知识图谱技术综述

徐增林[1]，盛泳潘[1]，贺丽荣[1]，王雅芳[2]

(1. 电子科技大学统计机器智能与学习实验室 成都 611731；2. 山东大学计算机科学与技术学院 济南 250101)

【摘要】知识图谱技术是人工智能技术的重要组成部分，其建立的具有语义处理能力与开放互联能力的知识库，可在智能搜索、智能问答、个性化推荐等智能信息服务中产生应用价值。该文在全面阐述知识图谱定义、架构的基础上，综述知识图谱中的知识抽取、知识表示、知识融合、知识推理四大核心技术的研究进展以及一些典型应用。该文还将评论当前研究存在的挑战。

关 键 词 知识融合； 知识图谱技术； 知识表示； 开放互联； 语义处理

中图分类号 TP182 文献标志码 A doi:10.3969/j.issn.1001-0548.2016.04.012

Review on Knowledge Graph Techniques

XU Zeng-lin[1], SHENG Yong-pan[1], HE Li-rong[1], and WANG Ya-fang[2]

(1. Statistical Machine Intelligence & Learning, University of Electronic Science and Technology of China Chengdu 611731;
2. School of Computer Science and Technology, Shandong University Jinan 250101)

Abstract Knowledge graph technology is a critical part of artificial intelligence research. It establishes a knowledge base with the capacity of semantic processing and open interconnection in order to provide intelligent information service, such as search, question-answering, personalized recommendation, and so on. This article first presents a comprehensive study on definitions and architectures of knowledge graphs. Then we summarizes recent advances in knowledge graphs, including knowledge extraction, knowledge representation, knowledge fusion, and knowledge reasoning, with typical applications. Finally, this article concludes with future challenges of knowledge graphs.

Key words knowledge fusion; knowledge graph techniques; knowledge representation; open interconnection; semantic processing

伴随着Web技术的不断演进与发展，人类先后经历了以文档互联为主要特征的“Web 1.0”时代与数据互联为特征的“Web 2.0”时代，正在迈向基于知识互联的崭新“Web 3.0”时代[1]。知识互联的目标是构建一个人与机器都可理解的万维网，使得人们的网络更加智能化。然而，由于万维网上的内容多源异质，组织结构松散，给大数据环境下的知识互联带来了极大的挑战。因此，人们需要根据大数据环境下的知识组织原则[2]，从新的视角去探索既符合网络信息资源发展变化又能适应用户认知需求的知识互联方法[3]，从更深层次上揭示人类认知的整体性与关联性[4]。知识图谱(knowledge graph)以其强大的语义处理能力与开放互联能力，可为万维网上的知识互联奠定扎实的基础，使Web 3.0提出的“知识之网”愿景成为了可能。

知识图谱并非是一个全新的概念，早在2006年，文献[5]就提出了语义网的概念，呼吁推广、完善使用本体模型来形式化表达数据中的隐含语义，RDF (resource description framework)模式(RDF schema)和万维网本体语言(Web ontology language，OWL)的形式化模型就是基于上述目的产生的。随后掀起了一场语义网研究的热潮，知识图谱技术的出现正是基于以上相关研究，是对语义网标准与技术的一次扬弃与升华。

知识图谱于2012年5月17日被Google正式提出[6]，其初衷是为了提高搜索引擎的能力，增强用户的搜索质量以及搜索体验。目前，随着智能信息服务应用的不断发展，知识图谱已被广泛应用于智能搜索、智能问答、个性化推荐等领域。尤其是在智能搜索中，用户的搜索请求不再局限于简单的关键词匹配，

收稿日期：2016－05－15
基金项目：国家自然科学基金(61572111)；中央高校基础科研经费(ZYGX2014J058)
作者简介：徐增林(1980－)，男，博士，教授，主要从事机器学习及其在社会网络分析、互联网、计算生物学、信息安全等方面的研究.

搜索将根据用户查询的情境与意图进行推理，实现概念检索。与此同时，用户的搜索结果将具有层次化、结构化等重要特征。例如，用户搜索的关键词为梵高，引擎就会以知识卡片的形式给出梵高的详细生平、艺术生涯信息、不同时期的代表作品，并配合以图片等描述信息。知识图谱能够使计算机理解人类的语言交流模式，从而更加智能地反馈用户需要的答案[7]。与此同时，通过知识图谱能够将Web上的信息、数据以及链接关系聚集为知识，使信息资源更易于计算、理解以及评价，并且形成一套Web语义知识库。

本文的第一部分将沿着前面叙述，进一步剖析知识图谱的定义与架构；第二部分将以开放链接知识库、垂直行业知识这两类主要的知识库类型为代表，简要介绍其中的几个知名度较高的大规模知识库；第三部分将以知识图谱中的关键技术为重点，详细阐述知识获取、知识表示、知识融合、知识推理技术中的相关研究以及若干技术细节；第四部分将介绍知识图谱在智能搜索、深度问答、社交网络以及垂直行业中的典型应用；第五部分将介绍知识图谱所面临的一些困难与挑战；第六部分将对全文的内容进行总结。

1 知识图谱的定义与架构

1.1 知识图谱的定义

在维基百科的官方词条中：知识图谱是Google用于增强其搜索引擎功能的知识库[8]。本质上，知识图谱是一种揭示实体之间关系的语义网络，可以对现实世界的事物及其相互关系进行形式化地描述。现在的知识图谱已被用来泛指各种大规模的知识库。

三元组是知识图谱的一种通用表示方式，即$G=(E,R,S)$，其中$E=\{e_1,e_2,\cdots,e_{|E|}\}$是知识库中的实体集合，共包含$|E|$种不同实体；$R=\{r_1,r_2,\cdots,r_{|E|}\}$是知识库中的关系集合，共包含$|R|$种不同关系；$S\subseteq E\times R\times E$代表知识库中的三元组集合。三元组的基本形式主要包括实体1、关系、实体2和概念、属性、属性值等，实体是知识图谱中的最基本元素，不同的实体间存在不同的关系。概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等；属性主要指对象可能具有的属性、特征、特性、特点以及参数，例如国籍、生日等；属性值主要指对象指定属性的值，例如中国、1988-09-08等。每个实体(概念的外延)可用一个全局唯一确定的ID来标识，每个属性-属性值对(attribute-value pair，AVP)可用来刻画实体的内在特性，而关系可用来连接两个实体，刻画它们之间的关联。

就覆盖范围而言，知识图谱也可分为通用知识图谱和行业知识图谱。通用知识图谱注重广度，强调融合更多的实体，较行业知识图谱而言，其准确度不够高，并且受概念范围的影响，很难借助本体库对公理、规则以及约束条件的支持能力规范其实体、属性、实体间的关系等。通用知识图谱主要应用于智能搜索等领域。行业知识图谱通常需要依靠特定行业的数据来构建，具有特定的行业意义。行业知识图谱中，实体的属性与数据模式往往比较丰富，需要考虑到不同的业务场景与使用人员。

1.2 知识图谱的架构

知识图谱的架构主要包括自身的逻辑结构以及体系架构，分别说明如下。

1) 知识图谱的逻辑结构

知识图谱在逻辑上可分为模式层与数据层两个层次，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。如果用(实体1，关系，实体2)、(实体、属性，属性值)这样的三元组来表达事实，可选择图数据库作为存储介质，例如开源的Neo4j[9]、Twitter的FlockDB[10]、sones的GraphDB[11]等。模式层构建在数据层之上，主要是通过本体库来规范数据层的一系列事实表达。本体是结构化知识库的概念模板，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小。

2) 知识图谱的体系架构

知识图谱的体系架构是其指构建模式结构，如图1所示。其中虚线框内的部分为知识图谱的构建过程，该过程需要随人的认知能力不断更新迭代。

知识图谱主要有自顶向下(top-down)与自底向上(bottom-up)两种构建方式。自顶向下指的是先为知识图谱定义好本体与数据模式，再将实体加入到知识库。该构建方式需要利用一些现有的结构化知识库作为其基础知识库，例如Freebase项目就是采用这种方式，它的绝大部分数据是从维基百科中得到的。自底向上指的是从一些开放链接数据中提取出实体，选择其中置信度较高的加入到知识库，再构建顶层的本体模式[12]。目前，大多数知识图谱都采用自底向上的方式进行构建，其中最典型就是Google的Knowledge Vault[13]。

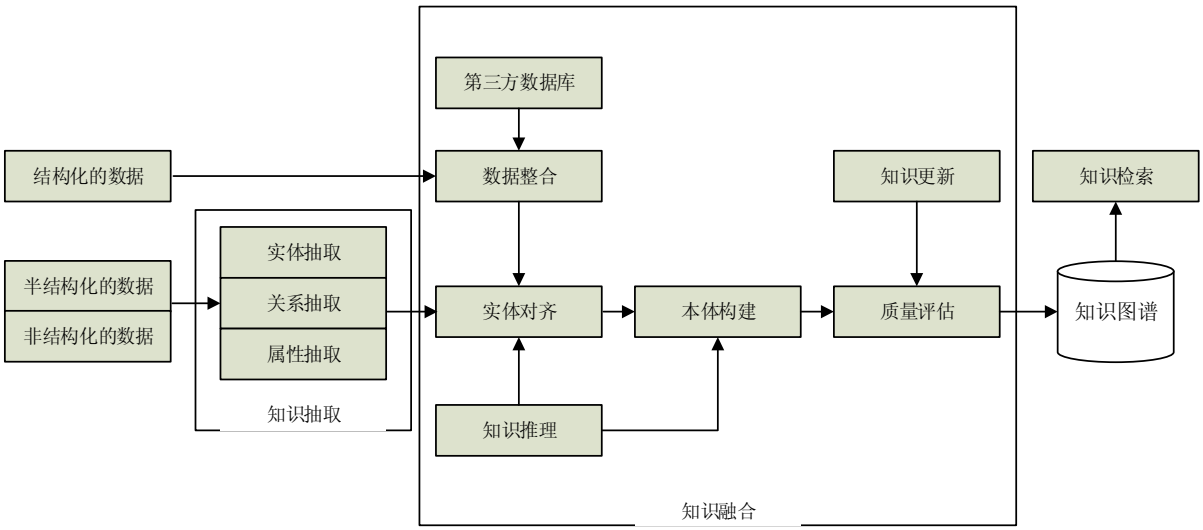


图1　知识图谱的体系架构

2　大规模知识库

随着语义Web资源数量激增、大量的RDF数据被发布和共享、LOD(linked open data)等项目的全面展开[14]，学术界与工业界的研究人员花费了大量的精力构建各种结构化的知识库。下面将以开放链接知识库、行业知识库这两类主要的知识库类型为代表，详细说明其中的几个知名度较高的大规模知识库。

2.1　开放链接知识库

在LOD项目的云图中，Freebase、Wikidata、DBpedia、YAGO这4个大规模知识库处于绝对核心的地位，它们中不仅包含大量的半结构化、非结构化数据，是知识图谱数据的重要来源。而且具有较高的领域覆盖面，与领域知识库存在大量的链接关系。

1) Freebase

Freebase知识库[15]早期由Metaweb公司创建，后来被Google收购，成为Google知识图谱的重要组成部分。Freebase中的数据主要是由人工构建，另外一部分数据则主要来源于维基百科、IMDB、Flickr等网站或语料库。截止到2014年年底，Freebase已经包含了6 800万个实体，10亿条关系信息，超过24亿条事实三元组信息，在2015年6月，Freebase整体移入至WikiData。

2) Wikidata

Wikidata[16]是维基媒体基金会主持的一个自由的协作式多语言辅助知识库，旨在为维基百科、维基共享资源以及其他的维基媒体项目提供支持。它是Wikipedia、Wikivoyage、Wikisource中结构化数据的中央存储器，并支持免费使用[17]。Wikidata中的数据主要以文档的形式进行存储，目前已包含了超过1 700万个文档。其中的每个文档都有一个主题或一个管理页面，且被唯一的数字标识。

3) DBpedia

DBpedia[18]是由德国莱比锡大学和曼海姆大学的科研人员创建的多语言综合型知识库，在LOD项目中处于最核心的地位。DBpedia是从多种语言的维基百科中抽取结构化信息，并且将其以关联数据的形式发布到互联网上，提供给在线网络应用、社交网站以及其他在线知识库。由于DBpedia的直接数据来源覆盖范围广阔，所以它包含了众多领域的实体信息。截止至2014年年底，DBpedia中的事实三元组数量已经超过了30亿条。除上述优点外，DBpedia还能够自动与维基百科保持同步，覆盖多种语言。

4) YAGO

YAGO[19]是由德国马普所(max planck institute，MPI)的科研人员构建的综合型知识库。YAGO整合了维基百科、WordNet[20]以及GeoNames等数据源，特别是将维基百科中的分类体系与WordNet的分类体系进行了融合，构建了一个复杂的类别层次结构体系。第一个版本包含了超过100万的实体以及超过500万的事实。2012年，发布了它的第二个版本，在YAGO的基础上进行了大规模的扩展，引入了一个新的数据源GeoNames[21]，被称为YAG02s。包含了超过1 000万的实体以及超过1.2亿的事实。

2.2　垂直行业知识库

行业知识库也可称为垂直型知识库，这类知识库的描述目标是特定的行业领域，通常需要依靠特定行业的数据才能构建，因此其描述范围极为有限。

下面将以MusicBrainz、IMDB、豆瓣等为代表进行说明。

1) IMDB

IMDB(internet movie database)[22]是一个关于电影演员、电影、电视节目、电视明星以及电影制作的资料库。截止到2012年2月，IMDB共收集了2 132 383部作品资料和4 530 159名人物资料。IMDB中的资料是按类型进行组织的。对于一个具体的条目，又包含了详细的元信息[23]。

2) MusicBrainz

MusicBrainz[24]是一个结构化的音乐维基百科，致力于收藏所有的音乐元数据，并向大众用户开放。任何注册用户都可以向网站中添加信息或投稿。由于Last.fm、GrooveShark、Pandora、Echonest等音乐服务网站的数据均来自于MusicBrainz，故MusicBrainz可通过数据库或Web服务两种方式将数据提供给社区。对于商业用户而言，MusicBrainz提供的在线服务可为用户提供本地化的数据库与复制包[25]。

3) ConceptNet

ConceptNet[26]是一个语义知识网络，主要由一系列的代表概念的结点构成，这些概念将主要采用自然语言单词或短语的表达形式，通过相互连接建立语义联系。ConceptNet包含了大量计算机可了解的世界的信息，这些信息将有助于计算机更好地实现搜索、问答以及理解人类的意图。ConceptNet 5[27]是基于ConceptNet的一个开源项目，主要通过GPLv3协议进行开源。

3 知识图谱的关键技术

大规模知识库的构建与应用需要多种智能信息处理技术的支持。通过知识抽取技术，可以从一些公开的半结构化、非结构化的数据中提取出实体、关系、属性等知识要素。通过知识融合，可消除实体、关系、属性等指称项与事实对象之间的歧义，形成高质量的知识库。知识推理则是在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库。分布式的知识表示形成的综合向量对知识库的构建、推理、融合以及应用均具有重要的意义。接下来，本文将以知识抽取、知识表示、知识融合以及知识推理技术为重点，详细说明其中的相关研究。

3.1 知识抽取

知识抽取主要是面向开放的链接数据，通过自动化的技术抽取出可用的知识单元，知识单元主要包括实体(概念的外延)、关系以及属性3个知识要素，并以此为基础，形成一系列高质量的事实表达，为上层模式层的构建奠定基础。

3.1.1 实体抽取

早期的实体抽取也称为命名实体学习(named entity learning)或命名实体识别(named entity recognition)，指的是从原始语料中自动识别出命名实体。由于实体是知识图谱中的最基本元素，其抽取的完整性、准确率、召回率等将直接影响到知识库的质量。因此，实体抽取是知识抽取中最为基础与关键的一步。

文献[28]将实体抽取的方法分为3种：基于规则与词典的方法、基于统计机器学习的方法以及面向开放域的抽取方法。基于规则的方法通常需要为目标实体编写模板，然后在原始语料中进行匹配；基于统计机器学习的方法主要是通过机器学习的方法对原始语料进行训练，然后再利用训练好的模型去识别实体；面向开放域的抽取将是面向海量的Web语料[12,29]。

1) 基于规则与词典的实体抽取方法

早期的实体抽取是在限定文本领域、限定语义单元类型的条件下进行的，主要采用的是基于规则与词典的方法，例如使用已定义的规则，抽取出文本中的人名、地名、组织机构名、特定时间等实体[30]。文献[31]首次实现了一套能够抽取公司名称的实体抽取系统，其中主要用到了启发式算法与规则模板相结合的方法。然而，基于规则模板的方法不仅需要依靠大量的专家来编写规则或模板，覆盖的领域范围有限，而且很难适应数据变化的新需求。

2) 基于统计机器学习的实体抽取方法

随后，研究者尝试将机器学习中的监督学习算法用于命名实体的抽取问题上。例如文献[32]利用KNN算法与条件随机场模型，实现了对Twitter文本数据中实体的识别。单纯的监督学习算法在性能上不仅受到训练集合的限制，并且算法的准确率与召回率都不够理想。相关研究者认识到监督学习算法的制约性后，尝试将监督学习算法与规则相互结合，取得了一定的成果。例如文献[33]基于字典，使用最大熵算法在Medline论文摘要的GENIA数据集上进行了实体抽取实验，实验的准确率与召回率都在70%以上。

3) 面向开放域的实体抽取方法

针对如何从少量实体实例中自动发现具有区分力的模式，进而扩展到海量文本去给实体做分类与

聚类的问题，文献[34]提出了一种通过迭代方式扩展实体语料库的解决方案，其基本思想是通过少量的实体实例建立特征模型，再通过该模型应用于新的数据集得到新的命名实体。文献[35]提出了一种基于无监督学习的开放域聚类算法，其基本思想是基于已知实体的语义特征去搜索日志中识别出命名的实体，然后进行聚类。

3.1.2 关系抽取

关系抽取的目标是解决实体间语义链接的问题，早期的关系抽取主要是通过人工构造语义规则以及模板的方法识别实体关系。随后，实体间的关系模型逐渐替代了人工预定义的语法与规则。但是仍需要提前定义实体间的关系类型。文献[36]提出了面向开放域的信息抽取框架(open information extraction,OIE)，这是抽取模式上的一个巨大进步。但OIE方法在对实体的隐含关系抽取方面性能低下，因此部分研究者提出了基于马尔可夫逻辑网、基于本体推理的深层隐含关系抽取方法[37]。

1) 开放式实体关系抽取

开放式实体关系抽取可分为二元开放式关系抽取和n元开放式关系抽取。在二元开放式关系抽取中，早期的研究有KnowItAll[38]与TextRunner[37]系统，在准确率与召回率上表现一般。文献[39]提出了一种基于Wikipedia的OIE方法WOE，经自监督学习得到抽取器，准确率较TextRunner有明显的提高。针对WOE的缺点，文献[40]提出了第二代OIE ReVerb系统，以动词关系抽取为主。文献[41]提出了第三代OIE系统OLLIE(open language learning for information extraction)，尝试弥补并扩展OIE的模型及相应的系统，抽取结果的准确度得到了增强。

然而，基于语义角色标注的OIE分析显示：英文语句中40%的实体关系是n元的[42]，如处理不当，可能会影响整体抽取的完整性。文献[43]提出了一种可抽取任意英文语句中n元实体关系的方法KPAKEN，弥补了ReVerb的不足。但是由于算法对语句深层语法特征的提取导致其效率显著下降，并不适用于大规模开放域语料的情况。

2) 基于联合推理的实体关系抽取

联合推理的关系抽取中的典型方法是马尔可夫逻辑网MLN(Markov logic network)[44]，它是一种将马尔可夫网络与一阶逻辑相结合的统计关系学习框架，同时也是在OIE中融入推理的一种重要实体关系抽取模型。基于该模型，文献[45]提出了一种无监督学习模型StatSnowball，不同于传统的OIE，该方法可自动产生或选择模板生成抽取器。在StatSnowball的基础上，文献[37,46]提出了一种实体识别与关系抽取相结合的模型EntSum，主要由扩展的CRF命名实体识别模块与基于StatSnowball的关系抽取模块组成，在保证准确率的同时也提高了召回率。文献[37,47]提出了一种简易的Markov逻辑TML(tractable Markov logic)，TML将领域知识分解为若干部分，各部分主要来源于事物类的层次化结构，并依据此结构，将各大部分进一步分解为若干个子部分，以此类推。TML具有较强的表示能力，能够较为简洁地表示概念以及关系的本体结构。

3.1.3 属性抽取

属性抽取主要是针对实体而言的，通过属性可形成对实体的完整勾画。由于实体的属性可以看成是实体与属性值之间的一种名称性关系，因此可以将实体属性的抽取问题转换为关系抽取问题。文献[37,48]提出的基于规则与启发式算法的属性抽取方法能够从Wikipedia及WordNet的半结构化网页中自动抽取相应的属性名称与属性值，还可扩展为一套本体知识库。实验表明：该算法的抽取准确率可达到95%。

大量的属性数据主要存在于半结构化、非结构化的大规模开放域数据集中。抽取这些属性的方法，一种是将上述从百科网站上抽取的结构化数据作为可用于属性抽取的训练集，然后再将该模型应用于开放域中的实体属性抽取[49]；另一种是根据实体属性与属性值之间的关系模式，直接从开放域数据集上抽取属性。但是由于属性值附近普遍存在一些限定属性值含义的属性名等，所以该抽取方法的准确率并不高[50]。

3.2 知识表示

虽然，基于三元组的知识表示形式受到了人们广泛的认可，但是其在计算效率、数据稀疏性等方面却面临着诸多问题。近年来，以深度学习为代表的表示学习技术取得了重要的进展，可以将实体的语义信息表示为稠密低维实值向量，进而在低维空间中高效计算实体、关系及其之间的复杂语义关联，对知识库的构建、推理、融合以及应用均具有重要的意义[51-53]。本文将重点介绍知识表示学习的代表模型、复杂关系翻译模型、多源异质信息融合模型方面的研究成果。

3.2.1 应用场景

分布式表示旨在用一个综合的向量来表示实体对象的语义信息，是一种模仿人脑工作的表示机制[53]，

通过知识表示而得到的分布式表示形式在知识图谱的计算、补全、推理等方面将起到重要的作用：

1) 语义相似度计算。由于实体通过分布式表示而形成的是一个个低维的实值向量，所以，可使用熵权系数法[54]、余弦相似性[55]等方法计算它们间的相似性。这种相似性刻画了实体之间的语义关联程度，为自然语言处理等提供了极大的便利。

2) 链接预测。通过分布式表示模型，可以预测图谱中任意两个实体之间的关系，以及实体间已存在的关系的正确性。尤其是在大规模知识图谱的上下文中，需要不断补充其中的实体关系，所以链接预测又被称为知识图谱的补全[53]。

3.2.2 代表模型

知识表示学习的代表模型主要包括距离模型、双线性模型、神经张量模型、矩阵分解模型、翻译模型等。

1) 距离模型

文献[56]提出了知识库中实体以及关系的结构化表示方法(structured embedding，SE)，其基本思想是：首先将实体用向量进行表示，然后通过关系矩阵将实体投影到与实体向量同一维度的向量空间中，最后通过计算投影向量之间的距离来判断实体间已存在的关系的置信度。由于距离模型中的关系矩阵是两个不同的矩阵，故实体间的协同性较差，这也是该模型本身的主要缺陷。

2) 单层神经网络模型

文献[57]针对上述提到的距离模型中的缺陷，提出了采用单层神经网络的非线性模型(single layer model，SLM)，模型为知识库中每个三元组 (h,r,t) 定义了以下形式的评价函数：

$$f_r(h,t)=\boldsymbol{\mu}_t^{\mathrm{T}} g(\boldsymbol{M}_{r,1} l_h+\boldsymbol{M}_{r,2}\boldsymbol{l}_t)$$

式中，$\boldsymbol{\mu}_r^{\mathrm{T}}\in\mathbb{R}^k$ 为关系 r 的向量化表示；$g(\cdot)$ 为 tanh 函数；$\boldsymbol{M}_{r,1}$、$\boldsymbol{M}_{r,2}\in\mathbb{R}^{d\times k}$ 是通过关系 r 定义的两个矩阵。单层神经网络模型的非线性操作虽然能够进一步刻画实体在关系下的语义相关性，但在计算开销上却大大增加。

3) 双线性模型

双线性模型又叫隐变量模型(latent factor model，LFM)，由文献[58-59]首先提出。模型为知识库中每个三元组 (h,r,t) 定义的评价函数具有如下形式：

$$f_r(h,t)=\boldsymbol{l}_h^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{l}_t$$

式中，$\boldsymbol{M}_r\in\mathbb{R}^{d\times d}$ 是通过关系 r 定义的双线性变换矩阵；$\boldsymbol{l}_h$、$\boldsymbol{l}_t\in\mathbb{R}^d$ 是三元组中头实体与尾实体的向量化表示。

双线性模型主要是通过基于实体间关系的双线性变换来刻画实体在关系下的语义相关性。模型不仅形式简单、易于计算，而且还能够有效刻画实体间的协同性[53]。基于上述工作，文献[60]尝试将双线性变换矩阵 $\boldsymbol{M}_r$ 变换为对角矩阵，提出了DISTMULT模型，不仅简化了计算的复杂度，并且实验效果得到了显著提升。

4) 神经张量模型

文献[61]提出的神经张量模型，其基本思想是：在不同的维度下，将实体联系起来，表示实体间复杂的语义联系。模型为知识库中的每个三元组 (h,r,t) 定义了以下形式的评价函数：

$$f_r(h,t)=\boldsymbol{\mu}_r^{\mathrm{T}} g(\boldsymbol{l}_h\boldsymbol{M}_r\boldsymbol{l}_t+\boldsymbol{M}_{r,1}\boldsymbol{l}_h+\boldsymbol{M}_{r,2}\boldsymbol{l}_t+\boldsymbol{b}_r)$$

式中，$\boldsymbol{\mu}_r^{\mathrm{T}}\in\mathbb{R}^k$ 为关系 r 的向量化表示；$g(\cdot)$ 为 tanh 函数；$\boldsymbol{M}_r\in\mathbb{R}^{d\times d\times k}$ 是一个三阶张量；$\boldsymbol{M}_{r,1}$、$\boldsymbol{M}_{r,2}\in\mathbb{R}^{d\times k}$ 是通过关系 r 定义的两个投影矩阵。

神经张量模型在构建实体的向量表示时，是将该实体中的所有单词的向量取平均值，这样一方面可以重复使用单词向量构建实体，另一方面将有利于增强低维向量的稠密程度以及实体与关系的语义计算[53]。

5) 矩阵分解模型

通过矩阵分解的方式可得到低维的向量表示，故不少研究者提出可采用该方式进行知识表示学习，其中的典型代表是文献[62]提出的RESACL模型。

在RESCAL模型中，知识库中的三元组 (h,r,t) 集合被表示为一个三阶张量，如果该三元组存在，张量中对应位置的元素被置1，否则置为0。通过张量分解算法，可将张量中每个三元组 (h,r,t) 对应的张量值 X_{hrt} 分解为双线性模型中的知识表示形式 $\boldsymbol{l}_h^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{l}_t$，并使 $|\boldsymbol{X}_{\mathrm{hrt}}-\boldsymbol{l}_h^{\mathrm{T}}\boldsymbol{M}_r\boldsymbol{l}_t|_{L_2}$ 尽量小。

6) 翻译模型

文献[63]受到平移不变现象的启发，提出了TransE模型，即将知识库中实体之间的关系看成是从实体间的某种平移，并用向量表示。关系 $\boldsymbol{l}_r$ 可以看作是从头实体向量 $\boldsymbol{l}_h$ 到尾实体向量 $\boldsymbol{l}_t$ 的翻译。对于知识库中的每个三元组 (h,r,t)，TransE都希望满足以下关系：$\boldsymbol{l}_h+\boldsymbol{l}_r\approx\boldsymbol{l}_t$，其损失函数为：$f_r(h,t)=|\boldsymbol{l}_h+\boldsymbol{l}_r-\boldsymbol{l}_t|_{L_1/L_2}$，即向量 $\boldsymbol{l}_h+\boldsymbol{l}_r$ 和 $\boldsymbol{l}_t$ 的 $\boldsymbol{L}_1$ 或 $\boldsymbol{L}_2$ 距离。该模型的参数较少，计算的复杂度显著降低。与此同时，TransE模型在大规模稀疏知识库上也同样具

有较好的性能与可扩展性。

3.2.3 复杂关系模型

知识库中的实体关系类型也可分为1-to-1、1-to-*N*、*N*-to-1、*N*-to-*N*4种类型[63]，而复杂关系主要指的是1-to-*N*、*N*-to-1、*N*-to-*N*的3种关系类型。

由于TransE模型不能用在处理复杂关系上[53]，一系列基于它的扩展模型纷纷被提出，下面将着重介绍其中的几项代表性工作。

1) TransH模型

文献[64]提出的TransH模型尝试通过不同的形式表示不同关系中的实体结构，对于同一个实体而言，它在不同的关系下也扮演着不同的角色。模型首先通过关系向量 $\boldsymbol{l}_r$ 与其正交的法向量 $\boldsymbol{w}_r$ 选取某一个超平面 F，然后将头实体向量 $\boldsymbol{l}_h$ 和尾实体向量 $\boldsymbol{l}_t$ 沿法向量 $\boldsymbol{w}_r$ 的方向投影到 F，最后计算损失函数。TransH使不同的实体在不同的关系下拥有了不同的表示形式，但由于实体向量被投影到了关系的语义空间中，故它们具有相同的维度。

2) TransR模型

由于实体、关系是不同的对象，不同的关系所关注的实体的属性也不尽相同，将它们映射到同一个语义空间，在一定程度上就限制了模型的表达能力。所以，文献[65]提出了TransR模型。模型首先将知识库中的每个三元组 (h,r,t) 的头实体与尾实体向关系空间中投影，然后希望满足 $\boldsymbol{l}_{h_r}+\boldsymbol{l}_r\approx\boldsymbol{l}_{t_r}$ 的关系，最后计算损失函数。

文献[65]提出的CTransR模型认为关系还可做更细致的划分，这将有利于提高实体与关系的语义联系。在CTransR模型中，通过对关系 r 对应的头实体、尾实体向量的差值 $\boldsymbol{l}_h-\boldsymbol{l}_t$ 进行聚类，可将 r 划分为若干个子关系 r_c。

3) TransD模型

考虑到在知识库的三元组中，头实体和尾实体表示的含义、类型以及属性可能有较大差异，之前的TransR模型使它们被同一个投影矩阵进行映射，在一定程度上就限制了模型的表达能力。除此之外，将实体映射到关系空间体现的是从实体到关系的语义联系，而TransR模型中提出的投影矩阵仅考虑了不同的关系类型，而忽视了实体与关系之间的交互。因此，文献[66]提出了TransD模型，模型分别定义了头实体与尾实体在关系空间上的投影矩阵。

4) TransG模型

文献[67]提出的TransG模型认为一种关系可能会对应多种语义，而每一种语义都可以用一个高斯分布表示。TransG模型考虑到了关系 r 的不同语义，使用高斯混合模型来描述知识库中每个三元组 (h,r,t) 的头实体与尾实体之间的关系，具有较高的实体区分度。

5) KG2E模型

考虑到知识库中的实体以及关系的不确定性，文献[68]提出了KG2E模型，其中同样是用高斯分布来刻画实体与关系。模型使用高斯分布的均值表示实体或关系在语义空间中的中心位置，协方差则表示实体或关系的不确定度。

知识库中，每个三元组 (h,r,t) 的头实体向量 $\boldsymbol{l}_h$ 与尾实体向量 $\boldsymbol{l}_t$ 之间的关系可表示为：

$$P_e=\boldsymbol{l}_h-\boldsymbol{l}_t\sim N(\mu_h-\mu_t,\Sigma_h+\Sigma_r)$$

关系 r 可表示为：

$$P_r\sim N(\mu_r,\Sigma_r)$$

由此，可以通过 P_e 与 P_r 两个相似度的评价给三元组打分。用于对分布相似度进行评价的方法主要是KL散度与期望概率。

3.2.4 多源信息融合

三元组作为知识库的一种通用表示形式，通过表示学习，能够以较为直接的方式表示实体、关系及其之间的复杂语义关联。然而，互联网中仍蕴含着大量与知识库实体、关系有关的信息未被考虑或有效利用，如充分融合、利用这些多源异质的相关信息，将有利于进一步提升现有知识表示模型的区分能力以及性能[53]。

目前，多源异质信息融合模型方面的研究尚处于起步阶段，涉及的信息来源也极为有限，具有较为广阔的研究前景。下面将主要介绍其中通过融合本文信息进行知识表示的代表性工作。

文献[69]提出的DKRL(description-embodied knowledge representation learning)，模型将Freebase知识库中的实体描述文本数据作为其主要数据来源，通过CBOW模型[70]，将文本中多个词对应的词向量加起来表示文本；其中的另一个CNN模型[71]则利用模型中层间的联系和空域信息的紧密关系来做文本的处理与特征提取，除此之外，CNN模型中还充分考虑到了文本中不同单词的次序问题。

DKRL模型在新实体的表示能力方面较强，它能根据新实体的简短描述产生对应的表示形式，这对于知识融合以及知识图谱补全等具有重要的意义。

文献[64]选择维基百科知识库，并通过word2vec将知识库中的正文词语表示为向量，同时使用TransE模型[63]对该知识库进行表示学习。目标是使

通过word2vec表示的实体与知识库中学习到的实体尽可能接近，从而使文本能够与知识库相互融合。

3.3 知识融合

由于知识图谱中的知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、知识间的关联不够明确等问题，所以必须要进行知识的融合。知识融合是高层次的知识组织[72]，使来自不同知识源的知识在同一框架规范下进行异构数据整合、消歧、加工、推理验证、更新等步骤[73]，达到数据、信息、方法、经验以及人的思想的融合，形成高质量的知识库。

3.3.1 实体对齐

实体对齐(entity alignment)也称为实体匹配(entity matching)或实体解析(entity resolution)，主要是用于消除异构数据中实体冲突、指向不明等不一致性问题，可以从顶层创建一个大规模的统一知识库，从而帮助机器理解多源异质的数据，形成高质量的知识。

在大数据的环境下，受知识库规模的影响，在进行知识库实体对齐时，主要会面临以下3个方面的挑战[74]：1) 计算复杂度。匹配算法的计算复杂度会随知识库的规模呈二次增长，难以接受；2) 数据质量。由于不同知识库的构建目的与方式有所不同，可能存在知识质量良莠不齐、相似重复数据、孤立数据、数据时间粒度不一致等问题[75]；3) 先验训练数据。在大规模知识库中想要获得这种先验数据却非常困难。通常情况下，需要研究者手工构造先验训练数据。

基于上述，知识库实体对齐的主要流程将包括[74]：1) 将待对齐数据进行分区索引，以降低计算的复杂度；2) 利用相似度函数或相似性算法查找匹配实例；3) 使用实体对齐算法进行实例融合；4) 将步骤2)与步骤3)的结果结合起来，形成最终的对齐结果。

对齐算法可分为成对实体对齐与集体实体对齐两大类，而集体实体对齐又可分为局部集体实体对齐与全局集体实体对齐。

1) 成对实体对齐方法

① 基于传统概率模型的实体对齐方法

基于传统概率模型的实体对齐方法主要就是考虑两个实体各自属性的相似性，而并不考虑实体间的关系。文献[76]将基于属性相似度评分来判断实体是否匹配的问题转化为一个分类问题，建立了该问题的概率模型，缺点是没有体现重要属性对于实体相似度的影响。文献[77]基于概率实体链接模型，为每个匹配的属性对分配了不同的权重，匹配准确度有所提高。文献[78]还结合贝叶斯网络对属性的相关性进行建模，并使用最大似然估计方法对模型中的参数进行估计。

② 基于机器学习的实体对齐方法

基于机器学习的实体对齐方法主要是将实体对齐问题转化为二分类问题。根据是否使用标注数据可分为有监督学习与无监督学习两类，基于监督学习的实体对齐方法主要可分为成对实体对齐、基于聚类的对齐、主动学习。

通过属性比较向量来判断实体对匹配与否可称为成对实体对齐。这类方法中的典型代表有决策树[79]、支持向量机[80]、集成学习[81]等。文献[82]使用分类回归树、线性分析判别等方法完成了实体辨析。文献[83]基于二阶段实体链接分析模型，提出了一种新的SVM分类方法，匹配准确率远高于TAILOR中的混合算法。

基于聚类的实体对齐算法，其主要思想是将相似的实体尽量聚集到一起，再进行实体对齐。文献[84]提出了一种扩展性较强的自适应实体名称匹配与聚类算法，可通过训练样本生成一个自适应的距离函数。文献[85]采用类似的方法，在条件随机场实体对齐模型中使用监督学习的方法训练产生距离函数，然后调整权重，使特征函数与学习参数的积最大。

在主动学习中，可通过与人员的不断交互来解决很难获得足够的训练数据问题，文献[86]构建的ALIAS系统可通过人机交互的方式完成实体链接与去重的任务。文献[87]采用相似的方法构建了Active Atlas系统。

2) 局部集体实体对齐方法

局部集体实体对齐方法为实体本身的属性以及与它有关联的实体的属性分别设置不同的权重，并通过加权求和计算总体的相似度，还可使用向量空间模型以及余弦相似性来判别大规模知识库中的实体的相似程度[88]，算法为每个实体建立了名称向量与虚拟文档向量，名称向量用于标识实体的属性，虚拟文档向量则用于表示实体的属性值以及其邻居节点的属性值的加权和值[74]。为了评价向量中每个分量的重要性，算法主要使用TF-IDF为每个分量设置权重，并为分量向量建立倒排索引，最后选择余弦相似性函数计算它们的相似程度[74]。该算法的召回率较高，执行速度快，但准确率不足。其根本原

因在于没有真正从语义方面进行考虑。

3) 全局集体实体对齐方法

① 基于相似性传播的集体实体对齐方法

基于相似性传播的方法是一种典型的集体实体对齐方法，匹配的两个实体与它们产生直接关联的其他实体也会具有较高的相似性，而这种相似性又会影响关联的其他实体[74]。

相似性传播集体实体对齐方法最早来源于文献[89-90]提出的集合关系聚类算法，该算法主要通过一种改进的层次凝聚算法迭代产生匹配对象。文献[91]在以上算法的基础上提出了适用于大规模知识库实体对齐的算法SiGMa，该算法将实体对齐问题看成是一个全局匹配评分目标函数的优化问题进行建模，属于二次分配问题，可通过贪婪优化算法求得其近似解。SiGMa方法[74]能够综合考虑实体对的属性与关系，通过集体实体的领域，不断迭代发现所有的匹配对。

② 基于概率模型的集体实体对齐方法

基于概率模型的集体实体对齐方法主要采用统计关系学习进行计算与推理，常用的方法有LDA模型[92]、CRF模型[93]、Markov逻辑网[94]等。

文献[92]将LDA模型应用于实体的解析过程中，通过其中的隐含变量获取实体之间的关系。但在大规模的数据集上效果一般。文献[85]提出了一种基于图划分技术的CRF实体辨析模型，该模型以观察值为条件产生实体判别的决策，有利于处理属性间具有依赖关系的数据。文献[93]在CRF实体辨析模型的基础上提出了一种基于条件随机场模型的多关系的实体链接算法，引入了基于canopy的索引，提高了大规模知识库环境下的集体实体对齐效率。文献[94]提出了一种基于Markov逻辑网的实体解析方法。通过Markov逻辑网，可构建一个Markov网，将概率图模型中的最大可能性计算问题转化为典型的最大化加权可满足性问题，但基于Markov网进行实体辨析时，需要定义一系列的等价谓词公理，通过它们完成知识库的集体实体对齐。

3.3.2 知识加工

通过实体对齐，可以得到一系列的基本事实表达或初步的本体雏形，然而事实并不等于知识，它只是知识的基本单位。要形成高质量的知识，还需要经过知识加工的过程，从层次上形成一个大规模的知识体系，统一对知识进行管理。知识加工主要包括本体构建与质量评估两方面的内容。

1) 本体构建

本体是同一领域内不同主体之间进行交流、连通的语义基础[95]，其主要呈现树状结构，相邻的层次节点或概念之间具有严格的“IsA”关系，有利于进行约束、推理等，却不利于表达概念的多样性。本体在知识图谱中的地位相当于知识库的模具，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小[96]。

本体可通过人工编辑的方式手动构建，也可通过数据驱动自动构建，然后再经质量评估方法与人工审核相结合的方式加以修正与确认[12]。在海量的实体数据面前，人工编辑构建的方式工作量极其巨大，故当前主流的本体库产品，都是面向特定领域，采用自动构建技术而逐步扩展形成的。例如Microsoft的Probase本体库就是采用数据驱动的方法，利用机器学习算法从网页文本中抽取概念间的“IsA”关系，然后合并形成概念层次结构。目前，Probase所包含的概念总数已达到千万级别，准确率高达92.8%，是目前为止包含概念数量最多，同时也是概念可信程度最高的知识库[97]。

数据驱动的本体自动构建过程主要可分为以下3个阶段[98]：① 纵向概念间的并列关系计算。通过计算任意2个实体间并列关系的相似度，可辨析它们在语义层面是否属于同一个概念。计算方法主要包括模式匹配与分布相似度两种[12,99]。② 实体上下位关系抽取。上下位关系抽取方法包括基于语法的抽取与基于语义的抽取两种方式，例如目前主流的信息抽取系统KnowltAll[38]、TextRunner[37]、NELL[100]等，都可以在语法层面抽取实体的上下位关系，而Probase则是采用基于语义的抽取模式[101]。③ 本体生成。对各层次得到的概念进行聚类，并为每一类的实体指定1个或多个公共上位词。文献[102]基于主题层次聚类的方法构建了本体结构。与此同时，为了解决主题模型不适用于短文本的问题，提出了基于单词共现网络的主题聚类与上下位词抽取模型。

2) 质量评估

对知识库的质量评估任务通常是与实体对齐任务一起进行的，其意义在于，可以对知识的可信度进行量化，保留置信度较高的，舍弃置信度较低的，有效确保知识的质量。

文献[103]基于LDIF框架，提出了一种新的知识质量评估方法，用户可根据业务需求来定义质量评估函数，或者通过对多种评估方法的综合考评来确定知识的最终质量评分。例如在对REVERRB系统的

信息抽取质量进行评估时，文献[104]采用人工标注的方式对1 000个句子中的实体关系三元组进行了标注，并以此作为训练集，使用logistic回归模型计算抽取结果的置信度。例如Google的Knowledge Vault项目则根据指定数据信息的抽取频率对信息的可信度进行评分，然后利用从可信知识库中得到的先验知识对可信度进行修正。实验结果表明：该方法可以有效地降低对数据信息正误判断的不确定性，提高知识的质量[105]。

3.3.3 知识更新

人类的认知能力、知识储备以及业务需求都会随时间而不断递增。因此，知识图谱的内容也需要与时俱进，不论是通用知识图谱，还是行业知识图谱，它们都需要不断地迭代更新，扩展现有的知识，增加新的知识。

根据知识图谱的逻辑结构，其更新主要包括模式层的更新与数据层的更新。模式层的更新是指本体中元素的更新，包括概念的增加、修改、删除，概念属性的更新以及概念之间上下位关系的更新等。其中，概念属性的更新操作将直接影响到所有直接或间接属性的子概念和实体[106]。通常来说，模式层的增量更新方式消耗资源较少，但是多数情况下是在人工干预的情况下完成的，例如需要人工定义规则，人工处理冲突等。因此，实施起来并不容易[107]。数据层的更新指的是实体元素的更新，包括实体的增加、修改、删除，以及实体的基本信息和属性值。由于数据层的更新一般影响面较小，因此通常以自动的方式完成。

3.4 知识推理

知识推理则是在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库。在推理的过程中，往往需要关联规则的支持。由于实体、实体属性以及关系的多样性，人们很难穷举所有的推理规则，一些较为复杂的推理规则往往是手动总结的。对于推理规则的挖掘，主要还是依赖于实体以及关系间的丰富同现情况。知识推理的对象可以是实体、实体的属性、实体间的关系、本体库中概念的层次结构等。

知识推理方法主要可分为基于逻辑的推理与基于图的推理两种类别。

3.4.1 基于逻辑的推理

基于逻辑的推理方式主要包括一阶谓词逻辑(first order logic)、描述逻辑(description logic)以及规则等。一阶谓词逻辑推理是以命题为基本进行推理，而命题又包含个体和谓词。逻辑中的个体对应知识库中的实体对象，具有客观独立性，可以是具体一个或泛指一类，例如奥巴马、选民等；谓词则描述了个体的性质或个体间的关系。文献[108]针对已有一阶谓词逻辑推理方法中存在的推理效率低下等问题，提出了一种基于谓词变迁系统的图形推理法，定义了描述谓词间与/或关系的谓词，通过谓词图表示变迁系统，实现了反向的推理目标。实验结果表明：该方法推理效率较高，性能优越。

描述逻辑是在命题逻辑与一阶谓词逻辑上发展而来，目的是在表示能力与推理复杂度之间追求一种平衡。基于描述逻辑的知识库主要包括Tbox(terminology box)与ABox(assertion box)[109]。通过TBox与ABox，可将关于知识库中复杂的实体关系推理转化为一致性的检验问题，从而简化并实现推理[110]。

通过本体的概念层次进行推理时，其中概念主要是通过OWL(Web ontology language)本体语义进行描述的。OWL文档可以表示为一个具有树形结构的状态空间，这样一些对接结点的推理算法就能够较好地应用起来，例如文献[111]提出了基于RDF和PD*语义的正向推理算法，该算法以RDF蕴涵规则为前提，结合了sesame算法以及PD*的语义，是一个典型的迭代算法，它主要考虑结点与推理规则的前提是否有匹配，由于该算法的触发条件导致推理的时间复杂度较高，文献[112]提出了ORBO算法，该算法从结点出发考虑，判断推理规则中第一条推理关系的前提是否满足，不仅节约了时间，还降低了算法的时间复杂度。

3.4.2 基于图的推理

在基于图的推理方法中，文献[113]提出的path-constraint random walk，path ranking等算法较为典型，主要是利用了关系路径中的蕴涵信息，通过图中两个实体间的多步路径来预测它们之间的语义关系。即从源节点开始，在图上根据路径建模算法进行游走，如果能够到达目标节点，则推测源节点和目标节点间存在联系。关系路径的建模方法研究工作尚处于初期，其中在关系路径的可靠性计算、关系路径的语义组合操作等方面，仍有很多工作需进一步探索并完成。

除上述两种类别的知识推理方法外，部分研究人员将研究重点转向跨知识库的推理方法研究，例如文献[75]提出的基于组合描述逻辑的Tableau算法，该方法主要利用概念间的相似性对不同知识库

中的概念进行关联、合并，通过已有的知识完成跨知识库的推理。

4 知识图谱的典型应用

知识图谱为互联网上海量、异构、动态的大数据表达、组织、管理以及利用提供了一种更为有效的方式，使得网络的智能化水平更高，更加接近于人类的认知思维。目前，知识图谱已在智能搜索、深度问答、社交网络以及一些垂直行业中有所应用，成为支撑这些应用发展的动力源泉。

4.1 智能搜索

基于知识图谱的智能搜索是一种基于长尾的搜索，搜索引擎以知识卡片的形式将搜索结果展现出来。用户的查询请求将经过查询式语义理解与知识检索两个阶段：1) 查询式语义理解。知识图谱对查询式的语义分析主要包括：① 对查询请求文本进行分词、词性标注以及纠错；② 描述归一化，使其与知识库中的相关知识进行匹配[114]；③ 语境分析。在不同的语境下，用户查询式中的对象会有所差别，因此知识图谱需要结合用户当时的情感，将用户此时需要的答案及时反馈给用户；④ 查询扩展。明确了用户的查询意图以及相关概念后，需要加入当前语境下的相关概念进行扩展。2) 知识检索。经过查询式分析后的标准查询语句进入知识库检索引擎，引擎会在知识库中检索相应的实体以及与其在类别、关系、相关性等方面匹配度较高的实体[115]。通过对知识库的深层挖掘与提炼后，引擎将给出具有重要性排序的完整知识体系。

智能搜索引擎主要以3种形式展现知识：1) 集成的语义数据。例如当用户搜索梵高，搜索引擎将以知识卡片的形式给出梵高的详细生平，并配合以图片等信息；2) 直接给出用户查询问题的答案。例如当用户搜索“姚明的身高是多少？”，搜索引擎的结果是“226 cm”；3) 根据用户的查询给出推荐列表[7]等。

国外的搜索引擎以谷歌的Google Search[6]、微软的Bing Search[116]最为典型。谷歌的知识图谱相继融入了维基百科、CIA世界概览等公共资源以及从其他网站搜集、整理的大量语义数据[117]，微软的Bing Search[116]和Facebook[117]、Twitter[118]等大型社交服务站点达成了合作协议，在用户个性化内容的搜集、定制化方面具有显著的优势。

国内的主流搜索引擎公司，如百度、搜狗等在近两年来相继将知识图谱的相关研究从概念转向产品应用。搜狗的知立方[119]是国内搜索引擎行业的第一款知识图谱产品，它通过整合互联网上的碎片化语义信息，对用户的搜索进行逻辑推荐与计算，并将最核心的知识反馈给用户。百度将知识图谱命名为知心[120]，主要致力于构建一个庞大的通用型知识网络，以图文并茂的形式展现知识的方方面面[7]。

4.2 深度问答

问答系统是信息检索系统的一种高级形式，能够以准确简洁的自然语言为用户提供问题的解答。之所以说问答是一种高级形式的检索，是因为在问答系统中同样有查询式理解与知识检索这两个重要的过程，并且与智能搜索中相应过程中的相关细节是完全一致的。多数问答系统更倾向于将给定的问题分解为多个小的问题，然后逐一去知识库中抽取匹配的答案，并自动检测其在时间与空间上的吻合度等，最后将答案进行合并，以直观的方式展现给用户。

目前，很多问答平台都引入了知识图谱，例如华盛顿大学的Paralex系统[121]和苹果的智能语音助手Siri[122]，都能够为用户提供回答、介绍等服务；亚马逊收购的自然语言助手Evi[123]，它授权了Nuance的语音识别技术，采用True Knowledge引擎进行开发，也可提供类似Siri的服务。国内百度公司研发的小度机器人[124]，天津聚问网络技术服务中心开发的大型在线问答系统OASK[125]，专门为门户、企业、媒体、教育等各类网站提供良好的交互式问答解决方案。

4.3 社交网络

社交网站Facebook于2013年推出了Graph Search[126]产品，其核心技术就是通过知识图谱将人、地点、事情等联系在一起，并以直观的方式支持精确的自然语言查询，例如输入查询式：“我朋友喜欢的餐厅”“住在纽约并且喜欢篮球和中国电影的朋友”等，知识图谱会帮助用户在庞大的社交网络中找到与自己最具相关性的人、照片、地点和兴趣等[7]。Graph Search提供的上述服务贴近个人的生活，满足了用户发现知识以及寻找最具相关性的人的需求。

4.4 垂直行业应用

下面将以金融、医疗、电商行业为例，说明知识图谱在上述行业中的典型应用。

1) 金融行业

在金融行业中，反欺诈是一个重要的环节。它

的难点在于如何将不同税务子系统中的数据整合在一起。通过知识图谱，一方面有利于组织相关的知识碎片，通过深入的语义分析与推理，可对信息内容的一致性充分验证，从而识别或提前发现欺诈行为；另一方面，知识图谱本身就是一种基于图结构的关系网络，基于这种图结构能够帮助人们更有效地分析复杂税务关系中存在的潜在风险[127]。在精准营销方面，知识图谱可通过链接的多个数据源，形成对用户或用户群体的完整知识体系描述，从而更好地去认识、理解、分析用户或用户群体的行为。例如，金融公司的市场经理用知识图谱去分析待销售用户群体之间的关系，去发现他们的共同爱好，从而更有针对性地对这类用户人群制定营销策略[127]。

2) 医疗行业

耶鲁大学拥有全球最大的神经科学数据库Senselab[128]，然而，脑科学研究还需要综合从微观分子层面一直到宏观行为层面的各个层次的知识。因此，耶鲁大学的脑计划研究人员将不同层次的，与脑研究相关的数据进行检索、比较、分析、整合、建模、仿真，绘制出了描述脑结构的神经网络图谱，从而解决了当前神经科学所面临的海量数据问题，从微观基因到宏观行为，从多个层次上加深了人类对大脑的理解，达到了“认识大脑、保护大脑、创造大脑”的目标。

4) 电商行业

电商网站的主要目的之一就是通过对商品的文字描述、图片展示、相关信息罗列等可视化的知识展现，为消费者提供最满意的购物服务与体验。通过知识图谱，可以提升电商平台的技术性、易用性、交互性等影响用户体验的因素[129]。

阿里巴巴是应用知识图谱的代表电商网站之一，它旗下的一淘网不仅包含了淘宝数亿的商品，更建立了商品间关联的信息以及从互联网抽取的相关信息，通过整合所有信息，形成了阿里巴巴知识库和产品库，构建了它自身的知识图谱[7,130]。当用户输入关键词查看商品时，知识图谱会为用户提供此次购物方面最相关的信息，包括整合后分类罗列的商品结果、使用建议、搭配等[7,130]。

除此之外，另外一些垂直行业也需要引入知识图谱，如教育科研行业、图书馆、证券业、生物医疗以及需要进行大数据分析的一些行业[131]。这些行业对整合性和关联性的资源需求迫切，知识图谱可以为其提供更加精确规范的行业数据以及丰富的表达，帮助用户更加便捷地获取行业知识[7]。

5 知识图谱的挑战

知识图谱技术是对语义网标准与技术的一次扬弃与升华。自Google提出之初至今，其热度依然有增无减，并随着深度学习、类脑科学等领域的发展，有逐步演进并发展为智能机器的大脑知识库之趋势。

在关注到知识图谱在自然语言处理、人工智能等领域展现巨大潜力的同时，也不难发现知识图谱中的知识获取、知识表示、知识推理等技术依然面临着一些困难与挑战，很多重要的开放问题急待学术界与工业界协力来解决。在未来的几年时间内，知识图谱仍将是大数据智能的前沿研究问题。

5.1 知识获取

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。一些传统的知识元素(实体、关系、属性)抽取技术与方法，它们在限定领域、主题的数据集上获得了较好的效果，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。目前，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。

KnowItAll、TextRunner、WOE、ReVerb、R2A2、KPAKEN这些系统已为开放域环境下，实体关系抽取中的二元关系抽取、*n*元关系抽取发展开创了先河，具有广阔的研究前景。再者，对于隐含关系的抽取，目前主流的开放式信息抽取方法性能低下或尚无法实现。因此，以马尔可夫逻辑网、本体推理的联合推理方法将成为学术界的研究热点。联合推理方法不仅能够推断文本语料所不能显示的深层隐含信息，还能够综合信息抽取各阶段的子任务，像杠杆一样在各方面之间寻求平衡，以趋向整体向上的理想效果[37]，为大规模开放域下的知识抽取提供了一种新的思路。

除上述外，跨语言的知识抽取方法也成为了当前的研究热点，对于我国的研究者而言，更应发挥自身在中文信息处理方面的天然优势，面对挑战与机遇，做出应有的贡献。

5.2 知识表示

知识表示对知识图谱的构建、推理、融合以及应用均具有重要的意义。目前存在的表示方式仍是基于三元组形式完成的语义映射，在面对复杂的知识类型、多源融合的信息时，其表达能力仍然有限。因此有研究者提出，应针对不同的应用场景设计不

同的知识表示方法。下面将具体说明知识表示在复杂关系、多源信息融合中遇到的挑战以及未来的研究方向。

1) 复杂关系中的知识表示

已有的工作将知识库中的实体关系类型分为1-to-1、1-to-*N*、*N*-to-1、*N*-to-*N*这4种，这种划分方法无法直观地解释知识的本质类型特点，也无法更有针对性地表示复杂关系中的知识。但发现分布式的知识表示方法来源于认知科学，具有灵活的可扩展能力。基于上述，对认知科学领域人类知识类型的探索将有助于知识类型的划分、表示以及处理，是未来知识表示研究的重要发展方向。

2) 多源信息融合中的知识表示

对于多源信息融合中的知识表示研究尚处于起步阶段，涉及的信息来源也极为有限，已有的少数工作都是围绕文本与知识库的融合而展开的。另外，文献[132]将注意力转向面向关系表示的多源信息融合领域，并已在CNN上进行了一定的实现。在知识融合表示中，融合是最关键的前期步骤，如能有机的融合多源异质的实体、关系等信息，将有利于进一步提升知识表示模型的区分能力以及性能。基于实体的、关系的、Web文本的、多知识库的融合均具有较为广阔的研究前景。

5.3　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。实体对齐是知识融合中的关键步骤，虽然相关研究已取得了丰硕的成果，但仍有广阔的发展空间。下面将具体说明实体对齐在大规模知识库环境下所遇到的挑战以及未来的研究方向。

1) 并行与分布式算法

大规模的知识库不仅蕴含了海量的知识，其结构、数据特征也极其复杂，这些对知识库实体对齐算法的准确率、执行效率提出了一定的挑战。目前，不少研究者正着力研究对齐算法的并行化或分布式版本，在兼顾算法准确率与召回率的同时，将进一步利用并行编程环境MPI，分布式计算框架Hadoop、Spark等平台，提升知识库对齐的整体效果。

2) 众包算法

人机结合的众包算法可以有效地提高知识融合的质量[74]。众包算法的设计讲求数据量、知识库对齐质量以及人工标注三者的权衡。将众包平台与知识库对齐模型有机结合起来，并且能够有效判别人工标注的质量，这些均具有较为广阔的研究前景[74]。

3) 跨语言知识库对齐

多语言的知识库越来越多，多语言知识库的互补能力将为知识图谱在多语言搜索、问答、翻译等领域的实际应用提供更多的可能。文献[133]已在这方面取得了一定的进展，但知识库对齐的质量不高，这方面仍有广阔的研究空间。

知识加工是形成高质量知识的重要途径，其中本体自动构建、本体抽取、本体聚类等问题是目前的研究热点。在知识质量评估方面，构建完善的质量评估技术标准或指标体系是该领域未来的研究目标。随着人类认知能力、知识储备以及业务需求的不断递增，知识图谱也需要不断地迭代更新。然而现有的更新技术均过多依赖于人工的干预，增量更新技术将是知识图谱未来实现自动化更新的重要研究方向。如何确保自动化更新的有效性，是更新过程中面临的又一重大挑战。

5.4　知识应用

目前，大规模知识图谱的应用场景和方式还比较有限，其在智能搜索、深度问答、社交网络以及其他行业中的使用也只是处于初级阶段，仍具有广阔的可扩展空间。人们在挖掘需求、探索知识图谱的应用场景时，应充分考虑知识图谱的以下优势：1) 对海量、异构、动态的半结构化、非结构化数据的有效组织与表达能力；2) 依托于强大知识库的深度知识推理能力；3) 与深度学习、类脑科学等领域相结合，逐步扩展的认知能力。在对知识图谱技术有丰富积累的基础上，敏锐的感知人们的需求，可为大规模知识图谱的应用找到更宽广、更合适的应用之道。

6　结 束 语

本文在对知识图谱的定义、架构、大规模知识库等全面阐述的基础上，较为深入地研究了知识图谱中知识抽取、知识表示、知识融合以及知识推理4大核心技术，并就当前产业界的需求介绍了它在智能搜索、深度问答、社交网络以及一些垂直行业中的实际应用。总结了目前知识图谱面临的主要挑战，并对其未来的研究方向进行了展望。

知识图谱的重要性不仅在于它是一个拥有强大语义处理能力与开放互联能力的知识库，并且它还是一把开启智能机器大脑的钥匙，能够打开Web 3.0时代的知识宝库，为相关学科领域开启新的发展方向。

在未来的几年时间内，知识图谱仍将是大数据智能的前沿研究问题。期待更多的研究者能够加入

到知识图谱研究的行列中来，也希望本文能够为知识图谱技术在国内的研究发展提供一些帮助。

参 考 文 献

[1] SHETH A, THIRUNARAYAN K. Semantics empowered Web 3.0:managing enterprise, social, sensor, and cloud-based data and service for advanced applications[M]. San Rafael, CA: Morgan and Claypool, 2013.

[2] 王知津, 王璇, 马婧. 论知识组织的十大原则[J]. 国家图书馆学刊, 2012, 21(4): 3-11.
WANG Zhi-jin, WANG Xuan, MA Jing. The ten principles of knowledge organization[J]. Journal of The National Library of China, 2012, 21(4): 3-11.

[3] 索传军. 网络信息资源组织研究的新视角[J]. 图书馆情报工作, 2013, 57(7): 5-12.
SUO Chuan-jun. A new perspective for web resource organization research[J]. Library and Information Service, 2013, 57(7): 5-12.

[4] 钟翠娇. 网络信息语义组织及检索研究[J]. 图书馆学研究, 2010, 75(17): 68-71.
ZHONG Cui-jiao. Research on semantic organization of web information and retrieval[J]. Research on Library Science, 2010, 75(17): 68-71.

[5] BERNERS-LEE T, HENDLER J, LASSILA O. The semantic Web[J]. Scientific American Magazine, 2008, 23(1): 1-4.

[6] AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.

[7] 曹倩, 赵一鸣. 知识图谱的技术实现流程及相关应用[J]. 情报理论与实践(ITA), 2015, 12(38): 127-132.
CAO Qian, ZHAO Yi-ming. The realization process and related applications of knowledge graph[J]. Information Studies: Theory & Application(ITA), 2015, 12(38): 127- 132.

[8] Wikipedia. Knowledge graph[EB/OL]. [2016-05-09]. https://en.wikipedia.org/wiki/Knowledge_Graph.

[9] Shenshouer. Neo4j[EB/OL]. [2016-05-09]. http://neo4j.com/.

[10] FlockDB Official. FlockDB[EB/OL]. [2016-05-09]. http://webscripts.softpedia.com/script/Database-Tools/FlockDB-66248.html.

[11] Graphdb Official. Graphdb[EB/OL]. [2016-05-09]. http://www.graphdb.net/.

[12] 刘峤, 李杨, 杨段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.
LIU Qiao, LI yang, YANG Duan-hong, et al. Knowledge graph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.

[13] DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]//Proc of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014.

[14] BIZER C, Al E. Linked data-the story so far[J]. International Journal on Semantic Web & Information System, 2009, 5(3): 1-22.

[15] BOLLACKER K, COOK R, TUFTS P. Freebase: a shared database of structured general human knowledge[C]//Proc of the 22nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2007: 1962-1963.

[16] WMF. Wikidata[EB/OL]. [2015-11-11]. https://www.wikidata.org/wiki/Wikidata:Main_Page.

[17] Wikipedia. Data revolution for Wikipedia[EB/OL]. (2012-03-30). https://www.wikimedia.org/.

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

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

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

[21] MAEDCHE A, STAAB S. The text-to-onto ontology learning environment[C]//Software Demonstration at ICCS-2000-Eight International Conference on Conceptual Structures. [S.l.]: [s.n.], 2000.

[22] IMDB Official. IMDB[EB/OL]. [2016-02-27]. http://www.imdb.com.

[23] 百度百科. IMDB[EB/OL]. [2016-02-27]. http://baike.baidu.com/view/785720.htm?fromtitle=IMDB&fromid=925061&type=syn.
Baidu Bake. IMDB[EB/OL]. [2016-02-27]. http://baike.baidu.com/view/785720.htm?fromtitle=IMDB&fromid=925061&type=syn.

[24] MetaBrainz Foundation. Musicbrainz[EB/OL]. [2016-06-06]. http://musicbrainz.org/.

[25] 全球网站库. Musicbrainz[EB/OL]. (2013-05-20). http://www.0430.com/us/web7028.
Global Web Sites. Musicbrainz[EB/OL]. (2013-05-20). http://www.0430.com/us/web7028.

[26] OSCHINA. ConceptNet[EB/OL]. [2016-01-09]. http://www.oschina.net/p/conceptnet.

[27] CONCEPTNET5. ConceptNet5[EB/OL]. [2014-04-06]. http://conceptnet5.media.mit.edu/.

[28] 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010(6): 42-47.
SUN Zhen, WANG Hui-lin. Overview on the advance of the research on named entity recognition[J]. New Technology of Library and Information Service, 2010(6): 42-47.

[29] 赵军, 刘康, 周光有, 等. 开放式文本信息抽取[J]. 中文信息学报, 2011, 25(6): 98-110.
ZHAO Jun, LIU Kang, ZHOU Guang-you, et al. Open information extraction[J]. Journal of Chinese Information Processing, 2011, 25(6): 98-110.

[30] CHINCHOR N, MARSH E. Muc-7 information extraction task definition[C]//Proc of the 7th Message Understanding Conf. Philadelphia: Linguistic Data Consortium, 1998: 359-367.

[31] RAU L F. Extracting company names from text[C]//Proc of the 7th IEEE Conf on Artificial Intelligence Applications. Piscataway, NJ: IEEE, 1991: 29-32.

[32] LIU Xiao-hua, ZHANG Shao-dian, WEI Fu-ru, et al. Recognizing named entities in tweets[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 359-367.

[33] LIN Yi-feng, TSAI T, CHOU Wen-chi, et al. A maximum entropy approach to biomedical named entity recognition[C]//Proc of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics. New York: ACM, 2004.

[34] WHITELAW C, KEHLENBECK A, PETROVIC N, et al. Web-scale named entity recognition[C]//Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008.

[35] JAIN A, PENNACCHIOTTI M. Open entity extraction from web search query logs[C]//Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2010: 510-518.

[36] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction for the Web[C]//Proc of the 20th Int Joint Conf on Artificial Intelligence. New York: ACM, 2007: 2670-2676.

[37] 杨博, 蔡东风, 杨华. 开放式信息抽取研究进展[J]. 中文信息学报, 2014, 4: 1-11.
YANG Bo, CAI Dong-feng, YANG Hua. Progress in open information extraction[J]. Journal of Chinese Information Processing, 2014, 4: 1-11.

[38] ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the Web: an experimental study[J]. Artificial Intelligence, 2005, 165(1): 91-134.

[39] WU F, WELD D S. Open information extraction using Wikipedia[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. Sweden: ACL, 2010: 118-127.

[40] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Washington: EMNLP, 2015.

[41] SCHMITZ M M, BART R, SODERLAND S, et al. Open language learning for information extraction[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CONLL). New York: ACM, 2012.

[42] JANARA C M, STEPHEN S, OREN E. An analysis of open information extraction based on semantic role labeling[C]//Proceedings of K-CAP. New York: ACM, 2011: 113-120.

[43] AKBIK A, LOSER A. KRAKEN: N-ary facts in open information extraction[C]//Proceedings of AKBC-WEKEX at NAACL. New York: ACM, 2012: 52-56.

[44] DOMINGOS P, LOWD D. Markov logic: an interface layer for artificial intelligence[M]. San Rafael, CA: Morgan & Claypool, 2009.

[45] ZHU Jun, NIE Zai-qing, LIU Xiao-jiang, et al. StatSnowball: a statistical approach to extracting entity relationships[C]//Proceedings of the 18th International Conference on World Wide Web. Switzerland: WWW, 2009: 101-110.

[46] LIU Xiao-jiang, YU Neng-hai. People summarization by combining named entity recognition and relation extraction[J]. Journal of Convergence Information Technology, 2010, 5(10): 233-241.

[47] DOMINGOS P, WEBB A. A tractable first-order probabilistic logic[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence. San Francisco, CA: AAAI, 2012.

[48] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 697-706.

[49] WU Fei, WELD D S. Antonomously semantifying wikipedia[C]//Proc of the 16th ACM Conf on Information and Knowledge Management. New York: ACM, 2007: 41-50.

[50] 王宇, 谭松波, 廖祥文, 等. 基于扩展领域模型的有名属性抽取[J]. 计算机研究与发展, 2010, 47(9): 1567-1573.
WANG Yu, TAN Song-bo, LIAO Xiang-wen, et al. Extracted domain model based on named attribute extraction[J]. Journal of Computer Research and Development, 2010, 47(9): 1567-1573.

[51] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-7.

[52] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: a review and new perspectives[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.

[53] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 1-16.
LIU Zhi-yuan, SUN Mao-song, LIN Yan-kai, et al. Knowledge representation learning: a review[J]. Journal of Computer Research and Development, 2016, 53(2): 1-16.

[54] 中国知网. 应用熵权系数法评标[EB/OL]. (2014-10-11). http://mall.cnki.net/magazine/Article/TSGD200302003.htm.
China National Knowledge Infrastructure. Bidding evaluation based on entropy weight method[EB/OL]. (2014-10-11). http://mall.cnki.net/magazine/Article/TSGD200302003. htm.

[55] 5lulu技术库. 余弦相似性[EB/OL]. (2015-03-16). http://tec.5lulu.com/computer/detail/k0n1g5e1hj8i7f.html.
5lulu Technology Library. Cosine similarity[EB/OL]. (2015-03-16). http://tec.5lulu.com/computer/detail/k0n1g5e1hj8i7f.html.

[56] BORDES A, WESTON J, COLLOBERT R, et al. Learnting structured embeddings for knowledge bases[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2011: 301-306.

[57] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 926-934.

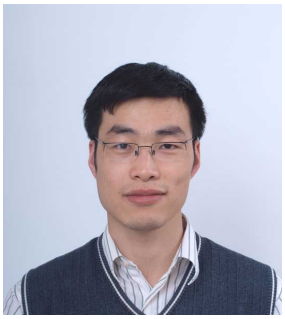
[58] JENATTON R, ROUX N L, BORDES A, et al. A latent factor model for highly multi-relational data[C]//Proc of

NIPS. Cambridge, MA: MIT Press, 2012: 3167-3175.
[59] SUTSKEVER I, TENENBAUM J B, SALAKHUTDINOV R. Modelling relational data using Bayesian clustered tensor factorization[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2009: 1821-1828.
[60] YANG B, YIH W, HE X, et al. Embedding enties and relations for learning and inference in knowledge bases[C]//Proc of Int Conf on Learning Representations (ICLR). France: ICLR Press, 2015.
[61] SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 926-934.
[62] NICKEL M, TRESP V, KRIEGEL H. A three-way model for collective learning on multi-relational data[C]//Proc of ICML. New York: ACM, 2011: 809-816.
[63] BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 2787-2795.
[64] WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2014: 1112-1119.
[65] LIN Y, LIU Z, SUN M, et al. Learning entity and relation embedding for knowledge graph completion[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2015.
[66] JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proc of ACL. Stroudsburg, PA: ACL, 2015: 687-696.
[67] XIAO H, HUANG M, HAO Y, et al. TransG: a generative mixture model for knowledge graph embedding[J]. Arxiv Preprint ArXiv, 2015, 1509: 05488.
[68] HE S, LIU K, JI J, et al. Learning to represent knowledge graphs with Gaussian embedding[C]//Proc of CIKM. New York: ACM, 2015: 623-632.
[69] XIE R, LIU Z, JIA J, et al. Representation learning of knowledge graphs with entity descriptions[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2016.
[70] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 3111-3119.
[71] COLLOBERT R, WESTON J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proc of ICML. New York: ACM, 2008: 160-167.
[72] 徐绪堪, 房道伟, 蒋勋, 等. 知识组织中知识粒度化表示和规范化研究[J]. 图书情报知识, 2014(6): 101-106, 90.
XU Xu-kan, FANG Dao-wei, JIANG Xun, et al. Research on knowledge granularity representation and standardization during knowledge organization[J]. Document, Information & Knowledge, 2014(6): 101-106, 90.
[73] 张坤. 面向知识图谱的搜索技术(搜狗)[EB/OL]. (2015-02-18). http://www.cipsc.org.cn/kg1/.
ZHANG Kun. The search technology oriented knowledge graph(Sogou)[EB/OL]. (2015-02-18). http://www.cipsc.org.cn/kg1/.
[74] 庄严, 李国良, 冯建华. 知识库实体对齐技术综述[J]. 计算机研究与发展, 2016, 01: 165-192.
ZHUANG Yan, LI Guo-liang, FENG Jian-hua. A survey on entity alignment of knowledge base[J]. Journal of Computer Research and Development, 2016, 1: 165-192.
[75] 蒋勋, 徐绪堪. 面向知识服务的知识库逻辑结构模型[J]. 图书与情报, 2013(6): 23-31.
JIANG Xun, XU Xu-kan. Knowledge service-oriented model of knowledge base logical structure research[J]. Library and Information, 2013(6): 23-31.
[76] NEWCOMBE H B, KENNEDY J M, AXFORD S J, et al. Automatic linkage of vital records[J]. Science, 1959, 130(3381): 954-959.
[77] HERZOG T N, SCHEUREN F J, WINKLER W E. Data quality and record linkage techniques[M]. Berlin: Springer, 2007.
[78] WINKLER W E. Methods for record linkage and Bayesian networks, RRS2002/05[R]. Washington DC: US Bureau of the Census, 2001.
[79] HAN J W, KAMBE M. Data mining: Concepts and techniques[M]. San Francisco, CA: Morgan Kaufmann, 2006.
[80] VAPNIK V. The nature of statistical learning theory[M]. Berlin: Springer, 2000.
[81] KANTARDZIC M. Data mining[M]. Hoboken, NJ: John Wiley & Sons, 2011.
[82] COCHINWALA M, KURIEN V, LALK G, et al. Efficient data reconciliation[J]. Information Sciences, 2011, 137(14): 1-15.
[83] CHRISTEN P. Automatic training example selection for scalable unsupervised record linkage[C]//LNAI 5012: Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conf. Berlin: Springer, 2008.
[84] COHEN W W, RICHMAN J. Learning to match and cluster large high-dimensional data sets for data integration[C]//Proc of the 2002 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 475-480.
[85] MCCALLUM A, WELLNER B. Conditional models of identity uncertainty with application to noun coreference[C]//Proc of Advances in Neural Information Processing System. Cambridge, MA: MIT Press, 2005: 905-912.
[86] SARAWAGI S, BHAMIDIPATY A. Interactive deduplication using active learning[C]//Proc of the 2002 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 269-278.
[87] TEJADA S, KNOBLOCK C A, MINTON S. Learning domain independent string transformation weights for high accuracy object identification[C]//Proc of the 2002 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 350-359.
[88] LI Juan-zi, WANG Zhi-chun, ZHANG Xiao, et al. Large scale instance matching via multiple indexes and candidate selection[J]. Knowledge-Based Systems, 2013, 50: 112-120.

[89] DONG X. Reference reconciliation in complex information spaces[C]//Proc of the 2005 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2005: 85-96.
[90] BHATTACHARYA I, GETOOR L. Collective entity resolution in relational data[J]. ACM Trans on Knowledge Discovery from Data, 2007, 1(2): 9-15.
[91] LACOSTE-Julien S, PALLA K, DAVIES A, et al. SIGMA: Simple greedy matching for aligning large knowledge bases[C]//Proc of the 2013 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 572-580.
[92] BHATTACHARYA I, GETOOR L. Alatent dirichlet allocation model for unsupervised entity resolution[C]// Proc of the 6th SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2006: 47-58.
[93] DOMINGOS P. Multi-relational record linkage[C]//Proc of the KDD-2004 Workshop on Muti-Relational Data Mining. New York: ACM, 2004.
[94] SINGLA P, DOMINGOS P. Entity resolution with Markov logic[C]//Proc of 2006 IEEE Int Conf on Data Mining(ICDM 2006). Piscataway, NJ: IEEE, 2006.
[95] STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: Principles and methods[J]. Data & Knowledge Engineering, 1998, 25(1): 161-197.
[96] WONG W, LIU Wei, BENNAMOUN M. Ontology learning from text: a look back and into the future[J]. ACM Computing Surveys, 2012, 44(4): 18-24.
[97] WU Wen-tao, LI Hong-song, WANG Hai-xun, et al. Probase: a probabilistic taxonomy for text understanding [C]//Proc of the 31st ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012.
[98] 史树明. 自动和半自动知识提取[J]. 中国计算机学会通讯, 2013, 9(8): 65-73.
SHI Shu-ming. Automatic and semi-automatic knowledge extraction[J]. China Computer Federation Communication, 2013, 9(8): 65-73.
[99] HARRIS Z S. Distributional structure[J]. Word, 1954, 10(23): 146-162.
[100] Carnegie Mellon University. NELL[EB/OL]. [2016-06-08]. http://rtw.ml.cmu.edu/rtw/.
[101] ZENG Yi, WANG Dong-sheng, ZHANG Tie-lin, et al. CASIA-KB: a multi-source Chinese semantic knowledge base built from structured and unstructured Web data[C]//Semantic Technology. Berlin: Springer, 2014: 75-88.
[102] WANG C, DANILEVSKY M, DESAI N, et al. A phrase mining framework for recursive construction of a topical hierarchy[C]//Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 437-445.
[103] FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Proc of the Conf on Empirical Methods in Natual Language Processing. Stroudsburg, PA: ACL, 2011: 1535-1545.
[104] MENDES P N, MUHLEISEN H, BIZER C. Sieve: Linked data quality assessment and fusion[C]//Proc of the 2nd Int Workshop on Linked Web Data Management at Extending Database Technology. New York: ACM, 2012: 116-123.
[105] DONG Xin, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion[C]//Proc of the 20th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 601-610.
[106] TAN C H, AGICHTEIN E, IPEIROTIS P, et al. Trust, but verify: Predicting contribution quality for knowledge base construction and curation[C]//Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 553-562.
[107] 耿霞, 张继军, 李蔚妍. 知识图谱构建技术综述[J]. 计算机科学, 2014, 41(7): 148-152.
GENG Xia, ZHANG Ji-jun, LI Wei-yan. Knowledge graph construction techniques[J]. Computer Science, 2014, 41(7): 148-152.
[108] 描述逻辑. 描述逻辑基础知识[EB/OL]. (2014-02-24). http://www.2cto.com/database/201402/280927.html.
Description Logic. The foundation for description logic[EB/OL]. (2014-02-24). http://www.2cto.com/database/201402/280927.html.
[109] LEE T W, LEWICKI M S, GIROLAMI M, et al. Blind source separation of more sources than mixtures using overcomplete representation[J]. Signal Processing Letters, 1999, 6(4): 87-90.
[110] Ian Dickinson. Imp Iementation experience with the DIG 1.1specification[EB/OL]. (2004-05-10). http://www.hpl.hp.com/semweb/publications.html.
[111] 龚资. 基于OWL描述的本体推理研究[D]. 长春: 吉林大学, 2007.
GONG Zi. Research on ontology reasoning based on OWL[D]. Changchun: Jilin University, 2007.
[112] LIU Shao-yuan, HSU K H, KUO Li-jing. A semantic service match approach based on wordnet and SWRL rules[C]//Proc of the 10th IEEE Int Conf on E-Business Engineering. Piscataway, NJ: IEEE, 2013: 419-422.
[113] LAO N, MITCHELL T, COHEN W W. Random walk inference and learning in a large scale knowledge base[C]//Proc of EMNLP. Stroudsburg, PA: ACL, 2011: 529-539.
[114] 王志, 夏士雄, 牛强. 本体知识库的自然语言查询重写研究[J]. 微电子学与计算机, 2009, 26(8): 137-139.
WANG Zhi, XIA Shi-xiong, NIU Qiang. Research on natual language query rewriting for ontology-based knowledge base[J]. Microellectronics & Computer, 2009, 26(8): 137-139.
[115] BLANCO R, CAMBAZOGLU B B, MIKE P, et al. Entity recommendation in web search[C]//Pro of the 12th International Semantic Web Conference(ISWC). Berlin: Springer-Verlag, 2013: 33-48.
[116] BRACHMAN R J. What IS-A is and isn't: an analysis of taxonomic links in semantic networks[J]. Computer; (United States), 1983, 10(1): 5-13.
[117] Facebook. Facebook[EB/OL]. [2014-02-04]. https://www.facebook.com/.
[118] Twitter. Twitter[EB/OL]. [2016-05-08]. https://twitter.com/.
[119] 百度百科.搜狗知立方[EB/OL]. [2016-05-07]. http://baike.baidu.com/link?url=_J_2r2xYz0q-STwlYxqPZ00ZZuYyiA_kkZAohtC5EhmIzOjSwywKheEThY2gdXdzxS

8_8ickSMomF5t-M0qxMa.
Baidu Baike. Sogou zhi lifang[EB/OL]. [2016-05-07]. http://baike.baidu.com/link?url=_J_2r2xYz0q-STwlYxqPZ00ZZuYyiA_kkZAohtC5EhmIzOjSwywKheEThY2gdXdzxS8_8ickSMomF5t-M0qxMa.
[120] Baidu. Zhi xin[EB/OL]. [2016-06-08]. http://baike.baidu.com/link?url=V8IzyRd1vMlq5kq-BVMNf0yZHLmgXQ_SuPdjS3VW49NnzwurEQboXBTCQaEEa-SacwD2Emdn5wH6OFxQ5goU3a.
[121] Fader. Paralex[EB/OL]. [2016-05-08]. http://knowitall.cs.washington.edu/paralex.
[122] 百度百科. Siri[EB/OL]. [2016-05-02]. http://baike.baidu.com/subview/6573497/7996501.htm.
Baidu Baike. Siri[EB/OL]. [2016-05-02]. http://baike.baidu. com/subview/6573497/7996501.htm.
[123] 百度百科. Evi[EB/OL]. [2016-03-18]. http://baike.baidu.com/view/7574050.htm.
Baidu Baike. Evi[EB/OL]. [2016-03-18]. http://baike.baidu.com/view/7574050.htm.
[124] 百度. 度秘[EB/OL]. (2015-09-13). http://xiaodu.baidu.com/.
Baidu. Dumi[EB/OL]. (2015-09-13). http://xiaodu.baidu.com/.
[125] 百度百科. OASK 问答系统[EB/OL]. [2016-03-27]. http:// baike.baidu.com/view/8206827.htm.
Baidu Baike. OASK Question answering system[EB/OL]. [2016-03-27]. http://baike.baidu.com/view/8206827.htm.
[126] 百度百科. Graph search[EB/OL]. [2016-01-22]. http://baike.baidu.com/view/9966007.htm.
Baidu Baike. Graph search[EB/OL]. [2016-01-22]. http://baike.baidu.com/view/9966007.htm.
[127] 李文哲. 互联网金融, 如何用知识图谱识别欺诈行为[EB/OL]. [2016-01-09]. http://mp.weixin.qq.com/s?__biz=MjM5MTQzNzU2NA==&mid=401686695&idx=1&sn=aa7ca7f5c448075771ebd3533857b422&scene=23&srcid=0109l5viLtbzxKSzzw5hFdIg#rd.
LI Wen-zhe. Internet finance, how to identifiy fraud conduct using knowledge graph[EB/OL]. [2016-01-09]. http://mp.weixin.qq.com/s?__biz=MjM5MTQzNzU2NA==&mid=401686695&idx=1&sn=aa7ca7f5c448075771ebd3533857b422&scene=23&srcid=0109l5viLtbzxKSzzw5hFdIg#rd.
[128] Senselab. Center for medical informatics at yale university school of medicine yale university school of medicine [EB/OL]. [2016-01-08]. http://ycmi.med.yale.edu/.
[129] 田玲, 马丽仪. 基于用户体验的网站信息服务水平综合评价研究[J]. 生态经济, 2013(10): 160-162.
TIAN Ling, MA Li-yi. Research on comprehensive evaluation of website information services based on user experience[J]. Ecological Economy, 2013(10): 160-162.
[130] 一淘网.知识图谱[EB/OL]. (2014-12-12). https://www.aliyun.com/zixun/aggregation/13323.html.
ETao. Knowledge graph[EB/OL]. (2014-12-12). https://www. aliyun.com/zixun/aggregation/13323.html.
[131] 李涓子. 知识图谱: 大数据语义链接的基石[EB/OL]. (2015-02-20). http://www.cipsc.org.cn/kg2/.
LI Juan-zi. Knowledge graph: the foundation for big data semantic link[EB/OL]. (2015-02-20). http://www.cipsc.org. cn/kg2/.
[132] ZENG D, LIU K, CHEN Y, et al. Distant supervision for relation extraction via piecewise convolutional neural networks[C]//Proc of EMNLP. Stroudsburg, PA: ACL, 2015.
[133] WANG Z, LI J, WANG Z, et al. Cross-lingual knowledge linking across wiki knowledge bases[C]//Proc of the 2012 Int Conf on World Wide Web. New York: ACM, 2012: 459-468.

编 辑 税 红



徐增林(1980—), 教授, 博士生导师, 中组部“青年千人计划”入选者, 现任电子科技大学大数据研究中心数据挖掘与推理研究所轮值所长。主要研究兴趣为机器学习及其在社会网络分析、互联网、计算生物学、信息安全等方面的应用。他在包括IEEE TPAMI, IEEE TNN, NIPS, ICML, IJCAI, AAAI等顶级会议和刊物发表论文近30篇，引用近千次，发表专著2部，书籍章节2篇。

于2012年在多伦多召开的国际人工智能大会(AAAI)上做教学报告。是JMLR，IEEE TPAMI等机器学习与人工智能领域主要期刊的审稿人和香港教育资助局的基金评审人；多次担任人工智能领域的主要国际会议，如AAAI/IJCAI等会议的程序委员会成员；多次担任机器学习和大数据研究方面研讨会的组织委员会主席。