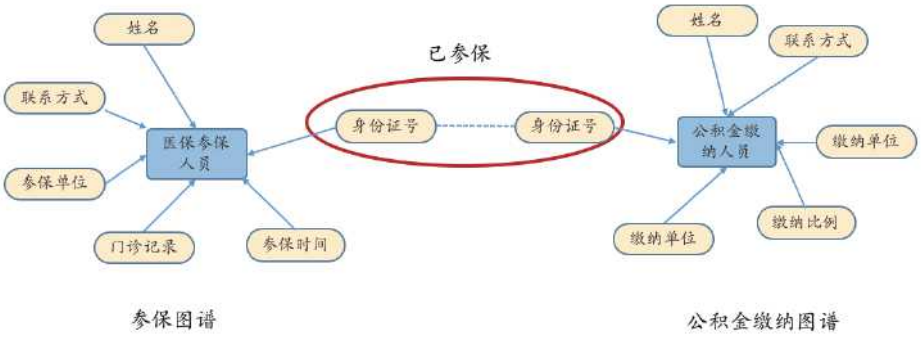

图5.19 医保审计知识图谱

3、场景三：知识图谱在辅助行政审批的使用

行政服务中心是集信息与咨询、审批与收费、管理与协调、投诉与监督于一体的综合性行政服务机构。目前，对于行政服务中心综合窗口而言，存在问题包括：办件量大，工作人员业务负荷重；工作人员审批事项多，工作人员学习成本高；审批过程仍存在大量重复性作，人员效率得不到充分利用等。通过业务材料、审批规则、法规等建立知识图谱，发生审批业务时扫描的材料通过人工智能识别，提取材料信息，并进行合规审核。最终，辅助工作人员收件与审批，为整体审批提速提效，在提高办事效率的同时降低退件率。实际业务与辅助审批系统运行过程如图5.20所示。

图5.20 政务服务审批系统运行流程

此外，知识图谱在社会管理方面的应用还包括智慧应急、智慧交通、智慧扶贫、智慧安全生产等。以智慧防汛抗旱为例，通过对水库、河道、湖泊、闸站等关键节点水位数据及雨情、气象、工情等防汛信息的智能提取与综合分析，构建防汛抗旱知识图谱，建立防汛抗旱各类实体间的链接关系、空间关系、时间关系及扩展关系等，实现灾害历史相似分析、灾害影响关联分析，为应急联动指挥提供智慧化辅助支撑和决策支持。

六、智慧司法

（一）背景介绍

智慧司法是综合运用人工智能、大数据、互联网、物联网、云计算等信息技术手段，遵循司法公开、公平、公正的原则，与司法领域业务知识经验深度融合，使司法机关在审判、检查、侦查、监管职能各方面得到全面的智慧提升，实现社会治理、公共法律服务等的智慧化。近年来，我国不断深入推进“智慧司法”建设. 随着越来越多的相匹配的司法信息平台的搭建运行，我国司法信息化走向前所未有的广度和深度。2018年上半年

全国法院新收案件1229.5万，人均新收案件99.2件；上半年全国法院结案952.8万件，人均结案76.8件。在巨大的案例压力下，政府基层工作人员数量不足40万，远远不能满足多达6亿人次/年的法律服务需求，并且法律服务专业性强，咨询、诉讼服务费用高，也是导致大部分企业和个人得不到健全法律服务的主要原因。知识图谱的构建是实现智慧司法不可逾越的建设基础，知识图谱能够表达法律知识体系间的逻辑关联，并显示被关联的体系内的知识。司法知识图谱可实现智慧司法的技术底层，找到对应的实体属性概念，触发相关的推送知识，还可以通过配对的规则，用概率来实现排名推荐，对类案进行分析，还可以广泛运用于要素式的审判，法律行为分析的预测，结果预判的分析等等，能很好地达到数据关联、知识拓展和应用支持。

图5.21 知识图谱在智慧司法中的应用

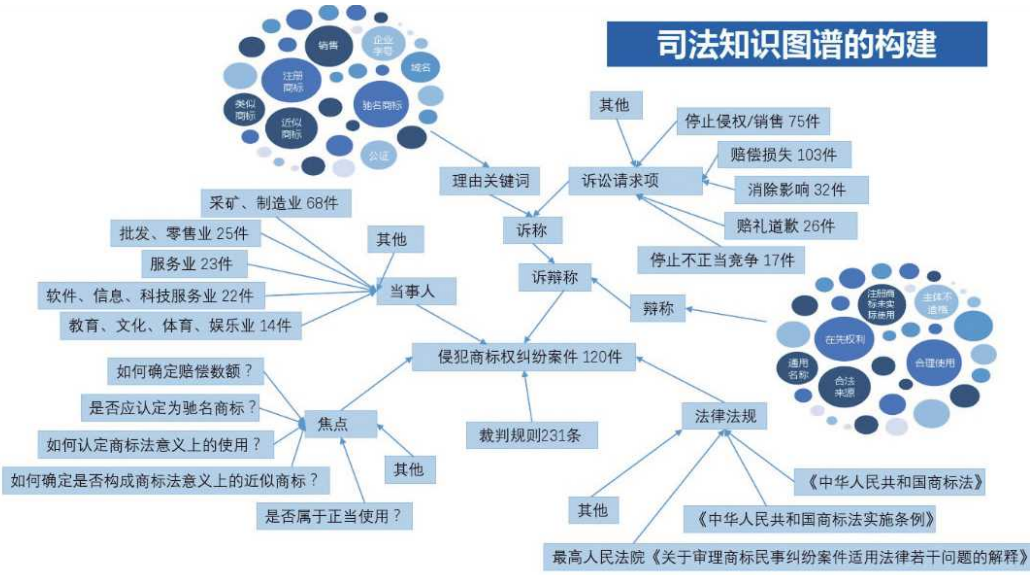

图5.22 司法知识图谱的构建

（二）典型应用

1、搜索引擎构建

基于传统数据库的搜索，对复杂信息或者复杂关系的表达比较难，通过既定的模式所花费的开发成本较高而且对复杂关系的多维度查询性能比较低下。基于知识图谱的搜索，结果指向有具体关系属性的相关实体，依托于其他各类数据，可以提供针对不同场景需求的法律数据关联和检索分析服务。目前已有很多面向司法领域从业者、公民的法律知识搜索工具和应用，如无讼案例、聚法案例、中国裁判文书网、OpenLaw、CaseShare、法信、北大法宝、最高人民法院裁判文书、知识产权裁判文书网和百度度小法等。

2、证据索引

将证据指引嵌入系统内，利用知识图谱对证据的合法性、关联性进行判断，进而识别是否达到确实、充分的标准。在这方面，刑事案件要相对

容易实现，因为只需要达到“排除合理怀疑”即可，而民事案件受利益衡量、价值判断、个案因素制约，其证据判断要更加复杂。

3、类案检索及推送

当前的类案推送系统以知识图谱和自然语言处理为核心，结合技术服务商的人工建模标注，可以达到一定程度的自动推送和检索，但仍然存在类案推送“有效性”、“有序性”及“完整性”不完善的困境。2017年10月，中共中央办公厅印发的《关于加强法官检察官正规化专业化职业化建设全面落实司法责任制的意见》，要求依托大数据技术，完善智能辅助办案系统的类案推送、结果比对、数据分析等功能，促进法律适用统一。

4、结果预判

根据类案进行推演，可以延伸到案件结果的预测，如量刑预测和经济成本预测等。除此以外，从监察管理的角度，管理者可以对判案结果进行类案判决的偏离预警，对于“同判度”较高的类案，基于预测性判断对法官制作的裁判文书判决结果与之发生的重大偏离情况，进行自动预警从而防止裁判尺度出现重大偏离。同时，对结果偏离大的案件则需进行重点监察，消除由于法官个人因素导致的问题，促进法律适用统一的推进。

5、证据分析

通过对知识图谱的构建，搭建论证案件事实的分析推理路径，结合案件事实、法律要素，为办案人识别、审查、运用证据，整理证据链的证明逻辑，并对证据的合法性、关联性进行判断，进而识别是否达到确实、充分的标准，为梳理案件事实提供智能化辅助。实际应用中，刑事案件要相对容易实现，因为只需要达到“排除合理怀疑”即可，而民事案件受利益衡量、价值判断、个案因素制约，其证据判断要更加复杂。

6、文书生成

以权威文书样式、法律法规和相关司法解释为基础，通过自然语言处理、机器学习、序列标注等技术手段对电子文件进行解析，学习不同类型

文书中针对法律事实的常用表述规则。经由法律专家校准结合知识图谱，固化形成专业的文书表述逻辑和规则库，应用到模板中的智能文书自动制作系统，实现流程如图5.23所示。

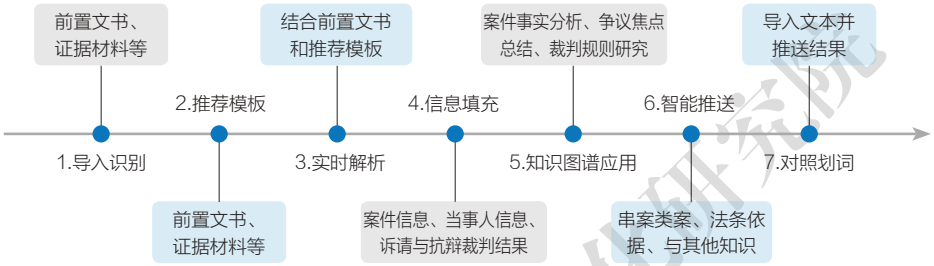

图5.23 文书生成流程

7、智慧调解

依据案件情况以及调解人员的技能知识图谱，为当事人在全社会范围智能推荐擅长的调解员。同时，以知识图谱为基础的调解智能服务，为各界调解员调解工作开展提供参考，根据记录的事件进展结合知识图谱、事理图谱自动生成调解协议，或者进入诉讼流程，相关的诉讼结果预测等也会推送给相应的判案法官，辅助调解员提高调解成功率，减轻法院的办案压力。

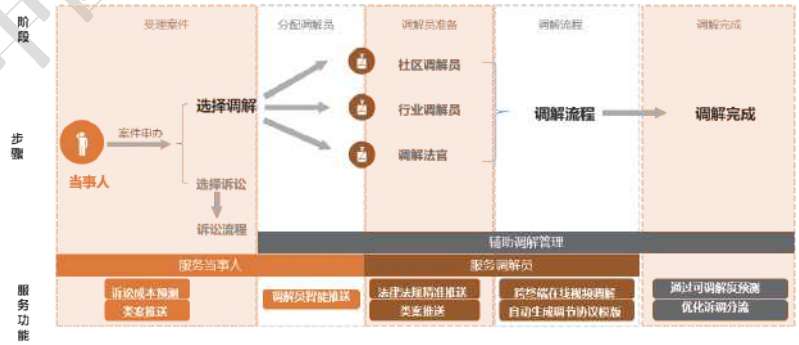

图5.24 智慧调节流程图

七、智慧交通

（一）背景介绍

现代城市发展过程中的一大问题是交通拥堵，为解决城市发展中的这一顽疾，有必要以现代化高科技技术为支撑，建造城市中的智慧交通系统，从源头入手缓解城市拥挤问题。当前，“智慧交通”更加注重人、车、路和环境的和谐协调的关系处理，使交通发展更加具有可持续的协调发展意识，以更好地改善交通秩序和交通环境、节约能源、降低环境污染，智慧交通中知识图谱应用框架如图5.25所示。

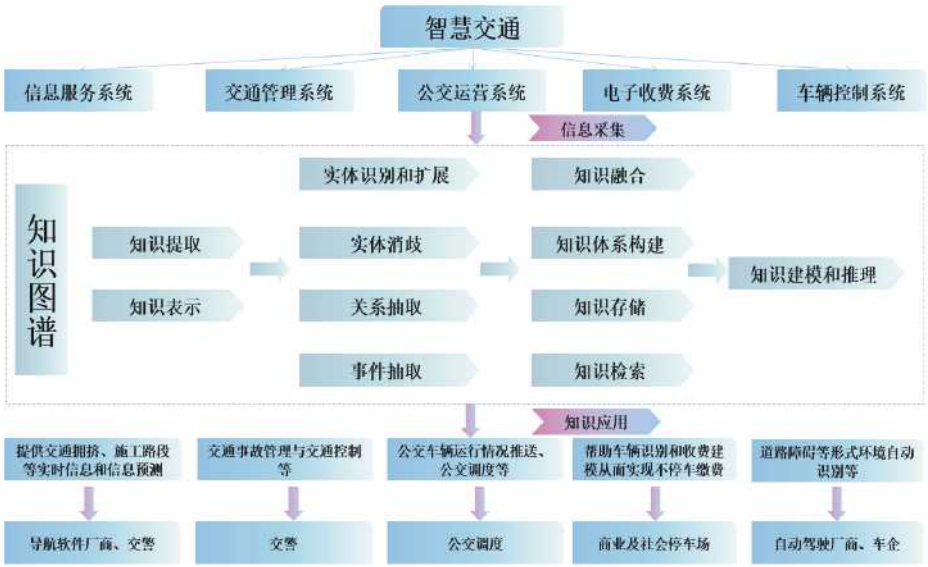

图5.25 智慧交通中知识图谱应用框架

（二）典型应用

根据“智慧交通”系统在城市中的作用，将其典型应用分为以下四个方面：城市安全及交通管理服务，如城市事件感知与智能处理、人车识

别、交通信号控制等；出行服务，如路径规划、停车诱导等；交通拥堵缓解，如交通诱导、不停车收费ETC等；交通规划决策支持，综合交通信息平台、交通数据采集等。

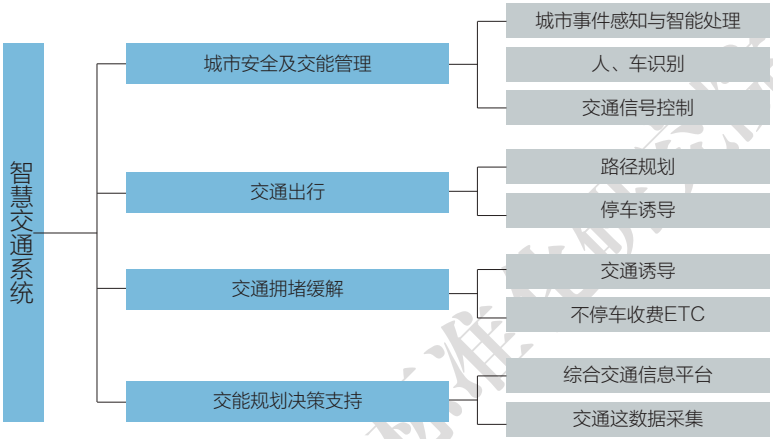

图5.26 “智慧交通系统”应用分类

1、知识图谱在人车识别中的应用

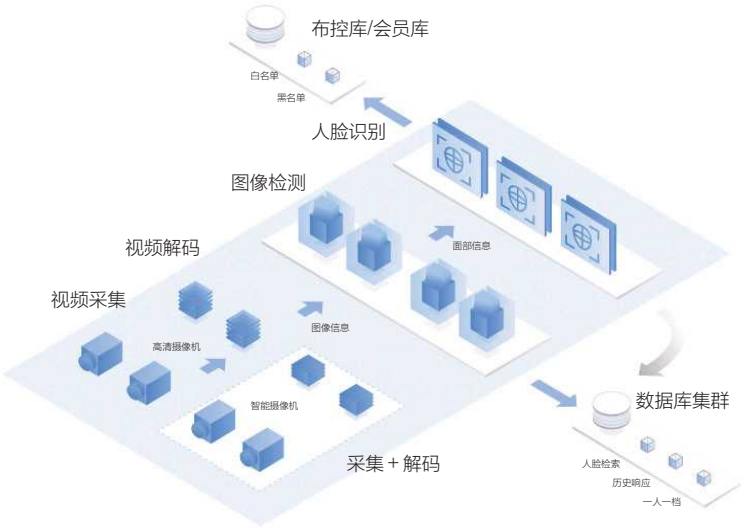

图5.27 利用知识图谱构建人脸天网的流程图

通过高清 IP 摄像机、视频平台等接入及流媒体解码和截帧图片质量判断，快速对检测人脸进行属性标记；结合依托知识图谱对城市居民构建的数据库集群，将特定集合的大规模人脸照片按特征值相似度进行动态分组归类，利用知识建模进行关系推理，进而快速检索居民本身以及和其有关系的其他居民的历史和实时交通动向。此外，对车辆也可进行同样的识别与检索操作。

图5.28 人、车识别与标签实例

2、知识图谱在路径规划中的应用

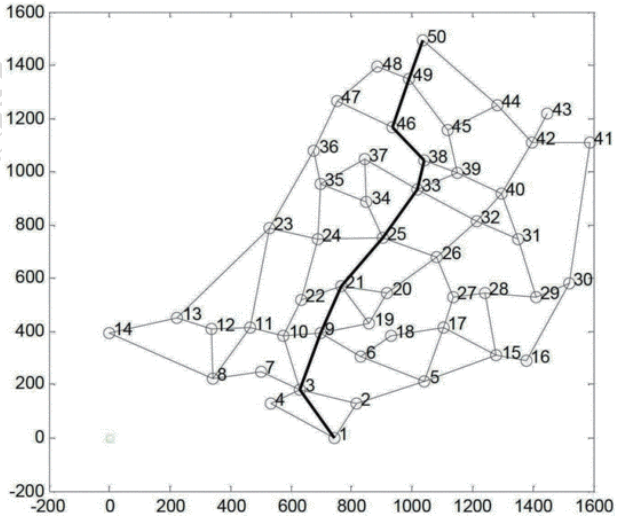

图5.29 路径规划实例

路径规划是指在具有障碍物的环境中，按照一定的评价标准寻找一条从起始状态到目标状态的无碰撞路径。在交通控制、资源配置、车辆导航等领域，最短路径算法逐渐成为研究的热点。随着人工智能技术的不断发展，知识图谱的引入使得类似路径规划得到了更快速的解决。基于出行事理的路径规划与推荐——图谱本身所刻画的是一种具有时空属性的关系型知识，还在具有时空属性(包括先后顺序，空间的先后顺序)的其他场景中推广应用。

3、知识图谱在交通诱导中的应用

城市大脑将所有人、车、道路数据都接入系统，通过人工智能分析技术，把庞大的数据转化为科学合理的业务模型，就此形成城市交通实时大视图，以此来完成城市交通系统的调度和管理。知识图谱可将摄像头与红绿灯采集的信息进行关联，通过对从摄像头得到的即时交通流量数据分

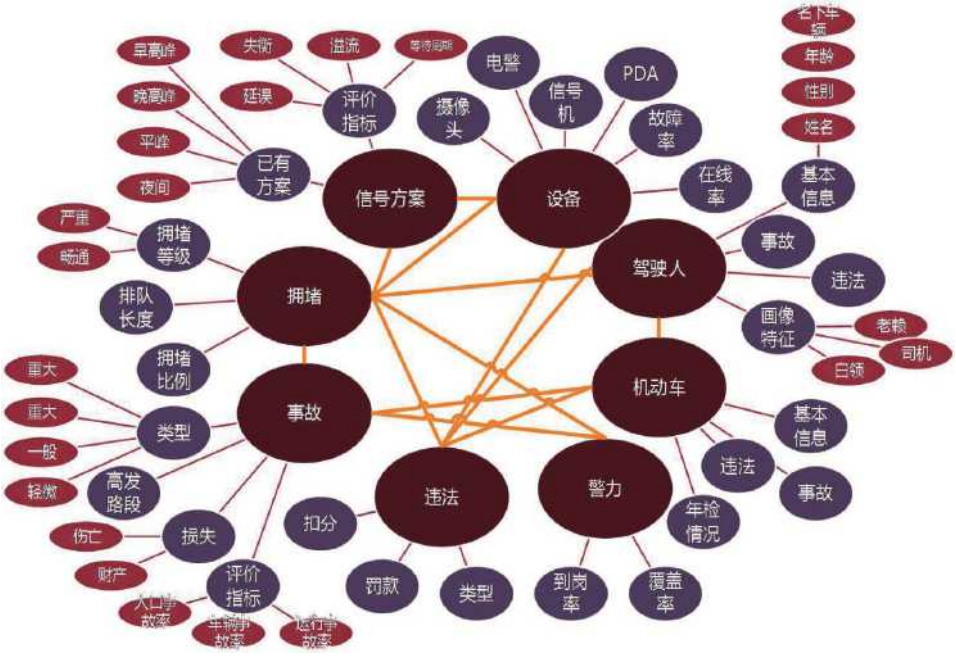

图5.30 城市交通知识图谱

析，优化路口的时间分配，提高交通效率。同时，与监控、监测、交通流量分布优化等技术的结合，完善公安、城管、公路等监控体系和信息网络系统，建立以交通诱导、应急指挥、智能出行、出租车和公交车管理等系统为重点的、统一的智能化城市交通综合管理和服务系统建设，实现交通信息的充分共享、公路交通状况的实时监控及动态管理。

八、智能电网

（一）背景介绍

智能电网以物理电网为基础，将现代先进的传感测量技术、通讯技术、信息技术、计算机技术和控制技术与物理电网高度集成而形成的新型电网。它以充分满足用户对电力的需求和优化资源配置、确保电力供应的安全性、可靠性和经济性、满足环保约束、保证电能质量、适应电力市场化发展等为目的，实现对用户可靠、经济、清洁、互动的电力供应和增值服务。分布式发电、储能技术和电动汽车的快速发展逐渐改变了传统的供用电模式，促使电力流、信息流、业务流不断融合。

（二）典型应用

当前知识图谱在智能电网领域的应用主要面向设备运维、客户服务、知识管理中心等方面，涉及电力设备缺陷记录检索、电网公司客户服务、智能变电站二次安全措施自动生成、全业务统一数据中心、设备故障诊断与管理等。

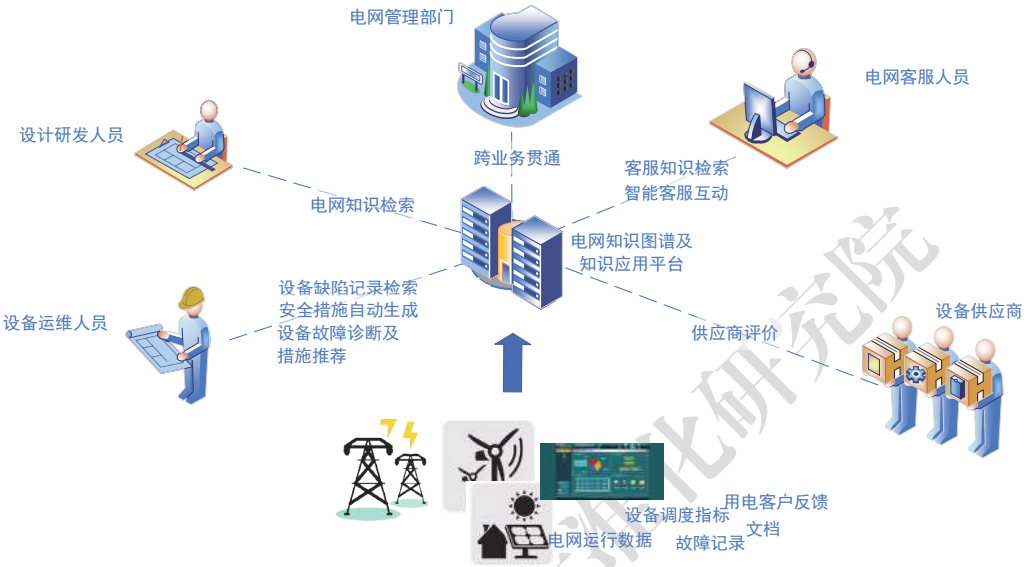

图5.31 知识图谱在智能电网中的应用分类

1、电力设备缺陷记录检索

在电力设备的日常巡检和试验中，积累了大量关于设备缺陷情况的缺陷记录文本，而在缺陷的分级、消缺等处理工作完成后，相应的缺陷记录和处理记录往往闲置于系统中。此外，由于缺陷情况复杂多变，目前很多缺陷处理工作依赖于处理人员的知识与经验。通过利用现有电力设备缺陷记录语料构建电力设备缺陷知识图谱，实现相近电力设备缺陷记录检索，从而借鉴前人经验处理设备缺陷，有效提升缺陷记录检索效率，对知识和经验相对不足的处理人员有着重要的指导意义。

2、电网公司客户服务

随着企业知识和信息量的爆炸式增长及企业对知识智能管理和应用需求的日益增加，为更方便地发现与应用知识，提高业务运营绩效，企业级知识已成为关注热点。通过构建电网企业知识图谱，建设电网公司客户服务可有效提升服务质量与效率，降低服务人员劳动强度。其中，基于海量工单数据与电力业务知识体系实现供电服务中知识、问题、经验的自动归

类，辅助使客服人员快速找到用户所需的问题与答案是其中的核心问题之一。以山东省为例，据统计，2013年至2015年通过95598客户服务热线受理的山东省内工单数量年均超过400万条，涉及业务咨询、故障报修、投诉、意见等多种供电服务的主要类型和超过300种子类型[1]。

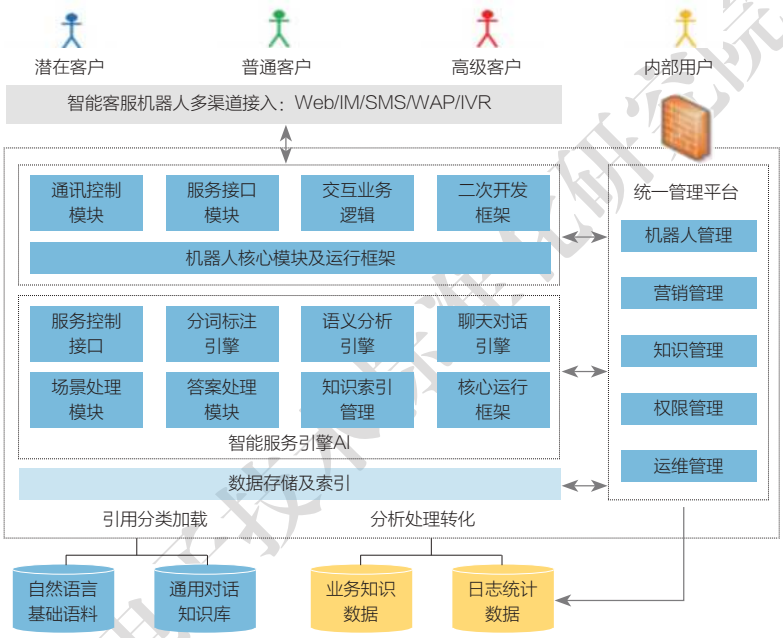

图5.32 智能客服问题系统的知识检索架构

3、智能变电站二次安全措施自动生成

智能变电站在扩建、改建及检修时，如何高效可靠地隔离相关的二次设备是当前智能电网建设过程中面临的一个难点。依托已有智能变电站信息收集和分析基础设施，通过构建知识图谱，可建立智能变电站搜索引擎，进而形成安全措施自动生成系统。二次安全措施有：投入装置检修压板；退出继电保护装置功能软压板以及拔掉装置之间通信光纤；线路保护检修安全措施；变压器保护检修安全措施等。

[1] 田晓, 刘勇超, 王婧, 等. 电网公司客户服务知识图谱构建的应用价值[J]. 山东电力技术, 2015, 42(12): 65–67.

4、全业务统一数据中心

全业务统一数据中心是指可对电网的结构化数据、量测采集数据和非结构化数据进行建模、采集、统一存储管理及分析应用的平台。然而，传统全业务统一数据中心虽可以是新啊统一采集数据，但未建立数据之间的关联性，无法跨专业贯通，数据资源有待进行智能分析与管理。通过构建知识图谱，而建立数据之间的关联性及语义级互操作的统一数据服务，实现电网数据的高效、智能化查询及保证企业数据的跨业务贯通，有利于公司经济效益提升，为公司全业务统一数据分析提供基础。

图5.33 知识图谱智能搜索视图[1]

5、设备故障诊断与管理

电力设备信息的准确检索对电力系统资产的全生命周期管理具有重要作用。由于数据重复、数据库分散、数据关系弱、数据更新缓慢,电力资产管理系统需要采取策略避免信息丢失、偏差,并提高数据存储和提取效率。通过合并现有的多源异构电力设备相关数据,构建电力设备知识图谱并配合检索方法，可有效提高对电力设备全生命周期管理的效率。此外，现场设备检修时，往往需要了解设备各个维度的信息，例如该设备上一次

[1] 王渊,彭晨辉,王志强,范强,姚一杨,华召云.知识图谱在电网全业务统一数据中心的应用[J].计算机工程与应用,2019,55(15):104-109.

检修的时间、故障原因及处置措施、设备问题的专题分析报告、设备实时参数的趋势曲线、与该设备有过相同缺陷或者被同时处置过的设备状态等，通过建立知识图谱并构建设备知识平台，可以将这些信息快捷、准确地提供给检修人员和管理人员。

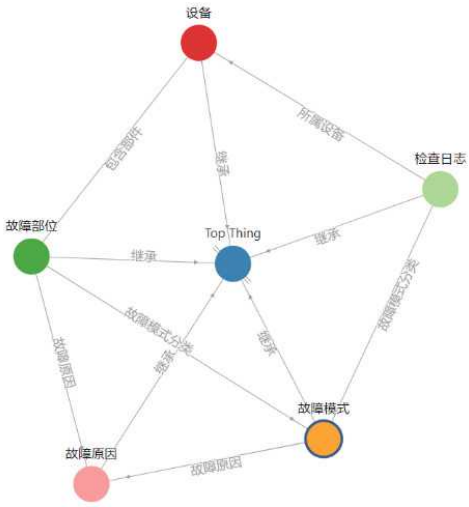

图5.34 设备故障检测知识图谱示意图

九、智能公安

（一）背景介绍

智能公安是利用互联网、物联网、人工智能、云计算、智能引擎、视频技术、知识图谱等技术为支撑，以公安信息化为核心，通过互联化、物联化、智能化的方式，促进公安系统各个功能模块高度集成、协调运作，实现警务信息“强度整合、高度共享、深度应用”之目标的警务发展新理念和新模式[1]。通过知识图谱和机器学习等相关的人工智能技术，全面整

[1] 张兆瑞. “智慧警务”:大数据时代的警务模式[J]. 公安研究, 2014(6): 19–26.

合、融合及关联各数据链路产生的数据信息，可更全面、更深刻把握犯罪形势、动态特征、局部特点、演变规律、发展趋势，为决策指挥提供动态的、系统的数据依据，实现传统决策向数据化、动态化、精细化决策转变，以达到实现智能公安的目的。智能公安的全景图如图5.35所示，其中包括了面向公安领域的基础资源、平台支撑、数据支撑、核心数据支撑、核心算法和智能应用等。

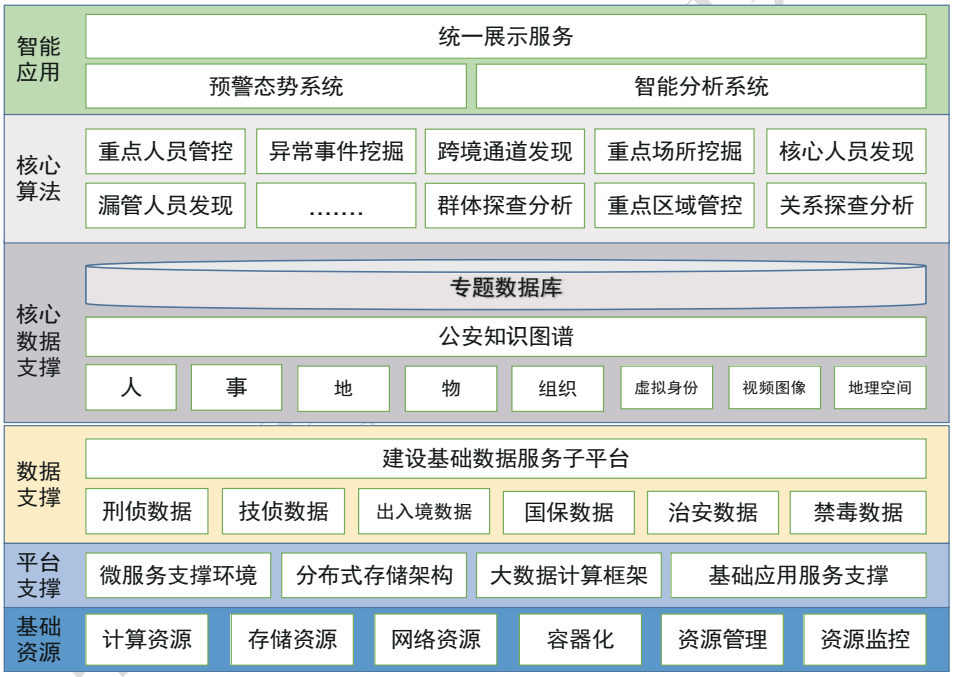

图5.35 智能公安全景图

（二）典型应用

图5.36展示了基于公安知识图谱的应用分类，可分为4大方向：全息档案研判、战法应用、情报检索与分析、事件预警等。其中，全息档案研判包括认为画像、案件画像等；战法应用包括高危人员分析研判、嫌疑人分析、串并案件分析、伴随分析及时空轨迹研判等；情报检索与分析包括

深度语义检索、警务知识问答、网络有害信息识别等；事件预警包括群体性事件预警、社会稳控事件预警等。可有效指导城市公共安全防控、警力资源调度、重大安保布防等应用，将极大地提供公安警力资源的利用率，降低城市案发率。

图5.36 基于公安知识图谱的应用分类图

1、案情辅助分析

案情辅助系统通过同时利用公安系统标签、轨迹、关系三大数据体系，充分开展图谱数据库存储模块、图谱分析模块和图谱算法模块间的统一协作，将人类智能和机器智能在交互式的产品使用模式下高效协同，创造出更大价值。由于现实情景中，疑犯为了降低被怀疑的概率，很大程度上会使用清白的车，但是要找到这样的犯罪工具并不容易，一个很常见的模式是拉一个没有前科的同乡好友一起作案。基于这样的场景，利用图谱算法对所有记录人员进行社群划分，快速地大幅缩小可疑范围，提升警务工作效率。图5.37展示了一种基于知识图谱产品进行犯罪分析的案例图。

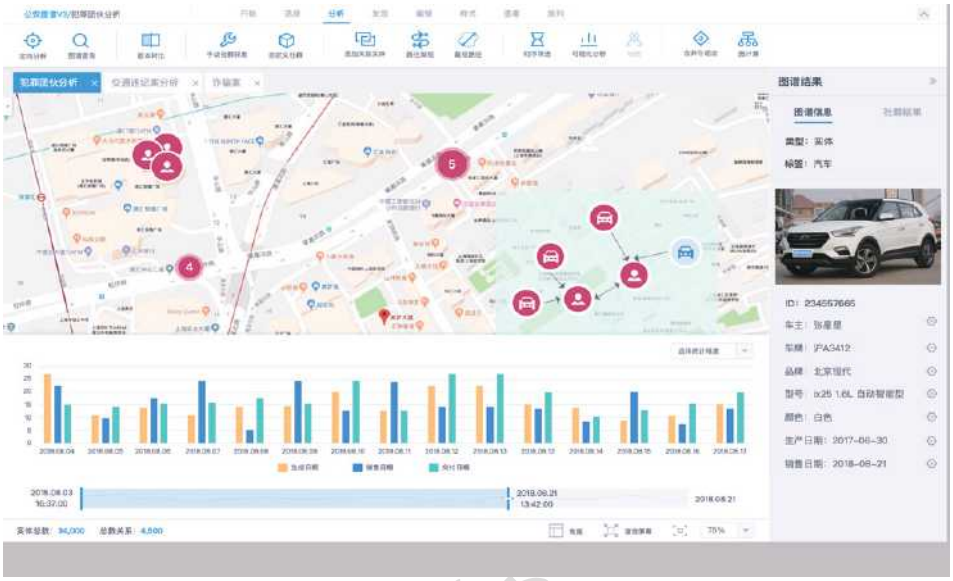

图5.37 基于知识图谱产品进行犯罪分析的案例图

2、关系网络挖掘分析模型

关系网络分析主要针对人、事、物、组织等对象要素的数据进行关系网络图谱挖掘分析，通过公安领域知识图谱，构建公安全量数据的关联图谱。在侦查办案过程中，对涉及的人、事、物、组织等对象，可基于知识图谱快速调出该对象的关系网络，并可根据公安业务侦查思路，不断进行关系网络的扩展。同时，支持采用图形算法展示关系网络中各实体之间的关系，帮助侦查人员快速梳理各类分散的、独立的情报线索。

3、亲密关系人图谱分析模型

结合知识图谱库中的各种人员关系数据、轨迹数据、关联数据等，实现对关注对象的亲密人员分析，结合各种同行的次数、亲属关系的层级和其它联系关系次数等，进行图谱关系网络算法综合计算，最终得出关系亲密度圈子，并输出关系亲密度前10的人员进行关系可视化展示。

4、物品关系图谱分析模型

物品关系图谱分析主要针对关注对象的车辆、车牌号、驾驶证、身份证、手机号、电子邮箱、虚拟身份账户、银行账户等对象，基于物品对象的关系图谱分析，通过图谱的可视化方式展示物品图谱关系网络，并可不断扩展物品关系网络进行物品关系挖掘分析。

5、团伙关系图谱分析模型

团伙关系图谱分析主要基于对象的基本属性信息进行分析，并在这些信息数据之间建立内在关联，通过关系推断和隐含关系挖掘，分析出目标对象群体的集群关系，分析指标包括目标对象群体的同类工作关系、同类犯罪倾向关系、同类居住区域关系、同类活动规律关系、同类网络关系等。

6、社交关系图谱分析模型

社交关系图谱分析主要基于对象的互联网、通讯网络等虚拟身份信息和动态关联信息进行分析，并在这些数据之间建立起关联关系图谱，通过知识图谱的关系推断、隐含关系挖掘以及虚实身份映射等，分析出目标对象的社交网络图谱关系，并展示目标对象不同的社会渠道的核心关系圈。

7、人员分类“冒烟指数”模型

人员分类“冒烟指数”模型指以人员为中心，构建人员分类预测“冒烟指数”模型，实现对涉恐类人员、涉毒类人员等的分类预测预警，输出人员分类的预警指数，辅助业务人员侦查。

8、类案分析思路案例推演模型

基于已构建的知识图谱体系，在不断支撑各类类案的侦查分析之后，最终可形成各警种类案的侦查分析思路模型，沉淀成为类案分析思路模型，在案件侦查过程可辅助各警种类案的侦查推演分析。

9、其它关系图谱挖掘分析模型

根据公安侦查办案的业务需要，可基于公安知识图谱平台，定制扩展

更多的关系图谱挖掘分析功能，包括涉毒人员关系图谱挖掘分析、关注人员活动轨迹图谱挖掘分析、前科人员影响力挖掘分析、城市犯罪人员传播力挖掘分析、犯罪团伙预测发现等关系图谱挖掘分析。

十、智慧农业

（一）背景介绍

智慧农业通过生产领域的智能化、经营领域的差异性以及服务领域的全方位信息服务，推动农业产业链改造升级;实现农业精细化、高效化与绿色化，保障农产品安全、农业竞争力提升和农业可持续发展。目前，我国的农业仍处于分散式、半机械化的阶段，存在多种资源利用率低的现象。而且大量的农业数据资料是很分散的、不集中的，很难形成一个相互联系的整体，因此农民及研究学者很难从中获取到直观有价值的信息。知识图谱可将这些离散的信息相互关联，把复杂的农业知识直观地展示给农民、农业技术人员和相关决策者，例如，中国知网推出的智慧农民云平台、IBM推出的Watson农业决策平台等。

（二）典型应用

1、农业信息服务

基于百科、农业专业网站、行业专家录入数据、知乎问答、书籍期刊论文等文献数据等数据源抽取的农业相关数据，可构建“农作物-栽培技术-农产品-营养”等实体关系和特征关系的语义网络，形成农业知识图谱，并进一步向农业相关从业者提供信息检索及问答相关的服务。

（1）信息检索

根据检索词，基于知识图谱将具有同一属性关系的农业知识单元有序联系在一起。如输入专家名字，可以呈现以该专家为中心的知识地

图，涵盖该专家发表的论文、合著者、工作单位、成果、专利等数据，通过点击论文合作者的发现，还能链接以合作者为中心的知识图谱，进而以检索词为中心获取周边，不仅扩大用户的知识检索范围，同时发掘用户的潜在需求。

图5.38 基于知识图谱的农业信息检索[1]

（2）信息问答

基于农业知识图谱建设智能问答机器人，可以实现与用户的智能交互，提升交互体验，降低农业这个专业领域知识获取的难度，例如让机器人回答“有什么”(西瓜有效的杀虫剂有哪些？)、“怎么做”(如何用南瓜嫁接黄瓜？)、“是什么”(玉米是温性还是凉性的食物？)、“什么时候”(胡萝卜什么时候种植？)、“概念判断”(土豆吃多了会上火吗？)。以病虫害知识问答系统为例，其分为前端展示模块、后台处理模块、知识库构建、问答模块四个功能模块。

[1] https://cloud.tencent.com/developer/news/179360

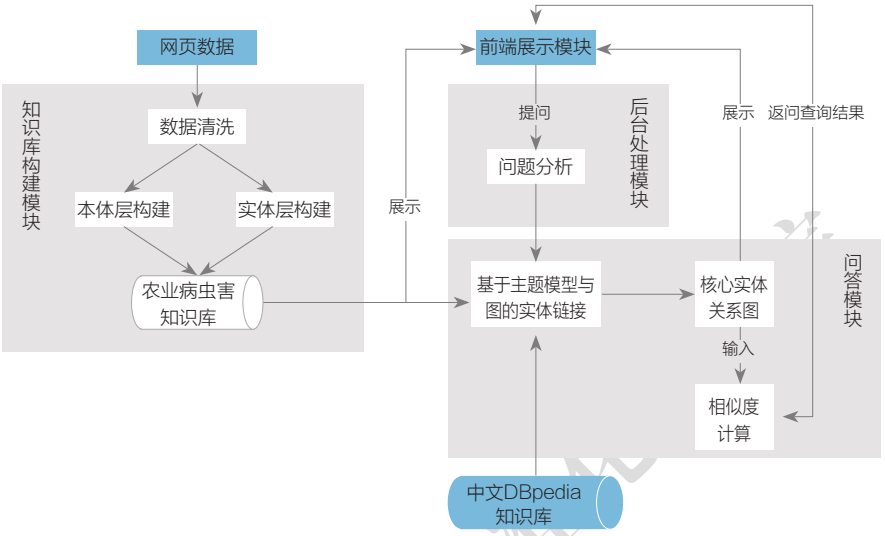

图5.39 农业病虫害知识问答系统架构图[1]

2、产业指导

以农业领域重点产业规模数据库、企业数据库、政策数据库、产业资源数据库、专利数据库、投资数据库等数据为核心，可以从产业、区域、企业等多角度绘制产业图谱并进行可视化分析，从而为农业相关从业者、企业、研究院所提供全产业链、全方位的产业发展现状和趋势，并给出产业发展相关的决策与指导，包括产业布局研究、生产服务和指导、产量统计、市场预警等。此外，知识图谱以知识为单元重新组合后，可发现深层次产业问题，同时能够将理论、实验数据、市场信息、统计数据等进行关联和统一，为重要的决策提供科学依据。以苹果产业为例，该产业本体类目结构图及实体关联结构图如图5.40所示。

[1] 夏迎春. 基于知识图谱的农业知识服务系统研究[D].安徽农业大学,2018.

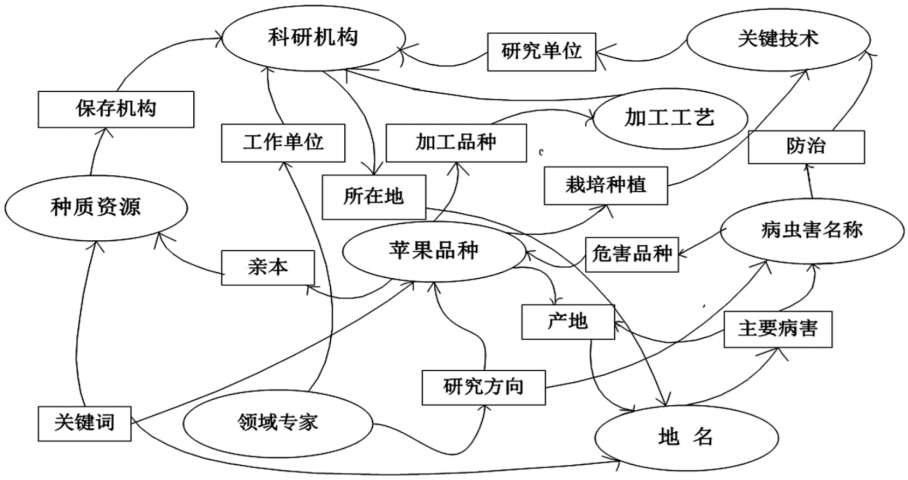

图5.40 苹果产业实体关联结构图[1]

3、农产品推荐

通过农产品知识图谱构建与机器视觉、自然语言处理等技术相结合，可根据客户需求推荐相应的农产品并推送各季节的饮食建议，提升用户饮食的合理性及安全性。例如，和而泰利用自主研发的“拍照荐果”功能，结合其开发的知识图谱平台依据用户拍摄的张照，可以分析出顾客的肤质数据，并基于环境数据、健康数据、果品数据，为顾客提供适合的水果购买建议。

十一、社交网络

（一）背景介绍

社交网络已成为发展最快的互联网应用，社交网络用户不仅仅是信

[1] 陈亚东, 鲜国建, 寇远涛, 等. 我国苹果产业知识图谱构建研究[J]. 中国农业资源与区划, 2017, 38(11): 40–45.

息的接受者，也成为了信息的制造、加工和传播者，通过关注、被关注的方式形成了一张张庞杂繁复的用户关系网。随着社交网络中用户及信息急速增长，依托挖掘社交网络中的海量信息构建知识图谱，推进数据挖掘、网络图论、Web、社交网络、搜索引擎理论等研究，促进它们朝着更智能化、语义化的方向发展已成为新趋势。同时，所构建的知识图谱能够广泛地应用到社交信息检索、社交朋友推荐、社交舆情分析等系统研发，提高这些系统的语义理解能力，向用户提供更全面、更准确信息。

（二）典型应用

知识图谱在社交网络常用于使社交网站、互联网应用等可以成为个性化用户社交环境并发挥价值（涉及搜索、推荐、娱乐、社交、商务等）的场景，并以数据信息、资料、图谱等形式开展商业应用，如数字社区个性化、社会化信息互荐、娱乐、消费等。

1、社交搜索

社交搜索是一种用户可以执行针对社会化媒体内目标联系人的搜索。社交搜索具有以下特点：注重社区效应；聚合话题；基于“情景搜索”；以用户体验为中心；完整的用户识别体系；超强的用户粘性；好友导向；个性化推荐；构建兴趣图谱；搜索结果的人性化与精准化。基于知识图谱的社交搜索除了具有社交搜索的特点外，还具有其自身的特点：

- 能够搜索特定爱好的人：通过条件细化，用户可以找到有特定特征的人或组织，认识到更多的人。
- 能够搜索指定的照片：基于知识图谱的搜索可以通过语义在Facebook等中搜索用户需要的照片。基于语义理解的图片搜索可以让用户在庞大的数据信息流中，快速找到用户需要的图片。
- 能够搜索具体产品的评价：社交图谱搜索可以给用户提供其所需要的某种具体产品的评价，为用户提供自己好友对某一产品的看

法或评价。

2、餐饮娱乐

在餐饮娱乐方面，可通过充分挖掘并关联各个场景数据，结合自然语言处理、计算机视觉等人工智能技术驱动机器解读用户评论和行为数据，理解用户在菜品、价格、服务、环境等方面的喜好，构建人、店、商品、场景之间的知识关联，从而形成餐饮娱乐知识图谱。同时，通过不断挖掘数据背后的知识，可以推动智能化的本地生活服务拓展，实现面向用户的餐饮店铺推荐及面向店铺管理者的辅助决策等应用。

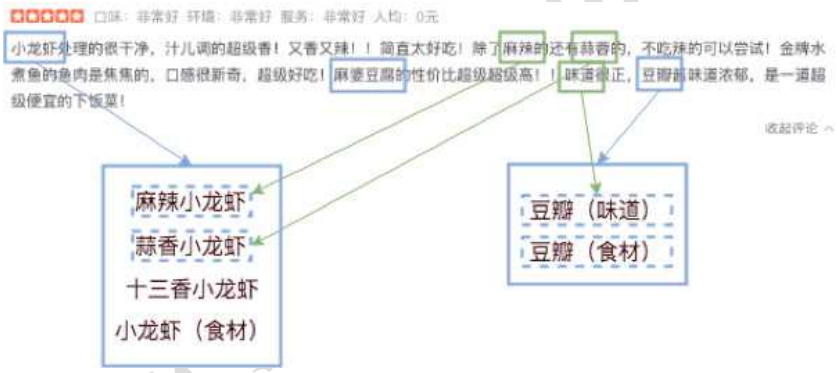

图5.41 知识图谱在餐饮娱乐中的应用示例[1]

（1）智能搜索

知识图谱可以从多维度精准地刻画商家，并在美食搜索、娱乐搜索、旅游搜索等搜索场景中应用，为用户搜索出更适合的店。基于知识图谱的搜索结果，不仅具有精准性，还具有多样性，例如：当用户在美食类目下搜索关键词“鱼”，通过图谱可以认知到用户的搜索词是“鱼”这种“食材”。因此搜索的结果不仅有“糖醋鱼”、“清蒸鱼”这样的精准结果，还有“赛螃蟹”这样以鱼肉作为主食材的菜品，大大增加了搜索结果的多

[1] https://tech.meituan.com/2018/11/22/meituan-brain-nlp-01.html

样性，提升用户的搜索体验。并且对于每一个推荐的商家，能够基于知识图谱找到用户最关心的因素，从而生成“千人千面”的推荐理由。

图5.42 基于知识图谱的智能检索结果示例[1]

（2）辅助决策

通过机器智能阅读和分析每个商家的每一条评论，可以充分理解每个用户对于商家的感受，针对每个商家将大量的用户评价进行归纳总结，从而可以发现商家在市场上的竞争优势/劣势、用户对于商家的总体印象趋势、商家的菜品的受欢迎程度变化等信息。进一步，通过对用户评论全方位细粒度分析， 可以细致刻画商家服务现状以及对商家提供前瞻性经营方向，最终智能化指导商家精准优化经营模式。

[1] https://tech.meituan.com/2018/11/22/meituan-brain-nlp-01.html

十二、新零售

（一）背景介绍

新零售，即个人、企业以互联网为依托，通过运用大数据、人工智能等先进技术手段并运用心理学知识，对商品的生产、流通与销售过程进行升级改造，进而重塑业态结构与生态圈，并对线上服务、线下体验以及现代物流进行深度融合的零售新模式 以阿里巴巴为例，立足于当下的电商场景从认知用户的需求出发，充分利用知识图谱技术，构建起了一个全新的电商知识图谱：新零售电商认知图谱。目前电商认知图谱是一个以用户需求为中心，连接商品、用户、购物需求，以及各类开放领域知识、常识的大规模语义网络。不仅包含了以商品为中心的知识图谱（Product Graph），还包含了以用户需求的显式节点概念为中心的知识图谱（Concept Net）。形成了以概念、商品、标准产品、标准品牌等为核心，利用实体识别、实体链指和语义分析技术，整合关联了例如舆情、百科、国家行业标准等9大类一级本体，包含了百亿级别的三元组，以人货场为核心形成了巨大的知识网。

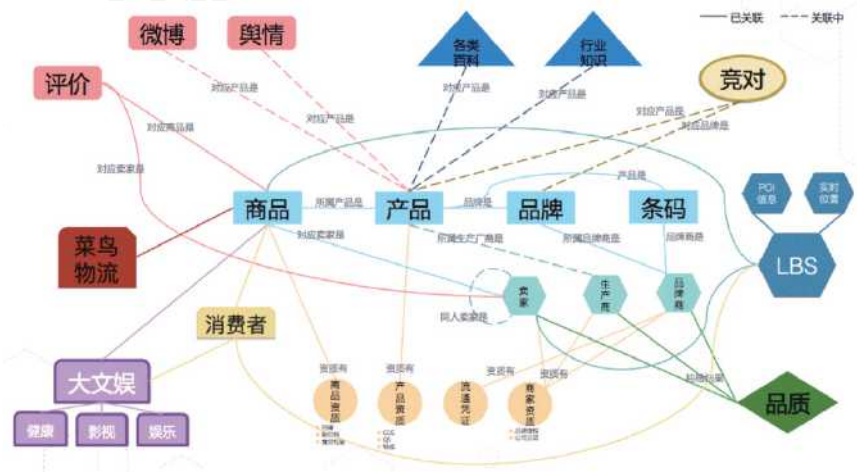

图5.43 电商知识图谱示意图

（二）知识图谱构建过程

电商平台最大的挑战是从日益增长的海量商品（数十亿）中挑选出的一个小的子集（几十或上百）展示给用户，以满足用户个性化的购物需求。近年来电商搜索、推荐算法已经取得了长足的进步，但这些算法依然存在一些问题而为人诟病，例如在商品搜索中“不智能”的体验时有发生。而在商品推荐中，重复推荐、缺少新意等也是经常被用户所诟病的。为了打破这个隔阂，让电商搜索、推荐算法更好地认知用户需求，阿里电商知识图谱将用户需求显式地表达成图中的节点，概览如图5.44所示。

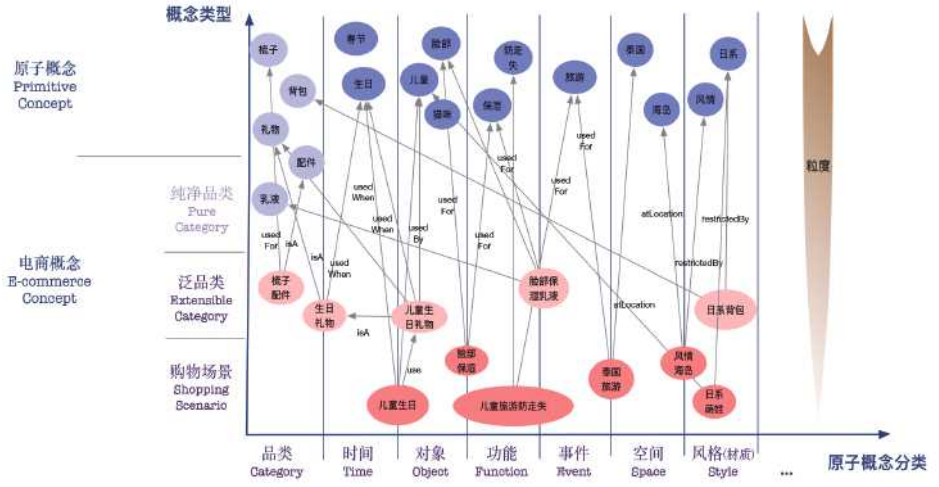

图5.44 阿里巴巴电商认知图谱概览

在该电商认知图谱中，目前一共定义了19种关系类型，并用三元组表示所有节点之间的关系。这些关系包括“is_related_to（相关）”、“isA（是一种）”、“has_instance（有实例）”、“is_part_of（是一部分）”等。其中对电商场景业务直接用途最大的关系是电商概念到商品之间的关联关系：例如一个购物场景“儿童防走失”所对应的商品到底是哪些；以及电商品类之间的上下位关系：例如“舞蹈裙”是一种“表演服”。

关系	名称	节点类型	举例
r/relatedTo	相关	电商概念 -> 电商概念 / 商品	烧烤叉 -relatedTo-> 烧烤架
r/use	使用	原子概念（受众 / 事件 / 功能）-> 电商概念	烧烤 -use-> 烧烤叉
r/usedFor	被用于	电商概念 -> 原子概念（事件 / 功能）	烧烤叉 -usedFor-> 烧烤
r/usedBy	被 .. 用	电商概念 -> 原子概念（受众）	儿童礼物 -usedBy-> 儿童
r/isA	是一种	任何 -> 任何	儿童袜子 -isA-> 儿童用品
r/partOf	是 ... 的一部分	原子概念 -> 原子概念	鸡翅 -partOf-> 鸡
r/atLocation	位于	任何 -> 原子概念（空间）	泰国旅游 -atLocation-> 泰国
r/synonym	同义词	任何 -> 任何	出行 -synonym-> 出门
…	…	…	…

表5.1 认知图谱关系举例（部分）

（三）典型应用

1、电商搜索中的应用

电商认知图谱在原有的电商知识体系的基础上融入了大量概念和知识，为商品搜索引擎的智能化升级带来了新的动力。以“国产冰箱”关键词搜索为例，传统搜索引擎中可能因为没有“国产”这个概念而无法得到全面的结果，而在知识图谱体系中“国产”可作为一个品牌类概念的一个属性，每一个冰箱的品牌都将能查询到这一属性，从而解决该问题。其他的应用包括：1）在搜索结果页中插入和搜索词相关的主题形式的卡片，猜测真正的用户需求，这里的主题即为认知图谱的电商概念。2）进行搜索词关联提示，例如，当用户搜索了烧烤架和木炭之后，因为烧烤架和木炭都是户外烧烤场景下所需要的商品，可提示“户外烧烤”，如下图所示。

图5.45 手机淘宝搜索框下搜索词提示

2、电商推荐中的应用

认知图谱在电商推荐也得到了广泛应用。下图是在手机淘宝app中首页“猜你喜欢”中的主题卡片推荐。主题推荐区别于商品推荐，将认知图谱中的电商概念包装成一个主题卡片的形式，穿插在商品信息流推荐页面中呈现给用户，例如“烘焙大全”，当用户点开这个主题，就会进入另一个页面，包含了烘焙所需的各类商品。如果推荐准确，将大大提升用户体验，仿佛淘宝是一个导购员，猜中了客户的需求，并提供一系列不同商品以供选择，会让用户觉得很舒心。另一个重要的应用是推荐理由，认知图谱的电商概念是用户需求的表达，又是两三个词组成的短语，本身就是一种简洁有力的推荐理由，在商品推荐中加入电商概念作为推荐理由，可以帮助提升用户体验，让用户更好地接受推荐的商品。

图5.46 手机淘宝首页猜你喜欢主题卡片推荐

十三、智慧外交

（一）背景介绍

智慧外交是指通过事件分析的手段，从历史、政治、经济、军事、文化等多个层面对各个国家的关系进行定量分析，提供智能化的外交关系研判和外交决策支撑。依托公开媒体、互联网及内部信息等海量资源数据，综合运用知识抽取、实体归一、知识计算与推理等技术，可以构建形成关于国家和事件的外交知识图谱，实现对事件的性质和影响进行分析，并在事件信息量化处理的基础上，提供复杂、高维时序数据的关联分析。此外，还可通过随机过程建模及基于机器学习的趋势预测，实现突破事件的热点动态跟踪、趋势预测和危机预警等功能，给出国家安全的量化风险评估，为战略层面安全形势提供模型评估支撑，有效提高国家安全、外交等部门在国家战略形势实时研判能力，知识图谱构建过程如图5.47所示。

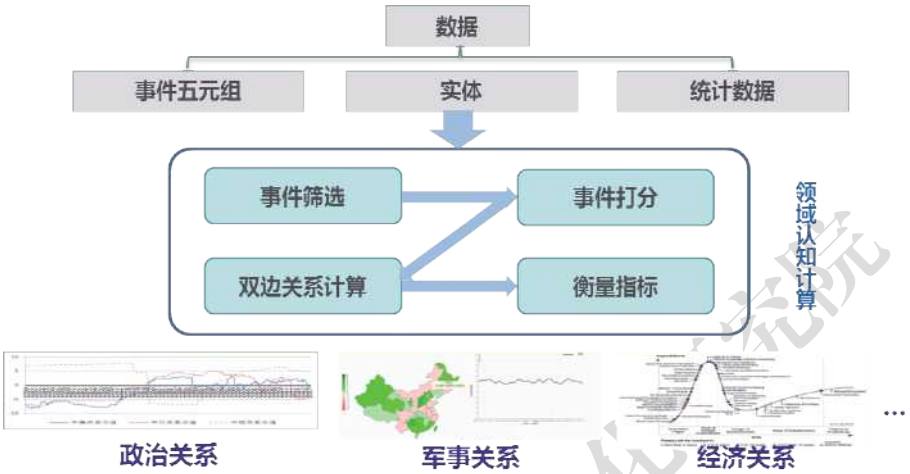

图5.47 国家间关系分析的一般流程

（二）典型应用

1、全球新闻事件整合

GDELT（Global Database of Events, Languages, and Tone，全球新闻整合数据库）作为该方面的代表系统，是美国佐治亚城大学建立的面向全球的、海量新闻信息分析项目，项目通过收集全世界的广播、报纸、网络新闻及社交媒体信息，建立了面向全球的海量新闻事件库及知识库；通过构建面向海量事件数据的多样化可视化服务，实现了海量情报信息的可视化展现，为理解海量情报信息提供有效支持；通过对海量政治事件的建模分析及关联挖掘，实现了对国际政治事件的定量研究、趋势预测、分析重演等。在学术界，GDELT被应用于“新加坡新闻事件可视化分析及预测”及“冲突预测”等。

2、全球多模套信息整合

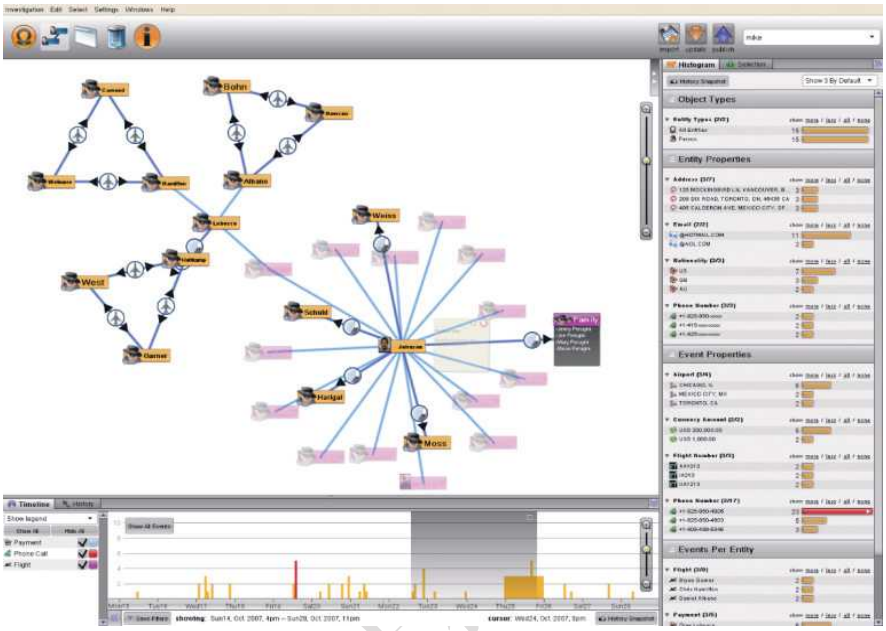

图5.48 Palantir系统锁定嫌疑人分析功能界面

Palantir系统作为该方面应用的代表，可以通过整合各类人员信息，包括地址、邮件、电话、消费、旅行等各类数据建立可视化关系网络，来分析各类人员信息和事件之间的关联关系，构造完整的证据链条。Palantir的核心技术一是全量的数据融合和关联存储，通过本体论抽象出各类数据资源的关系进行映射存储和关联索引；分析模型的动态组合和快速定制，针对不同分析场景可以快速组建分析模型；面向人机共生的大数据可视化技术，构造全局分析链，把人脑决策和数据可视化探索进行深度融合。

十四、智慧城建

（一）背景介绍

智慧城建是运用高新技术手段感测、分析、整合城市运行核心系统的

各项关键信息，从而对包括民生、环保、公共安全、城市服务、工商业活动在内的各种需求做出智能响应。当前，我国正处于快速城市化的阶段，伴随着城市规模的扩大和实力的增强，城市病也如影随形。特别是大城市、超大城市这些人口高度聚集的城市区域，资源紧张、环境污染、交通拥堵、住房拥挤、公共产品及服务短缺等社会问题已成为城市进一步发展的瓶颈和短板，综合分析大城市的综合承载力现状及增强大城市承载力的途径逐渐得到深入关注和研究。

城市建设信息涉及基础设施、交通出行、医疗服务、公共安全、教育服务、生产制造等公众生活相关的方方面面，具有高维度、异构性、多样性、稀疏性和海量性等特征。整合上述城市运行核心系统的各项关键数据是智慧城建必须面临和解决的问题，知识图谱对于海量高维复杂数据的处理能力则能很好地为该问题提供解决方案，知识融合技术可对大量数据进行动态分析和整合，同时利用知识建模和知识计算可将散布在城市各个角落的数据连接起来，对城市进行全域的即时分析、指挥、调动、管理，从而实现对城市的精准分析、整体研判、协同指挥。

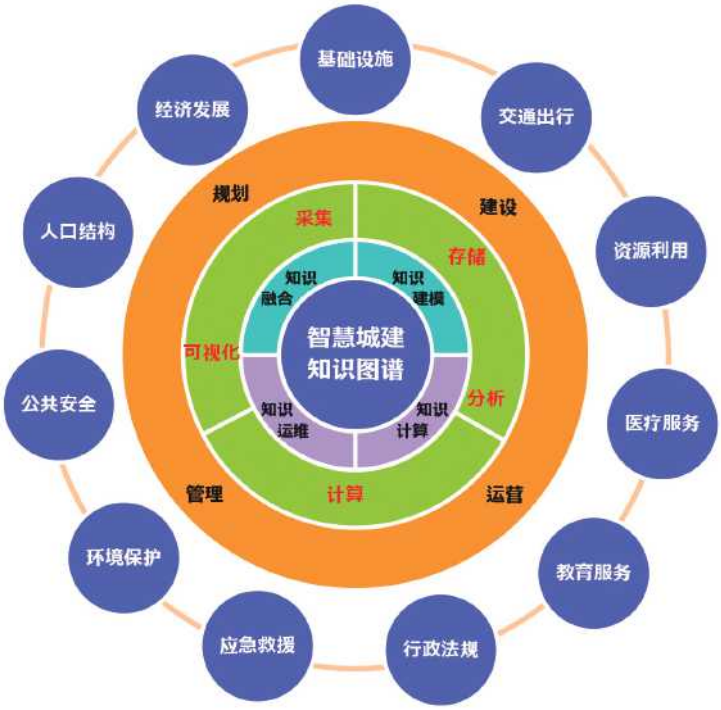

图5.49 智慧城建结构概念图

（二）典型应用

1、城市人口承载力知识图谱

城市人口承载力从两方面分析，一是建设用地供给规模以及建设用地人口数量限制，通常从建设用地人口总量和人均占用建设用地指标进行分析，它确定的是既定建设用地规模下，适宜的人口容量限度；二是城市社会就业保障能力所承载的人口限度。通过知识获取、知识表示、知识存储等技术，从海量、多维、异构数据中汇集上述城市人口承载力相关的实体以及属性关系，构建城市人口承载力知识图谱模型，进而探索影响城市人口承载力的核心因素，构建最优城市人口结构模型，并对未来城市人口态势进行推理和感知。

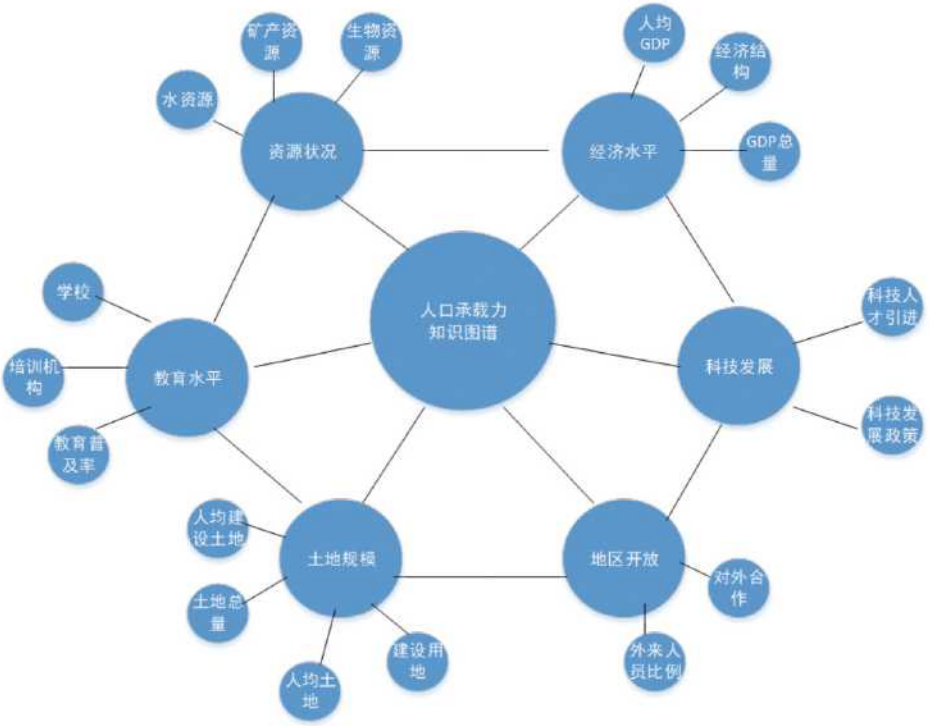

图5.50 人工承载力知识图谱示意图

2、城市市政公用设施知识图谱

以集合城市市政公用设施属性信息为目标，以监管城市公用设施水平为中心，利用知识图谱技术建立围绕城市人口、面积、维护建设投资、城市供水、燃气、供热、道路和桥梁、排水与排污、园林绿化、环境卫生等实体关系，抽取上述关系之间的关联性，搭建层层公共设施知识图谱。通过刻画全面的城市市政公用设施画像，推算用水普及率、燃气普及率、人均道路、人均公共交通、污水处理率、园林绿化率等数据，并通过汇集市政公用设施的基本信息等静态信息和公共设施更新、规划等新闻、公告等动态信息，从城市市政公用设施水平分析出发，探索如何调整、优化现有公共设施布局，更好地服务于城市发展规划。

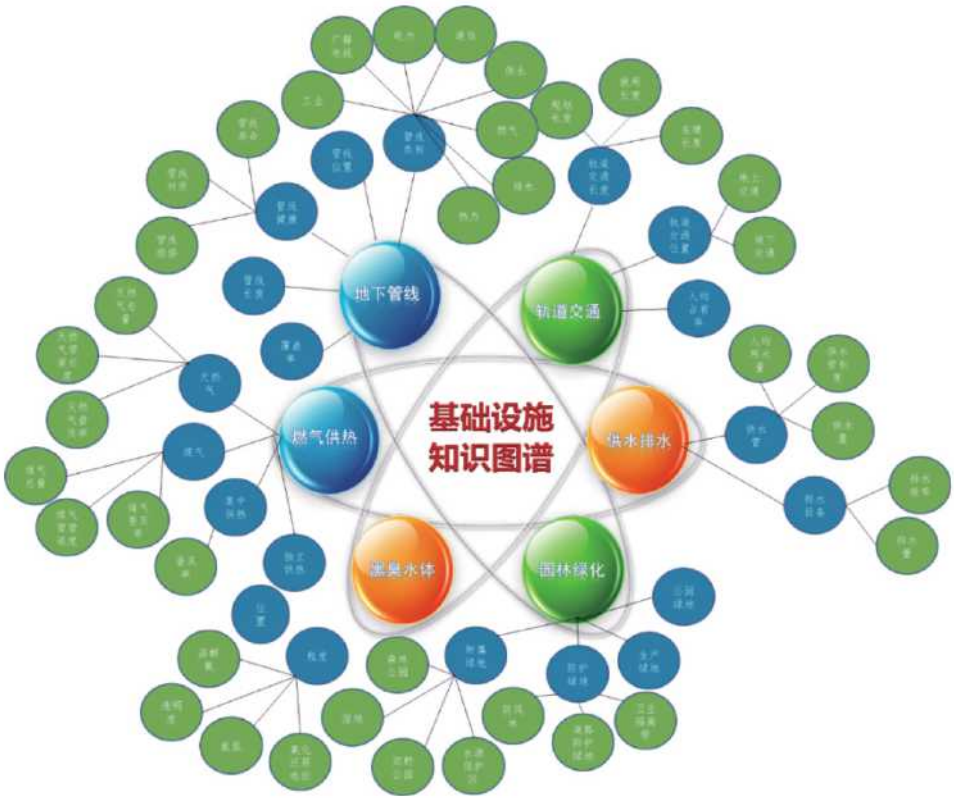

图5.51 城市市政公用设施知识图谱

3、城市经济承载力分析

城市资源经济承载力表达的是在一定的经济技术条件和城市区位条件下，城市的经济价值产出能力，反映城市的经济规模和增值潜力。知识图谱从产业链上下游关系、投融资关系对城市经济发展进行全面、动态的监控，能够实时对城市经济形势进行评估。此外，知识图谱可总结城市经济发展各领域的梯度构成，依据经济增长速度对城市经济未来规划作出相应调整，并分析、提炼推动城市经济发展的重要因素，依据历史数据预测未来城市经济高速增长的发展方向。

十五、其他

目前，语音识别、视频识别、图像识别等感知智能相关技术已逐渐成熟，知识图谱通过与上述技术的融合应用在智能家居、智慧传媒、智慧军工等领域也取得了一定的实践成果。例如，知识图谱作为智能家居的大脑，可以实现用户和服务的连接及设备和服务的连接，让智慧家庭场景中用户获取服务更智能。通过在场景生态中集成所有关联家电的知识图谱、海量家庭用语资料库代表可为未来人与设备、服务的和谐共存创造条件。

第六章 知识图谱存在的挑战

一、数据相关的挑战

数据是知识图谱的基石,知识图谱的数据来源主要有两种：自有数据（自身采集 或拥有等）和外源数据。自有数据由建设者因特定的业务需求而采集获取或经过特定加工处理形成且被自己所拥有的数据；外源数据是通过互联网爬取、开放共享或和数据交易所获得的数据。实际应用中，数据来源多样性造成数据标准不统一、数据质量差，导致多源数据歧义、噪声大、数据间关联关系不明确等问题。数据歧义和关系不明确对知识图谱构建和推理形成了巨大的阻碍，对知识图谱应用成效的提升和技术的进步提出了巨大的挑战。

1、多源数据的歧义、噪声大

多源数据依托于不同的业务目需求和工具产生，行业跨度大且缺乏统一的业务数据模型标准和描述规范，异构、歧义、噪声大特征显著，尤其是通过互联网获取的网络知识。对于知识图谱而言，数据歧义体现为实体、概念的属性描述缺乏精确性、一致性或者实体间关系错置、扭曲，导致图谱结构混乱及语义描述相互冲突、推理机制紊乱、推理结果扭曲失真，最终对图谱应用成效产生巨大副作用。噪声大是互联网数据一大显著特征，海量的噪声数据稀释了数据信息内容和价值密度。数据充斥大量的噪声，将会在图谱中产生大量孤立、无用的实体和属性关系，这些实体和关系还会占据大量的存储空间和关系路径，使图谱实体层变得冗余、繁杂，不仅不能加快图谱推理速率、提升推理精度，反而导致存储资源的浪费和图谱推理应用的低效。

2、数据关联性不明确

关联是知识图谱的核心和本质所在，知识图谱依赖于关系连接的方式

将一个个实体、概念进行关联形成关系网以实现知识的表示、推理，完成知识服务引擎。数据间的关联性是否明确、完备是决定图谱实体间关系是否清晰、完整的关键。然而，在实际应用当中，多源异构数据之间的关联性并非严格明确的。其一，数据间的关联性指向类型不明确，具体表现为关联性作用为相关关系还是因果关系，有方向性还是无方向性。其二，即使明确关联性存在，但现有关联性的作用有多强，是单独体现还是联合其他关系共同体现也很难明确。

二、算法相关的挑战

知识图谱技术是对语义网标准与技术的一次扬弃与升华，但知识图谱中的知识抽取、知识融合及知识计算等技术依然面临着许多困难与挑战，亟待学术界与工业界协力去解决。在未来几年时间内，知识图谱仍将是人工智能前沿研究问题。

1、现有技术存在的算法挑战

知识图谱系统从获取、建模、融合、计算等，每个环节均涉及不同的算法，目前各步骤所用到的算法根据现状和需求不同存在不同挑战。

（1）知识抽取

知识抽取是知识图谱组织构建、进行问答检索的主要任务，对于深层语义的理解以及处理具有重要的意义。现有的知识元素（实体、关系）抽取技术与方法往往是在限定领域、主题的数据集上进行，虽然取得了较好的效果，但由于制约条件较多方法的可扩展能力不够强，未能很好地适应大规模、领域独立、高效的开放式信息抽取要求。而且，基于大规模开放域的知识抽取研究仍处于起步阶段，尚需研究者努力去攻关开垦。

（2）知识融合

知识融合对于知识图谱的构建、表示均具有重要的意义。指代消解和实体对齐均是知识融合中的关键步骤，虽然相关研究已取得了较为丰硕的

成果，但仍有广阔的发展空间。

A、指代消解

对指代消解而言，虽然英文指代消解的研究取得了很大的成就，研究方法从基于规则的方法到基于统计的方法，但由于汉语和英文在构词、语法和句型等许多方面都存在较大的差异，因此，汉语指代消解的问题不能完全照搬英文中相关的消解技术，必须针对汉语自身的特点利用汉语有关的知识资源进行指代消解。

B、实体对齐

对实体对齐而言，首先，经过命名实体识别流程获取的实体名通常是有歧义的，也就是说一个实体名通常会指代知识库中的多个实体，比如：网球运动员李娜与《青藏高原》的演唱者李娜并不属于同一个实体。如何将实体名链接到知识库中的正确实体是实体链接需要解决的问题。其次，对通用领域知识库进行的实体链接还会带来额外的挑战，例如实体量巨大对候选实体增加了大量噪音，而且也需要对实体进行相关领域判断。

C、实体链接

在对短文本输入进行的实体链接也会带来额外的挑战。在大多数情况下，输入文本只是输入一个句子，甚至是一个词组。与针对长文本或者文档的实体识别与链接方法不同，短文本输入的上下文信息非常缺乏并且几乎没有共现实体的信息。但是现实生活中，大部分的文本信息都是以短文本的方式存在，这需要着重发掘词语与实体的关系，从而从极少的文本中提取到语义信息。

D、其他

从实体识别到关系抽取、共指消解、实体对齐整个流程的误差会从第一个环节开始层层传播，使得实体对齐最终获得的实体质量急剧衰减。除此之外，各环节的预测任务相互独立且鲜有互动，无法处理全局的依赖关系。而且，目前尚无受到广泛认可的实体链接评测框架，不同实体链接研

究在问题定义、基本假设、评测数据集等多个维度均存在较大差异，难以进行有效的比较。

（3）知识计算

现有的知识推理模型往往需要大量高质量的样本进行训练学习，这需要耗费很大的代价去获取样本。然而，人往往凭借相关先验知识只需少量样本就能快速学习推理。在此过程中，大脑感知外部环境，对感兴趣或待学习的信息保持关注，并通过与已有先验知识的结合快速建立起新的知识，而后经过神经元的加工整理，形成难以被遗忘的长时记忆。目前已有一些用神经网络模拟人脑的学习和推理的研究，但复杂度较高且需要大量训练样本的支撑。另一方面，现有的知识推理实现方式，可以解决大部分领域和业务场景的复杂计算需求，但对于计算资源和存储资源的占用较大，且现有算法效率存在瓶颈，应用到大规模数据集时往往需要较长时间，无法满足许多领域实时、准实时或及时的计算决策要求。

2、算法性能的挑战

（1）算法泛化能力差

目前的算法仍然是基于（实体1、关系、实体2）三元组形式的语义解析，算法严重依赖现有的数据集。当客户有新需求、数据变更和扩充时需要修改数据结构及业务逻辑，因此算法的扩展性差、对客户响应慢、维护成本高已成为算法所面临的一系列问题。同时，随着时间和空间的发展，知识图谱的数据规模不断的增大，数据会动态进行增加、修改和删除，算法若对数据的动态变化不能进行很好的复用和拓展，修改数据结构和算法逻辑会对整个知识图谱的构建过程造成不可估量的影响。

（2）算法鲁棒性差

算法的鲁棒性是指在异常和危险的情况下算法生存的关键，指在被测数据受到干扰时算法得出的结果是否相对稳定。知识图谱是需要建立在海量数据之上的一种应用，在构建知识图谱的关系时数据来自不同的数据

源，常见的数据源包括：公开信息、业务数据、三方数据、用户授权数据等。在构建过程中，从不同来源的数据中抽取的知识可能存在大量噪声和冗余，甚至算法输入的数据可能恶意生成样本和训练数据污染，面对噪声数据的干扰、多源异构数据的复杂性及AI安全所带来的恶意增加的干扰数据，现有算法模型的鲁棒性还有待进一步提高。

（3）算法多样化，缺乏统一的评测指标

现阶段，随着知识图谱构建和应用需求的不断增加，针对知识图谱的各算法技术也在不断的更新和发展，针对知识图谱的歧义性、动态性和非规范性等问题，机器学习、深度学习、自然语言处理等领域都提出了针对性的算法技术。现阶段，在学术研究中上，众多文献中对同一任务不同算法的评测指标都是针对开源数据集的研究，但在实际算法需求中数据的来源、格式、任务需求等存在差异。目前大规模知识图谱的应用场景都是针对特定的任务需求，如何选择恰当的算法并对不同的算法进行统一的评定，进而实现图谱的智能化应用仍然是一个研究重点。因此，对不同的算法进行统一的评测是现阶段存在的挑战之一。

3、对算法可解释性的挑战

可解释性是指判别过程是否可以转化成具备逻辑关系的规则。目前对深度学习领域中的一个显著的问题：随着模型复杂性的增加，模型可解释性按照同样的速度降低。此外，对算法可解释性的要求，各行业、各情况下也有不同。在不同的领域下，对算法可解释性的要求不同；在样本集数量不同时，对可解释性的要求也不同。而且，对于算法可解释性的标准，目前各领域也缺乏一致意见。所以，对算法可解释性的要求、定义和标准细则都有待进一步研究和明确。

三、基础知识库相关的挑战

1、基础知识库融合挑战

基础知识库是构建知识图谱的主要数据来源，包括维基百科、Freebase、YAGO、及国内的Zhishi.me、CN-Dbpedia、X-lore等。但在构建大规模知识图谱时将涉及不同基础知识库数据，而且在不同基础知识库的融合过程中会出现数据不一致、格式不统一、数据质量参差不齐等问题。应该如何对不同基础数据库的数据进行判断，如何建立判断标准，若涉及的相关基础数据库对该信息的描述都不完整，又该如何抉择？若同一个问题，不同基础数据库从不同维度进行描述，又该进行怎样选取，是选择其中认为较为认可的描述知识信息，还是整合两者描述信息？这些问题都是基础知识库融合将面临的问题。

2、垂直领域知识库构建挑战

目前基础知识库多集中在通用百科领域，缺乏垂直领域的知识库。虽然百科知识库会涉及部分垂直领域相关知识，但缺乏专业、成体系的描述，如专业性要求较高的医疗领域、司法领域等。如果想构建比较完整的知识图谱，需要丰富知识的广度和深度，则亟待各构建垂直领域的专业知识库，实现对现有百科知识库的补充和完善。而且，在垂直领域知识库构建过程中，会面临知识获取问题、知识库边界问题及专业知识的正确性验证问题。

3、基础知识库不愿开放的挑战

目前个别企业或机构已建立了自有的领域知识库，由于投入了大量的人力和资金，导致不愿意对社会免费开放，进而形成了封闭域数据，影响了该领域知识的有效利用率。此外，该知识库知识的补充仅限于该机构本身，无法发挥众包的优势，不利于知识的丰富和发展。甚至可能由于该机构的资金、人力等原因，导致该知识库的停止使用和维护，这些因素都将影响垂直领域知识库构建的进程。

上述问题都将影响基础知识库的构建和开放利用，进而影响知识图谱的构建，如何去克服这些问题，打造在百科知识库的基础上建立各垂直领域知识库，形成开放的知识库生态，是知识库构建及知识图谱深远发展面临的重要挑战。

四、开发工具相关的挑战

随着业内构建知识图谱的需求增多，很多信息化厂商和大学开始通过抽象知识图谱构建过程，总结归纳图谱构建相关方法学，尝试搭建图谱构建工具。当前业内这类工具通常会覆盖从Schema构建到构建知识图谱及最后的图谱融合和实体对齐等功能，辅助业务人员构建知识图谱，减轻从业者对图谱业务知识的依赖。但由于知识图谱本身属于人工智能领域较为新兴的技术方向，使得该领域内目前市场上各类知识图谱开发工具的水平参差不齐。对于如何构建出完整、易用、高效的知识图谱构建工具还存在和诸多挑战。

1、贯穿知识图谱全生命周期的平台缺失

而对于业内知识图谱技术有关的厂商还没有建设出来覆盖知识图谱全生命周期的平台类产品来支持知识图谱技术的发展与应用。目前实际构建知识图谱的过程中，对于多数知识图谱生命周期的流程，还处于基于非专业工具，甚至没有可利用的工具的阶段。由于功能模块的短板及专业工具的匮乏，使得贯穿知识图谱全生命周期的平台严重缺失，这不仅降低了知识图谱构建以及搭建应用的效率，也大大增加了由于工具的不专业性使得构建出的知识图谱中存在极大的风险。如何构建出可以贯穿知识图谱全生命周期的平台型产品成为业内非常重要的挑战。

2、算法、工具、专家间人机协同有待提升

随着人工智能技术的发展，目前行业大部分知识谱图相关的厂商声称自己利用机器学习技术在数据中自动抽取并构建知识图谱。但是对于特定

垂直行业来说，由于需要非常缜密的知识逻辑结构及丰富的业务知识尚无法实现，通常会选择算法、工具、领域专家三者进行深度人机协同的方式来进行。由算法将知识从数据中进行抽取和剥离，并将这些知识通过工具进行可视化，并且可供使用者在工具上对知识进行编辑，最后由业务领域专家对已经抽取出来的知识进行二次编辑与业务知识摄入。而如何构建出提高知识编辑效率的工具，其既能深入融合算法与数据接口满足高效地抽取出精确知识的需求，又能为业务领域专家提供良好用户体验，对于构建此类工具的厂商而言，也是一个充满挑战的任务。

3、基于文本的知识图谱构建工具性能弱

大多数实体关系的均蕴含在文本数据中，如何利用好这部分数据成为能否构建出高质量知识图谱的关键。虽然随着人工智能技术的发展，越来越多的算法可以从文本数据中抽取实体，并构建实体间对应的关系，但是通常这类计算需要人工对数据进行大量标注来支撑监督学习的模型构建。为了解决这类问题，最近衍生出远程监督学习的概念来利用Freebase等知识库已经标注好的知识来构建模型，但是对于大多数垂直领域，此类知识库目前还非常匮乏。而且中文领域目前没有类似的开源知识库供中文文本进行类似的计算与建模。如何让图谱构建工具更能利用中文文本数据中汲取知识来构建知识图谱是图谱构建工具的一大挑战。

4、跨语言语系的知识图谱构建工具有待突破

由于语言语系种类的关系，在利用文本数据进行知识抽取时会存在不小的差异。从各语言知识储量来看，英文领域中已经存在的可用于知识图谱的数据相对较多，例如通用知识领域可Wiki、Freebase等。但是由于机器翻译的精度问题及手工实体对齐的工作量巨大，从语序上来看中文和英文通常是“主-谓-宾”的语序结构，而日文是“主-宾-谓”，这就让关系识别的计算方法在各语言间也不能通用。如何针对不同语言构建更适合的知识抽取算法集成到工具中，以及如何构建跨语言语系的知识图谱构建也

急需建设。

5、知识图谱中间件有待建设

在构建好知识图谱之后，如何基于知识图谱构建出符合业务需求的工具也是非常重要的过程。目前基于图谱的诸多应用，例如基于图谱的检索系统、KB-QA等应用的构建都需要很多的人工参与及查询语言规则的指定。如何整理出一个可以基于知识图谱构建应用的方法学，将之利用到一个平台类产品中，并能通过平台以低人工投入并能构建出高质量知识图谱相关应用，将成为知识图谱中间件的建设关键。

五、隐私、安全和伦理相关的挑战

从部署拓扑结构而言，知识图谱技术以数据为核心、数据库为载体的方式来存储，有单机、云平台、集群及其组合的部署方式，结合大数据平台、云平台、业务系统、灾备、网络系统及其与知识图谱之间的通信接口。知识图谱的安全问题，是为了保护其数据内容、存储载体、能够访问知识图谱的系统、平台、网络及之间的接口安全。

1、应用/接口数据安全风险管理

内部业务人员、外部合作伙伴、第三方合作业务系统通过利用相应的系统权限违规窃取/滥用数据。一般来说，内部业务人员拥有比外部、第三方更高的系统访问权限，内部人员对数据的访问和使用虽然经过了账号授权，但没有对其访问和传输内容进行审查容易造成数据流失和敏感信息泄露的问题。外部合作伙伴虽然拥有较低的系统权限，但是通过漏洞或钓鱼进行渗透网络和系统渗透、数据库注入、账号提权、病毒植入等方式也可以访问到数据库中的核心信息，从而产生数据丢失、窃取甚至拖库等风险。第三方系统一般通过接口访问知识图谱存储的数据库，在接口本身安全性、三方系统安全性方面无法做到有效管理。其次，敏感应用/接口缺少监控管理手段，容易造成数据泄露和资源占用。对于这类场景，需要从

流量中知识图谱相关的协议解析开始，审查各个账号、接口获取的数据内容，防止数据泄露和未授权流量的产生。

2、敏感数据发现和风险评估

根据国家《数据安全管理办法（征求意见稿）》第六条，数据安全的风险评估势在必行。在风险评估之前，必要的一步就是对当前的数据进行梳理，分类分级、打标签，并识别敏感数据。然后，对已识别数据的驻留和流转风险进行评估。梳理数据的一种手段就是资产扫描，通过对数据库、数据内容及相应的大数据平台组件进行扫描，识别其中的敏感组件和数据并分类。其次，还要扫描数据传输路径和驻留节点，对流程可视化之后更便于识别风险。

3、数据安全异常行为分析

行为异常一般包含的场景：内部业务人员违规访问内部系统造成数据泄露/滥用事件；运维人员违规访问数据库/服务器，造成数据泄露事件；第三方通过接口违规提取截留敏感数据，造成数据泄露。从安全运营和运维角度来说，希望能够主动发现潜在的风险，但无奈系统节点繁多、流量庞杂、日志种类多数量大。因此，该类场景下需要从用户行为分析和日志审计的角度来切入，以可视化的方式呈现安全隐患。

4、数据安全事件溯源取证

网络攻击的纵深防御，可以分为事前、事中、事后三个阶段，事前感知与预防，事中拦截并阻断，事后加固和溯源。在进行攻击溯源时，安全或IT部门在遇到数据安全事件时由于攻击手段的多样性经常会缺少有效溯源手段；而各种设备、系统、数据库日志各自独立，无法关联分析，无法定位事件源头；并且对于海量日志的搜索，传统安全产品检索分析性能无法满足需要。因此在溯源取证环节，对日志关联分析、攻击者画像、发现攻击链条等方面，由于知识图谱的数据结构、日志以及系统拓扑有别于传统业务系统，都将对安全系统的构建提出挑战。

5、知识图谱的4A安全管理

知识图谱作为底层的数据服务，为多种应用（如各类搜索引擎、对话系统）和各类接口提供服务或数据。在认证、账号、权限和审计方面均需统一管理，也就是4A安全管理,其中审计管理（Audit）全面记录用户在知识图谱相关系统与接口的登录行为和操作行为。基于图谱的异常行为定义和规则，实现知识内容获取、数据库操作、相关组件使用的有效审计。

6、模型与图谱内容质量安全

知识图谱构建技术中最重要的一个环节就是知识获取，包括实体抽取和关系抽取等，而且最终提供服务的知识图谱也高度依赖于这两个技术要素，这两类技术决定了知识图谱内容质量的好坏。为了保证输出高质量的图谱，并且维持这一质量要求，不仅需要从系统、平台、数据库、网络维度保护模型本身的安全性，防止训练好的模型文件被破坏，而且还需要保证输入数据的安全性，从模型的训练数据到抽取好待入库的知识结构，避免如恶意代码或原有数据、关系的替换，以免篡改行为导致知识图谱的质量下降。

7、其他挑战

数据时代的安全架构：以数据为中心的审计与保护（Data-Centric Audit and Protection,简称“DCAP”）是由Gartner提出的术语。它强调特定数据本身的安全性，弱化了周边环境的安全考虑。DCAP主要的优点之一是将数据安全应用于待保护的特定数据片段但不影响正常业务，数据保护与企业战略保持一致。而目前的安全系统与建设方案都是全方位的重型防护体系，势必对业务流畅性产生一定影响，其中包括数据的分类发现与安全策略、审计和行为分析与告警、数据的保护等。

六、测试认证相关的挑战

知识图谱相关技术及系统的测试评估作为知识图谱发展中的重要环节，国内外不同机构正在推动该方面的研究工作，其中全国知识图谱与语义计算大会作为中国中文信息学会语言与知识计算专委会定期举办的全国年度学术会议每年发布主题测评任务，促进国内知识图谱领域的技术发展，以及学术成果与产业需求的融合和对接。此外，美国伦斯勒理工学院Tetherless World Constellation研究所围绕知识图谱质量评估正在研制知识图谱测评系统，以检测和评估大规模异构知识图谱中存在的不连续性及潜藏错误标签。目前，该测评系统架构图如图6.1所示，并在一项大规模生物学知识图谱上完成了测试。整体而言，知识图谱的测试认证相关研究仍处于起步阶段，面临以下挑战：

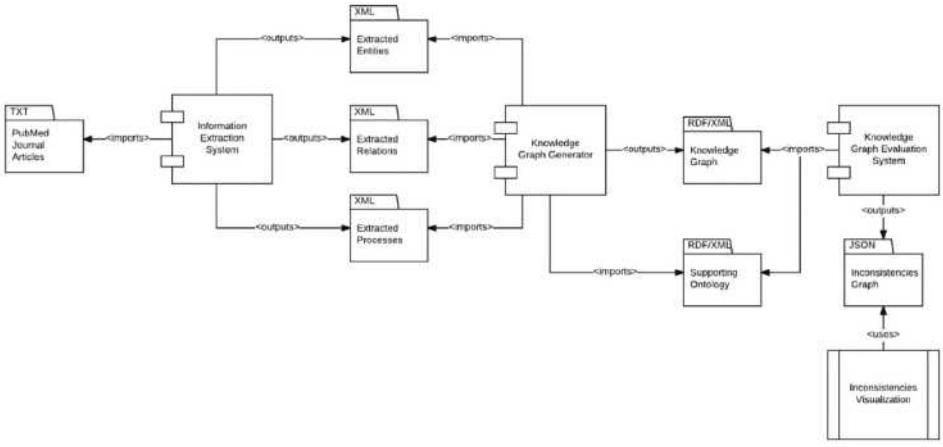

图6.1 美国伦斯勒理工学院知识图谱测评系统架构图

1、知识图谱测试与评估涉及环节众多

知识图谱测试与评估不仅涉及知识图谱输入数据、整体性能、平台功能等，还涉及知识图谱构建过程中知识获取、知识表示、知识存储、知识融合、知识建模、知识计算等各环节。其中，知识获取与知识融合环节测

试评估获得的关注度较高，知识融合环节中的质量评估主要是对知识的可信度进行量化，保留置信度较高的并舍弃置信度较低的，有效确保知识的质量。此外，2019年全国知识图谱与语义计算大会就面向中文短文本的实体链指任务、人物关系抽取、面向金融领域的事件主体抽取、公众公司公告信息抽取等多个方面发布了评测任务。

2、知识图谱测试与评估方法缺失

知识图谱测试与评估方法较少，且集中在知识图谱构建的特定环节或特定问题，尚未形成完整测评体系，无法指导当前各企业及相关用户的使用。美国伦斯勒理工学院发布的研究报告中针对实体类型不匹配、事件类型不匹配、实体事件二元性、上下文约束等知识图谱不一致性测评进行了探讨；德国Philipp Cimiano等专家针对知识图谱优化的现有评估方法从回顾性评估、计算性能等方面进行了总结，但数量有限；针对知识融合中质量评估问题，Mendes等人在LDIF框架基础上提出了一种新的质量评估方法（Sieve方法），支持用户根据自身业务需求灵 活定义质量评估函数，也可以对多种评估方法的结果进行综合考评以确定知识的最终质量评分。

3、知识图谱测试与评估标准空白

当前知识图谱相关标准较少，且尚无测试与评估标准发布，缺少获得业内一致认可的共性测评指标与方法。德国Philipp Cimiano等专家在文献中对知识图谱相关优化算法性能基于DBpedia、Zhishi.me、Open Cyc等数据库从精度、召回率、准确性、精度与召回率曲线下面积、ROC曲线下面积、均方根误差等指标进行了比较。在全国知识图谱与语义计算大会组织的评测任务中，增加了F1-Measure相关指标，并给出了具体的计算公式。但各项指标是否能够覆盖知识图谱测评需求还有待论证，而且知识图谱构建过程中各环节指标也待明确。

4、标准测试集建设迟滞

标准测试数据集作为知识图谱测评的重要基础，高质量的测试数据集

不仅有利于降低知识图谱相关系统的开发成本，也有利于多知识图谱产品间的横向对比，提升测评结果的公平性。DBpedia、Open Cyc、NELL等国外开源数据库及Zhishi.me、PKU-PIE、THUOCL、CN-DBpedia等国内开源数据库对知识图谱的发展起到了重要支撑作用。此外，TAC-KBP、MUC、全国知识图谱与语义计算大会等发布的测评任务中也会附相应的测评数据集。同时，OpenKG作为中国中文信息学会语言与知识计算专业委员会所倡导的开放知识图谱项目目前也已公布92项开源数据集。但各项数据集多是相关公司或组织独立开发所得，而且并非面向知识图谱测试而开发，有待进一步融合与发展。而且随着知识图谱相关产品在各领域的逐步落地与应用，未来面向特定领域的测试数据集需求将不断提升，如何在现有基础上构建相应领域的标准测试数据集也将是一大挑战。

5、测试人员及机构资质要求不明确

知识图谱测试人员需要同时掌握良好的知识图谱构建相关知识及软件测试相关能力。对于第三方测试机构而言，在具备相应测试人员基础上，还需要配备良好的测试环境和检测设备、完备的管理机制，而且测试实验室及测试人员需获得检测认证相关资质，才能够确保出具有公信力的检测报告。目前，由于知识图谱测评相关测试床及测试用例匮乏，而且现有测试人员及机构资质和测试环境多针对其他技术领域，其能否完全覆盖知识图谱测试中的特殊需求缺乏验证基础，未来有待进一步加强该方面技术突破及研究。

七、商业模式相关的挑战

商业模式是通过协调关系，利用资源，获得价值和利益，将这些内容连接起来，从而形成一种企业满足消费者需求的动态系统。一种商业模式包含的要素很多导致商业模式的种类也很多，但一个好的商业模式应具有自身的、别人不能复制的优势。对于知识图谱领域的创业公司，其在商业

模式主要为以定制化解决方案的形式与金融、医疗、工业等领域的客户进行深度绑定，建设知识图谱并积累与客户磨合中的行业经验，不断完善和提高所建立知识图谱的广度与深度，逐步建立技术壁垒并完成项目交付。

1、客户内生驱动力不足，大规模盈利模式不清晰

对于部分行业而言，知识图谱建设主要通过知识工具，提高研究过程和形成决策过程的效率，并不能带来直接的收入增长或者成本下降，也无法直接给企业指明解决问题的具体办法。这导致知识图谱应用无法像自然语言处理、计算机视觉等人工智能技术解决直观的明确目标，得到快速部署于应用，一定程度影响了客户的付费意愿。此外，在客户内部所需知识图谱构建过程中，往往需要技术提供方跟客户进行深度沟通，在综合考虑客户的需求及实施路径基础上完成整体项目，企业成本往往随着项目的增多而呈现线性增长，进而对企业的收入造成影响。

2、知识图谱构建与应用标准化模块推广受限

客户建设知识图谱项目倾向于优化决策，其中的知识信息需求与服务需求较为定制化严重。目前，尽管有部分公司通过销售知识图谱标准化模块的形式试图获得更快的发展，但标准化的模块很难全面满足客户需求，一定程度上限制了模块的开发与推广。而且，当前企业对知识图谱项目建设周期及准确率均有较高要求，在构建过程中仍需要较多的人工干预以达到项目预期效果。未来，如何实现标准化模块与非标准化模块或组件间的协同开发与应用将是企业实现大规模知识图谱快速部署与应用的一大挑战。

3、数据获取受法规约束加大

知识图谱构建需要建立在大量数据的基础之上，不可避免地涉及到个人信息的合理使用问题。而且，实施方基于知识图谱可以很容易地导出数据被采集者不愿意泄露的隐私，例如从公共数据中推导出私人信息，从个人信息中推导出和个人有 关的其他人员（如朋友、亲人、同事）信息

（在线行为、人际关系等）。未来随着法律法规的不断完善，个人隐私将得到更有力的保护，如何在数据正当获取与个人隐私保护中取得平衡，并降低拟开放数据中人工审慎与判断强度，将是知识图谱商业化发展的一大难点。

八、人才相关的挑战

根据腾讯研究院和BOSS直聘联合发布的《全球人工智能人才白皮书》显示，全球AI领域的人才约30万，而市场需求在百万级。全球每年毕业于AI领域的高校学生约2万人，远远不能满足市场对人才的需求。然而，知识图谱跟行业的结合是非常紧密的，还需要领域专家一起合作构建垂直领域的知识图谱，因此对人才的提出了更高的要求和挑战。

1、人工智能相关人才整体缺口较大

《2017年中美人工智能创投现状与趋势研究报告》指出，目前我国人工智能的人才培养已成为一个关键问题，人才缺失可能会对未来AI产业发展产生牵制作用。在美国国家科技委员会发布的2017年人工智能全球大学排名中，前50 名均位于欧美地区，我国的大学无一上榜。而且，我国缺乏人工智能与传统行业的跨界人才，不利于AI在各垂直行业的应用推广。同时，知识图谱的应用还处于探索阶段，没有成熟的系统可以借鉴，也没有经验丰富人才的储备，开发人员更多需要具备快速的学习和业务场景理解能力，结合用户场景和技术实现的难易程度，以较低的成本快速交付。

2、多技术交叉复合型人才缺口大

知识图谱的应用不仅仅是一套算法，而是包括知识获取、知识提取、知识表示、知识存储、知识融合、知识推理、知识检索、知识问答等一系列技术。需要多个技术领域的工程师来协同工作，比如数据库工程师、知识工程师、NLP工程师、开发工程师等。这种协同工作的组织需要对整套技术有了解的高级技术或产品人才进行顶层设计和协调。然而，这种人才

的培养周期长，更多需要从实战中锻炼出来。

3、具备深厚行业背景的人才缺口大

知识图谱构建过程中，有不少场景可能大部分工作是在进行数据源梳理和清洗等数据治理工作，然后再进行本体的建模和图谱的构建。这个过程由于知识图谱应用和行业关联度较高，需要开发人员能迅速地了解业务并和业务人员高效沟通完成客户需求的理解，如果涉及到文本抽取的话还需要熟悉业务领域知识的专家来进行数据标注的工作。这对开发人员和业务人员都提出了更高的要求。而且，当前人工智能高端人才通常集中于软件和互联网行业，而其他行业中部门负责信息化的人员对人工智能概念和技术的理解与掌握总体上看还不准确、不全面，难以支撑企业智能化改造升级。综合而言，现阶段既了解各行业专业技术和发展规律又掌握人工智能关键技术，还能够进行应用开发的复合型人才严重缺乏。

第七章 知识图谱的标准化现状与需求

一、知识图谱标准化现状

（一）国际与国外标准化现状

1、W3C

W3C，全称World Wide Web Consortium，中文名称万维网联盟，是万维网主要的国际标准化组织机构，同时也是万维网领域最具有权威性和影响力的国际中立性技术标准化组织。W3C标准化组织建立成立于1994年，主要宗旨是通过促进通用协议的发展并确保其通用性，对web关键技术进行标准化工作。

在知识图谱领域，W3C相关标准化工作主要集中在语义网知识描述体系方面，研制与发布XML、RDF、SPARQL、RDF Schema、OWL等系列标准，形成了一系列知识图谱中知识表示关键技术标准，如图7.1所示。语义网知识描述技术栈涵盖了知识表示、知识查询、知识推理三部分标准。在知识表示方面，W3C理事会推荐了XML、RDF、RDFS、OWL四项主要技术标准，其中RDF系列标准包括RDF Primer、RDF Test Cases、RDF Concept、RDF Syntax。同时W3C理事会提议的SPARQL Requirements与SPARQL Language标准成为检索和操作基于RDF存储知识图谱。

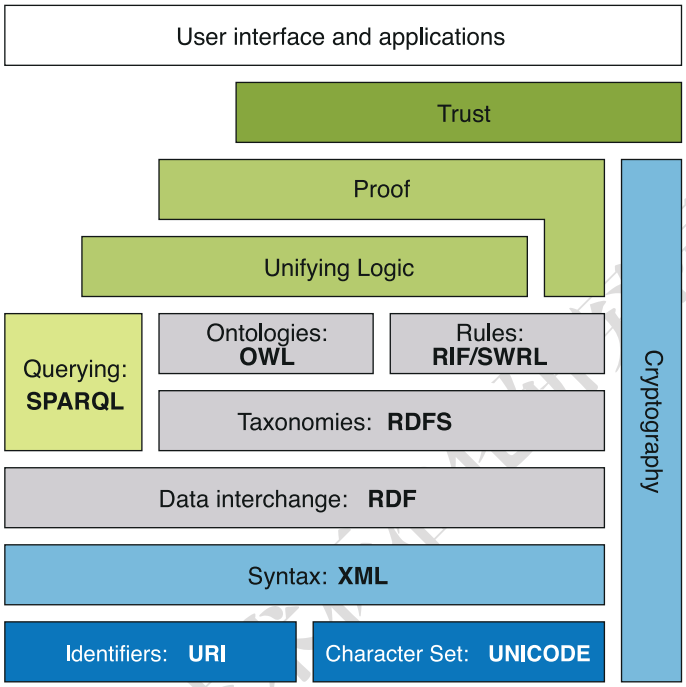

图7.1 知识表示关键技术标准

2、ISO/IEC JTC 1

ISO/IEC JTC 1（国际标准化组织/国际电工委员会的第一联合技术委员会）是信息技术领域的国际标准化委员会，已经在人工智能领域进行了二十多年的标准化研制工作，主要集中在人工智能词汇、计算机图像处理、云计算、大数据等人工智能关键技术领域。于2017年10月批准并成立的JTC 1/SC 42人工智能分技术委员会，主要围绕基础标准(Foundational standards)、计算方法（Computational methods）、可信赖性(Trustworthiness)和社会关注(Societal concerns)等方面开展国际标准化工作。

JTC 1/SC 42 人工智能分技术委员会在2018年8月23日发布了《计算方法与人工智能系统研究报告》第二版，并在其中对知识图谱系统以及知识图谱计算方法与特点、知识图谱行业应用进行了论述，同时分析了知识图

谱系统标准化需求与标准化可能存在的问题。此外，JTC 1/SC 42 在2018年7月3日提出了一项知识图谱相关提案，以期帮助企业开发知识图谱应用，特别是中小企业和初创企业遵循该框架的知识图谱构建方法，帮助知识图谱的数字基础设施供应商了解知识图谱并提供有效的知识图谱工具。

3、IEEE

IEEE标准协会的标准制定内容涵盖信息技术、通信、电力和能源等多个领域。中国电子技术标准化研究院向IEEE标准协会提报的标准提案《知识图谱架构》（Framework of Knowledge Graph，项目编号：P2807）于2019年3月20日正式获批立项，并于2019年8月20日至21日召开IEEE知识图谱工作组（IEEE/C/ SAB/KG_WG）暨标准启动会。中国电子技术标准化研究院物联网研究中心应用技术研究室主任韦莎担任工作组主席，清华大学人工智能研究院知识智能研究中心主任李涓子教授和阿里巴巴集团高级标准化专家王昊分别担任工作组副主席与秘书。

4、MUC

消息理解会议(Message Under-standing Conference, MUC)，主要针对关系抽取概念发布MUC-6、MUC-7评测标准。MUC要求从非结构化文本中抽取信息填入预定义模板中，包括实体、实体属性、实体间关系、事件和充当事件角色的实体。

5、NIST

美国国家标准技术研究院（National Institute of Standards and Technology, NIST）直属美国商务部，主要从事物理、生物和工程方面的基础和应用研究。在MUC-7之后，MUC由美国国家标准技术研究院组织的自动内容抽取(Automatic Content Extraction Evaluation, ACE)评测取代，ACE评测标准从1999年开始筹划，2000年正式启动，其中关系识别和检测任务定义了较为详细的关系类别体系，用于两个实体间的语义关系抽取。ACE-2008包括了7大类和18个子类的实体关系，从2004年开始，事件抽取

成为ACE评测的主要任务。

此外，国际电信联盟（International Telecommunications Union, ITU）2016年开始进行人工智能相关标准化研究。但前尚未发布知识图谱相关标准以及研制计划。

（二）国内标准化现状

全国信息技术标准化技术委员会（SAC/TC 28）负责对ISO/IEC JTC 1工作，在人工智能领域的标准化研制工作主要集中在术语词汇、大数据、云计算等方面。针对知识图谱术，TC28在相关国际标准的基础之上发布了《信息技术 词汇 第28部分:人工智能基本概念与专家系统》、《信息技术 词汇 第31部分:人工智能机器学习》、《信息技术 大数据 术语》三项项基础国家标准，主要定义了知识工程、知识表示、知识获取、本体等部分知识图谱领域专业术语。2019年7月8日，国家标准化管理委员会下达2019年第二批国家标准制修订计划（国标委发[2019]22号），其中由中国电子技术标准化研究院提出的《信息技术 人工智能 知识图谱技术框架》标准（计划号：20192137-T-469）获得立项，并由TC28归口。此外，第一研究所与明略数据联合编写了行业应用领域首个《公安知识图谱标准与白皮书》。

二、知识图谱标准体系

结合知识图谱技术及标准化研究现状，形成知识图谱标准体系框架，有利于围绕知识图谱技术研究及应用的标准化工作开展，面向市场和技术发展需求逐步开展相关标准化工作。同时，做好总体设计和布局，加强基础共性标准的研制，形成系列协调配套的关键标准，提升我国知识图谱标准的先进性和国际竞争力。

1、知识图谱标准体系结构

知识图谱标准体系结构包括“A 基础共性”、“B 数字基础设施”、“C 关键技术”、“D 产品/服务”、“E 行业应用”、“F 运维/安全”等六个部分，主要反映标准体系各部分的组成关系，如图7.2所示。

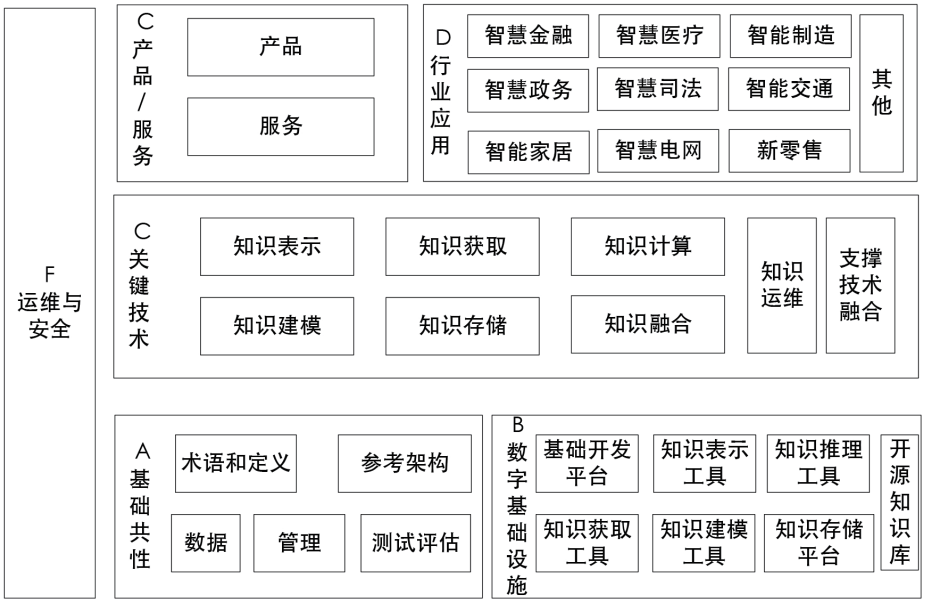

图7.2 知识图谱标准体系结构图

具体而言，基础共性标准包括术语和定义、参考架构、数据、管理、测试评估等标准，规范了知识图谱技术领域中的专业术语及术语间的关系，提出了参考架构，并明确了知识图谱输入数据、管理、测试评估等方面的要求。基础共性标准为体系结构中的其他部分提供支撑，有助于准确理解知识图谱技术内容并促进技术交流和研究。

数字基础设施标准主要针对平台、工具及开源知识库等基础设施3个方面展开，具体包括知识获取工具、知识建模工具、知识存储平台、知识推理工具、数据获取工具、基础开发平台，如基础云平台、基础大数据平台等，为知识图谱构建与应用提供的基础设施、网络系统、数据采集、数

据存储、数据处理等多层次支撑。

关键技术标准主要针对知识获取、知识表示、知识建模、知识融合、知识存储、知识计算、知识运维及自然语言处理等相关支撑技术融合等方面展开，覆盖知识图谱生命周期全链条，为技术开发与应用提供保障。

产品/服务标准主要从产品和服务2个维度展开，主要包括知识图谱领域形成的系统或平台、产品、新服务模式等标准。行业应用标准位于知识图谱标准体系结构的最顶层，面向不同行业的不同需求，对其他部分标准进行细化和落地支撑各行业推进知识图谱技术发展。运维/安全标准包括质量保障和安全保障，为知识图谱安全运行与隐私保护、运行维护与风险管控等提供标准支撑。

2、标准体系框架

依托知识图谱技术特点、标准化需求、当前各行业应用的初步实践及未来技术发展趋势，结合知识图谱标准体系结构图，知识图谱标准体系框架如图7.3所示。

（1）基础共性标准

该类标准主要用于统一知识图谱相关概念，解决知识图谱基础共性关键问题，包括术语和定义、参考架构、数据、管理、测试评估等。针对知识图谱术语相关标准，围绕知识图谱发展现状开展标准制修订工作，统一规范领域内的技术术语和定义，便于领域内做进一步技术沟通；深入研究知识图谱构建相关技术及产业链，开展知识图谱参考架构等标准研制工作；结合知识图谱领域发展需求，开展用于知识库构建、知识图谱构建的数据类型规范、数据标注规范等数据资源相关标准的研制工作；针对知识图谱技术、行业发展较为成熟的领域，提出图谱存储、图谱查询等管理标准；开展知识图谱质量评估原则及模型、算法的泛化性等性能测试及知识图谱成熟度分级要求等标准研制工作。

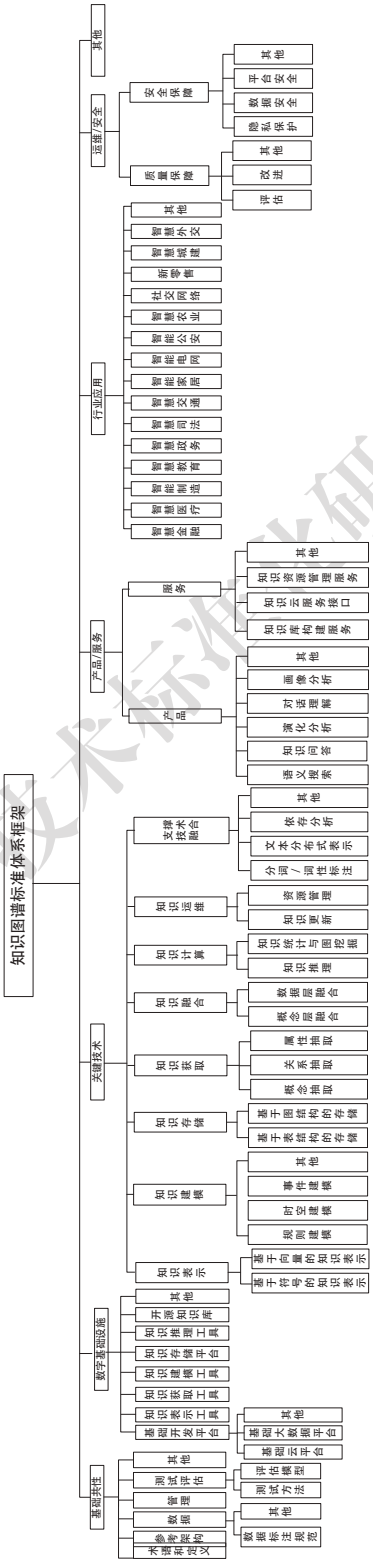

图7.3 知识图谱标准体系框架

（2）数字基础设施标准

该类标准主要针对知识图谱底层平台、工具等数字基础设施进行规范，平台类包括基础开发平台，如基础云平台、基础大数据平台等，工具类包括知识获取工具、知识建模工具、知识存储平台、知识推理工具、数据获取工具等。云平台、大数据平台等技术已有较完善的标准体系，具备一定的研究基础。知识获取工具、知识建模工具、知识推理工具等方面的标准研制，重点围绕工具的架构、功能要求、接口规范、性能测试要求与测试方法、计算能力要求等方面开展研制。

（3）关键技术标准

知识图谱关键技术标准用于对构建知识图谱过程中各环节关键技术进行规范和统一，包括知识获取、知识表示、知识建模、知识融合、知识存储、知识计算、知识运维及自然语言处理等相关支撑技术融合等。其中，由于国内外自然语言处理技术应用及产业发展仍处于起步阶段，自然语言处理与知识图谱融合相关的标准化工作方向需考虑分词/词性标注、文本分布式表示、依存分析、文本内容准确率评估方法、内容相关度分析方法等。

（4）产品/服务标准

产品/服务标准包括语义搜索、演化分析、知识问答、对话理解、关联推荐等通用知识图谱应用系统产品标准以及知识库构建、知识云接口等服务标准。针对系统产品标准的制定，可从系统产品功能性、性能效率、安全性、可维护性等方面进行质量评估。

（5）行业应用标准

行业应用类标准主要是针对知识图谱为各行业所能提供的知识服务角度出发制定相关标准。其中包括各领域根据其特性产生的专用知识图谱，涉及智慧金融、智慧医疗、智能制造、智慧教育、智慧政务、智慧司法、智慧交通、智能家居、智能电网、智能公安、智慧农业、社交网络、新零售等领域。

（6）运维／安全标准

运维/安全标准包括质量保障和安全保障标准，质量保障包括评估、改进等；安全保障包括数据安全、平台安全等。质量评估标准的研究重点围绕知识库构建质量、知识图谱构建质量、平台工具质量、知识获取质量、知识建模质量，以及改进措施、改进情况监督等方面开展；数据安全、隐私保护、平台安全等标准的研究是支撑知识图谱相关系统正常运行的基础，重点集中于安全管理体系、安全监测预警等标准的研制。

三、知识图谱标准化需求

知识图谱的标准化对提升构建效率、确保数据在多领域复用并发挥知识图谱分析和技术价值有重要意义。为了满足各领域复杂业务需求，高效、快速地构建科学、合理、可扩展性强、业务支撑能力强的知识图谱系统及解决方案需要一系列标准化方案支撑，具体包括：

知识图谱构建过程标准化：形成科学、统一的本体建模标准、数据治理标准和数据处理流程，保证数据准确、完整、实时地融合到知识图谱系统中。

知识图谱数据存储标准化：知识图谱的数据存储标准通过明确定义图谱存储能力、存储架构、数据接入标准及查询索引标准，从技术的角度确保知识图谱相关系统拥有强大的知识接入、知识存储与知识检索能力。

知识图谱服务标准化：规范各领域的知识组织方式、知识服务基本内容，保证系统能够全方位地提供知识图谱中有效信息，实现对知识图谱应用的支撑。

知识图谱分析标准化：在充分理解知识图谱分析技术及价值、深入理解数据及业务的基础上，知识图谱可视化分析标准，使图谱中的知识能够以更全面、更丰富、更符合业务需要且便于业务人员理解的方式提供出来，充分发挥知识价值，延伸业务专家智慧。

知识图谱服务安全化：在知识图谱构建及运维过程中，保障数据安全，通过权限控制等保护图谱数据不被篡改、非法增删、复制、解密、显示、使用等。

（一）知识图谱标准制定需求

1、知识图谱架构标准

知识图谱架构是对知识图谱领域所包含的技术、规范、产品及系统工具进行系统性描述的高层概念模型，是对知识图谱相关技术组件范围和相关关系的整体性界定。知识图谱架构标准的制定有助于帮助组织机构、知识图谱基础设施开发商（如云服务开发者、工程师等）更好地理解知识图谱模式架构，同时提供高效的开发工具，有助于垂直行业提取并形成大规模标准知识库，衍生更多知识图谱应用。

2、知识表示与建模标准

从1960年科学家们提出语义网的知识表示方法以来，知识表示与建模成为知识工程领域研究的重要基础问题，早期常用的知识表示有基于框架的表示语言和产生式规则等。为了解决语义理论模型和形式化的语义定义等问题提出了OWL、RDF等描述逻辑语言，这些表示方法被W3C发布成为行业标准。W3C相关标准化工作主要集中在语义网知识描述体系方面，研制与发布RDF、RDF Schema、OWL系列标准，当前大规模知识存储、知识计算、复杂关系建模、时空建模成为知识表示研究的巨大挑战。另一方面随着知识图谱的发展，越来越多的开放图谱和领域图谱开始出现，多个图谱和跨语言图谱之间的融合缺少统一的标准规范。因此应研究更加高效知识表示方法，形成统一完善的开放域和特定域本体构建规范，促进知识图谱得到更好应用和发展。

3、数据标注规范

当前以深度学习为代表的人工智能方法得到快速发展，大规模数据集

的构建和计算能力的提升起到了重要作用，然而大多数计算模型严重依赖于大规模高质量的学习样本。在实际应用中，大规模高质量样本的获取会耗费巨大的人力和时间，针对知识图谱数据集构建，一方面，基于小样本学习的算法模型成为了技术发展的重要趋势，另一方面，如何通过技术标准统一数据标注方法流程、标注质量成为了深入研究的重要课题。通过对知识获取、知识融合、知识表示和知识计算等知识图谱构建重要环节进行标准化建模，定义相关问题数据标注规范，进而降低数据标注的难度并提高数据集的可复用性。

4、测试方法与评估指标体系

针对知识图谱相关技术构建标准的算法测试方法和评估指标体系，有助于学者和行业之间相关成果的横向对比，有助于推动技术发展与创新，例如自2010年以来，在每年度ImageNet大规模视觉识别挑战赛中，来自全球范围的研究团队在给定的数据集上评估其算法，并在几项视觉识别任务中争夺更高的准确性，有力地推动了计算机视觉技术的学术研究和工业应用。然而当前知识图谱领域的大规模数据集以及测试方法与评估指标体系还比较缺乏，因此有待加强研究制定统一知识获取、知识表示、知识融合和知识计算等相关测试方法与评估指标体系。

（二）知识图谱标准化工作建议

结合知识图谱技术研究现状和在各行业的实际应用需求，借鉴人工智能、大数据等领域现有标准化工作实践经验，提出以下知识图谱标准化工作建议。

1、加强知识图谱重点标准研制

知识图谱作为战略性新兴产业的重要组成部分，对贯彻实施《新一代人工智能发展规划》和“互联网+”行动计划具有重要意义。为加快推进知识图谱技术标准工作，提升标准对建设基于知识图谱应用的支撑作用，

建议强化知识图谱领域重点标准的研制工作，包括知识图谱参考框架、知识表示与建模标准、知识图谱相关数据标注规范、知识图谱测评与评估规范等。

2、推进知识图谱标准在各行业中的应用示范

技术标准的制定需要以应用实施为前提，需要针对重点产业、地区、领域开展知识图谱标准试验验证和试点应用工作。建议围绕知识图谱已有标准化工作基础，建立知识图谱构建质量评估体系，优先在知识图谱技术产品服务集聚行业建立重点标准示范，开展标准应用示范工作。另一方面，加强标准对产品质量、服务能力提升和行业管理的支撑作用，开展标准符合性评估、测试工作。

3、参与知识图谱国际标准化工作

组织标准化核心机构和重点企业，跟踪研究知识图谱相关国际标准化进展，深度参与国际标准制定工作，积极贡献国际标准提案，提升自主标准国际化水平；支持相关单位参与国际标准化工作，努力争取工作组召集人等国际标准化组织职务；承办知识图谱相关国际标准化活动，加强我国知识图谱技术标准化组织与相关国际组织的交流与合作，提升国际话语权。

第八章 结论与展望

一、趋势展望

1、技术发展趋势展望

知识图谱主要技术包括知识获取、知识表示、知识存储、知识建模、知识融合、知识计算、知识运维等七个方面，尽管目前已取得了很多成就，但仍在快速演进当中。例如，在知识获取方面，资源缺乏、面向开放域、跨语言及跨媒体等方向的知识抽取正在成为未来的研究方向；在知识表示方面，符号与表示学习的融合统一、面向事理逻辑的知识表示、融合时空间维度的知识表示、融合跨媒体元素的知识表示正在成为未来的研究方向；在知识存储方面，基于RDF知识表示的分布式存储、涉及高适应性的知识存储、基于LOD（Linked Open Data）的知识存储、Hyper Graph的进一步研究和应用正在成为未来的研究方向。

2、应用趋势展望

目前，大规模知识图谱的应用场景还比较有限，其在智能语义搜索、深度问答（包括基于信息检索的问答系统、基于语义分析的问答系统）、演化分析、对话理解等方面的应用也处于初级阶段，仍具有广阔的应用与推广前景。从知识图谱应用发展趋势来看，当前正在从通用知识图谱应用向领域或行业知识图谱应用拓展，如金融、医疗、公安、医疗、司法、电商等，依托知识图谱强大知识库的深度知识推理能力和逐步扩展的认知能力，帮助相关行业从业者对特定的问题进行分析、推理、辅助决策。

3、标准化趋势展望

随着ISO/IEC JTC1/SC42、W3C、IEEE、全国信息技术标准化技术委员会、国家人工智能标准化总体组等国内外标准化组织或机构对知识图谱

标准化的关注与推动，《知识图谱技术架构》等多项知识图谱相关国际、国家标准获得立项或提出讨论。未来，知识图谱领域基础共性及关键技术标准将不断涌现，依托正在研制的知识图谱技术架构等标准，通过聚焦核心标准化需求逐步建立基本的知识图谱标准体系并孵化典型行业中的知识图谱应用标准，形成国际标准、国家标准、行业标准和团体标准良性互动的局面。

二、支持与监管政策相关建议

1、加强知识图谱标准化顶层设计与统筹推进机制

标准化工作是知识图谱技术与产业发展的基础和前提。现有的知识图谱标准化工作还处于起步阶段，涉及领域众多，工作协调难度大，因此首先需要加强标准化顶层设计，以“基础统领、应用牵引”为原则，基于国内外知识图谱技术和应用现状、知识图谱标准化现状，梳理知识图谱产业生态体系脉络，把握技术演进趋势和产业未来重点发展方向，扎实构建满足产业发展需求、先进适用的知识图谱标准体系。建议充分整合领域优质产学研资源，探索建立以企业为主体、产学研相结合的技术创新和标准制定体系，科学谋划、适度超前布局知识图谱标准化工作，营造开放合作的标准化工作氛围。

2、加强知识图谱重点领域标准研制和应用示范

近年来，知识图谱技术进展迅速，可规模化、可商业化产品和应用层出不穷，在很多领域或行业取得了显著落地效果，同时也在过程中涌现出一大批理论与工程问题，迫切需要围绕知识图谱标准化需求，按照“急用先行、成熟先上”的原则，开展术语、参考框架、算法模型、技术平台等重点急需标准的研制，切实发挥知识图谱标准对产业和技术发展的支撑保障作用。鼓励相关行业协会、重点企业参与知识图谱标准宣传、意见征集和试验验证与应用，形成工作合力，建议优先在国家新一代人工智能创新

发展试验区建立知识图谱标准示范基地，提升标准孵化和研制质量，增加标准与技术环境的适应能力，提升重点领域知识图谱标准实施应用成效，探索建立标准研制与科技研发、行业应用高度融合的长效机制并逐渐在全国推广。

3、加快知识图谱标准符合性测试工具和平台建设

知识图谱标准的实际应用效果需要制定科学的评判依据，建立健全知识图谱标准试验与符合性测试评估体系，明确测试范围和评估标准，形成知识图谱标准符合性测试规范与工具，并搭建知识图谱标准的符合性测试平台，提高测试执行的准确率和效率。加强专业化、专职化的标准符合性测试机构建设，鼓励适应知识图谱技术和产业发展且具有领域影响力和公信力的第三方检测认证服务机构发展，重点支持1-2家检测机构成为国家级标准符合性检测单位，提供知识图谱标准化符合性评测服务和技术指导，提升标准符合性检测服务供给能力和专业化水平。

4、加快知识图谱公共服务平台建设

知识图谱的构建离不开数据的积累和优秀的运算性能，然而数据和算力所需大量资金投入也给很多企业设置了一道不低的门槛，对公共服务平台的建设需求也一日显著。加快知识图谱公共服务平台建设，围绕知识图谱技术验证、标准测试数据集开发、构建工具研制、数据开放与共享等需求，加快通用领域及典型行业知识图谱公共服务平台建设，探索知识图谱用户、产品供应商、工具开发方、第三方服务机构间的协同交流机制，鼓励各方基于现有基础不断加强相关服务能力，加深横、纵向产业链的交流合作与需求对接，加快集聚全球知识图谱领域的高端创新与服务要素，实现全方位的产业升级。

5、成立知识图谱产业联盟

为更好地推动知识图相关机构间的接通互通，解决和破解发展难题，推动全产业链整体跃升，建议以政府引导、企业为主、开放合作、公平竞

争为原则，汇聚知识图谱领域优势产业资源，建立知识图谱产业联盟。积极探索联合开发、收益共享、风险共担的契约关系，聚焦产业技术创新发展需求，以共性关键技术研发、重点技术标准研制、核心应用研发为重点，支撑和引领知识图谱领域技术进步和产业发展，提升知识图谱产业核心竞争力。依托产业联盟平台，进一步加强知识图谱产学研用整体产业的技术交流与合作，建立引领行业发展的合作研发平台和科技成果转化基地，打造具有自主知识产权、知名品牌和国际影响力的产品与技术，推动知识图谱技术与实体经济融合发展。

三、技术开发与应用相关建议

1、加强知识图谱核心关键技术支持与突破

突破知识图谱基础理论及关键核心技术瓶颈，以算法为核心，以数据和硬件为基础，以大规模知识库的构建与应用为导向，实施重大关键技术攻关工程，制定知识图谱共性技术开发路线图，重点提升知识获取、知识表示、知识建模、知识融合、知识存储、知识计算、知识运维等多种智能信息处理技术，形成开放兼容、稳定成熟的技术体系。同时，通过梳理和细化标准化需求，以技术突破带动核心技术标准突破。

具体开发建议包括：研究以隐含关系抽取为目标的马尔可夫逻辑网、本体推理的联合推理方法以及跨语言的知识抽取方法，提升知识抽取技术；研究基于实体的、关系的、Web文本的、多知识库的多源异质信息融合，加强认知科学领域人类复杂知识类型的探索，提升知识表示能力；研究基于大规模知识库环境下的并行编程环境MPI、分布式计算框架Hadoop等平台的对齐算法并行化或分布式算法，研究基于数据量、知识库对齐质量以及人工标注三者权衡的众包算法，并将其与跨语言知识库对齐模型有效结合，以此提升知识融合能力；研究知识加工中本体自动构建、本体抽取、本体聚类等问题；研究知识质量评估技术标准或指标体系的构建；研

究知识图谱自动化更新涉及的增量更新技术等;研究面向事件与事件间关系的“事理图谱”构建相关技术等。

2、加强知识图谱优秀解决方案/产品展示与推广

知识图谱当前已在部分企业取得了良好的应用成效，而且形成了一定的实践积累和平台化产品。然而，传统企业受限于内部本领域专业人才对知识图谱构建理论及难点了解有限，无法准确反馈和把握企业对知识图谱构建的需求及意义，有必要通过梳理知识图谱在典型行业的优秀案例并形成案例集，建设开放性实验室，推出优质培训课程等方式加强知识图谱优秀平台或产品的展示与推广，打破知识图谱开发企业、研究院所、高校与各领域企业间的沟通屏障。

3、加强通用和领域知识图谱开放平台建设

开放的通用知识图谱和领域知识图谱平台是推动知识图谱技术在各行业融合应用的重要基础设施，能够避免企业在建设知识图谱过程中从零开始或重复建设，也可降低知识图谱项目实施方的设计开发成本。此外，通过知识图谱在各企业项目的落地实施还可不断反哺通用知识图谱和领域知识图谱的完善，提升实体覆盖率及典型领域模式的完备性，进而实现平台质量与影响力的提升。目前，通用知识图谱尽管已得到了清华大学、复旦大学等高校的大力投入，但领域知识图谱依然匮乏，而且常识知识图谱由于构建难度大依然有待突破。

4、加强知识图谱与其他人工智能技术的融合

自然语言处理、深度学习等技术在构建知识图谱和应用知识图谱时均具有重要意义。然而，无论是基于深度学习的序列标注算法进行高质量的实体识别，还是近些年诞生的基于开放式数据库的远程监督学习，尚未在知识图谱构建及应用领域实现深度融合。此外，机器在知识图谱构建和应用过程中的角色虽然已有所提升，但整体发展速度依然有限。有待加速推动拟在知识图谱方向上投入资源的企业、高效和研究院探索和研究更符合

知识图谱相关业务的计算过程及方法学，以支撑知识图谱相关技术更广泛的落地与应用。

5、加强知识图谱相关人才培养

知识图谱和其他人工智能领域相关技术不同，在知识图谱构建的各环节都需要大量业务知识的摄入及业务专家的参与，例如从实体属性间的关联关系辅助预判实体间的关系种类、通过学习用户在文本数据中抽取实体属性构建三元组数据的抽取逻辑，而自动的从其他新文本中模仿人类动作来抽取新的三元组数据、甚至是通过各实体的属性以及关系的值来自动预判实体对齐或者是实体融合。有待加强同时具备知识图谱与其他领域知识储备的交叉复合型人才，为知识图谱应用与企业需求及业务场景的高度匹配提供保障。