



# 知识图谱 发展报告(2018)

KNOWLEDGE GRAPH

DEVELOPMENT REPORT

中国·北京  
2018.08

语言与知识计算专委会  
中国中文信息学会



# 前言

## 1. 知识图谱的研究目标与意义

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。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。

知识图谱技术是指知识图谱建立和应用的技术，是融合认知计算、知识表示与推理、信息检索与抽取、自然语言处理与语义 Web、数据挖掘与机器学习等方向的交叉研究。知识图谱于 2012 年由谷歌提出并成功应用于搜索引擎，知识图谱属于人工智能重要研究领域——**知识工程的研究范畴**，是利用知识工程建立大规模知识资源的一个杀手锏应用。94 年图灵奖获得者、知识工程的建立者费根鲍姆给出的知识工程定义——将知识集成到计算机系统从而完成只有特定领域专家才能完成的复杂任务。在大数据时代，知识工程是从大数据中自动或半自动获取知识，建立基于知识的系统，以提供互联网智能知识服务。大数据对智能服务的需求，已经从单纯的搜集获取信息，转变为自动化的知识服务。我们需要利用知识工程为大数据添加语义/知识，使数据产生智慧（smart data），完成从数据到信息到知识，最终到智能应用的转变过程，从而实现对大数据的洞察、提供用户关心问题的答案、为决策提供支持、改进用户体验等目标。知识图谱在下面应用中已经凸显出越来越重要的应用价值：

- 知识融合：当前互联网大数据具有分布异构的特点，通过知识图谱可以对这些数据资源进行语义标注和链接，建立以知识为中心的资源语义集成服务；
- 语义搜索和推荐：知识图谱可以将用户搜索输入的关键词，映射为知识图谱中客观世界的概念和实体，搜索结果直接显示出满足用户需求的结构化信息内容，而不是互联网网页；
- 问答和对话系统：基于知识的问答系统将知识图谱看成一个大规模知识库，通过理解将用户的问题转化为对知识图谱的查询，直接得到用户关心问题的答案；
- 大数据分析与决策：知识图谱通过语义链接可以帮助理解大数据，获得对大数据的洞察，提供决策支持。

## 2. 知识工程的发展历程

知识图谱的发展是人工智能重要分支知识工程在大数据环境中的成功应用。回顾知识工程四十年来发展历程，总结知识工程的演进过程和技术进展，体会知识工程为人工智能所做出的贡献和未来面临的挑战，可以将知识工程分成五个标志性的阶段，前知识工程时期、专家系统时期、万维网 1.0 时期，群体智能时期以及知识图谱时期。

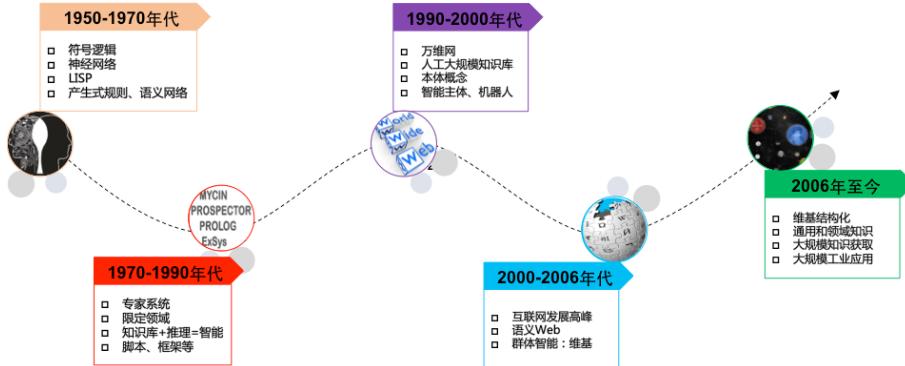


图 1. 知识工程发展历程

### 1950-1970 时期：图灵测试

人工智能旨在让机器能够像人一样解决复杂问题，图灵测试是评测智能的手段。这一阶段主要有两个方法：符号主义和连结主义。符号主义认为物理符号系统是智能行为的充要条件，连结主义则认为大脑（神经元及其连接机制）是一切智能活动的基础。这一阶段具有代表性的工作是通用问题求解程序（GPS）：将问题进行形式化表达，通过搜索，从问题初始状态，结合规则或表示得到目标状态。其中最成功应用是博弈论和机器定理证明等。这一时期的知识表示方法主要有逻辑知识表示、产生式规则、语义网络等。这一时代人工智能和知识工程的先驱 Minsky, McCarthy 和 Newell 以 Simon 四位学者因为在感知机、人工智能语言和通用问题求解和形式化语言方面的杰出工作分别获得了 1969 年、1971 年、1975 年的图灵奖。

### 1970-1990 时期：专家系统

通用问题求解强调利用人的求解问题的能力建立智能系统，而忽略了知识对智能的支持，使人工智能难以在实际应用中发挥作用。70 年开始，人工智能开始转向建立基于知识的系统，通过知识库+推理机实现智能，这一时期涌现出很多成功的限定领域专家系统，如 MYCIN 医疗诊断专家系统、识别分子结构的 DENRAL 专家系统以及计算机故障诊断 XCON 专家系统等。94 年图灵奖获得者 Feigenbaum 教授在 70 年代提出知识工程的定义，确立了知识工程在人工智能中的核心地位。这一时期知识表示方法有新的演进，包括框架和脚本等。80 年代后

期出现很多专家系统的开发平台，可以帮助将专家的领域知识转变成计算机可以处理的知识。

### **1990-2000 时期：万维网 Web 1.0**

在 1990 年代到 2000 年，出现了很多人工构建大规模知识库，包括广泛应用的英文 WordNet，采用一阶谓词逻辑知识表示的 Cyc 常识知识库，以及中文的 Hownet。Web1.0 万维网的产生为人们提供了一个开放平台，使用 HTML 定义文本的内容，通过超链接把文本连接起来，使得大众可以共享信息。W3C 提出的可扩展标记语言 XML，实现对互联网文档内容的结构通过定义标签进行标记，为互联网环境下大规模知识表示和共享奠定了基础。这一时期还提出了本体的知识表示方法。

### **2000-2006 时期：群体智能 Web 2.0**

Web1.0 万维网的出现使得知识从封闭知识走向开放知识，从集中知识成为分布知识。原来专家系统是系统内部定义的知识，现在可以实现知识源之间相互链接，可以通过关联来产生更多的知识而非完全由固定人生产。这个过程中出现了群体智能，最典型的代表就是维基百科，实际上是用户去建立知识，体现了互联网大众用户对知识的贡献，成为今天大规模结构化知识图谱的重要基础。也是在 2001 年，万维网发明人、2016 年图灵奖获得者 Tim Berners-Lee 提出语义 Web 的概念，旨在对互联网内容进行结构化语义表示，并提出互联网上语义标识语言 RDF（资源描述框架）和 OWL（万维网本体表述语言），利用本体描述互联网内容的语义结构，通过对网页进行语义标识得到网页语义信息，从而获得网页内容的语义信息，使人和机器能够更好地协同工作。

### **2006 年至今：知识图谱**

“知识就是力量”，将万维网内容转化为能够为智能应用提供动力的机器可理解和计算的知识是这一时期的目标。从 2006 年开始，大规模维基百科类富结构知识资源的出现和网络规模信息提取方法的进步，使得大规模知识获取方法取得了巨大进展。与 Cyc、WordNet 和 HowNet 等手工研制的知识库和本体的开创性项目不同，这一时期知识获取是自动化的，并且在网络规模下运行。当前自动构建的知识库已成为语义搜索、大数据分析、智能推荐和数据集成的强大资产，在大型行业和领域中正在得到广泛使用。典型的例子是谷歌收购 Freebase 后在 2012 年推出的知识图谱（Knowledge Graph），Facebook 的图谱搜索，Microsoft Satori 以及商业、金融、生命科学等领域特定的知识库。

最具代表性大规模网络知识获取的工作包括 DBpedia, Freebase, KnowItAll, WikiTaxonomy 和 YAGO，以及 BabelNet, ConceptNet, DeepDive, NELL, Probable, Wikidata, XLore, Zhishi.me 等。这些知识图谱遵循 RDF 数据模型，包含数以千

万级或者亿级规模的实体，以及数十亿或百亿事实（即属性值和其他实体的关系），并且这些实体被组织在成千上万的由语义类体现的客观世界的概念结构中。

现在我们看知识图谱的发展和应用状况，除了通用的大规模知识图谱，各行业也在建立行业和领域的知识图谱，当前知识图谱的应用包括语义搜索、问答系统与聊天、大数据语义分析以及智能知识服务等，在智能客服、商业智能等真实场景体现出广泛的应用价值，而更多知识图谱的创新应用还有待开发。

### 3. 知识图谱技术

人们通过概念掌握对客观世界的理解，概念是对客观世界事物的抽象，是将人们对世界认知联系在一起的纽带。知识图谱以结构化的形式描述客观世界中概念、实体及其关系。实体是客观世界中的事物，概念是对具有相同属性的事物的概括和抽象。本体是知识图谱的知识表示基础，可以形式化表示为， $O=\{C, H, P, A, I\}$ ， $C$  为概念集合，如事物性概念和事件类概念， $H$  是概念的上下位关系集合，也称为 Taxonomy 知识， $P$  是属性集合，描述概念所具有的特征， $A$  是规则集合，描述领域规则， $I$  是实例集合，用来描述实例-属性-值。Google 于 2012 年提出知识图谱，并在语义搜索中取得成功应用。知识图谱可以看做是本体知识表示的一个大规模应用，Google 知识图谱的知识表示结构主要描述客观存在实体和实体的关系，对于每个概念都有确定的描述这个概念的属性集合。

知识图谱技术是知识图谱建立和应用的技术，是语义 Web、自然语言处理和机器学习等的交叉学科。我们将知识图谱技术分为三个部分：知识图谱构建技术、知识图谱查询和推理技术，以及知识图谱应用。在大数据环境下，从互联网开放环境的大数据中获得知识，用这些知识提供智能服务互联网/行业，同时通过互联网可以获得更多的知识。这是一个迭代的相互增强过程，可以实现从互联网信息服务到智能知识服务的跃迁。

#### 3.1 知识图谱构建

##### 知识表示与建模

知识表示将现实世界中的各类知识表达成计算机可存储和计算的结构。机器必须要掌握大量的知识，特别是常识知识才能实现真正类人的智能。从有人工智能的历史开始，就有了知识表示的研究。知识图谱的知识表示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，为理解互联网内容提供了基础支撑。

##### 知识表示学习

随着以深度学习为代表的表示学习的发展，面向知识图谱中实体和关系的表示学习也取得了重要的进展。知识表示学习将实体和关系表示为稠密的低维向量，

实现了对实体和关系的分布式表示，可以高效地对实体和关系进行计算，缓解知识稀疏、有助于实现知识融合，已经成为知识图谱语义链接预测和知识补全的重要方法。由于知识表示学习能够显著提升计算效率，有效缓解数据稀疏，实现异质信息融合，因此对于知识库的构建、推理和应用具有重要意义，值得广受关注、深入研究。

## 实体识别与链接

实体是客观世界的事物，是构成知识图谱的基本单位（这里实体指个体或者实例）。实体分为限定类别的实体（如常用的人名、地名、组织机构等）以及开放类别实体（如药物名称、疾病等名称）。实体识别是识别文本中指定类别的实体。实体链接是识别出文本中提及实体的词或者短语（称为实体提及），并与知识库中对应实体进行链接。

实体识别与链接是知识图谱构建、知识补全与知识应用的核心技术。实体识别技术可以检测文本中的新实体，并将其加入到现有知识库中。实体链接技术通过发现现有实体在文本中的不同出现，可以针对性的发现关于特定实体的新知识。实体识别与链接的研究将为计算机类人推理和自然语言理解提供知识基础。

## 实体关系学习

实体关系描述客观存在的事物之间的关联关系，定义为两个或多个实体之间的某种联系，实体关系学习就是自动从文本中检测和识别出实体之间具有的某种语义关系，也称为关系抽取。实体关系抽取分类预定义关系抽取和开放关系抽取。预定义关系抽取是指系统所抽取的关系是预先定义好的，比如知识图谱中定义好的关系类别，如上下位关系、国家—首都关系等。开放式关系抽取。开放式关系抽取不预先定义抽取的关系类别，由系统自动从文本中发现并抽取关系。实体关系识别是知识图谱自动构建和自然语言理解的基础。

## 事件知识学习

事件是促使事物状态和关系改变的条件，是动态的、结构化的知识。目前已存在的知识资源（如谷歌知识图谱）所描述多是实体以及实体之间的关系，缺乏对事件知识的描述。针对不同领域的不同应用，事件有不同的描述范畴。一种将事件定义为发生在某个特定的时间点或时间段、某个特定的地域范围内，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。一种将事件认为是细化了的主题，是由某些原因、条件引起，发生在特定时间、地点，涉及某些对象，并可能伴随某些必然结果的事情。事件知识学习，即将非结构化文本文本中自然语言所表达的事件以结构化的形式呈现，对于知识表示、理解、计算和应用意义重大。

知识图谱中的事件知识隐含互联网资源中，包括已有的结构化的语义知识、

数据库的结构化信息、半结构化的信息资源以及非结构化资源，不同性质的资源有不同的知识获取方法。

## 3.2 知识图谱查询和推理计算

### 知识存储和查询

知识图谱以图（Graph）的方式来展现实体、事件及其之间的关系。知识图谱存储和查询研究如何设计有效的存储模式支持对大规模图数据的有效管理，实现对知识图谱中知识高效查询。因为知识图谱的结构是复杂的图结构，给知识图谱的存储和查询带来了挑战。当前目前知识图谱多以三元存在的 RDF 形式进行存储管理，对知识图谱的查询支持 SPARQL 查询。

### 知识推理

知识推理从给定的知识图谱推导出新的实体跟实体之间的关系。知识图谱推理可以分为基于符号的推理和基于统计的推理。在人工智能的研究中，基于符号的推理一般是基于经典逻辑(一阶谓词逻辑或者命题逻辑)或者经典逻辑的变异(比如说缺省逻辑)。基于符号的推理可以从一个已有的知识图谱推理出新的实体间关系，可用于建立新知识或者对知识图谱进行逻辑的冲突检测。基于统计的方法一般指关系机器学习方法，即通过统计规律从知识图谱中学习到新的实体间关系。知识推理在知识计算中具有重要作用，如知识分类、知识校验、知识链接预测与知识补全等。

## 3.3 知识图谱应用

### 通用和领域知识图谱

知识图谱分为通用知识图谱与领域知识图谱两类，两类图谱本质相同，其区别主要体现在覆盖范围与使用方式上。通用知识图谱可以形象地看成一个面向通用领域的结构化的百科知识库，其中包含了大量的现实世界中的常识性知识，覆盖面广。领域知识图谱又叫行业知识图谱或垂直知识图谱，通常面向某一特定领域，可看成是一个基于语义技术的行业知识库，因其基于行业数据构建，有着严格而丰富的数据模式，所以对该领域知识的深度、知识准确性有着更高的要求。

### 语义集成

语义集成的目标就是将不同知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序间的交互提供语义互操作性。常用技术方法包括本体匹配(也称为本体 映射)、实例匹配(也称为实体对齐、对象共指消解)以及知识融合等。语义集成是知识图谱研究中的一个核心问题，对于链接数据和知识融合至关重要。语义集成研究对于提升基于知识图谱的信息服务水平和智能化程度，

推动语义网以及人工智能、数据库、自然语言处理等相关领域的研究发展，具有重要的理论价值和广泛的应用前景，可以创造巨大的社会和经济效益。

### 语义搜索

知识图谱是对客观世界认识的形式化表示，将字符串映射为客观事件的事务（实体、事件以及之间的关系）。当前基于关键词的搜索技术在知识图谱的知识支持下可以上升到基于实体和关系的检索，称之为语义搜索。语义搜索利用知识图谱可以准确地捕捉用户搜索意图，借助于知识图谱，直接给出满足用户搜索意图的答案，而不是包含关键词的相关网页的链接。

### 基于知识的问答

问答系统(Question Answering, QA)是指让计算机自动回答用户所提出的问题，是信息服务的一种高级形式。不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准的自然语言形式的答案。华盛顿大学图灵中心主任 Etzioni 教授 2011 年曾在 Nature 上发表文章《Search Needs a Shake-Up》，其中明确指出：“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态” [Etzioni O., 2011]。因此，问答系统被看做是未来信息服务的颠覆性技术之一，被认为是机器具备语言理解能力的主要验证手段之一。

## 4. 报告的宗旨和组织结构

中国中文信息学会语言与知识计算专委会旨在为学术界和工业界提供在知识图谱、语义计算和语言理解等方面的产学研用的交流平台，提升语言与知识计算学术方向在国家科学的研究和国际学术方面的影响力，促进研究成果应用和向产品的转化。

《知识图谱发展报告》是语言与知识计算专委会邀请知识图谱技术领域专家对本学科方向和前沿技术的一次梳理，并在今后定期更新最新进展。我们的定位是深度科普，旨在向政府、企业、媒体等对知识图谱感兴趣的研究机构和企业界介绍相关领域的基本概念、研究和应用方向，向高校、科研院所和高技术企业中从事相关工作的专业人士介绍相关领域的前沿技术和发展趋势。

为此根据知识图谱技术研究内容，我们邀请各个方向的学者撰写了各个方向的报告：

- ◆ 前言：李涓子（清华大学）、赵军（中国科学院自动化研究所）
- ◆ 知识表示与建模：陈华钧（浙江大学）
- ◆ 知识表示学习：刘知远、林衍凯（清华大学）
- ◆ 实体识别与链接：韩先培、孙乐（中国科学院软件研究所）

- ◆ 实体关系学习：陈玉博（中国科学院自动化研究所）
- ◆ 事件知识学习：侯磊、许斌（清华大学）、胡琳梅（北京邮电大学）
- ◆ 知识存储和查询：邹磊（北京大学）、彭鹏（湖南大学）
- ◆ 知识推理：漆桂林（东南大学）、王泉（中国科学院信息工程研究所）、季秋（南京邮电大学）
- ◆ 通用与领域知识图谱：王昊奋（上海乐言科技）、丁军（上海海义知信息科技有限公司），张伟（阿里巴巴网络技术有限公司）
- ◆ 语义集成：胡伟（南京大学）
- ◆ 语义搜索：秦兵、刘铭（哈尔滨工业大学）
- ◆ 基于知识的问答：刘康、何世柱（中国科学院自动化研究所）

最后由韩先培(中国科学院软件研究所)、刘康(中国科学院自动化研究所)、侯磊(清华大学)等对初稿反馈意见，校对统一成文。

由于时间仓促，难免有疏漏，甚至错误的地方，仅供有志于语言与知识计算研究和知识图谱研究和开发的同仁参考，激发更广泛的思考和讨论，期待在我们的共同努力下知识图谱以及语义计算技术能够取得更辉煌的成绩！

中国中文信息学会  
语言与知识计算专委会  
2018年8月

# 目录

第一章 知识表示与建模.....	1
第二章 知识表示学习.....	12
第三章 实体识别与链接.....	21
第四章 实体关系学习.....	29
第五章 事件知识学习.....	45
第六章 知识存储与查询.....	65
第七章 知识推理.....	83
第八章 通用和领域知识图谱.....	98
第九章 语义集成.....	124
第十章 语义搜索.....	134
第十一章 基于知识的问答.....	145

# 第一章 知识表示与建模

## 1. 什么是知识表示

尽管人工智能依靠机器学习技术的进步取得了巨大的进展，例如，AlphaGo Zero 不依赖人类知识的监督，通过自我强化学习击败获得极高的棋力，但人工智能在很多方面，如语言理解、视觉场景理解、决策分析等，仍然举步维艰。一个关键的问题就是，机器必须要掌握大量的知识，特别是常识知识才能实现真正类人的智能。

哲学家柏拉图把知识（Knowledge）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。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。人工智能的核心也是研究怎样用计算机易于处理的方式表示、学习和处理各种各样的知识。知识表示是现实世界的可计算模型（Computable Model of Reality），广义的讲，神经网络也是一种知识表示形式。

上世纪 90 年代，MIT AI 实验室的 R. Davis 定义了知识表示的五大用途或特点：

- (1) 客观事物的机器标示（A KR is a Surrogate），即知识表示首先需要定义客观实体的机器指代或指称。
- (2) 一组本体约定和概念模型（A KR is a Set of Ontological Commitments），即知识表示还需要定义用于描述客观事物的概念和类别体系。
- (3) 支持推理的表示基础（A KR is a Theory of Intelligent Reasoning），即知识表示还需要提供机器推理的模型与方法。
- (4) 用于高效计算的数据结构（A KR is a medium for Efficient Computation），即知识表示也是一种用于高效计算的数据结构。
- (5) 人可理解的机器语言（A KR is a Medium of Human Expression），即知识表示还必须接近于人认知，是人可理解的机器语言。

有关知识表示的研究可以追溯到人工智能的早期研究。例如，1960 年，认知科学家 Allan M. Collins 提出了 Semantic Network（语义网络）的知识表示方法，以网络的方式来描述概念之间的语义关系。典型的语义网络如 WordNet 属于词典类的知识库，主要定义名词、动词、形容词和副词之间的语义关系。1970 年，

随着专家系统的提出和商业化发展，知识库构建和知识表示更加得到重视。传统的专家系统通常包含知识库（Knowledge Base）和推理引擎（Inference Engine）两个核心模块。早期专家系统最常用的知识表示方法包括基于框架的语言（Frame-based Languages）和产生式规则（Production Rules）等。框架语言主要用于描述客观世界的类别、个体、属性及关系等，较多的被应用于辅助自然语言理解。产生式规则主要用于描述类似于 IF-THEN 的逻辑结构，适合于刻画过程性知识。

不论是语义网络，还是框架语言和产生式规则都缺少严格的语义理论模型和形式化的语义定义。为了解决这一问题，人们开始研究具有较好的理论模型基础和算法复杂度的知识表示框架。比较有代表性的是描述逻辑语言（Description Logic）。描述逻辑是目前大多数本体语言（如 OWL）的理论基础。第一个描述逻辑语言是 1985 年由 Ronald J. Brachman 等提出的 KL-ONE。描述逻辑主要用于刻画概念（Concepts）、属性（Roles）、个体（Individual）、关系（Relationships）、元语（Axioms，即逻辑描述 Logic Statement）等知识表达要素。与传统专家系统的知识表示语言不同，描述逻辑家族更为关心知识表示能力和推理计算复杂性之间关系，并深入研究了各种表达构件的组合所带来的查询、分类、一致性检测等推理计算的计算复杂度问题。

1998 年，Web 之父 Tim Berners Lee 提出了 Semantic Web 的概念。其早期理想是希望把传统基于超文本链接的 Web 逐步转化为基于实体链接的语义网。语义网的基础数据模型 RDF 受到了元数据模型、框架系统和面向对象语言等多方面的影响，其最初目的是为人们在 Web 上发布结构化数据提供一个标准的数据描述框架。RDF 最基本的表达构件是被称为三元组的（Subject, Predicate, Object）。与此同时，语义网进一步吸收描述逻辑的研究成果，发展出了用 OWL 系列标准化本体语言。现代知识图谱如 DBpedia、Yago、Freebase、Schema.ORG、Wikidata 等大都以语义网的表达模型为基础进行扩展或删减。

不论是早期专家系统时代的知识表示方法，还是语义网时代的知识表示模型，都属于以符号逻辑为基础的知识表示方法。符号知识表示的特点是易于刻画显性、离散的知识，因而具有内生的可解释性。但由于人类知识还包含大量不易于符号化的隐性知识，完全基于符号逻辑的知识表示通常由于知识的不完备而失去鲁棒性，特别是推理很难达到实用。由此催生了采用连续向量方式来表示知识的研究。

基于向量的方式表示知识的研究由来已有。表示学习的发展，以及自然语言处理领域词向量等嵌入（Embedding）技术手段的出现，启发了人们用类似于词向量的低维稠密向量的方式表示知识的研究。通过嵌入（Embedding）将知识图谱中的实体和关系投射到一个低维的连续向量空间，可以为每一个实体和关系学习出一个低维度的向量表示。这种基于向量的知识表示可以实现通过数值运算来

发现新事实和新关系，并能更有效的发现更多的隐性知识和潜在假设，这些隐性知识通常是人的主观不易于观察和总结出来的。更为重要的是，知识图谱嵌入也通常作为一种类型的先验知识辅助输入到很多深度神经网络模型中，用来约束和监督神经网络的训练过程。

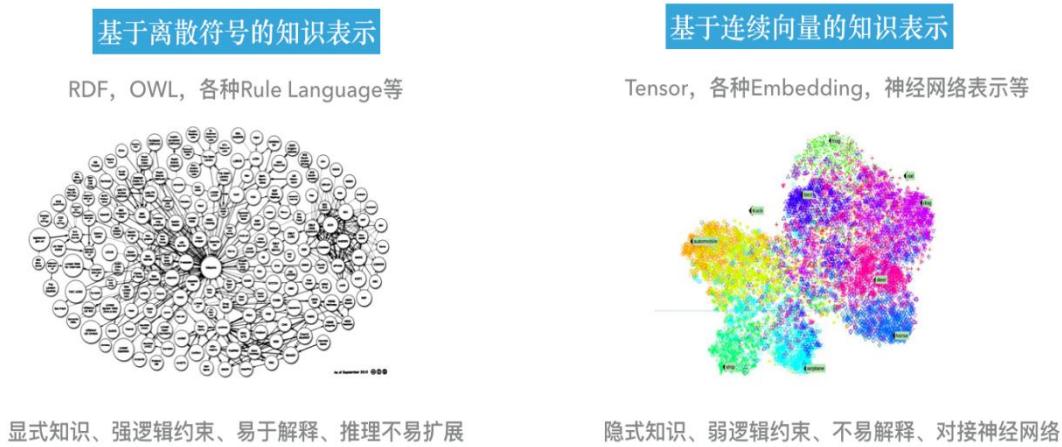


图 1. 基于离散符号的知识表示与基于连续向量的知识表示

综上所述，知识图谱时代的知识表示方法与传统人工智能相比，已经发生了很大的变化。一方面，现代知识图谱受到规模化扩展的要求，通常采用以三元组为基础的较为简单实用的知识表示方法，并弱化了对强逻辑表示的要求；另外一方面，由于知识图谱是很多搜索、问答和大数据分析系统的重要数据基础，基于向量的知识图谱表示使得这些数据更加易于与深度学习模型集成，使得基于向量空间的知识图谱表示得到越来越多的重视。

由于知识表示涉及大量传统人工智能的内容，并有其明确、严格的内涵及外延定义，为避免混淆，本文主要侧重于知识图谱的表示方法的介绍，因此将“知识表示”和“知识图谱的表示方法”加以了区分。

## 2. 知识图谱的表示方法

与传统专家系统时代的知识库不同，现代知识图谱通常规模巨大，这导致知识图谱的表示方法也与传统的知识表示有所不同。下面从知识图谱的规模化发展对知识表示带来的挑战出发，分别介绍了基于符号和基于向量的知识表示方法。

### 2.1 知识图谱的规模化带给知识表示的挑战

与传统专家系统时代主要依靠专家手工获取知识不同，现代知识图谱的显著特点是规模巨大，无法单一依靠人工和专家构建。传统的知识库，如由 Douglas Lenat 从 1984 年开始创建的常识知识库 Cyc 仅包含 700 万条<sup>1</sup>的事实描述

<sup>1</sup> 下文有关知识图谱规模的描述都以三元组（Triple）为计算单元，一个元组对应一条事实描述（Fact or

(Assertion)。Wordnet 主要依靠语言学专家定义名词、动词、形容词和副词之间的语义关系，目前包含大约 20 万条的语义关系。由著名人工智能专家 Marvin Minsky 于 1999 年起开始构建的 ConceptNet 常识知识库依靠了互联网众包、专家创建和游戏三种方法，但早期 ConceptNet 规模在百万级别，最新的 ConceptNet 5.0 也仅包含 2800 万 RDF 三元组关系描述。现代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已经达到百亿级别。DBpedia 已经包含约 30 亿 RDF 三元组，多语种的大百科语义网络 BabelNet 包含 19 亿的 RDF 三元组，Yago3.0 包含 1.3 亿元组，Wikidata 已经包含 4265 万条数据条目，元组数目也已经达到数十亿级别。截至目前，开放链接数据项目 Linked Open Data<sup>2</sup> 统计了其中有效的 2973 个数据集，总计包含大约 1494 亿三元组。

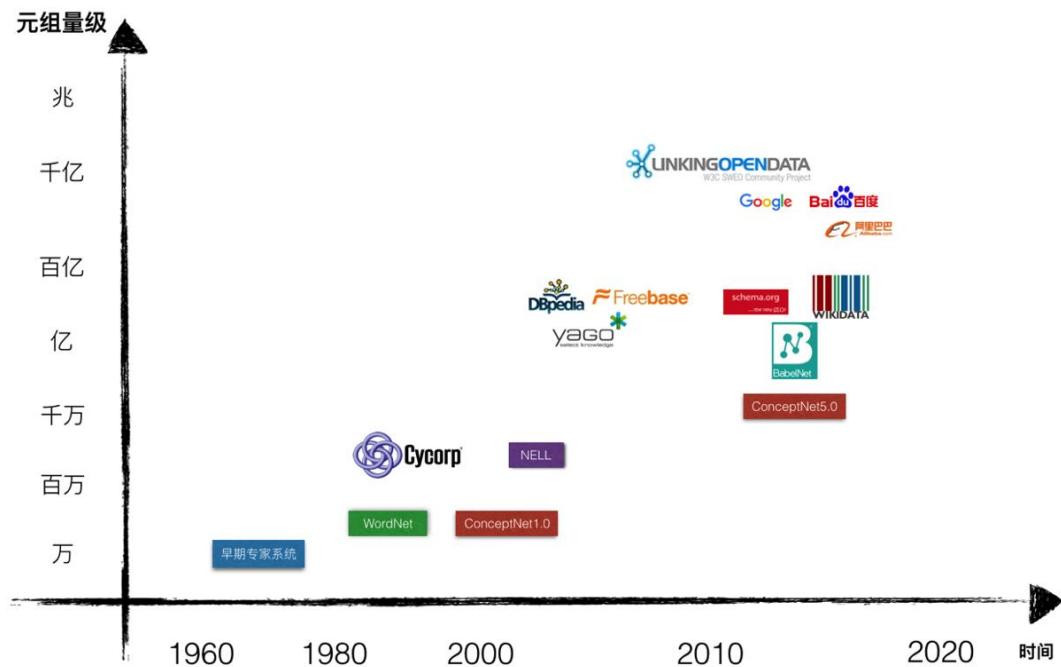


图 2. 知识库及知识图谱的规模发展趋势

现代知识图谱对知识规模的要求源于“知识完备性”难题。冯诺依曼曾估计单个个体的大脑中的全量知识需要  $2.4 \times 10^{20}$  个 bits 来存储。客观世界拥有不计其数的实体，人的主观世界更加包含有无法统计的概念，这些实体和概念之间又具有更多数量的复杂关系，导致大多数知识图谱都面临知识不完全的困境。在实际的领域应用场景中，知识不完全也是困扰大多数语义搜索、智能问答、知识辅助的决策分析系统的首要难题。

知识图谱对规模的扩展需求使得知识表示方法逐渐发生了四个方面的变化：  
 (1) 从强逻辑表达转化为轻语义表达；(2) 从较为注重 TBox 概念型知识转化

Assertion)

<sup>2</sup> <http://lod-cloud.net>

为更加注重 ABox 事实型知识；（3）从以推理为主要应用目标转化为综合搜索、问答、推理、分析等多方面的应用目标；（4）从以离散的符号逻辑表示向以连续的向量空间表示方向发展。

传统常识知识库如 Cyc 的知识表示语言主要以一阶谓词逻辑（FOPC）为基础，扩展了等价(Equality)、缺省推理(Default reasoning)、斯科林化(Skolemization)和部分二阶谓词逻辑等知识表示能力。基于描述逻辑（Description Logic）的本体语言（Ontology），如 EL++，为可判定可扩展的自动推理提供了知识表示理论基础，并更加侧重于 TBox 概念型知识。

而现代知识图谱如 Freebase、Wikidata、Yago、Schema.Org 等都在逻辑的语义表达方面降低了要求，并以事实型知识为主。例如，Freebase 的知识表示框架只包含如下几个要素：对象-Object，事实-Facts，类型-Types 和属性-Properties。“Object”代表实体；一个“Object”可以有一个或多个“Types”；“Properties”用来描述“Facts”；并使用复合值类型（CVT: Compound Value Types）来处理多元关系。Schema.Org 只定义轻量的 Schema，突出 ABox 事实型数据的重要性。

此外，随着表示学习与深度神经网络的发展，一个重要的发展趋势是基于向量的知识表示方法得到越来越多的重视。传统基于逻辑的符号知识表示的优点是基于显性知识表示，因而表示能力强，能处理较为复杂的知识结构，具有可解释性，并支持复杂的推理。基于表示学习的连续向量表示优点是易于捕获隐性知识，并易于与深度学习模型集成，缺点是对复杂知识结构的支持不够，可解释性差，不能支持复杂推理。目前，基于符号和基于向量的知识图谱表示并存并逐步相互融合。

## 2.2 基于符号的知识图谱表示方法

目前大多数知识图谱的实际存储方式都是以传统符号化的表示方法为主。大多数开放域的知识图谱都是基于语义网的表示模型进行了扩展或删改。下面主要以语义网的知识表示框架为例简要介绍基于符号的知识图谱表示方法。当然，语义网只是符号知识表示框架和方法的一种。

### ● RDF

RDF是最常用的符号语义表示模型。RDF的基本模型是有向标记图(Directed Labeled Graph)。图中的每一条边对应于一个三元组（Subject-主语,Predicate-谓语, Object-宾语）。一个三元组对于一个逻辑表达式或关于世界的陈述(Statement)。

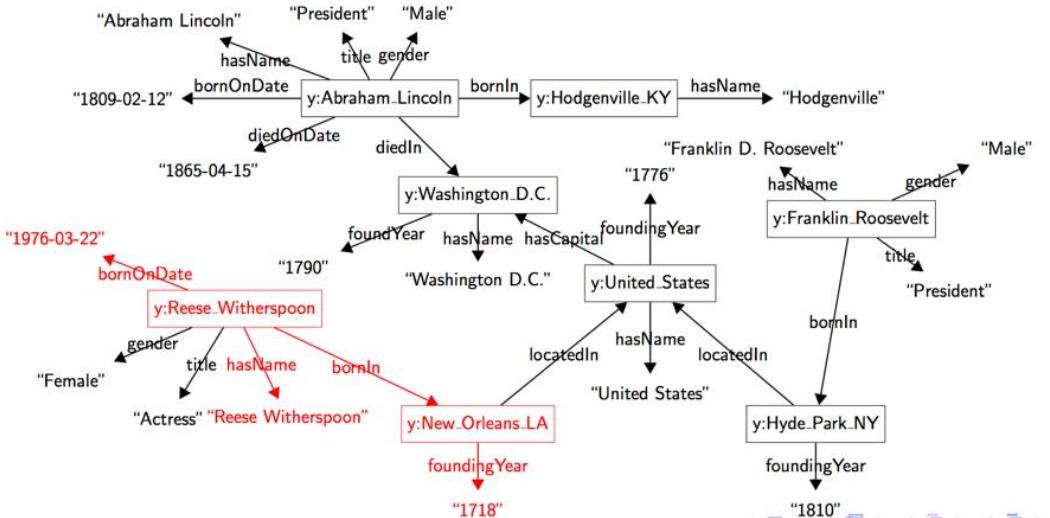


图 3. RDF 基于有向标记图模型

### ● RDFS

RDF 提供了描述客观世界事实的基本框架，但缺少类、属性等 Schema 层的定义手段。RDFS（RDF Schema）主要用于定义术语集、类集合和属性集合，主要包括如下元语： Class, subClassOf, type, Property, subPropertyOf, Domain, Range 等。基于这些简单的表达构件可以构建最基本的类层次体系和属性体系。

### ● OWL

OWL 主要在 RDFS 基础之上扩展了表示类和属性约束的表示能力，这使得可以构建更为复杂而完备的本体。这些扩展的本体表达能力包括：

- 1) 复杂类表达 Complex Classes, 如: intersection, union 和 complement 等;
- 2) 属性约束 Property Restrictions, 如: existential quantification, universal quantification, hasValue 等;
- 3) 基数约束 Cardinality Restrictions , 如 : maxQualifiedCardinality, minQualifiedCardinality, qualifiedCardinality 等;
- 4) 属性特征 Property Characteristics , 如: inverseOf, SymmetricProperty, AsymmetricProperty, propertyDisjointWith, ReflexiveProperty, FunctionalProperty 等。

OWL 以描述逻辑为主要理论基础，在很多领域知识图谱的构建，如医疗、金融、电商等有实际应用的价值。

## 2.3 基于向量的知识图谱表示学习模型

依据知识图谱嵌入表示模型建模原理将基于向量的知识表示模型划分为翻译模型、组合模型、神经网络模型。翻译模型的灵感来自 word2vec 中词汇关系的平移不变性，典型的方法包括基于向量的三角形法则和范数原理的 TransE 模型，通过超平面转化或线性变换处理多元关系的 TransH、TransR 和 TransD 模型，

通过增加一个稀疏度参数向量解决异构多元关系的 TranSparse 模型等。

组合模型采用的是向量的线性组合和点积原理，典型特征是将实体建模为列向量、关系建模为矩阵，然后通过头实体向量与关系矩阵的线性组合，再与尾实体进行点积来计算打分函数。经典成员包括采用普通矩阵的 RESCAL、采用低秩矩阵的 LFM、采用对角矩阵的 DistMult 和采用循环矩阵的 Hole。神经网络模型采用神经网络拟合三元组，典型模型包括采用单层线性或双线性网络的 SME、采用单层非线性网络的 SLM、NTN 和 MLP，以及采用多层网络结构的 NAM。下一章对知识图谱的表示学习模型进行了详细介绍，此处不再赘述。

### 3. 常见知识库及知识图谱的知识表示方法

从人工智能的概念被提出开始，构建大规模的知识库一直都是人工智能、自然语言理解等领域的核心任务之一。下面分别介绍了早期知识库和以语义网为基础构建的知识图谱项目所采用的知识表示方法。不同的知识图谱项目都会根据实际的需要选择不同的知识表示框架。这些框架有着不同的描述术语、表达能力、数据格式等方面的考虑，但本质上具有相似之处。

#### 3.1 早期的知识库项目

Cyc 是持续时间最长，影响范围较广，争议也较多的知识库项目。Cyc 是在 1984 年由 Douglas Lenat 开始创建。最初的目标是要建立人类最大的常识知识库。典型的常识知识如 “Every tree is a plant”， “Plants die eventually” 等。Cyc 知识库的知识表示框架主要由术语 Terms 和断言 Assertions 组成。Terms 包含概念、关系和实体的定义。Assertions 用来建立 Terms 之间的关系，这既包括事实 Fact 描述，也包含规则 Rule 的描述。最新的 Cyc 知识库已经包含有 50 万条 Terms 和 700 万条 Assertions。Cyc 的主要特点是基于形式化的知识表示方法来刻画知识。形式化的优势是可以支持复杂的推理。但过于形式化也导致知识库的扩展性和应用的灵活性不够。Cyc 提供开放版本 OpenCyc。

WordNet 是最著名的词典知识库，主要用于词义消歧。WordNet 由普林斯顿大学认识科学实验室从 1985 年开始开发。WordNet 的表示框架主要定义了名词、动词、形容词和副词之间的语义关系。例如名词之间的上下位关系（如：“猫科动物”是“猫”的上位词），动词之间的蕴含关系（如：“打鼾”蕴含着“睡眠”）等。WordNet3.0 已经包含超过 15 万个词和 20 万个语义关系。

ConceptNet 是常识知识库。最早源于 MIT 媒体实验室的 Open Mind Common Sense (OMCS) 项目。OMCS 项目是由著名人工智能专家 Marvin Minsky 于 1999 年建议创立。ConceptNet 主要依靠互联网众包、专家创建和游戏三种方法来构建。ConceptNet 知识库以三元组形式的关系型知识构成。ConceptNet 5 版本已经包含有 2800 万关系描述。与 Cyc 相比，ConceptNet 采用了非形式化、更加接近自然

语言的描述，而不是像 Cyc 那样采用形式化的谓词逻辑。与链接数据和谷歌知识图谱相比，ConceptNet 比较侧重于词与词之间的关系。从这个角度看，ConceptNet 更加接近于 WordNet，但是又比 WordNet 包含的关系类型多。此外，ConceptNet 完全免费开放，并支持多种语言。

ConceptNet5 的知识表示框架主要包含如下要素：概念-Concepts、词-Words、短语-Phrases、断言 Assertions、关系-Relations、边-Edges。Concepts 由 Words 或 Phrases 组成，构成了图谱中的节点。与其它知识图谱的节点不同，这些 Concepts 通常是从自然语言文本中提取出来的，更加接近于自然语言描述，而不是形式化的命名。Assertions 描述了 Concepts 之间的关系，类似于 RDF 中的 Statements。Edges 类似于 RDF 中的 Property。一个 Concepts 包含多条边，而一条边可能有多个产生来源。例如，一个“化妆 Cause 漂亮”的断言可能来源于文本抽取，也可能来源于用户的手工输入。来源越多，该断言就越可靠。ConceptNet 根据来源的多少和可靠程度计算每个断言的置信度。ConceptNet5 中的关系包含 21 个预定义的、多语言通用的关系（如：IsA、UsedFor 等）和从自然语言文本中抽取的更加接近于自然语言描述的非形式化的关系（如：on top of, caused by 等）。ConceptNet5 对 URI 进行了精心的设计。URI 同时考虑了类型（如，是概念还是关系）、语言、正则化后的概念名称、词性、歧义等因素。例如“run”是一个动词，但也可能是一个名词（如 basement 比赛中一个“run”），其 URI 为：“/c/en/run/n/basement”。其中，n 代指这是一个名词，basement 用于区分歧义。在处理表示“x is the first argument of y”这类多元关系的问题上，ConceptNet5 把所有关于某条边的附加信息增加为边的属性。

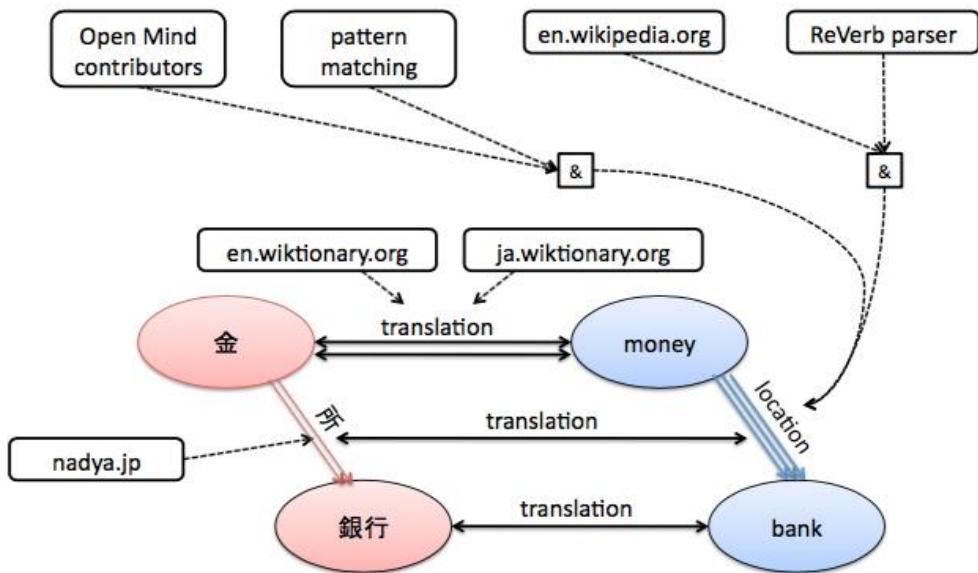


图 4. ConceptNet 5 的知识表示框架

### 3.2 语义网与知识图谱

互联网的发展为知识工程提供了新的机遇。在一定程度上，是互联网的出现帮助突破了传统知识工程在知识获取方面的瓶颈。从 1998 年 Tim Berners Lee 提出语义网至今，涌现出大量以互联网资源为基础的新一代知识库。这类知识库的构建方法可以分为三类：互联网众包、专家协作和互联网挖掘。

Freebase 是一个开放共享的、协同构建的大规模链接数据库。Freebase 是由硅谷创业公司 MetaWeb 于 2005 年启动的一个语义网项目。2010 年，谷歌收购了 Freebase 作为其知识图谱数据来源之一。Freebase 主要采用社区成员协作方式构建。其主要数据来源包括维基百科 Wikipedia、世界名人数据库 NNDB、开放音乐数据库 MusicBrainz，以及社区用户的贡献等。Freebase 基于 RDF 三元组模型，底层采用图数据库进行存储。Freebase 的一个特点是不对顶层本体做非常严格的控制，用户可以创建和编辑类和关系的定义。2016 年，谷歌宣布将 Freebase 的数据和 API 服务都迁移至 Wikidata，并正式关闭了 Freebase。

Freebase 的知识表示框架主要包含以下几个要素：对象-Object，事实-Facts，类型-Types 和属性-Properties。“Object”代表实体。每一个“Object”有一个唯一的 ID，称为 MID(Machine ID)。一个“Object”可以有一个或多个“Types”。“Properties”用来描述“Facts”。例如：“Barack Obama”是一个 Object，并拥有一个唯一的 MID：“/m/02mjmr”。这个 Object 的一个 type 是“/government/us\_president”，并有一个称为“/government/us\_president/presidency\_number”的 Property，其数值是“44”。Freebase 使用复合值类型（CVT: Compound Value Types）来处理多元关系。

例如下面这个例子中的 CVT 描述了关于 Obama 的任职期限的多元关系“government\_position\_held”。这个多元关系包含多个子二元关系：“office\_holder”，“office\_position”，“from”，“to”等。一个 CVT 就是一个有唯一 MID 的 Object，也可以有多个 Types。为了以示区别，Freebase 把所有非 CVT 的 Object 也称为“Topic”。

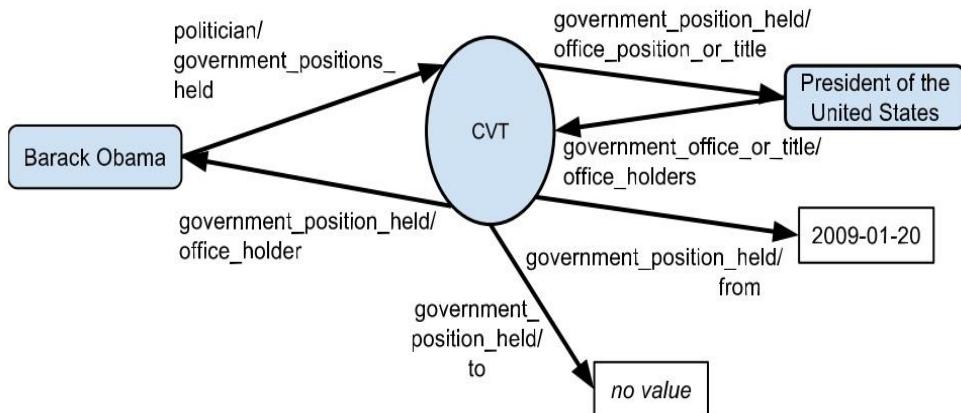


图 5. Freebase 的知识表示框架示例

DBpedia 是早期的语义网项目。DBpedia 意指数据库版本的 Wikipedia，是从 Wikipedia 抽取出来的链接数据集。DBpedia 采用了一个较为严格的本体，包含人、地点、音乐、电影、组织机构、物种、疾病等类定义。此外，DBpedia 还与 Freebase，OpenCYC、Bio2RDF 等多个数据集建立了数据链接。DBpedia 采用了 RDF 语义数据模型，总共包含 30 亿 RDF 三元组。

Schema.org: Schema.org 是 2011 年起，由 Bing、Google、Yahoo 和 Yandex 等搜索引擎公司共同支持的语义网项目。Schema.org 支持各个网站采用语义标签（Semantic Markup）的方式将语义化的链接数据嵌入到网页中。搜索引擎自动搜集和归集这些，快速的从网页中抽取语义化的数据。Schema.org 提供了一个词汇本体用于描述这些语义标签。截止目前，这个词汇本体已经包含 600 多个类和 900 多个关系，覆盖范围包括：个人、组织机构、地点、时间、医疗、商品等。谷歌于 2015 年推出的定制化知识图谱支持个人和企业在其网页中增加包括企业联系方法、个人社交信息等在内的语义标签，并通过这种方式快速的汇集高质量的知识图谱数据。

WikiData: WikiData 的目标是构建一个免费开放、多语言、任何人或机器都可以编辑修改的大规模链接知识库。WikiData 由维基百科于 2012 年启动，早期得到微软联合创始人 Paul Allen、Gordon Betty Moore 基金会以及 Google 的联合资助。WikiData 继承了 Wikipedia 的众包协作的机制，但与 Wikipedia 不同，WikiData 支持的是以三元组为基础的知识条目（Items）的自由编辑。一个三元组代表一个关于该条目的陈述（Statements）。例如可以给“地球”的条目增加“<地球，地表面积是，五亿平方公里>”的三元组陈述。截止 2016 年，WikiData 已经包含超过 2470 多万个知识条目。

WikiData 的知识表示框架主要包含如下要素：页面-Pages，实体-Entities，条目-Items，属性-Properties，陈述-Statements，修饰-Qualifiers，引用-Reference 等。WikiData 起源于 Wikipedia，因此，与 Wikipedia 一样，是以页面“Page”为基本组织单元。Entities 类似于 OWL: Things，代指最顶层的对象。每一个 Entity 都有一个独立的维基页面。主要有两类 Entities：Items 和 Properties。Item 类似于 RDF 中的 Instance，代指实例对象。Properties 和 Statement 分别等价于 RDF 中的 Property 和 Statement。通常一个 Item 的页面还包含有多个别名-aliases 和多个指向维基百科的外部链接-Sitelinks。每个 Entities 有多个 Statements。一个 Statement 包含：一个 Property、一个或多个 Values、一个或多个 Qualifiers，一个或多个 References、一个标示重要性程度的 Rank。修饰-Qualifiers 用于处理复杂的多元表示。如下图中的一个陈述“spouse: Jane Belson”描述了一个二元关系。我们可以使用 Qualifiers 给这个陈述增加多个附加信息来刻画多元关系，如：“start date: 25 November 1991” and “end date: 11 May 2011,”等。引用-References 用于标识每个

陈述的来源或出处，如来源于某个维基百科页面等。引用也是一种 Qualifiers，通常添加到 Statements 的附加信息中。WikiData 支持多种数值类型，包括：其自有的 Item 类型、RDF Literal、URL、媒体类型 Commons Media 和三种复杂类型：Time、Globe coordinates 和 Quantity。WikiData 允许给每个 Statement 增加三种权重：normal（缺省），preferred 和 deprecated。WikiData 定义了三种 Snacks 作为 Statement 的具体描述结构：PropertyValueSnack、PropertyNoValueSnack、PropertySomeValueSnack。“PropertyNoValueSnack”类似于 OWL 中的“Negation”，用于表示类似于“Elizabeth I of England had no spouse.”的知识。“PropertySomeValueSnack”类似于 OWL 中的存在量词“someValuesFrom”，用于表示类似于“Pope Linus had a date of birth, but it is unknown to us”这样的知识。WikiData 的 URI 机制遵循了 Linked Open Data 的 URI 原则，采用统一的 URI 机制：<http://www.wikidata.org/entity/<id>>。其中<id>可以是一个 Item，如 Q49，或者是一个 Property，如 P234。

## 4. 总结

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。人工智能的一个核心也是研究怎样用计算机易于处理的方式表示、学习和处理各种各样的人类知识。知识表示是现实世界的可计算模型，广义的讲，神经网络也是一种知识表示形式。

与传统专家系统时代的知识库表示方法不同，现代知识图谱由于要满足规模化的扩建需求，大多降低了对强逻辑表达的要求，并以三元组为基础的关系型知识为主。并更多的关注实例层面的知识构建。另外一方面，由于知识图谱是很多搜索、问答和大数据分析系统的重要数据基础，基于向量的知识图谱表示使得这些数据更加易于与深度学习模型集成，使得基于向量空间的知识图谱表示得到越来越多的重视。

此外，围绕知识图谱表示的一个重要的研究趋势和动态是：把符号逻辑与表示学习结合起来研究更加鲁棒、易于捕获隐含知识、易于与深度学习集成、并适应大规模知识图谱应用的新型表示框架。这需要较好的平衡符号逻辑的表示能力和表示学习模型的复杂性。一方面要能处理结构多样性、捕获表达构件的语义和支持较为复杂的推理，另一方面又要求学习模型的复杂性较低。这些新型的表示框架的研究对于知识图谱构建所涉及的抽取、融合、补全、问答和分析等任务都具有重要的基础性研究意义。

## 第二章 知识表示学习

### 1. 任务定义、目标和研究意义

知识表示是知识获取与应用的基础，因此知识表示学习问题，是贯穿知识库的构建与应用全过程的关键问题。人们通常以网络的形式组织知识库中的知识，网络中每个节点代表实体（人名、地名、机构名、概念等），而每条连边则代表实体间的关系。然而，基于网络形式的知识表示面临诸多挑战性难题，主要包括如下两个方面：

(1) 计算效率问题。基于网络的知识表示形式中，每个实体均用不同的节点表示。当利用知识库计算实体间的语义或推理关系时，往往需要人们设计专门的图算法来实现，存在可移植性差的问题。更重要的，基于图的算法计算复杂度高，可扩展性差，当知识库规模达到一定规模时，就很难较好地满足实时计算的需求。

(2) 数据稀疏问题。与其他类型的大规模数据类似，大规模知识库也遵守长尾分布，在长尾部分的实体和关系上，面临严重的数据稀疏问题。例如，对于长尾部分的罕见实体，由于只有极少的知识或路径涉及它们，对这些实体的语义或推理关系的计算往往准确率极低。

近年来，以深度学习[Bengio, et al., 2009]为代表的表示学习[Bengio, et al., 2013]技术异军突起，在语音识别、图像分析和自然语言处理领域获得广泛关注。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，两个对象距离越近，则说明其语义相似度越高。

知识表示学习，则是面向知识库中的实体和关系进行表示学习。知识表示学习实现了对实体和关系的分布式表示，它具有以下主要优点：

(1) 显著提升计算效率。知识库的三元组表示实际就是基于独热表示的。如前所分析的，在这种表示方式下，需要设计专门的图算法计算实体间的语义和推理关系，计算复杂度高，可扩展性差。而表示学习得到的分布式表示，则能够高效地实现语义相似度计算等操作，显著提升计算效率。

(2) 有效缓解数据稀疏。由于表示学习将对象投影到统一的低维空间中，使每个对象均对应一个稠密向量，从而有效缓解数据稀疏问题，这主要体现在两个方面。一方面，每个对象的向量均为稠密有值的，因此可以度量任意对象之间的语义相似程度。另一方面，将大量对象投影到统一空间的过程，能够将高频对象的语义信息用于帮助低频对象的语义表示，提高低频对象的语义表示的精确性。

(3) 实现异质信息融合。不同来源的异质信息需要融合为整体，才能得到有效应用。例如，人们构造了大量知识库，这些知识库的构建规范和信息来源均

有不同，例如著名的世界知识库有 DBpedia、YAGO、Freebase 等。大量实体和关系在不同知识库中的名称不同。如何实现多知识库的有机融合，对知识库应用具有重要意义。通过设计合理的表示学习模型，将不同来源的对象投影到同一个语义空间中，就能够建立统一的表示空间，实现多知识库的信息融合。此外，当我们在信息检索或自然语言处理中应用知识库时，往往需要计算查询词、句子、文档和知识库实体之间的复杂语义关联。由于这些对象的异质性，在往常是棘手问题。而知识表示学习亦能为此提供统一表示空间，轻而易举实现异质对象之间的语义关联计算。

综上，由于知识表示学习能够显著提升计算效率，有效缓解数据稀疏，实现异质信息融合，因此对于知识库的构建、推理和应用具有重要意义，值得广受关注、深入研究。

## 2. 研究内容和关键科学问题

知识表示学习是面向知识库中实体和关系的表示学习。通过将实体或关系投影到低维向量空间，我们能够实现对实体和关系的语义信息的表示，可以高效地计算实体、关系及其之间的复杂语义关联。这对知识库的构建、推理与应用均有重要意义。目前，已经在知识图谱补全、关系抽取等任务中取得了瞩目成果。但是，知识表示学习仍然面临很多挑战。

### 2.1 复杂关系建模

现有知识表示学习方法无法有效地处理知识图谱中的复杂关系。这里的复杂关系定义如下。按照知识库中关系两端连接实体的数目，可以将关系划分为 1-1、1-N、N-1 和 N-N 四种类型。例如 N-1 类型关系指的是，该类型关系中的一个尾实体会平均对应多个头实体，即我们将 1-N、N-1 和 N-N 称为复杂关系。研究发现，各种知识获取算法在处理四种类型关系时的性能差异较大，在处理复杂关系时性能显著降低。如何实现表示学习对复杂关系的建模成为知识表示学习的一个难点。

### 2.2 多源信息融合

知识表示学习面临的另外一个重要挑战如何实现多源信息融合。现有的知识表示学习模型仅利用知识图谱的三元组结构信息进行表示学习，尚有大量与知识有关的其他信息没有得到有效利用，例如：

- (1) 知识库中的其他信息，如实体和关系的描述信息、类别信息等；
- (2) 知识库外的海量信息，如互联网文本蕴含了大量与知识库实体和关系有关的信息。

如何充分融合这些多源异质信息，实现知识表示学习，具有重要意义，可以

改善数据稀疏问题，提高知识表示的区分能力。

### 2.3 关系路径建模

在知识图谱中，多步的关系路径也能够反映实体之间的语义关系。Lao 等人曾提出 Path-Constraint Random Walk[Lao, et al., 2010]、Path Ranking Algorithm[Lao, et al., 2010]等算法，利用两实体间的关系路径信息，预测它们的关系，取得显著效果，说明关系路径蕴含着丰富的信息。如何突破知识表示学习孤立学习每个三元组的局限性，充分考虑关系路径信息是知识表示学习的关键问题。

## 3. 技术方法和研究现状

知识表示学习是近年来的研究热点，研究者提出了多种模型，学习知识库中的实体和关系的表示。本节将主要介绍其中的代表方法。

结构表示[Bordes, et al. 2011]（Structured Embedding, SE）是较早的几个知识表示方法之一。对于一个事实三元组，SE 将头实体向量和尾实体向量通过关系的两个矩阵投影到关系的对应空间中，然后在该空间中计算两投影向量的距离。这个距离反映了两个实体在该关系下的语义相关度，它们的距离越小，说明这两个实体存在这种关系。然而，SE 模型有一个重要缺陷：它对头、尾实体使用两个不同的矩阵进行投影，协同性较差，往往无法精确刻画两实体与关系之间的语义联系。因此，单层神经网络模型[Socher, et al. 2013]（Single Layer Model, SLM）尝试采用单层神经网络的非线性操作，来减轻 SE 无法协同精确刻画实体与关系的语义联系的问题。虽然 SLM 是 SE 模型的改进版本，但是它的非线性操作仅提供了实体和关系之间比较微弱的联系。与此同时，却引入了更加高的计算复杂度。此外，语义匹配能量模型[Bordes, et al., 2012; Bordes, et al., 2014]（Semantic Matching Energy, SME）提出更复杂的操作，寻找实体和关系之间的语义联系。在 SME 中，每个实体和关系都用低维向量表示。在此基础上，SME 定义若干投影矩阵，利用双线性函数来刻画实体与关系的内在联系。同样利用双线性函数的还有隐变量模型[Sutskever, et al., 2009; Jenatton, et al., 2012]（Latent Factor Model, LFM），该模型提出利用基于关系的双线性变换，刻画实体和关系之间的二阶联系。与以往模型相比，LFM 取得巨大突破：通过简单有效的方法刻画了实体和关系的语义联系，协同性较好，计算复杂度低。后来的 DISTMULT 模型[Yang, et al., 2015]还探索了 LFM 的简化形式：将关系矩阵设置为对角阵。实验表明，这种简化不仅极大降低了模型复杂度，模型效果还得到显著提升。在 LFM 的基础上，张量神经网络模型[Socher, et al. 2013]（Neural Tensor Network, NTN）进一步利用关系的双线性变换来刻画实体与关系之间的联系，其基本思想是用双线性张量取代传统神经网络中的线性变换层，在不同的维度下将头、尾实体向量联系起来。由于 NTN 引入了张量操作，虽然能够更精确地刻画实体和关系的复杂语

义联系，但是计算复杂度非常高，需要大量三元组样例才能得到成分学习。实验表明，NTN 在大规模稀疏知识图谱上的效果较差。此外，矩阵分解同样是得到低维向量表示的重要途径。因此，也有研究者提出采用矩阵分解进行知识表示学习。这方面的代表方法是 RESACL 模型[Nickel, et al., 2011; Nickel, et al., 2012]。RESACL 的基本思想与前述 LFM 类似。不同之处在于，RESACL 会优化张量中的所有位置，包括值为 0 的位置；而 LFM 只会优化知识库中存在的三元组。

最近，Bordes 等人受到词向量空间对于词汇语义与句法关系存在有趣的平移不变现象的启发，提出了 TransE 模型[Bordes, et al., 2013]，将知识库中的关系看作实体间的某种平移向量。与以往模型相比，TransE 模型参数较少，计算复杂度低，却能直接建立实体和关系之间的复杂语义联系。Bordes 等人在 WordNet 和 Freebase 等数据集上进行链接预测等评测任务，实验表明 TransE 的性能较以往模型有显著提升。特别是在大规模稀疏知识图谱上，TransE 的性能尤其惊人。由于 TransE 简单有效，自提出以来，有大量研究工作对 TransE 进行扩展和应用。可以说，TransE 已经成为知识表示学习的代表模型。在 TransE 的基础上，研究者提出了众多改进模型来解决 TransE 中仍无法处理的问题。

### 3.1 复杂关系建模

TransE 由于模型简单，在大规模知识图谱上效果明显。但是也由于过于简单，导致 TransE 在处理前面提到的知识库的复杂关系时捉襟见肘。例如，假如知识库中有两个三元组，分别是(美国, 总统, 奥巴马)和(美国, 总统, 布什)。这里的关系“总统”是典型的 1-N 的复杂关系。如果用 TransE 从这两个三元组学习知识表示，将会使奥巴马和布什的向量变得相同。

为了解决 TransE 模型在处理 1-N、N-1、N-N 复杂关系时的局限性，TransH 模型[Wang, et al., 2014]提出让一个实体在不同的关系下拥有不同的表示。TransR 模型[Lin, et al., 2015]进一步认为不同的关系拥有不同的语义空间。对每个三元组，首先应将实体利用矩阵投影到对应的关系空间中，然后再建立从头实体到尾实体的翻译关系。针对在知识库中实体的异质性和不平衡性，还有 TransR 模型中矩阵参数过多的问题，TransD 模型[Ji, et al., 2015]和 TranSparse 模型[18]对 TransR 模型中的投影矩阵进行了进一步的优化。此外，TransG 模型[Xiao, et al., 2015]和 KG2E 模型[He, et al. 2015]提出了利用高斯分布来表示知识库中的实体和关系，可以在表示过程中考虑实体和关系本身语义上的不确定性。可以看到，在 TransE 之后，在如何处理复杂关系建模的挑战问题上，提出了 TransH、TransR、TransD、TranSparse、TransG 和 KG2E 等多种模型，从不同角度尝试解决复杂关系建模问题，可谓百花齐放。在相关数据集合上的实验表明，这些方法均较 TransE 有显著的性能提升，验证了这些方法的有效性。

### 3.2 多源信息融合

知识表示学习面临的另外一个重要挑战如何实现多源信息融合。现有的知识表示学习模型如 TransE 等，仅利用知识图谱的三元组结构信息进行表示学习，尚有大量与知识有关的其他信息没有得到有效利用。如何充分融合这些多源异质信息，实现知识表示学习，具有重要意义，可以改善数据稀疏问题，提高知识表示的区分能力。

在融合上述信息进行知识表示学习方面，已经有一些研究工作，但总体来讲还处于起步状态，这里简单介绍其中几个代表性工作。考虑实体描述的知识表示学习模型 (Description- Embodied Knowledge Representation Learning, DKRL) [Xie, et al., 2016]。DKRL 模型提出在知识表示学习中考虑 Freebase 等知识库中提供的实体描述文本信息。在文本表示方面，DKRL 考虑了两种模型：一种是 CBOW，将文本中的词向量简单相加作为文本表示；一种是卷积神经网络，能够考虑文本中的词序信息。DKRL 的优势在于，除了能够提升实体表示的区分能力外，还能实现对新实体的表示。当新出现一个未曾在知识库中的实体时，DKRL 可以根据它的简短描述产生它的实体表示，用于知识图谱补全等任务。这对于不断扩充知识图谱具有重要意义。

此外，Wang 等人提出在表示学习中考虑文本数据，利用 word2vec 学习维基百科正文中的词表示，利用 TransE 学习知识库中的知识表示。然后利用维基百科正文中的链接信息（锚文本与实体的对应关系），让文本中实体对应的词表示与知识库中的实体表示尽可能接近，从而实现文本与知识库融合的表示学习。Zhong 等人还将类似的想法用于融合实体描述信息 [Zhong, et al., 2015]。

已有工作表明，多源信息融合能够有效提升知识表示的性能，特别是可以有效处理新实体的表示问题。但是，也可以看出，多源信息融合的知识表示学习仍处于非常起步的阶段，相关工作较少，考虑的信息源非常有限，有大量的信息（如音频、图片、视频等）未被考虑，具有广阔的研究前景。

### 3.3 关系路径建模

在知识图谱中，多步的关系路径也能够反映实体之间的语义关系。为了突破 TransE 等模型孤立学习每个三元组的局限性，Lin 等人提出考虑关系路径的表示学习方法，以 TransE 作为扩展基础，提出 Path-based TransE (PTransE) 模型 [Lin, et al., 2015]。几乎同时，也有其他研究团队在知识表示学习中成功考虑了关系路径的建模 [Alberto, et al., 2015]。关系路径的表示学习也被用来进行基于知识库的自动问答 [Gu, et al., 2015]。

PTransE 等研究的实验表明，考虑关系路径能够极大提升知识表示学习的区分性，提高在知识图谱补全等任务上的性能。关系路径建模工作还比较初步，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，还有很多细致的考察工作需要完成。

## 4. 技术展望与发展趋势

近年来知识表示学习已经崭露头角，在很多任务上展现了巨大的应用潜力。对于 TransE 等模型面临的挑战，也已经提出了很多改进方案。然而，知识表示学习距离真正实用还很远，本节将对知识表示学习的未来方向进行展望。

- 面向不同知识类型的知识表示学习

已有工作将知识库的关系划分为 1-1、1-N、N-1 和 N-N 四类，这种关系类型划分略显粗糙，无法直观地解释知识的本质类型特点。近期发表在 Science 等权威期刊的认知科学研究成果[Kemp, et al., 2009; Tenenbaum, et al., 2011] 总结认为，人类知识包括以下几种结构：(1) 树状关系，表示实体间的层次分类关系；(2) 二维网格关系，表示现实世界的空间信息；(3) 单维顺序关系，表示实体间的偏序关系；(4) 有向网络关系，表示实体间的关联或因果关系。认知科学对人类知识类型的总结，有助于对知识图谱中知识类型的划分和处理。未来有必要结合人工智能和认知科学的最新研究成果，有针对性地设计知识类型划分标准，开展面向不同复杂关系类型的知识表示学习研究。

- 多源信息融合的知识表示学习

在多源信息融合的知识表示学习方面，相关工作还比较有限，主要是考虑实体描述的知识表示学习模型，以及文本与知识库融合的知识表示学习，这些模型无论是信息来源，还是融合手段都非常有限。我们认为在多源信息融合的知识表示学习方面，我们还可以对下列方面进行探索：(1) 融合知识库中实体和关系的其他信息，知识库中拥有关于实体和关系的丰富信息，如描述文本、层次类型等。有机融合这些信息，将显著提升知识表示学习的表示能力；(2) 融合互联网文本、图像、音频、视频信息，互联网海量文本、音频、视频数据是知识库的重要知识来源，有效地利用这些信息进行知识表示可以极大地提升现有知识表示方法的表示能力；(3) 融合多知识库信息，人们利用不同的信息源构建了不同的知识库。如何对多知识库信息进行融合表示，对于建立统一的大规模知识库意义重大。

- 考虑复杂推理模式的知识表示学习

考虑关系路径的知识表示学习，实际上是充分利用了两实体间的关系和关系路径之间的推理模式，来为表示学习模型提供更精确的约束信息。例如，根据三元组（康熙，父亲，雍正）和（雍正，父亲，乾隆）构成的“康熙”和“乾隆”之间“父亲+父亲”的关系路径，再结合三元组（康熙，祖父，乾隆），PTransE 实际上额外提供了“父亲+父亲=祖父”的推理模式，从而提升知识表示的精确性。

实际上，关系路径只是复杂推理模式中的一种特殊形式，它要求头实体和尾

实体必须保持不变。但实际上，知识库中还有其他形式的推理模式，例如三元组（美国，总统，奥巴马）和（奥巴马，是，美国人）之间就存在着推理关系，但是两者的头、尾实体并不完全一致。如果能将这些复杂推理模式考虑到知识表示学习中，将能更进一步提升知识表示的性能。

在该问题中，如何总结和表示这些复杂推理模式，是关键难题。目前来看，一阶逻辑（First-Order Logic, FOL）是对复杂推理模式的较佳表示方案，未来我们需要探索一阶逻辑的分布式表示，及其融合到知识表示学习中的技术方案。

- 面向大规模知识库的在线学习和快速学习

大规模知识库稀疏性很强。初步实验表明，已有表示学习模型在大规模知识库上性能堪忧，特别是对低频实体和关系的表示效果较差。而且知识库规模不断扩大，我们需要设计高效的在线学习方案。除了充分融合多源信息降低稀疏性之外，我们还可以探索如何优化表示学习的方式，借鉴课程学习和迁移学习等算法思想，进一步改善知识表示的效果。

- 基于知识分布式表示的应用

知识表示学习还处于起步阶段，在知识获取、融合和推理等方向均有广阔的应用空间。我们需要在若干重要任务上探索和验证知识表示学习的有效性。例如，关系抽取任务如果能够基于知识表示学习有效利用知识库信息，将能够极大提升抽取性能和覆盖面。再如，我们可以充分利用表示学习在信息融合上的优势，实现跨领域和跨语言的知识库融合。目前，知识分布式表示的作用已经在信息抽取、自动问答、信息检索、推荐系统中得到初步验证，未来还需在更多任务上进行更加深入的探索。

## 参考文献

- [Alberto, et al., 2015] Alberto G, Bordes A, Usunier N. Composing Relationships with Translations[C], in Proceedings of EMNLP 2015.
- [Bengio, et al., 2009] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends in Machine Learning. 2009, 2(1): 1-127.
- [Bengio, et al., 2013] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2013, 35(8): 1798-1828.
- [Bordes, et al. 2011] Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases[C], in Proceedings of AAAI 2011. 301-306.
- [Bordes, et al., 2012] Bordes A, Glorot X, Weston J, et al. Joint learning of words and meaning representations for open-text semantic parsing[C], in Proceedings of AISTATS 2012, 127-135.

- [Bordes, et al., 2013] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data[C], in Proceedings of NIPS 2013, 2787-2795.
- [Bordes, et al., 2014] Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data[J]. Machine Learning. 2014, 94(2): 233-259.
- [Gu, et al., 2015] Gu K, Miller J, Liang P. Traversing Knowledge Graphs in Vector Space[C], in Proceedings of EMNLP 2015.
- [He, et al., 2015] He S, Liu K, Ji G, et al. Learning to Represent Knowledge Graphs with Gaussian Embedding[C], in Proceedings of CIKM 2015, 623-632.
- [Jenatton, et al., 2012] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data[C] //Proceedings of NIPS 2012, 3167-3175.
- [Ji, et al., 2015] Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix[C], in Proceedings of ACL 2015, 687-696.
- [Ji, et al., 2016] Ji G, Liu K, He S, et al. Knowledge graph completion with adaptive sparse transfer matrix[J], in Proceedings of ACL 2016.
- [Kemp, et al., 2009] Kemp C, Tenenbaum J B. Structured statistical models of inductive reasoning[J]. Psychological review. 2009, 116(1): 20.
- [Lao, et al., 2010] Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks[J]. Machine learning. 2010, 81(1): 53-67.
- [Lao, et al., 2011] Lao N, Mitchell T, Cohen W W. Random walk inference and learning in a large scale knowledge base[C], in Proceedings of EMNLP 2011: 529-539.
- [Lin, et al., 2015] Lin Y, Liu Z, Luan H, Sun M, Rao S, Liu S. Modeling Relation Paths for Representation Learning of Knowledge Bases[C], in Proceedings of EMNLP 2015.
- [Lin, et al., 2015] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion[C], in Proceedings of AAAI, 2015.
- [Nickel, et al., 2011] Nickel M, Tresp V, Kriegel H. A three-way model for collective learning on multi-relational data[C], in Proceedings of ICML 2011, 809-816.
- [Nickel, et al., 2012] Nickel M, Tresp V, Kriegel H. Factorizing YAGO: scalable machine learning for linked data[C], in Proceedings of WWW 2012, 271-280.
- [Socher, et al. 2013] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion[C], in Proceedings of NIPS 2013: 926-934.
- [Sutskever, et al., 2009] Sutskever I, Tenenbaum J B, Salakhutdinov R. Modelling

- relational data using Bayesian clustered tensor factorization[C], in Proceedings of NIPS. 2009: 1821-1828.
- [Tenenbaum, et al., 2011] Tenenbaum J B, Kemp C, Griffiths T L, et al. How to grow a mind: Statistics, structure, and abstraction[J] science. 2011, 331(6022): 1279-1285.
- [Wang, et al., 2014] Wang Z, Zhang J, Feng J, et al. Knowledge Graph Embedding by Translating on Hyperplanes[C], in Proceedings of AAAI 2014, 1112-1119.
- [Xiao, et al., 2015] Xiao H, Huang M, Hao Y, et al. TransG: A Generative Mixture Model for Knowledge Graph Embedding[J]. arXiv preprint arXiv:1509.05488. 2015.
- [Xie, et al., 2016] Xie R, Liu Z, Jia J, et al. Representation Learning of Knowledge Graphs with Entity Descriptions[C], in Proceedings of AAAI, 2016.
- [Yang, et al., 2015] Yang B, Yih W, He X, et al. Embedding Entities and Relations for Learning and Inference in Knowledge Bases[C], in Proceedings of International Conference on Learning Representations (ICLR). 2015.
- [Zhong, et al., 2015] Zhong H, Zhang J, Wang Z, et al. Aligning knowledge and text embeddings by entity descriptions[C], in Proceedings of EMNLP 2015, 267-272.

## 第三章 实体识别与链接

### 1. 任务定义、目标和研究意义

实体是文本中承载信息的重要语言单位，一段文本的语义可以表述为其包含的实体及这些实体相互之间的关联和交互。实体识别也就成为了文本意义理解的基础。例如，“26日下午，一架叙利亚空军 L-39 教练机在哈马省被 HTS 使用的肩携式防空导弹击落”中的信息可以通过其包含的时间实体“26号下午”，机构实体“叙利亚空军”、“HTS”，地点实体“哈马省”和武器实体“L-39 教练机”、“肩携式防空导弹”有效描述。实体也是知识图谱的核心单元，一个知识图谱通常是一个以实体为节点的巨大知识网络，包括实体、实体属性以及实体之间的关系。例如，一个医学领域知识图谱的核心单元是医学领域的实体，如疾病、症状、药物、医院、医生等。

命名实体识别是指识别文本中的命名性实体，并将其划分到指定类别的任务[Chinchor & Robinson, 1997]。常用实体类别包括人名、地名、机构名、日期等。实体链接主要解决实体名的歧义性和多样性问题，是指将文本中实体名指向其所代表的真实世界实体的任务，也通常被称为实体消歧[Ji et al., 2010]。例如，给一句话“苹果发布了最新产品 iPhone X”，实体链接系统需要将文本中的“苹果”与其真实世界所指的“苹果公司”进行对应。

实体识别与链接是海量文本分析的核心技术，为解决信息过载提供了有效手段。一方面，互联网文本数据的爆炸式增长带来了严重的“信息过载”问题。互联网数据中海量冗余信息、虚假信息和噪音信息导致查找和浏览有用信息变得愈发困难。命名实体识别技术通过将文本结构化为以实体为中心的语义表示，为分析非结构化文本提供了核心技术手段，是实现大数据资源化、知识化和普适化的核心技术，已被广泛应用于舆情监控、网络搜索、智能问答等多个重要领域。

作为知识图谱的基本单元，实体识别与链接是知识图谱构建和补全的核心技术。人工智能的长久核心目标之一是构建可支撑类人推理和自然语言理解的大规模常识知识库。然而，由于人类知识的复杂性、开放性、多样性和巨大的规模，当前依旧不存在满足上述需求的大规模知识库。实体识别技术检测文本中的新实体，并将其加入到现有知识库中。实体链接技术通过发现现有实体在文本中的不同出现，可以针对性的发现关于特定实体的新知识。实体识别与链接的研究将为计算机类人推理和自然语言理解提供知识基础。

### 2. 研究内容和挑战问题

实体识别与链接处理各种非结构化/半结构化的输入（如文本、新闻网页、商

品页面、微博、论坛页面等)，使用多种技术（统计方法、深度学习方法、知识挖掘方法），提取各种类型的实体（如人名、地名、商品、药物等），并将这些信息与现有知识图谱进行集成（实体链接）。以下分别介绍具体内容。

**实体识别。**命名实体识别的目的是识别文本中指定类别的实体，主要包括人名、地名、机构名、专有名词等的任务。例如，识别“*2016 年 6 月 20 日，骑士队在奥克兰击败勇士队获得 NBA 冠军*”这句中的地名（奥克兰）、时间（2016 年 6 月 20 日）、球队（骑士队、勇士队）和机构（NBA）。命名实体识别系统通常包含两个部分：实体边界识别和实体分类，其中实体边界识别判断一个字符串是否组成一个完整实体，而实体分类将识别出的实体划分到预先给定的不同类别中去。命名实体识别是一项极具实用价值的技术，目前中英文上通用命名实体识别（人名、地名、机构名）的 F1 值都能达到 90% 以上。命名实体识别的主要难点在于表达不规律、且缺乏训练语料的开放域命名实体类别（如电影、歌曲名）。

**实体链接。**实体链接的目的是将实体提及与知识库中对应实体进行链接。给定一段文本（如“*在旧金山的发布会上，苹果为开发者推出新编程语言 Swift*”），一个实体链接系统包括如下研究内容：

1. 识别文档中的目标提及(mention)。所谓提及，就是我们想要链接的对象，例如上述例子文本中的提及{“旧金山”，“苹果”，“Swift”}；
2. 针对每一个提及，识别该提及在知识图谱中可能指向的候选目标实体。例如，上述文本中的提及“苹果”可能指向的目标实体包括 {苹果(水果)，苹果公司，苹果(电影)，苹果(银行), ...}；
3. 基于提及的上下文等信息对目标实体进行排序。例如，系统需要根据“苹果”的上下文词语{发布会，编程语言，开发者，...}识别出该段文本中“苹果”指的是苹果公司，而不是苹果(水果)或者苹果(电影)；
4. 空提及检测与聚类。考虑到知识的规模和更新速度，知识库往往不能覆盖所有真实世界实体。为了解决上述问题，需要识别出知识库尚未包含其目标实体的提及，并将这些提及按其指向的真实世界实体进行聚类。例如，由于现有知识库没有包含上文中提及“Swift”指向的目标实体 Swift（编程语言），实体链接系统需要将“Swift”的目标实体设置为空实体“NIL”，表示该提及在知识库中没有链接对象。

实体链接是一项极具实用性的技术，目前面向 Wikipedia 知识库的实体链接准确率可以达到 90% 以上，F1 值在 0.85 以上。目前实体链接技术已在实际应用中得到广泛使用。

概况说来，上述实体分析任务主要面临以下几个关键科学问题：

1. **实体名的歧义性和多样性。**歧义性和多样性是自然语言的固有属性，也是实体识别和链接技术要解决的根本问题。在实体识别中，实体可以有各种各样

不同的表达，导致除了少数规范性实体（如电话号码，email 地址）之外，大部分实体都无法使用名字规则来捕捉其规律，而是需要构建统计上下文模型来进行识别。在实体链接中，实体的歧义导致一个实体名有许多可链接的对象，这使得如何挖掘更多的消歧证据、设计更高性能的消歧算法、构建覆盖度更高的实体引用表仍然是实体链接系统的核心问题。

**2. 资源缺乏(Low Resource)问题。**目前绝大部分的实体分析算法都依赖于有监督模型，需要大量的训练语料来达到实用性能。然而，考虑到标注语料的成本，在绝大部分情况下都不可能获得足够的训练语料来处理不同的领域、面向不同风格的文本（规范、非规范）、不同的语言（中文、英文、一带一路小语种等）等多种多样的情况。无需大量训练语料的无监督/半监督技术，资源自动构建技术，以及迁移学习等技术是解决上述问题的核心研究问题。

**3. 实体的开放性问题。**实体具有复杂性和开放性的特点。实体的复杂性指的是实体的类型多种多样，同时类型之间具有复杂的层次结构。实体的开放性指实体并不是一个封闭的集合，而是随着时间增加、演化和失效。实体的开放性和复杂性给实体分析带来了巨大的挑战：开放性使得现有有监督方法无法适应开放知识的抽取；实体的巨大规模使得无法使用枚举或者人工编写的方式来进行处理，同时随着时间变化现有模型的性能会下降。

### 3. 技术方法和研究现状

实体识别最早在上世纪 80 年代的 MUC 任务中被提出，一直是自然语言处理的研究热点。实体链接最早以实体共指消解、实体消歧等名字在数据库领域、Web 领域、自然语言领域被提出，并在 TAC 评测中被标准化。目前已经有许多方法被提出用于实体识别和链接。根据模型的不同，实体分析方法可以分为基于统计模型的方法、基于深度学习的方法和基于文本挖掘的方法；根据对监督知识的依赖，可以划分为无监督方法、弱监督方法、知识监督方法和有监督方法。以下按照模型的维度介绍目前的技术方法和研究现状。

#### 3.1 传统统计模型方法

**实体识别。**自 90 年代以来，统计模型一直是实体识别的主流方法。有非常多的统计方法被用来抽取文本中的实体识别，如最大熵分类模型、SVM、隐马尔可夫模型、条件随机场模型等等[Sundheim, 1996]。基于统计模型的方法通常将实体识别任务形式化为从文本输入到特定目标结构的预测，使用统计模型来建模输入与输出之间的关联，并使用机器学习方法来学习模型的参数。例如，最大熵分类模型将命名实体识别转换为子字符串的分类任务，实体识别的代表性统计模型条件随机场模型（CRF）它将实体识别问题转化为序列标注问题[Lafferty et al., 2001]。

**实体链接。**实体链接的核心是计算实体提及(mention)和知识库中实体的相似度，并基于上述相似度选择特定实体提及的目标实体[Ji et al., 2010]。上述过程的核心在于挖掘可用于识别提及目标实体相互关联的证据信息，将这些证据表示为供计算机处理的形式，并构建高性能的算法来综合不同证据进行链接决策。目前主要使用的证据信息包括实体统计信息、名字统计信息、上下文词语分布、实体关联度、文章主题等信息[Milne & Witten, 2008] [Han & Sun, 2011] [Zhang et al., 2011]。同时，考虑到一段文本中实体之间的相互关联，相关的全局推理算法也被提出来寻找全局最优决策[Han et al., 2011] [Ji & Chen, 2011]。

传统统计模型的主要缺点在于需要大量的标注语料来学习，这导致构建开放域或 Web 环境下的信息抽取系统时往往会遇到标注语料瓶颈。为解决上述问题，近年来已经开始研究高效的弱监督或无监督策略，如半监督算法、远距离监督算法、基于海量数据冗余性的自学习方法，等等[Agichtein & Gravano, 2000] [Etzioni et al., 2011] [Shi et al, 2014]。传统统计模型的另外一个缺点是其需要人工构建大量的特征，其训练并非一个端到端的过程。为解决上述问题，越来越多深度学习模型被用于实体识别和链接[Pan et al., 2017]。

### 3.2 深度学习方法

**实体识别。**随着深度学习在不同领域的火爆，越来越多的深度学习模型被提出用于解决实体识别问题。目前存在两类用于命名实体识别的典型深度学习架构，一种是 NN-CRF 架构[Lample et al, 2016]，在该架构中，CNN/LSTM 被用来学习每一个词位置处的向量表示，基于该向量表示，NN-CRF 解码该位置处的最佳标签。第二种是采用滑动窗口分类的思想，使用神经网络学习句子中的每一个 ngram 的表示，然后预测该 ngram 是否是一个目标实体[Xu et al., 2017]。

**实体链接。**实体链接的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。通过将不同类型的信息映射到相同的特征空间，并提供高效的端到端训练算法，深度学习方法给上述任务提供了强有力的新工具。目前的相关工作包括多源异构证据的向量表示学习、以及不同证据之间相似度的学习等工作[Ganea & Hofmann, 2017] [Gupta et al., 2017] [Sil et al., 2018]。

相比传统统计方法，深度学习方法的主要优点是其训练是一个端到端的过程，无需人工定义相关的特征。另外一个优点是深度学习可以学习任务特定的表示，建立不同模态、不同类型、不同语言之间信息的关联，从而取得更好的实体分析性能。目前，如何在深度学习方法中融入知识指导(如语言学结构约束、知识结构)、考虑多任务之间的约束、以及如何将深度学习用于解决资源缺乏问题(如构建语言无关的命名实体识别)是当前的工作的热点。

### 3.3 文本挖掘方法

传统统计方法和深度学习方法都需要大量训练语料和预先明确定义的目标实体类别，无法处理大数据环境下的开放实体分析任务。除非结构化文本之外，Web 中往往还存在大量的半结构高质量数据源，如维基百科、网页中的表格、列表、搜索引擎的查询日志等等。这些结构往往蕴含有丰富的语义信息。因此，半结构 Web 数据源上的语义知识获取（knowledge harvesting），如大规模知识共享社区（如百度百科、互动百科、维基百科）上的实体知识抽取，往往采用文本挖掘的方法。代表性文本挖掘抽取系统包括 DBpedia[Auer et al., 2007]、Yago[Suchanek & Kasneci, 2008、BabelNet、NELL 和 Kylin 等等。文本挖掘方法的核心是构建从特定结构（如列表、Infobox）构建实体挖掘的特定规则。由于规则本身可能带有不确定性和歧义性，同时目标结构可能会有一定的噪音，文本挖掘方法往往基于特定算法来对语义知识进行评分和过滤。

此外，人们发现结构化数据源只包含有限类别的实体，对长尾类别覆盖不足，另一方面的实体获取技术往往采用 Bootstrapping 策略，充分利用大数据的冗余性，开放式的从 Web 中获取指定类型的实体。该部分的代表性的工作包括 TextRunner 系统和 Snowball 系统[Agichtein & Gravano, 2000]。开放式实体集合扩展的主要问题是语义漂移问题，近年来的主要工作集中在解决该问题。具体技术包括互斥 Bootstrapping 技术、Co-Training 技术和 Co-Bootstrapping 技术。

文本挖掘方法只从容易获取且具有明确结构的语料中抽取知识，因此抽取出来的知识质量往往较高。然而，仅仅依靠结构化数据挖掘无法覆盖人类的大部分语义知识：首先，绝大部分结构化数据源中的知识都是流行度高的知识，对长尾知识的覆盖不足；此外，人们发现现有结构化数据源只能覆盖有限类别的语义知识，相比人类的知识仍远远不够。因此，如何结合文本挖掘方法（面向半结构化数据，抽取出的知识质量高但覆盖度低）和文本抽取方法（面向非结构化数据，抽取出的知识相比文本挖掘方法质量低但覆盖度高）的优点，融合来自不同数据源的知识，并将其与现有大规模知识库集成[Nakashole et al., 2012]，是文本挖掘方法的研究方向之一。

## 4. 技术展望与发展趋势

纵观实体识别研究发展的态势和技术现状，我们认为其发展方向如下：

### 1. 融合先验知识的深度学习模型

近年来，深度学习模型已经在实体识别和链接任务上取得了长足的进展，并展现了相当的技术潜力和优势。但是目前的深度学习模型的成功仍然依赖于大量的训练语料，缺乏面向任务特点的针对性设计。

之前的传统统计模型中已经证明许多先验知识对于实体识别和链接任务的

有效性，如句法结构、语言学知识、任务本身约束、知识库知识和特征结构等。如何在深度学习模型中融合上述先验知识并进行针对性的设计是提升现有深度模型的有效手段之一。

另一方面，现有深度模型在进行实体分析时仍然是一个黑箱模型，导致其可解释性不强，且难以采用增量的方式构建模型。如何构建可解释、增量式的深度学习模型也是未来值得解决的一个问题。

## 2. 资源缺乏环境下的实体分析技术

目前，绝大部分实体分析研究集中在构建更精准的模型和方法，这些方法通常面向预先定义好的实体类别，使用标注语料训练模型参数。然而，在构建真实环境下的信息抽取系统时，这些有监督方法往往具有如下不足：1) 现有监督模型在更换语料类型之后，往往会有大幅度的性能下降；2) 现有监督模型无法分析目标类别之外的实体；3) 现有监督模型依赖于大规模的训练语料来提升模型性能。

为解决上述问题，如何构建资源缺乏环境下的实体分析系统是相关技术实用化的核心问题。相关研究方向包括：构建迁移学习技术，充分利用已有的训练语料；研究自学习技术，在极少人工干预下构建高性能的终生学习信息抽取系统；研究增量学习技术，自动的重用之前的信息抽取模块，使得不同资源可以逐步增强，而不是每次都重头开始训练；研究无监督/半监督/知识监督技术，探索现有有监督学习技术之外的有效手段，解决标注语料瓶颈问题。

## 3. 面向开放域的可扩展实体分析技术

由于实体分析任务的基础性，越来越多的任务和应用需要实体识别和链接技术的支撑。这就要求实体分析技术能够处理各种不同的情境带来的挑战，在开放环境下取得良好性能。然而，现有实体分析系统往往针对新闻文本，对其它情境下（如不同文本类型微博、评论、列表页面等，不同上下文如多模态上下文、短文本上下文和数据库上下文）的研究不足。

因此，实体分析的发展方向之一是构建面向开放域的可扩展实体分析技术。具体包括：1) 数据规模上的可扩展性：信息抽取系统需要能够高效的处理海量规模的待抽取数据；2) 数据源类型上的可扩展性：信息抽取系统需要能够在面对不同类型数据源时取得鲁棒的性能；3) 领域的可扩展性：信息抽取系统需要能够方便的从一个领域迁移到另一个领域；4) 上下文的可扩展性：实体分析系统需要能够处理不同的上下文，并针对不同上下文的特定自适应的改进自身。

## 参考文献

- [Agichtein & Gravano, 2000] Agichtein, E., and Gravano, L., 2000. Snowball: Extracting relations from large plain-text collections. In: Proceedings of the fifth ACM conference on Digital libraries 2000.

- [Auer et al., 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R. and Ives, Z., 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722-735). Springer, Berlin, Heidelberg.
- [Chinchor & Robinson, 1997] Chinchor, N. and Robinson, P., 1997, September. MUC-7 named entity task definition. In *Proceedings of the 7th Conference on Message Understanding* (Vol. 29).
- [Etzioni et al., 2011] Etzioni, O., Fader, A., Christensen, J., Soderland, S. and Mausam, M., 2011, July. Open Information Extraction: The Second Generation. In *IJCAI* (Vol. 11, pp. 3-10).
- [Ganea & Hofmann, 2017] Ganea, O.E. and Hofmann, T., 2017. Deep Joint Entity Disambiguation with Local Neural Attention. In *Proceedings of EMNLP 2017*.
- [Gupta et al., 2017] Gupta, N., Singh, S. and Roth, D., 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2681-2690).
- [Han et al., 2011] Han, X., Sun, L. and Zhao J. 2011. Collective Entity Linking in Web Text: A Graph-Based Method. In: *Proceedings of 34th Annual ACM SIGIR Conference*.
- [Han & Sun, 2011] Han, X. and Sun, L. 2011. A Generative Entity-Mention Model for Linking Entities with Knowledge Base. In: *Proceedings of ACL-HLT*.
- [Ji & Chen, 2011] Ji, H. and Chen, Z. 2011. Collaborative Ranking: A Case Study on Entity Linking. In: *Proceedings of EMNLP*.
- [Ji et al., 2010] Ji, H., et al. 2010. Overview of the TAC 2010 knowledge base population track. In: *Proceedings of Text Analysis Conference*.
- [Kataria et al., 2011] Kataria, S. S., Kumar, K. S. and Rastogi, R. 2011. Entity Disambiguation with Hierarchical Topic Models. In: *Proceedings of KDD*.
- [Kulkarni et al., 2009] Kulkarni, S., Singh, A., Ramakrishnan, G. & Chakrabarti, S. 2009. Collective annotation of Wikipedia entities in web text. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 457-466.
- [Lafferty et al., 2001] Lafferty, J., McCallum, A. and Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML 2001*.
- [Lample et al, 2016] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K. and Dyer, C., 2016. Neural architectures for named entity recognition. arXiv preprint

arXiv:1603.01360.

- [Mihalcea & Csomai, 2007] Mihalcea, R. & Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 233-242.
- [Milne & Witten, 2008] Milne, D. & Witten, I. H. 2008. Learning to link with Wikipedia. In: Proceedings of the 17th ACM conference on Conference on information and knowledge management.
- [Nakashole et al., 2012] Nakashole, N., G. Weikum and F. Suchanek. 2012. PATTY: A Taxonomy of Relational Patterns with Semantic Types. In: EMNLP 2012.
- [Pan et al., 2017] Pan, X., Zhang, B., May, J., Nothman, J., Knight, K. and Ji, H., 2017. Cross-lingual name tagging and linking for 282 languages. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1946-1958).
- [Shi et al, 2014] Shi, B., Zhang, Z., Sun, L., & Han, X. A Probabilistic Co-Bootstrapping Method for Entity Set Expansion. In: COLING 2014.
- [Sil et al., 2018] Sil, A., Kundu, G., Florian, R. and Hamza, W., 2018. Neural Cross-Lingual Entity Linking. In Proceedings of AAAI 2018.
- [Suchanek & Kasneci, 2008] Suchanek, F. M. and Kasneci, G., et al. 2008. Yago: A large ontology from Wikipedia and Wordnet. In: Web Semantics: Science, Services and Agents on the World Wide Web 6(3): 203-217.
- [Sundheim 1996] Sundheim, B.M., 1996, May. Overview of results of the MUC-6 evaluation. In Proceedings of a workshop on held at Vienna, Virginia: May 6-8, 1996 (pp. 423-442). Association for Computational Linguistics.
- [Sun et al., 2015] Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., and Wang, X. 2015. Modeling mention, context and entity with neural networks for entity disambiguation. In IJCAI, pages 1333–1339
- [Xu et al., 2017] Xu, M., Jiang, H. and Watcharawittayakul, S., 2017. A Local Detection Approach for Named Entity Recognition and Mention Detection. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vol. 1, pp. 1
- [Zhang et al., 2011] Zhang, W. and Sim, Y. C., et al. 2011. Entity Linking with Effective Acronym Expansion, Instance Selection and Topic Modeling. In: Proceedings of IJCAI.

## 第四章 实体关系学习

### 1. 任务定义、目标和研究意义

关系定义为两个或多个实体之间的某种联系，实体关系学习就是自动从文本中检测和识别出实体之间具有的某种语义关系，也称为关系抽取。关系抽取的输出通常是一个三元组(实体 1, 关系, 实体 2)。例如，句子“北京是中国的首都、政治中心和文化中心”中表述的关系可以表示为 (中国, 首都, 北京), (中国, 政治中心, 北京) 和 (中国, 文化中心, 北京)。

关系抽取是知识图谱构建和信息抽取中的一个关键环节，具有重要的理论意义和广阔的应用前景，为多种应用提供重要的支持，主要表现在：

(1) 大规模知识图谱的自动构建。很多互联网应用都需要知识图谱的支撑，这个知识图谱不仅包含 WordNet、HowNet 等常识知识图谱中的通用语义知识，而且包含百科全书、领域知识图谱中的领域语义知识。如果能把多源异构知识集成为一个大的知识图谱，将可能提高很多互联网应用系统的性能，并开创语义网时代的很多应用。现有的知识图谱如 WordNet、HowNet 和 CYC 等大多数依靠专家人工编撰。随着互联网的发展，知识呈爆炸式增长，人工构建知识图谱特别是构建领域知识图谱时遇到了很大困难，不仅费时费力，而且存在知识覆盖率低、数据稀疏和更新缓慢等问题。利用关系抽取技术，知识图谱可以根据结构化的抽取结果自动生成。典型的例子有：Freebase、Yago 和 BDpedia。



图 1 百度搜索引擎对“身高 170 以上的中国射手座明星”查询返回的结果

(2) 为其它信息获取技术提供支持。1) 对信息检索提供支持：可以对复杂的查询进行关联搜索和推理，提供智能检索结果。例如，对于“身高 170 以上的中国射手座明星”，有了关系抽取技术的支持，通过知识图谱构建，就可以通过

推理的方法获得结果。图 1 展示了百度搜索引擎对“身高 170 以上的中国射手座明星”查询返回的结果，这个结果的背后就受到大规模知识图谱的支撑。2) 对问答系统提供支持：在问答中，关键步骤是建设一个领域无关的问答类型体系并找出与问答类型体系中每个问答类型相对应的答案模式，这就需要关系抽取技术的支持。

(3) 自然语言理解。目前深层的语言理解系统在正确率和性能方面还难以令人满意，关系抽取是篇章理解的关键技术，运用语言处理技术可以对文本的核心内容进行理解，因此语义关系抽取的研究将成为从简单的自然语言处理技术到真正的自然语言理解应用之间的一个重要纽带，能改进自然语言处理领域的很多任务的性能，如实体链接和机器翻译等。

## 2. 研究内容和挑战

关系抽取系统处理各种非结构化/半结构化的文本输入（如新闻网页、商品页面、微博、论坛页面等），使用多种技术（如规则方法、统计方法、知识挖掘方法），识别和发现各种预定义类别和开放类别的关系。根据关系类别是否预定义，目前关系抽取的核心研究内容可以划分为限定域关系抽取和开放域关系抽取。以下分别介绍具体的研究内容。

**限定域关系抽取。**限定域关系抽取是指系统所抽取的关系类别是预先定义好的，比如知识图谱中定义好的关系类别。在限定域关系抽取中关系的类别一般是人工定义或者从现有知识图谱中自动获取。由于关系类别已经预先定义，所以一般可以人工或者基于启发式地规则自动构建标注语料。因此限定域关系抽取中的主要研究内容是如何利用有监督或弱监督的方法抽取预定义的实体关系知识。在有监督的方法中主要的研究内容集中在如何挖掘更多能表征相应语义关系的特征上。在弱监督的方法中主要的研究内容集中在如何降低自动生成语料中的噪声。

**开放域关系抽取。**开放域关系抽取不预先定义抽取的关系类别，由系统自动从文本中发现并抽取关系。因此开放域关系抽取也称为开放式关系发现。由于没有事先定义关系的类别，因此开放域关系发现中利用关系指示词代表关系的类型。主要的研究内容是如何利用无监督的方法自动的抽取关系三元组。

关系抽取目前主要面临如下三个挑战：

- 自然语言表达的多样性

关系抽取的核心是将自然语言表达的关系知识映射到关系三元组上。然而，自然语言表达具有多样性和隐含性，导致关系抽取任务极具挑战性。自然语言表达的多样性指的是同一种关系可以有多种表达方式，例如“总部位置”这个语义关系可以用“X 的总部位于 Y”，“X 总部坐落于 Y”，“作为 X 的总部所在地，Y...”等等不同的文本表达方式。自然语言表达的多样性是关系抽取的一大挑战。

### ● 关系表达的隐含性

关系表达的隐含性是指关系有时候在文本中找不到任何明确的标识，关系隐含在文本中。例如：蒂姆·库克与中国移动董事长奚国华会面商谈“合作事宜”，透露出了他将带领苹果公司进一步开拓中国市场的讯号。在这一段文本中，并没有直接给出蒂姆·库克和苹果公司的关系，但是从“带领苹果公司”的表达，我们可以推断出蒂姆·库克是苹果公司的首席执行官(CEO)。关系表达的隐含性是关系抽取的一大挑战。

### ● 实体关系的复杂性

关系抽取的目标是抽取实体之间的语义关系，然而，真实世界中同一对实体之间可能有多个关系，而且有的关系可以同时存在，而有的关系是具有时间特性的。比如：中国和北京的关系有多个，北京坐落于中国，北京是中国的首都，北京是中国的政治中心，北京是中国的文化中心。这些关系是可以同时存在的。但是如果两个人本来是夫妻关系，后来离婚了，他们就不是夫妻关系了，是前妻或者前夫的关系，这个类关系具有时空性，不能单独存在，实体关系的复杂性是关系抽取的又一挑战。

## 3. 技术方法和研究现状

自上世纪 90 年代被提出以来，关系抽取一直是自然语言处理的研究热点。现有关系抽取方法可以从不同的维度进行划分。例如，根据关系的类型，关系抽取可以分为限定域关系抽取和开放域关系抽取；根据关系抽取的方法可以分为基于规则的方法和基于机器学习的方法；根据对监督知识的依赖，关系抽取可以分为有监督关系抽取、无监督关系抽取和弱监督关系抽取。下面分别从不同的维度对现有关系抽取的技术方法和研究现状就行介绍

### 3.1 限定域关系抽取和开放域关系抽取

限定域关系抽取是指系统所抽取的关系是预先定义好的，比如从句子中抽取知识图谱中定义好的关系。因为预定义关系的个数是有限的，部分文献[Zeng et al.,2014] [Lin et al.,2016] [Jiang et al.,2016] 将关系抽取任务视为多分类任务，其中每个关系为一个类别。开放域关系抽取是指不预先定义关系，由系统自动从文本中发现、抽取关系。由于限定域关系抽取可以抽取语义化的实体关系三元组，可以方便的用于辅助其它任务，而开放域关系抽取难以抽取语义化三元组，近年来，越来越多的研究者关注限定域关系抽取。限定域关系抽取的工作将在下一小节中详细介绍，本小结主要介绍开放域关系抽取。

开放域关系抽取是为了处理大量异构数据而设计的，其所抽取的关系类型不受限制，数量也不定。开放域关系抽取的目的是处理单个句子，将其变成三元组样式的结构化表示。华盛顿大学在这方面做了大量代表性的工作，如 TextRunner，

Kylin, WOE, ReVerb 等。TextRunner[Banko et al.,2007]能够直接从网页纯文本中抽取实体关系，在这一过程中只考虑文本中词与词之间的关系特征，而不考虑网页内部的结构特征。TextRunner 首先利用简单的启发式规则，在宾州树库上产生训练语料，提取一些浅层句法特征，训练一个分类器，用来判断两个实体间是否存在语义关系；然后在海量网络数据上，找到候选句子，提取浅层句法特征，利用分类器判断所抽取的关系对是否可信；最后利用网络数据的冗余信息，对初步认定可信的关系进行评估。其过程类似于语义角色标注。它把动词作为关系名称，通过动词链接两个实体。Textrunner 的平均错误率在 12% 左右，但是，TextRunner 得到的大部分关系词往往没有意义，而且容易把关键论元信息丢失。Kylin 和 WOE 的数据都是基于距离监督的方式产生的，但是由于它们将 Wikipedia 中的信息框作为距离监督中的结构化知识库，并利用其进行回标，而 wikipedia 中的信息框结构和内容是不断更新的，所以这里将 Kylin 和 WOE 归类到华盛顿大学在开放域关系抽取的系列工作。ReVerb[Fader et al.,2011]是在分析 TextRunner 和 WOE 中普遍的错误之后提出的基于句法和词汇约束的实体关系识别器。它主要解决了以前系统中普遍存在实体关系识别错误三元组和无信息量三元组的问题。上述系统都采用简单的启发方法识别论元，例如抽取简单的名词性短语或维基百科实体作为论元。但是这些启发式方法不能适应语言的复杂性。R2A2[Etzioni et al.,2011] 系统通过增加一个论元识别器，大大改善了论元识别的准确性。此外，斯坦福的 OpenIE 工具包 [Angeli et al.,2015] 也是一种典型的开放域抽取系统，其算法思想是通过将长句子切分为各种关联的片段，然后从这些子片段中抽取出三元组信息。上述的方法都是开放域的不限制关系的类别，因此抽出的关系缺乏语义信息，同一类关系会出现多种不同的抽取结果。

## 3.2 基于规则的关系抽取和基于机器学习的关系抽取

### 3.2.1 基于规则的关系抽取

所谓基于规则的关系抽取方法是指首先由通晓语言学知识的专家根据抽取任务的要求设计出一些包含词汇、句法和语义特征的手工规则（或称为模式），然后在文本分析的过程中寻找与这些模式相匹配的实例，从而推导出实体之间的语义关系。如 [Fukumoto et al.,1998] 依据两个实体之间相关联的特定谓词来判断它们之间的关系，不过其召回率太低，导致在 MUC-7 测试中 F 指数只有 39.1%。[Humphreys et al.,1998] 在篇章解释器（Discourse Interpreter）中利用一系列句法和语义规则识别出实体间的关系，其输入序列来源于增加了语义和指代等信息的句法分析器。虽然它们的结构较为复杂，但在 MUC-7 的模板任务中 F 指数也分别只有 23.7 和 54.7。[Aone et al.,1998] [Aone et al.,2000] 则充分利用语义关系的局部性特点，在名词短语标注的过程中识别出短语的中心词和它的修饰词之间可

能存在的关系，在 MUC-7 的模板关系任务中取得了 75.6 的最高 F 指数。

基于手工规则的方法需要领域专家构筑大规模的知识库，这不但需要有专业技能的专家，也需要付出大量劳动，因此这种方法的代价很大。知识库构建完成后，对于特定的领域的抽取具有较好的准确率，但移植到其他领域十分困难，效果往往较差。因此这种方法在可移植性方面存在着明显的不足。

### 3.2.2 基于机器学习的关系抽取

按照机器学习方法对语料库的不同需求大致可分成三大类：无监督关系抽取，有监督关系抽取、弱监督关系抽取。无监督关系抽取希望把表示相同关系的模版聚合起来，不需要人工标注的数据。有监督关系抽取使用人工标注的训练语料进行训练。有监督关系抽取目前可以取得最好的抽取效果，但是由于其需要费时费力的人工标注，因此难以扩展到大规模的场景下。因此有学者提出了利用知识库回标文本来自动获得大量的弱监督数据。目前弱监督关系抽取是关系抽取领域的一大热点。

#### 3.2.2.1 无监督关系抽取

无监督关系抽取方法主要基于分布假设（Distributional Hypothesis[Harris et al.,1954]理论，分布假设的核心思想是：如果两个词的用法相似及出现在相同上下文中，那么这两个词就意思相近。相应的，在实体关系抽取中，如果两个实体对具有相似的语境，那么这两个实体对倾向于具有相同的语义关系，基于此理论，无监督关系抽取将两个实体的上下文作为表征语义关系的特征。

基于分布假设理论，首先由 [Hasegawa et al.,2004]提出了一种基于无监督学习的实体关系发现方法，该方法将大量文本中同一实体对的所有上下文收集起来，并把这些上下文作为表示语义关系的特征，然后采用层次聚类的方法将特征相似度较高的实体对聚集在一起，最后从一个聚类中挑选出频率最高的词作为该类关系的名称。[Chen et al.,2005] 对 Hasegawa 的方法进行了改进，他们的方法将每个实体对的上下文，而不是所有相同实体对的上下文，作为实体之间的语义关系特征。在聚类时，先采用基于熵的方法对词汇特征进行排序，以提高特征集的空间搜索效率。最后使用 Discriminative Category Matching (DCM) 理论[Fung et al.,2002]来衡量特征在某个聚类中的重要性。

无监督关系抽取的核心是选取表示实体之间关系的特征，然后再聚类。上文介绍的方法主要选取上下文特征，与之不同的是，[Bollegala et al.,2010]利用关系的对偶性（Relation Duality），提出实体对空间和模板空间可以相互表示，基于这个理论，使用协同聚类来发现实体对及其关系模板的聚类簇，从每个聚类的簇里面选择代表性的模板当作此簇对应的关系。另外，无监督关系抽取面临着关系聚类簇中的多义问题，即同一个模板可能表达不同的关系，针对此问题，[Yao et al.,2012]使用主题模型（Topic Model）将实体对及其对应的关系模板分配到不同

的语义类别上，然后再使用聚类的方法将这些语义类别映射到语义关系。

无监督关系抽取方法可以发现新的关系，但其发现的新关系往往是相似模板的聚类，其缺点是得到的关系不具语义信息，难以规则化，很难被用来构建知识库，如果需要得到语义关系，需要通过将其同现有知识库的关系进行对齐，或者通过人工的方式来给每个聚类关系簇赋予语义信息。

### 3.2.2.2 有监督关系抽取

在使用有监督的方法解决关系抽取问题时，一般将关系抽取看作是一个多分类问题，提取特征向量后再使用有监督的分类器进行关系抽取，有监督的方法性能较好[Zhou et al.,2005] [Mooney et al.,2006] [Bunescu et al.,2005]，目前占据主导地位，研究人员在这方面做了大量的工作。有监督关系抽取可以分为：基于特征向量的方法、基于核函数的方法和基于神经网络的方法。

基于特征向量的方法特点是需要显式地将关系实例转换成分类器可以接受的特征向量，其研究重点在于怎样提取具有区分性的特征，通过获取各种有效的词汇、句法和语义等特征，然后有效地集成起来，从而产生描述关系实例的各种局部和全局特征。文献[Kambhatla et al.,2004]中的方法综合考虑实体本身、实体类型、依存树和解析树等特征，然后使用最大熵分类器判断实体间的关系。[Zhao et al.,2005]进一步将这些特征按照实体属性、二元属性、依存路径等类别进行划分。[Zhou et al.,2005]系统地研究了如何把包括基本词组块（Chunk）在内的各种特征组合起来，探讨了各种语言特征对关系抽取性能的贡献，特别研究了WordNet 和 Name List 等语义信息的影响。[Wang et al.,2006]又进一步加入了从句子的简化准逻辑形式（SQLF, Simplified Quasi Logical Form）导出的谓词语义属性，该属性定义了连接两个实体之间路径上的谓词序列，并且使用了多达 94 种语言特征。[Jiang et al.,2007]通过统一的特征空间表达形式来研究不同特征对关系抽取性能的影响，其中特征空间可划分为序列、句法树和依存树等特征子空间。实验结果表明，从三个子空间中提取出的基本单元特征能取得较好的性能，而再加入复杂的特征所带来的性能提升很小，只有当不同子空间和不同复杂度的特征结合起来时，才能取得最好的性能。基于特征向量的方法尽管速度很快，也比较有效，但其缺点是在转换结构化特征时需要显式地给出一个特征集合，由于实体间语义关系表达的复杂性和可变性，要进一步提高关系抽取的性能已经很困难了，因为很难再找出适合语义关系抽取的新的有效的词汇、句法或语义特征。

不同于特征向量的方法，基于核函数的方法不需要构造固有的特征向量空间，能很好地弥补基于特征向量方法的不足。在关系抽取中，基于核函数的方法直接以结构树为处理对象，在计算关系之间的距离的时候不再使用特征向量的内积而是用核函数，核函数可以在高维的特征空间中隐式地计算对象之间的距离，不用枚举所有的特征也可以计算向量的点积，表示实体关系很灵活，可以方便地利用

多种不同的特征，使用支持核函数的分类器进行关系抽取。基于核函数的关系抽取最早由[Zelenko et al.,2003]，他们在文本的浅层句法树的基础上定义了树核函数，并设计了一个计算树核函数相似度的动态规划算法，然后通过支持向量机（SVM）和表决感知器（Voted Perceptron）等[Grishman et al.,2005]分类算法来抽取实体间语义关系。[Culotta et al.,2004]提出基于依存树核函数的关系抽取，他们使用一些依存规则将包含实体对的句法分析树转换成依存树，并在树节点上依次增加词性、实体类型、词组块、WordNet 上位词等特征，最后使用 SVM 分类器进行关系抽取。Mooney 和 Bunescu[Bunescu et al.,2005]进一步使用最短依存树核函数，该核函数计算在依存树中两个实体之间的最短路径上的相同节点的数目，要求对于具有相同关系的实体对，其对应的最短依存树具有相同的高度且达到根节点的路径相同。为解决最短依存树核函数召回率较低的问题，Bunescu 和 Mooney[Mooney et al.,2006]又提出基于字符串序列核函数的关系抽取，首先提取出两个实体之间和前后一定数量的单词组成字符串并把其作为关系实例的表达形式，规定子序列中允许包含间隔项，进而实现关系抽取。[Zhou et al.,2007]提出最短路径包含树核，将语义关系实例表示为上下文相关的最短路径包含树，能根据句法结构动态扩充与上下文相关的谓词部分，并采用上下文相关的核函数计算方法，即在比较子树相似度时也考虑根结点的祖先结点，将该核函数同基于特征的方法结合起来，充分考虑结构化信息和平面特征的互补性。然而该类方法依赖传统的句法分析等复杂的自然语言处理工具，很多语言没有此类工具，即使有的语言有此类工具但是也会造成误差传递，影响最终的性能。

近年来，深度学习的方法在有监督关系抽取任务中占据了主导地位。[Zeng et al.,2014] 首先将卷积神经网络应用在了有监督关系抽取任务中。他们主要是应用了词向量将句子表示成了矩阵，再利用卷积神经网络和 maxpooling 得到句子的向量表示。最后用 softmax 分类器对该向量进行分类，得到句子的关系类别。同时期还有 [Thien et al.,2015] [Santos et al.,2015] 等工作也是采用了相似的方法。还有 [Socher et al.,2012] 利用了长短时记忆网络（LSTM）和句子的依存句法路径来建模句子的表示，最后再用 softmax 分类器进行分类。为了更好的建模句子，[Zhou et al.,2016] 提出使用双向长短时记忆网络和关注机制。目前大部分学者关注于如何更好的用深度学习模型建模句子。此类方法一般默认句子中已经标记出了候选实体，但是实际任务中，需要系统自动发现实体。而且此类方法需要大量的人工标注的语料作为训练数据才能取得较好的性能。

### 3.2.2.3 弱监督关系抽取

有监督关系抽取需要大量的标注样本，而人工标注数据费时费力、一致性差，尤其是面向海量异构的网络数据时，问题就更加明显，为此，研究人员提出弱监督关系抽取。弱监督关系抽取主要有两种框架，一种是使用半监督学习和主动学

习等技术以尽可能少的代价提升抽取效果，如[Sun et al.,2011]通过大规模的词聚类作为额外的特征，以解决实体之间特征过于泛化的问题，从而帮助关系抽取；[Sun et al.,2012]利用主动学习的技术，通过少量的标注数据来发现分类面附近的未标注数据，对这些数据进行人工标注，从而以更少的标注代价获得更好的抽取效果。另外一种框架是使用回标的思想，利用现有知识库中的关系三元组，自动回标三元组中实体所在的文本作为训练数据，由于其训练数据产生过程不需要人工标注，所以这种方法代价很低，更加适合大规模多领域的网络文本，它在信息抽取领域近年来得到较广泛的应用。

弱监督回标思想最早由[Craven et al.,1999]提出，主要研究怎样在文本中抽取结构化数据建立生物学知识库（Biological Knowledge Bases），他们利用 Yeast Protein Database 自动产生标注数据，然后训练朴素贝叶斯分类器抽取结构化数据。紧接着，[Mintz et al.,2009]使用利用 Freebase 作为知识库，将其中的关系实例所包含的实体同维基百科文本中的实体对齐，以此产生训练数据，然后使用逻辑斯谛回归进行关系抽取。弱监督回标主要基于以下假设：如果两个实体在知识库中具有一定的关系，那么根据同时包含这两个实体的句子，就都能推断出实体对在知识库中具有的关系。由于语言表达的多样性，弱监督的这种假设往往太过强烈，两个实体出现在同一个句子中并不能表示它们就一定具有某种语义关系，有可能这两个实体只是属于同一个话题而已[Riedel et al.,2010]。因此，虽然弱监督方法克服了有监督方法需要人工标注数据的不足，但也带来了新问题——回标噪声问题。研究人员提出了一系列模型和方法来克服回标噪声问题，Riedel 等将弱监督关系抽取看作是一个多示例问题，他们的假设中，只需要在回标出来的所有句子中，有一个句子能表示两个实体间的关系。将所有回标的句子看作一个包，其中的每一个句子就是包中的一个示例，从而解决回标噪声的问题。[Hoffmann et al.,2011] 更进一步，在多实例模型中考虑实体对间可能不止有一种关系，取得了更好的效果。[Surdeanu et al.,2012] 不但对噪声训练数据进行建模，并对实体对可能属于多个关系类型这个问题进行建模，他们提出了基于概率图模型的多实例多标签模型，在以 Freebase 为知识库和纽约时报作为回标语料的数据上进行实验，结果表明其模型提升了原始方法的抽取效果。[Takamatsu et al.,2012] 发现多示例模型的“至少一句表达真实关系”的假设有可能失败，其通过抽样统计显示，Freebase 知识库中 91.7% 的实体对在英文维基百科文章中只能回标到一个句子，此时多示例模型的假设不能成立，他们的工作通过产生式图模型来预测可能具有噪声的特征模板，然后过滤包含这些模板的正样本，利用剩下的样本训练抽取模型。利用 Freebase 作为知识库，在 NYU 语料上进行回标，针对 15 类关系选取置信度最高的 50 个结果进行人工评测，平均准确率为 89%。

上述方法都是基于传统特征的，然而传统特征的设计耗时费力，扩展性差。

近些年，基于神经网络的方法占据主导地位。[Zeng et al.,2015] 首先提出了用卷积神经网络来建模句子，并依据“至少一个假设”，将整个学习过程视为多示例学习。利用 Freebase 作为知识库，在 NYU 语料上进行回标，针对 55 类关系选取置信度最高的 100 个结果进行人工评测，平均准确率为 86%。[Ji et al.,2017] [Lin et al.,2016] 认为 [Zeng et al.,2015] 的方法只能够利用包中的一个句子，提出了使用关注机制，自动学得包中每个句子的权重，然后将句子的表示按照权重加权表示为包的表示，最后对包进行分类，得到包的关系。其中 [Ji et al.,2017] 还使用了外部文本信息。[Jiang et al.,2016] 提出了不同于“至少一个”假设，他们认为很多情况下，要判断一个包的关系要同时使用多个句子的信息，因此提出了 Cross-sentence maxpooling 的方法。除此之外，最近有 [Luo et al.,2011] 提出使用动态矩阵来建模噪声，以此来增强弱监督关系抽取。[Lin et al.,2017] 还关注了跨语言的弱监督关系抽取。

目前，基于机器学习的关系抽取方法占据了主导地位。然而，无监督的关系抽取得到的知识缺乏语义信息、很难归一化；有监督关系抽取中需要大量人工标注的高质量数据作为训练语料，人工标注耗时费力成本高，所以很难大规模推广；弱监督关系抽取虽然可以自动生成大规模训练语料，但是自动生成训练语料的过程中需要大规模的已有知识图谱作为种子，而且生成的语料中会有噪音数据。

## 4. 技术展望与发展趋势

从 20 世纪 90 年代以来，关系抽取技术研究蓬勃发展，已经成为了自然语言处理和知识图谱等领域的重要分支。这一方面得益于系列国际权威评测和会议的推动，如消息理解系列会议（MUC，Message Understanding Conference），自动内容抽取评测（ACE，Automatic Content Extraction）和文本分析会议系列评测（TAC，Text Analysis Conference）。另一方面也是因为关系抽取技术的重要性和实用性，使其同时得到了研究界和工业界的广泛关注。关系抽取技术自身的发展也大幅度推进了中文信息处理研究的发展，迫使研究人员面向实际应用需求，开始重视之前未被发现的研究难点和重点。纵观关系抽取研究发展的态势和技术现状，我们认为关系抽取的发展方向如下：

- 面向开放域的可语文化的关系抽取技术

目前，绝大部分的关系抽取研究集中预定义的关系抽取上，并致力于构建更精准的有监督抽取模型和方法，使用标注语料训练模型参数。然而，在构建真实环境下的关系抽取系统时，这些有监督方法往往存在如下不足：1) 更换语料类型之后，现有模型往往会有大幅度的性能下降；2) 无法抽取目标关系类别之外的实体关系知识；3) 性能依赖于大规模的训练语料；4) 现有监督模型往往依赖于高复杂度的自然语言处理应用，如句法分析。

目前已经有很多机构和学者进行开放域的关系抽取的研究，但是目前的方法抽取的关系很难语义化，同一个实体对的同一关系会抽取出不同的表达，另外不同的数据来源其质量和可信度不同，如何整合不同数据源抽取的关系知识，并将同一关系的知识进行消歧进而语义化是一个迫切需要解决的问题。

- 篇章级的关系抽取

现有大多数的关系抽取集中在从包含两个指定实体的一个或者多个句子中抽取关系，很少有工作将抽取范围扩大到篇章级别。然而，真实环境下，如产品说明书等，一篇文章会描述多个实体的多个属性或者关系，而且文本中存在大量的零指代的语言现象，因此必须利用篇章级的信息进行关系和属性值的抽取。

- 具有时空特性的多元关系抽取

目前，绝大部分的关系抽取研究集中在二元关系抽取上，即抽取目标为三元组（实体 1，关系，实体 2），然而二元关系很难表达实体关系的时间特性和空间特性，而且很多关系是多元的，例如：NBA 球星勒布朗詹姆斯效力过的球队。这就是一个多元关系，首先他效力过的球队有多支，其次效力于每支球队的时间也不同，这就是关系的时空性和多元性。具有时空特性的多元关系能建模和表达更丰富的关系知识，是未来研究的一个方向。

最后，纵观近 30 余年来关系抽取的现状和发展趋势，我们有理由相信，随着海量数据资源（如 Web）、大规模深度机器学习技术（如深度学习）和大规模知识资源（如知识图谱）的蓬勃发展，关系抽取这一极具挑战性同时也极其实用性的问题将会得到相当程度的解决。同时，随着低成本、高适应性、高可扩展性、可处理开放域的关系抽取研究的推进，关系抽取技术的实用化和产业化将在现有的良好基础之上取得进一步的长足发展。

## 参考文献

- [Angeli et al.,2015] Gabor Angeli, Melvin Johnson Premkumar, and Christopher D Manning. Leveraging linguistic structure for open domain information extraction. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015), 2015.
- [Aone et al.,1998] Aone C., Halverson L., Hampton T., and Ramos-Santacruz M. SRA: Description of the IE System Used for MUC-7. In Proceedings of the 7th Message Understanding Conference (MUC-7). 1998.
- [Aone et al.,2000] Aone C. and Ramos-Santacruz M. REES: A large-scale relation and event extraction system In Proceedings of the 6th Applied Natural Language Processing Conference (ANLP00), 2000, pages 76-83.
- [Banko et al.,2007] Michele Banko; Michael J Cafarella; Stephen Soderland; Matthew Broadhead and Oren Etzioni. Open Information Extraction for the Web.

## IJCAI2007

- [Betteridge et al.,2009] Justin Betteridge, Andrew Carlson, Sue Ann Hong, Estevam R Hruschka Jr, Edith LM Law, Tom M Mitchell, and Sophie H Wang. Toward never ending language learning. In AAAI Spring Symposium: Learning by Reading and Learning to Read, pages 1–2, 2009.
- [Bollegrala et al.,2010] Danushka Tarupathi Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In Proceedings of the 19th International Conference on World Wide Web, WWW ’10, pages 151–160, New York, NY, USA, 2010. ACM.
- [Bunescu et al.,2005] Razvan C. Bunescu and Raymond J. Mooney. A shortest path dependency kernel for relation extraction. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT ’05, pages 724–731, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [Carlson et al.,2010] Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. Coupled semi-supervised learning for information extraction. In Proceedings of the third ACM international conference on Web search and data mining, pages 101–110. ACM, 2010.
- [Chen et al.,2005] Jinxiu Chen, Donghong Ji, Chew Lim Tan, and Zhengyu Niu. Unsupervised feature selection for relation extraction. In Proceedings of the Second International Joint Conference on Natural Language Processing, IJCNLP ’05, pages 262–267, Berlin, Heidelberg, 2005. Springer-Verlag.
- [Collins et al.,2002] Michael Collins and Nigel Duffy. Convolution kernels for natural language. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, Advances in Neural Information Processing Systems 14, pages 625–632. MIT Press, 2002.
- [Craven et al.,1999] Mark Craven and Johan Kumlien. Constructing biological knowledge bases by extracting information from text sources. In Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, pages 77–86. AAAI Press, 1999.
- [Culotta et al.,2004] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL ’04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [Etzioni et al.,2011] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen

- Soderland, and Mausam Mausam. Open information extraction: The second generation. In IJCAI, volume 11, pages 3–10, 2011.
- [Fader et al.,2011] Anthony Fader; Stephen Soderland and Oren Etzioni. Identifying relations for open information extraction. ACL2011
- [Fukumoto et al.,1998] Fukumoto J., Masui F., Shimohata M., and Sasaki M. Oki Eletricity Industry: Description of the Oki System as Used for MUC-7 . In Proceedings of the 7th Message Understanding Conference (MUC-7). 1998.
- [Fung et al.,2002] Gabriel Pui Cheong Fung, Jeffrey Xu Yu, and Hongjun Lu. Discriminative category matching: Efficient text classification for huge document collections. In Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on, pages 187–194. IEEE, 2002.
- [Grishman et al.,1996] Grishman R. and Sundheim B. Message Understanding Conference-6: a brief history. In Proceeding of the 16th Conference on Computational Linguistics, August 05-09, 1996.
- [Grishman et al.,2005] Ralph Grishman, David Westbrook, and Adam Meyers. Nyu’s english ace 2005 system description. In Proceedings of ACE 2005 Evaluation Workshop. Washington,2005.
- [Harris et al.,1954] Zellig S Harris. Distributional structure. Word, 10:146–162, 1954.
- [Hasegawa et al.,2004] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL ’04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [Hoffmann et al.,2011] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT ’11, pages 541–550, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
- [Humphreys et al.,1998] Humphreys H., R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. University of Sheffield: Description of the LaSIE-II System Used for MUC-7. In Proceedings of the 7th Message Understanding Conference (MUC-7). 1998.
- [Jiang et al.,2007] Jing Jiang and ChengXiang Zhai. A systematic exploration of the feature space for relation extraction. In Proceedings of Human Language Technologies 2007 and the North American Chapter of the Association for

- Computational Linguistics, HLT-NAACL '07, pages 113–120, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics
- [Jiang et al.,2016] Xiaotian Jiang; Quan Wang; Peng Li; Bin Wang. Relation Extraction with Multi-Instance Multi-Label Convolutional Neural Networks. COLING2016
- [Ji et al.,2017] Guoliang Ji; Kang Liu; Shizhu He; Jun Zhao. Distant Supervision for Relation Extraction with Sentence-level Attention and Entity Descriptions. AAAI2017
- [Kambhatla et al.,2004] Nanda Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics on Interactive Poster and Demonstration Sessions, ACLdemo'04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
- [Lin et al.,2016] Yankai Lin; Shiqi Shen; Zhiyuan Liu; Huanbo Luan; Maosong Sun. Neural Relation Extraction with Selective Attention over Instances. ACL2016
- [Lin et al.,2017] Yankai Lin; Zhiyuan Liu; Maosong Sun. Neural Relation Extraction with Multi-lingual Attention. ACL2017
- [Luo et al.,2011] Bingfeng Luo; Yansong Feng; Zheng Wang; Zhanxing Zhu; Songfang Huang; Rui Yan and Dongyan Zhao. Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix. ACL2011
- [Mintz et al.,2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pages 1003–1011. Association for Computational Linguistics, 2009.
- [Mitra et al.,2002] Pabitra Mitra, C. A. Murthy, and Sankar K. Pal. Unsupervised feature selection using feature similarity. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(3):301–312, March 2002.
- [Mooney et al.,2006] Raymond J. Mooney and Razvan C. Bunescu. Subsequence kernels for relation extraction. In Y. Weiss, B. Schölkopf, and J.C. Platt, editors, Advances in Neural Information Processing Systems 18, pages 171–178. MIT Press, 2006.
- [Riedel et al.,2010] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in

- Databases: Part III, ECML PKDD’10, pages 148–163, Berlin, Heidelberg, 2010. Springer-Verlag.
- [Santos et al.,2015] Cicero Nogueira Dos Santos; Bing Xiang; Bowen Zhou. Classifying Relations by Ranking with Convolutional Neural Networks. ACL2015
- [Socher et al.,2012] Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. Semantic compositionality through recursive matrix-vector spaces. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 1201–1211, 2012.
- [Sun et al.,2011] Ang Sun, Ralph Grishman, and Satoshi Sekine. Semi-supervised relation extraction with large-scale word clustering. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, ACL ’11, pages 521–529. Association for Computational Linguistics, 2011.
- [Sun et al.,2012] Ang Sun and Ralph Grishman. Active learning for relation type extension with local and global data views. In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM ’12, pages 1105–1112. ACM, 2012.
- [Surdeanu et al.,2012] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance multi-label learning for relation extraction. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL ’12, pages 455–465, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Takamatsu et al.,2012] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1, ACL ’12, pages 721–729, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Thien et al.,2015] Nguyen Thien and Grishman Ralph. Relation Extraction: Perspective from Convolutional Neural Networks. Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015
- [Wang et al.,2006] Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, and Ji Wang. Automatic extraction of hierarchical relations from text. In Proceedings of the 3rd European Conference on The Semantic Web: Research and

- Applications, ESWC'06, pages 215–229, Berlin, Heidelberg, 2006. Springer-Verlag.
- [Wu et al.,2010] Fei Wu and Daniel S Weld. Open information extraction using Wikipedia. ACL2010
- [Yao et al.,2012] Limin Yao, Sebastian Riedel, and Andrew McCallum. Unsupervised relation discovery with sense disambiguation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL'12, pages 712–720, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [Zelenko et al.,2003] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. *The Journal of Machine Learning Research*, 3:1083–1106, 2003.
- [Zeng et al.,2014] Daojian Zeng; Kang Liu; Siwei Lai; Guangyou Zhou; Jun Zhao. Relation Classification via Convolutional Deep Neural Network. COLING2014
- [Zeng et al.,2015] Daojian Zeng; Kang Liu; yubo Chen; Jun Zhao. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. EMNLP2015
- [Zhang et al.,2006] Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou. A composite kernel to extract relations between entities with both flat and structured features. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL '06, pages 825–832, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [Zhao et al.,2005] Shubin Zhao and Ralph Grishman. Extracting relations with integrated information using kernel methods. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 419–426, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [Zhou et al.,2005] GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.
- [Zhou et al.,2005] GuoDong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pages 427–434, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[Zhou et al.,2007] Guodong Zhou, Min Zhang, Donghong Ji, and Qiaoming Zhu. Tree kernelbased relation extraction with context-sensitive structured parse tree information. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLPCoNLL’07, pages 728–736, 2007.

[Zhou et al.,2016] Peng Zhou; Wei Shi; Jun Tian; Zhenyu Qi; Bingchen Li; Hongwei Hao; Bo Xu. Attention-based bidirectional long short-term memory networks for relation classification. ACL2016

# 第五章 事件知识学习

## 1. 任务定义、目标和研究意义

事件（Event）的概念起源于认知科学，广泛应用于哲学、语言学、计算机等领域[Quine, 1985;Trabasso, 1985;Zwaan, 1999;Chemero, 2000;Zacks, 2001;Glasbey, 2004;Fernando, 2007]。遗憾的是，目前学术界对此尚且没有公认的定义，针对不同领域的不同应用，不同学者对事件有不同的描述。在计算机科学的范畴内最常用的事件定义有如下两种：

- 第一种源自信息抽取领域，最具国际影响力的自动内容抽取评测会议（Automatic Content Extraction, ACE）对其定义为：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[Doddington et.al., 2004]。
- 第二种源自信息检索领域，事件被认为是细化的用于检索的主题。美国国防高级计划研究委员会主办的话题检测与追踪（Topic Detection and Tracking, TDT）评测指出：事件是由某些原因、条件引起，发生在特定时间、地点，涉及某些对象，并可能伴随某些必然结果的事情[Allan et.al., 1998a]。

虽然两种定义的应用场景和侧重点略有差异，但均认为事件是促使事物状态和关系改变的条件[Dong et.al., 2010]。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。此外，很多认知科学家们认为人们是以事件为单位来体验和认识世界的，事件符合人类正常认知规律，如维特根斯坦在《逻辑哲学论》中论述到“世界是所有事实，而非事物的总和” [Ludwig, 2001]。因此，事件知识学习，即将非结构化文本中自然语言所表达的事件以结构化的形式呈现，对于知识表示、理解、计算和应用均意义重大。接下来，本文将沿着上述两种定义对事件知识学习的任务、挑战、研究现状和趋势进行梳理和展望。

### 1.1 任务定义

为了方便叙述，本文称针对第一种定义的相关研究为事件识别和抽取，针对第二种定义的相关研究为事件检测与追踪。

事件识别和抽取研究如何从描述事件信息的文本中识别并抽取出事件信息并以结构化的形式呈现出来，包括其发生的时间、地点、参与角色以及与之相关的动作或者状态的改变，核心的概念有：

- **事件描述（Event Mention）：**客观发生具体事件的自然语言描述，通常

是一个句子或者句群。同一事件可以有很多不同的事件描述，可能分布在同一篇文档的不同位置或不同的文档中。

- **事件触发词 (Event Trigger):** 事件描述中最能代表事件发生的词，是决定事件类别的重要特征，在 ACE 评测中事件触发词一般是动词或名词。
- **事件元素 (Event Argument):** 事件的参与者，是组成事件的核心部分，与事件触发词构成了事件的整个框架。事件元素主要由实体、时间和属性值等表达完整语义的细粒度单位组成。
- **元素角色 (Argument Role):** 事件元素与事件之间的语义关系，也就是事件元素在相应的事件中扮演什么角色。
- **事件类型 (Event Type):** 事件元素和触发词决定了事件的类别。很多评测和任务均制定了事件类别和相应模板，方便元素识别及角色判定。

事件检测与追踪旨在将文本新闻流按照其报道的事件进行组织，为传统媒体多种来源的新闻监控提供核心技术，以便让用户了解新闻及其发展。具体而言，事件发现与跟踪包括三个主要任务：分割，发现和跟踪，将新闻文本分解为事件，发现新的（不可预见的）事件，并跟踪以前报道事件的发展。事件发现任务又可细分为历史事件发现和在线事件发现两种形式，前者目标是从按时间排序的新闻文档中发现以前没有识别的事件，后者则是从实时新闻流中实时发现新的事件。

## 1.2 公开评测和数据集

### 1.2.1 事件识别和抽取

事件识别和抽取最早可以追溯到 20 世纪 70 年代耶鲁大学 Roger 等开展的故事理解相关研究，他们应用故事脚本理论从新闻报道中抽取工人罢工、地震等事件 [Roger, 1978]。随着信息抽取技术的不断发展，事件识别和抽取也受到越来越多的关注，主要推动力是一系列相关国际评测会议的开展以及语料资源的丰富。

#### 1.2.1.1 公开评测

消息理解会议 (Message Understanding Conference, MUC) 是公认最早的信息抽取评测会议 [Chinchor and Marsh, 1998]，由美国国防高级研究计划委员会 (Defense Advanced Research Projects Agency, DARPA) 于 1987 年首次举办。MUC 要求从非结构化文本中抽取信息填入预定义模板中的槽，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。从 1987 年到 1997 年的 7 届 MUC 评测，除了任务种类和槽值数量，还增加了模版嵌套、多语言抽取等。

前文提到的 ACE 评测会议由美国国家标准技术研究所 (National Institute of Standards and Technology, NIST) 从 1999 年 7 月开始准备，2000 年首次召开，到 2008 年共召开了八次，后被并入文本分析会议 (Text Analysis Conference, TAC)。从 2004 年起，事件抽取成为 ACE 评测的主要任务，其中的事件是预定义类型的

句子级事件，每个事件都标注了事件触发词、事件类型、事件子类型、事件元素和元素角色信息，2007 年评测任务中增加了时间表达式的识别和归一化。ACE 吸引了很多学者参与并设计测评系统，现在很多流行的事件抽取工具都是针对 ACE 测评研发的，ACE 测评对事件识别和抽取技术的研究具有非常深远的影响。

知识库生成测评（Knowledge Base Population, KBP）隶属于 TAC 会议，主要研究从自然语言文本中抽取信息并链接到现有知识库的相关技术。事件抽取为 KBP 的一项重要任务，2014 年首次加入评测，目前已成功举办四届，2018 届正在筹备。KBP 中事件类型和实体角色遵从 ACE 2005 的定义，每类子事件类型都定义了各自的事件元素，并在此基础上极大丰富了事件抽取任务内容，主要包括：事件识别、事件消歧、事件元素抽取和链接、事件元素验证和链接。2016 年起，KBP 评测增加了以文档为单位的识别，语料也从英文扩展至汉语和西班牙语。

此外还有一些限定领域事件识别和抽取的公开评测，例如，东京大学组织的 BioNLP 是面向生物医学事件抽取的最权威的评测，2009 年到 2013 年共举办三届，目标是从生物医学文献中抽取出事件触发词、事件类型和事件元素等生物事件信息，其他类似评测不再赘述。

### 1.2.1.2 数据集

由于事件结构的复杂性和自然语言表达的灵活多样性，目前还没有形成统一的事件框架体系。另外，语料依赖人工标注，标注过程耗时、费力、成本高，因此事件类型较少、整体规模也不容易扩大。相关研究多是在各国际评测和公开语料的推动下展开，下面主要介绍 ACE 数据集、TimeBank 语料和中文事件语料库。

ACE 语料来源包括新闻专线、广播新闻、广播会话、网络日志、论坛数据及电话会话，美国宾夕法尼亚大学的语言数据组织（Linguistics Data Consortium, LDC）对源文本进行标注，生成的事件抽取数据集包含英文 599 篇，中文 633 篇。ACE 共定义了 8 大类 33 小类的事件。除基本的触发词、事件类型、事件子类型、事件元素和元素角色信息，ACE 还为每个事件标注了如下四种属性：

- 极性（Polarity），即肯定的事件和否定的事件。
- 时态（Tense），包括过去发生的事件，现在正在发生的事件，将来即将发生的事件以及无法确定时态的事件。
- 指属（Genericity），包括特指（Specific）事件和泛指（Generic）事件。
- 形态（Modality），包括语气非常肯定的事件（Asserted Event）和信念事件（Believed Event）、假设事件（Hypothetical Event）和其它事件。

在此基础上，KBP 英文语料同样包含 599 篇标注文档，由人工过滤来确保每种事件类型都有多个实例，并且针对长句子进行了截断。KBP 2016 提供了 200 篇标注的英文文档、20 万词的中文文档以及 12 万词的西班牙文文档用于评测，但并未提供训练语料。

TimeBank 语料[Pustejovsky et.al., 2003]是由面向问答系统的时间和事件识别会议 (Time and Event Recognition for Question Answering Systems, TERQAS) 提供，主要应用于识别和抽取事件的时间元素及事件之间的时序关系。语料包括来源于 DUC、ACE 和 PropBank 中的 300 篇新闻报道。TimeBank 不关心时间之外的事件元素，主要标注了事件、时间及其类型、时间信号词以及事件间的时序关系等。其中，事件通过事件类、事件时态以及事件状态三个属性来描述，事件类型分为 Occurrence、Perception、Reporting、Aspectual、State、Intensional State、Intensional Action 和 Modal 等 8 种。

中文事件语料库 (Chinese Event Corpus, CEC) 由上海大学语义智能实验室构建，旨在填补中文突发事件语料库的空白，包含 CEC-1 和 CEC-2 两个语料库 [Meng, 2015]。CEC-1 是针对 200 篇国内外突发事件的中文新闻报道的标注，包含了 1228 个句子、3133 个事件和 4878 个事件元素，但标注的粒度相对较大，尤其是对事件元素，且未对事件关系分类。为弥补上述不足，CEC-2 选取 333 篇关于地震、火灾、交通事故、恐怖袭击及食物中毒五类突发事件的互联网新闻报道作为待标注语料。标注过程中不仅覆盖了语料库中的所有事件，而且在中文句法分析和语义分析后进行标注，符合中文的特点，还能对标注后的语料进行一致性检查，保证语料标注的质量。除标注语料，CEC-2 还保留了未标注的原始语料，其中记录了语料来源、标题、主体等信息。

### 1.2.2 事件检测与追踪

借鉴 MUC 的成功经验，DARPA 主办了 TDT 评测，旨在以事件的形式组织新闻报道，对其进行研究和评测[Allan, 2012]。话题 (Topic) 是 TDT 中的最基本的概念，起初与事件具有相同的含义，即指由某种原因引起的，发生在特定时间点或者时间段，在某个地域范围内，并可能导致某些必然结果的一个事件；后来演变为包括一个核心事件以及与之直接相关的事件的集合。

1998 年举行的首届 TDT 主要是针对中英文两种语料进行新闻报道切分、话题识别和话题追踪三项评测，第二届增加了新事件识别和报道关系识别评测。五个子任务均与事件检测追踪研究密切相关，新事件识别就是从给定的大量文档中识别出首次报道或者以前没有识别出来的事件。对新闻报道的切分就是将大量的文档聚成不同类别，涉及新事件识别、历史事件识别和报道关系识别。随后的历届 TDT 评测（2000 年—2004 年）都包含上述五个子任务，且将评测语言扩展到中文、英文和阿拉伯文。

LDC 为 TDT 系列评测提供了 TDT-pilot 和 TDT-2 到 TDT-5 五种语料。需要指出的是，TDT 语料标注方法与 ACE 等评测的标注完全不同：TDT-2 和 TDT-3 采用 YES、BRIEF 和 NO 三类标签分别表示当前报道内容与事件绝对相关、部分相关和不相关，TDT-4 和 TDT-5 则简化为 YES 和 NO 两种[Hong et.al., 2007]。

## 2. 研究内容和关键科学问题

事件知识学习是一项综合研究，需要比较深入的自然语言处理方法和技术作为支撑。相对于其他抽取和识别任务（如实体识别、关系抽取），事件识别和抽取更加复杂且富有挑战性，其难点主要表现在以下几个方面：

**认知层面：**事件具有复杂的内部结构。事件抽取不仅要识别出事件触发词和事件类别，还要识别出事件所涉及的所有元素并判断其在事件中扮演的角色。相较于实体和关系，事件涉及更多的实体和值，而且事件中各个元素间具有复杂关系和结构。因此需要对事件描述文本更深层次的理解。

**语言层面：**事件的表述是灵活的、具有歧义的。同一事件会有不同的描述和报道，例如“离开”既可以触发移动事件，也可以触发离职事件。同一事件的元素也可能会出现在多个句子、段落或者篇章中，一个句子或者一篇文章会描述多个不同但是相关或者不相关的事件。因此自然语言的灵活多变和歧义性对面向非结构化文本的事件抽取提出了很大的挑战。

**方法层面：**事件抽取会遇到错误累积的问题。事件抽取一般依赖于词法、句法分析等基本的自然语言处理工具，但实际中许多自然语言处理工具性能并不高，低性能的工具引入的错误会降低事件抽取系统的性能。

**语料层面：**标注语料规模小、数据稀疏。事件结构的复杂性和表述方式的歧义性导致人工标注事件的成本高、一致性差、耗时费力。因此，现有事件抽取相关数据集普遍规模较小，数据稀疏问题严重，对抽取的性能造成了很大的影响。

对于事件检测和追踪，虽然着眼点比事件识别和抽取要稍显宏观，但二者在认知、语言、方法和语料层面的挑战是高度统一的。

## 3. 技术方法和研究现状

考虑到事件识别和抽取、事件检测和追踪两个任务的处理对象、着眼点和技术路线的差异，本节分别对其主流的方法和现状进行梳理。

### 3.1 事件识别和抽取

根据抽取方法，事件抽取可以分为基于模式匹配的事件抽取和基于机器学习的事件抽取。接下来首先依此分类梳理事件抽取的国内外相关研究工作，然后对目前关注度较高的中文事件抽取相关研究进行介绍。

#### 3.1.1 基于模式匹配的方法

基于模式匹配的方法是指对某种类型事件的识别和抽取是在一些模式的指导下进行的，模式匹配的过程就是事件识别和抽取的过程。采用模式匹配的方法进行事件抽取的过程一般可以分为两个步骤：模式获取和模式匹配。模式准确性是影响整个方法性能的重要因素，按照模式构建过程中所需训练数据的来源可细

分为基于人工标注语料的方法和弱监督的方法。

### 3.1.1.1 基于人工标注语料的方法

顾名思义，此类方法的模式获取完全基于人工标注的语料，学习效果高度依赖于人工标注质量。Ellen 等基于“事件元素首次提及之处即可确定该元素与事件间关系”和“事件元素周围的语句中包含了事件元素在事件中的角色描述”两个假设开发的事件模式抽取系统 AutoSlog 就属于这个范畴[Riloff, 1993]。Kim 和 Moldovan 开发的 PALKA 是另一个典型代表[Kim and Moldovan, 1995]，他们假设“特定领域中高频出现的语言表示方式是可数的”，提出用语义框架和短语模式结构来表示特定领域中的模式，用语义树来表示语义框架、用短语链模型来表示短语模式。通过融入 WordNet 的语义信息，PALKA 在特定领域可取得接近纯人工抽取的效果。

### 3.1.1.2 弱监督的方法

这类方法不需要对语料进行完全标注，只需人工对语料进行一定的预分类或制定种子模式，由机器根据预分类语料或者种子模式自动进行模式学习。例如 Ellen 等研发的 AutoSlog 升级版 AutoSlog-TS 系统[Riloff and Shoen, 1995]就只需在人工预分类的语料上进行训练，可以解决标注标准不一致的问题，同时也降低了模式训练的准备工作量。欧洲委员会联合研究中心研发的 NEXUS 系统则使用无监督聚类的方式对语料进行预处理[Piskorski et.al., 2001; Tanev et.al., 2008]。

Yangarber 等研发的 ExDisco 通过匹配优质模式获取与待抽取事件相关的语料，利用人工制定的种子模式和经过一定预处理语料迭代来寻找新的匹配模式，省去了对语料进行人工标注或者预分类，只需提供少量的模式种子，大大减少了工作量[Yangarber et.al., 2000]。Chai 等则在其提出的模式抽取系统 TIMES 中引入了领域无关的概念层次知识库 WordNet，提升模式学习的泛化能力，并通过人工或者规则进行词义消歧，使最终的模式更加准确[Chai and Biermann, 1998]。GenPAM 系统在由特例生成泛化模式的学习过程中，有效利用模式间的相似性实现词义消歧，最大限度地减少了人工的工作量和对系统的干预[Jiang, 2005]。

总体而言，基于模式匹配的方法在特定领域中性能较好，知识表示简洁，便于理解和后续应用，但对于语言、领域和文档形式等均有不同程度的依赖，覆盖度和可移植性较差。

### 3.1.2 基于机器学习的方法

基于机器学习的方法建立在统计模型基础上，一般将事件抽取建模成多分类问题，因此研究的重点在于特征和分类器的选择。根据利用信息的不同可以分为基于特征、基于结构和基于神经网络三类主要方法。

### 3.1.2.1 基于特征的方法

基于特征的方法研究重点在于如何提取和集成具有区分性的特征，从而产生描述事件实例的各种局部和全局特征，作为特征向量输入分类器。该类方法多用于阶段性的管道抽取，即顺序执行事件触发词识别和元素抽取，从特征类型（或来源）上又可细分为利用句子级信息的方法和利用篇章级信息的方法。

**句子级信息：** Chieu 等首次将最大熵模型应用于事件抽取，使用了 unigram、bigram、命名实体等简单词法特征[Chieu and Ng, 2002]。Ahn 提出在事件抽取过程中同时使用 Timbl 和 MegaM 两种模型，并抽取候选词相关的词法特征、上下文特征、实体特征、句法特征和语言学特征，在事件触发词识别和元素抽取两个阶段都取得了不错的效果[Ahn, 2006]。

**篇章级信息：** Ji 等提出了跨文档事件抽取框架[Ji and Grishman, 2008]，其主要思想是对于一个句子级的抽取结果不仅要考虑当前的置信度，还要考虑与待抽取文本相关的文本对它的影响。具体实现时通过人工设置的 9 条推理规则定量地度量相关文本对当前抽取结果的影响，从而帮助修正原有的句子级事件抽取结果。该方法的优秀表现使得后来很多学者借鉴其利用篇章信息和背景知识的思想，相继出现了跨文本事件抽取的改进[Liao and Grishman, 2010]和跨实体事件抽取系统[Hong et.al., 2011]等。为了能更好地应用全局信息，Liu 等提出了利用全局信息（如事件的相关性）和更精确的局部信息（如实体类型）相结合的基于概率软逻辑推断的方法用于事件分类[Liu et.al., 2016a]。该方法首先利用局部信息做出初步分类，进而收集全局信息，学习事件和话题间，事件与事件间的共现信息，最后结合局部信息给出的初步分类和全局信息进行全局推理。

### 3.1.2.2 基于结构预测的方法

与基于特征适用的阶段性的管道抽取不同，基于结构的方法将事件结构看作依存树，抽取任务则相应地转化为依存树结构预测问题，触发词识别和元素抽取可以同时完成。例如，Li 等考虑到传统管道式事件抽取方法中多个步骤会导致错误传递以及忽略了事件触发词与事件元素之间的相互影响，首次提出基于结构感知机的联合模型同时完成事件触发词识别和事件元素识别两个子任务，并通过 beam search 缩小搜索解空间[Li et.al., 2013a]。为了利用更多的句子级信息，Li 等提出利用结构预测模型将实体、关系和事件进行联合抽取[Li et.al., 2014]。

### 3.1.2.3 基于神经网络的方法

上述两种方法在特征提取的过程中都依赖依存分析、词性标注、句法分析等传统的自然语言处理工具，容易造成误差累积，而且有很多语言没有自然语言处理工具。2015 年起，如何利用神经网络直接从文本中获取特征进而完成事件抽取成为研究热点。Chen 等[Chen et.al., 2015]和 Nguyen 等[Nguyen and Grishman, 2015]相继提出利用卷积神经网络模型（Conventional Neural Networks, CNN）抽

取特征来完成两阶段的识别任务。Feng 等提出利用循环神经网络（Recurrent Neural Networks, RNN）进行事件检测，取得了很好的性能，但没有探索循环神经网络在事件元素抽取阶段的效果[Feng et.al., 2016]。为了更好地考虑事件内部结构和各个元素间的关系，Nguyen 等将联合抽取模型与 RNN 相结合，利用带记忆的双向 RNN 抽取句子中的特征，并联合预测事件触发词和事件元素，进一步提升了抽取效果[Nguyen et.al., 2016]。

### 3.1.2.4 弱监督的方法

上述方法无一例外地需要大量的标注样本，而人工标注数据耗时费力、一致性差，尤其是面向海量异构的网络数据时，问题就更加明显。而无监督方法得到的事件信息没有规范的语义标签（事件类别，角色名称等），很难直接映射到现有的知识库中。因此，弱监督方法也是事件抽取中的一个重要分支。Chen 等提出利用部分高质量的标注语料训练分类器，然后利用初步训练好的分类器判断未标注的数据，选取高置信度的分类样本作为训练样本，通过迭代自动扩充训练样本[Chen and Ji, 2009]。Liao 等在相关文档中使用自训练的（Self-Training）的半监督学习方法扩展标注语料，并利用全局推理的方法考虑样例的多样性进而完成事件抽取；进一步提出同时针对词汇和句子两个粒度训练最大熵分类器，并用协同训练（Co-training）的方法扩展标注数据，进而对分类器进行更充分的训练[Liao and Grishman, 2011a; 2011b]。Liu 等利用 ACE 语料训练的分类器去判定 FrameNet 中句子的事件类别，再利用全局推断将 FrameNet 的语义框架和 ACE 中的事件类别进行映射，进而利用 FrameNet 中人工标注的事件样例扩展训练数据以提升事件检测性能。目前基于弱监督的事件抽取方法还处于起步阶段，亟需能自动生成大规模的、高质量的标注数据的方法提升事件抽取的性能[Liu et.al., 2016b]。

## 3.1.3 中文事件抽取

目前国内外事件抽取相关的研究大部分都是面向英文文本的英文事件抽取，面向中文文本的中文事件抽取工作才刚刚起步，主要面临技术和数据两方面的挑战。技术层面，中文的词句是意合的，词语间没有显式分隔符，而且中文实词在时态和形态上也没有明显变化，因此面向中文的事件抽取研究在基础自然语言处理层面具有天然的劣势。数据层面，由于起步较晚，缺乏统一的、公认的语料资源和相关评测，极大制约了中文事件抽取的研究。尽管如此，近些年中文事件抽取在公开评测、领域扩展和跨预料迁移方面也取得一定进展。

### 3.1.3.1 公开评测

基于 ACE 人工标注的中文数据集，国内学者探索了中文事件抽取的若干关键研究点。除了在模型方面的创新[Chen and Ng, 2012; Li et.al., 2012a; 2013b]，在中文语言特性的利用方面，Li 等通过中文词语的形态结构、同义词等信息捕获更

多的未知触发词，进而解决中文事件抽取面临的分词错误和训练数据稀疏等问题；进一步细分中文事件触发词内部的组合语义（复合、附加和转化），进而提高系统的性能[Li et.al., 2012b]。Ding 等利用聚类的方法自动生成新事件类型的语料，在抽取过程中特别地考虑了待抽取文本的 HowNet 相似度[Ding et.al., 2013]。

### 3.1.3.2 领域扩展

除了公开评测，国内很多机构均面向实际应用展开特定领域的事件抽取研究，覆盖突发灾难、金融、军事、体育、音乐等多个领域。例如，Zhou 等针对金融领域事件中的收购、分红和贷款三个典型事件，提出自动构建抽取规则集的方法进行中文金融领域事件抽取 [Zhou, 2003]；Liang 等利用事件框架的归纳和继承特性实现对灾难事件的抽取[Liang and Wu, 2006]；其他不再一一详述。

### 3.1.3.3 跨语料迁移

由于目前中文事件抽取缺少公认语料，很多学者尝试利用现有大量的高质量英文标注语料辅助中文事件抽取。Chen 等首次提出该想法并利用跨语言协同训练的 Bootstrap 方法进行事件抽取[Chen and Ji, 2009]。Ji 提出基于中英文单语事件抽取系统和基于并行语料两种构建跨语言同义谓词集合的方法辅助进行中文事件抽取[Ji, 2009]，Zhu 等利用机器翻译同时扩大中文和英文训练语料，联合利用两种语料进行事件抽取[Zhu et.al., 2014]。Hsi 等联合利用符号特征和分布式特征的方法，利用英文事件语料提升中文事件抽取的性能[Hsi et.al., 2016]。

## 3.2 事件检测和追踪

事件检测和追踪研究的主流方法包括基于相似度聚类和基于概率统计两类。

### 3.2.1 相似度聚类法

基于相似度的方法首先需要定义相似度度量，而后基于此进行聚类或者分类。Yang 等提出在 TDT 中用向量空间模型（Vector Space Model, VSM）对文档进行表示，并提出了组平均聚类(Group Average Clustering, GAC)和单一通过法(Single Pass Algorithm, SPA)两种聚类算法[Yang et.al., 1998]。GAC 只适用于历史事件发现，它利用分治策略进行聚类。SPA 可以顺序处理文档并增量式产生聚类结果，能同时应用于历史事件发现和在线事件发现。在此基础上，Yang 等还提出利用衰减函数和时间窗口对事件聚类进行约束[Yang et.al., 1999]。Allan 等也在 TDT 的评测中尝试了 Single-Pass 的方法并取得较好的性能[Allan et.al., 1998b]。

后续研究中，有些工作尝试寻找新的距离度量方式，如 Hellinger 距离[Brants et.al., 2003]，但主要的还是尝试更改文档特征的提取方法来提升效果。如 Yang 等提出根据类别信息重新计算命名实体和非命名实体的权重[Yang et.al., 2002]。张等提出了一种基于 TF-IDF 的改进算法，利用 WordNet 中词汇关系对原文本向量进行补充 [Zhang, 2006]。Kumaran 等提出使用制定规则的方法为文档打分，复杂

度要小于分别计算文档向量，所以速度较快[Kumaran and Allan, 2004]。对于语法规则的规范文档的处理方面，上述算法已经可以取得比较不错的效果，但现实情况是很多文本是被大量普通用户创造，文本长短不一，且内容、格式和语法等方面均不规范，导致对这些不规范文档进行 TDT 十分有挑战性。Guille 等提出一种针对 Twitter 的事件检测与跟踪的算法，利用社交网站的评论和转发等特性尽可能的获取足够多的信息[Guille et.al., 2014]，类似的还有 Shamma 等[Shamma et.al., 2011]和 Benhardus 等[Benhardus and Kalita, 2013]的工作，均尝试解决社交平台上的短小不规范文档的 TDT 研究。

总体而言，基于相似度的模型用途广泛，计算速度通常比较快，但缺乏对于统计规律的利用。

### 3.2.2 概率统计法

概率统计方法通常使用生成模型，由于需要大量数据的支持，所以这种方法更加适用于历史事件检测。对比基于相似度聚类的模型，这类模型虽然复杂，但当数据量充足时，通常可以取得更好的准确率。基于概率的方法是目前 TDT 中的研究热点，主要分成两个方向，一是针对新闻等比较正式的规范文档，另一个则用于不规则或没有规律的非规范文档，下面分开阐述。

对新闻等规范文档，文中一般包含有完整的时间、地点、人物等信息，找出这些要素可以帮助建立新闻之间的关联。Li 等[Li et.al., 2012b]提出的生成模型将事件和文章均表示成<人物，地点，关键字，时间>，事件和文章的区别是文章的时间是时间点，而事件是时间段。人物，地点，关键字提取出来后，使用朴素贝叶斯的思想求出生成数据的分布。最近的相关研究是 Ge 等人提出的 BINets[Ge et.al., 2016a]。BINets 是一个边上带有权值的图，图中的每个结点代表一段时间反常出现的词，两个结点之间权重由统计规律得出，即两个单词一起反常出现的概率越高，它们对应结点之间的权重越高。利用 BINets，Ge 又提出了基于 BINets 进行聚类的方法[Ge et.al., 2016b]，利用 BINets 找到某特定事件的中心位置，再通过 BINets 的权值进行聚类。

不规范文档方面，算法经常是基于 LDA 等主题模型的变体建立文档间的联系，Blei 等对一些变体的特点进行了总结[Blei and Lafferty, 2006]。Griffith 提出了通过在一段时间窗口内的后验概率可以估计在这段时间内对事件的支持程度 [Griffiths and Steyvers, 2004]。Hall 等提出通过计算事件在以年为单位的离散时间上的分布的后验概率来计算事件分布的强度变化趋势[Hall et.al., 2008]。Mei 等在其提出的基于主题模型的算法中充分利用了时间信息，提高了事件追踪的准确率 [Mei and Zhai, 2005]。Hu 等引入突破点的概念（即事件突发或发生重大转折的时间点），通过新闻切分、分析、演化关系发现等步骤检测突破点，得出新闻事件的时序摘要[Hu et.al., 2011]。徐等基于 LDA 进行改进，通过对不同时间段分别建

模，可以分析该事件热度的变化并持续追踪该事件[Xu et.al., 2016]。

### 3.3 事件知识库构建

前文曾经提到，已有知识图谱，如 DBpedia, Yago 和 Wikidata 等均侧重于实体的客观属性及实体间的静态关联，缺乏结构化的事件数据。事件知识学习的最终目的就是从非结构化的文本数据中抽取结构化的事件表示，构建事件知识库弥补现有知识图谱的动态事件信息缺失问题。目前事件知识库构建的研究处于起步阶段，基础就是上述两方面研究，基于句子级的事件抽取和文档级的事件发现。

#### 3.3.1 基于句子级的事件抽取

Wang 等构造了一种基于本体的新闻事件模型 NOEM，利用事件的类型、时间、空间、结构、因果、媒体六个方面特征描述新闻事件的 5WIH<sup>3</sup>语义要素[Wang and Zhao, 2012]。将抽取的关键事件语义要素自动扩充到本体后，可构成事件知识库，支持事件语义层次的应用。与现有事件模型的比较以及实际应用结果显示，NOEM 能够有效描述单个新闻文档中的关键事件、语义要素以及它们之间的关联，具有很强的形式化知识表达、应用集成和扩展能力。NewsReader 中，Rospocher 等提出一个以事件为中心的知识图谱表示，利用基于深度学习的 NLP 技术包括实体链接、语义角色标注等抽取不同语言的新闻当中的事件，并且将实体链接到已有的知识库 DBpedia 中，自动构建事件知识图谱[Rospocher et.al., 2016]。Tao 等提出了事件立方体(Eventcube)的概念，他们从新闻中抽取关键词(包括时间，人物，地点和事件等)构建词网络，基于该网络提出一种话题生成模型构建层次话题，每个话题对应词网络结构中的子图，并且支持多维度的搜索[Tao et.al., 2013]。Rouces 等构建了 FrameBase，提出了 N 维语义框架表示，是一种新的事件的表示方法，解决了传统 RDF 表示中三元组只能包含两个实体的局限。他们将句子看作实例，句子中的时间，事件类型，人物，地点等都是属于和实例关联的元素[Rouces et.al., 2015]。

#### 3.3.2 基于文档级的事件发现

Event Registry[Rupnik et.al., 2016]从多语言的新闻文档中抽取事件，将相似的新闻文档聚类，从每个类中获取一个宏观的事件，抽取事件中的人物、时间、地点等要素。类似的研究工作还包括：Kuzey 等将每个新闻文档看成一个节点，并通过新闻之间的相似度建立节点之间的边，形成图的结构，基于图对新闻文档进行聚类，每个类作为一个事件，同时得到事件之间的时序和层次关系[Kuzey et.al., 2014]；Hoxha 等利用基于文档的词袋表示，对新闻按照所描述的事件进行聚类，识别新闻中的实体并和已有的知识库进行实体链接，形成事件知识库[Hoxha et.al., 2016]。NewsMiner 通过 LDA 模型将新闻按照事件组织，并分析新

---

<sup>3</sup> Who (何人)、When (何时)、Where (何地)、What (何事)、Why (何因)、How (何种方式)

闻和评论之间的联系，在对事件，话题，以及实体之间的关系深入分析的基础上提供新闻多刻面搜索。随着相似事件的不断重复发生，事件知识可以通过增量学习得到积累完善[Hou et.al., 2015a]。在相似事件增量学习中，Hu 等将已有相似事件作为先验指导新事件知识学习并保证新事件的相对独立性，提出基于先验的狄利克雷过程混合模型，模型鼓励但不强制先验知识相关话题的出现，且允许新话题的出现[Hu et.al., 2015a;2015b]。除了新闻报道，维基百科页面的目录表蕴含了层次话题，Hu 等针对不同事件维基页面的层次话题的多样性，提出基于概率的贝叶斯网络结构学习方法，将事件维基页面的话题结构化信息和文本描述都以概率的方式建模为网络图的边上的权重[Hu et.al., 2015c]。

## 4. 技术展望与发展趋势

### 4.1 事件识别和抽取的发展趋势

通过 3.1 节的综述可以发现，事件抽取在 2002 年前基本会被形式化为模式发现和匹配，2002 年至 2013 年间，基于机器学习方法成为了主流，极大地提高了准确度并且降低了邻域迁移成本。2013 年以来，随着神经网络在图像领域取得的巨大成功，越来越多的研究者开始转向基于神经网络的事件抽取，为事件抽取任务的提升，特别是预定义的从非结构化文本中进行事件抽取任务的提升带来了新的契机。

**分步抽取到联合抽取：**事件抽取的目标往往是很多样化的，通常均会将任务拆分为几个步骤完成，最普遍的分解方式是 ACE 在 2005 年测评中定义的事件触发词识别、事件触发词分类、事件元素识别和事件元素分类四个阶段。近年来，更多工尝试将四个传统过程整合成更少的步骤，如前文提到的 Chen 和 Nguyen 的工作[Nguyen et.al., 2016; Chen and Ng, 2012]。从更高层面上讲，其他信息抽取任务（如实体抽取、关系抽取）也可以和事件抽取进行联合学习，在之后的研究过程中，联合抽取以避免分步噪音积累的思路一定会更加普遍。

**局部信息到全局信息：**事件抽取研究初期更多的考虑是当前词自身的特征，但研究者逐渐开始利用不同词之间的联系，从而获取更多的全局信息来完成事件抽取任务，例如 Li 等提出利用整数线性规划的方法联合抽取方法和为解决中文事件抽取中的成员缺失问题而提出联合利用句子、上下文和相关文档的中的相关事件和共指事件信息的事件抽取方法[Li et.al., 2013b]。此外还有前文提到的 Ji 等 2008 年首次提出的跨文档事件抽取中借助篇章信息和背景知识的思想[Ji and Grishman, 2008]。可以看出事件抽取考虑的信息越来越多样化和全局化。

**人工标注到半自动生成语料：**目前的语料多是英文语料，中文和其他语言的语料非常稀少。且由于事件本身的复杂程度，人工标注大量的语料十分困难。因此，越来越多的学者开始思考如何利用现有的语料迭代生成更多语料。目前主流

的解决思路是利用英文语料辅助另一种语言语料的生成，做跨语言迁移学习。另一种可能的解决思路是借鉴外部知识来自动扩展语料，例如 Chen 等研究如何基于世界知识和语言学知识大规模自动生成事件语料[Chen and Ji, 2009]。不管是哪种途径，事件抽取肯定向如何减少人工参与即可取得良好效果的方向发展。

## 4.2 事件检测和追踪的发展趋势

事件检测和追踪方面，基于 LDA 等主题模型的研究逐渐成为主流，相关研究的主要发展趋势包括两个方向：一是非参数化，放宽对话题数目的限制；二是多数据流共同建模，有效利用不同数据间的互补信息。

**非参数化：**Ahmed 等人在动态话题模型基础上，利用层次化狄利克雷过程 (Hierarchical Dirichlet Process, HDP) 放宽对话题数目的限制，提出无限动态话题模型 (iDTM)，该模型理论上允许同一个时间片内生成无限个新闻话题，方便对事件发展过程中话题产生、话题重要程度变化以及话题消亡等情况的建模 [Ahmed and Xing, 2010]。Cui 等和 Gao 等在 iDTM 基础上增加了两种话题演化行为：分裂和融合，即一个话题可以分裂成多个子话题，多个话题也可能合并为一个大话题，通过对模型的调整使其能够在保证原有功能基础上对上述两种行为进行描述，最后根据话题的生命周期以及周期内强度变化生成时序结构化摘要，并提供可视化展示[Cui et.al., 2011; Gao et.al., 2011]。

**多流交互：**Hong 等扩展 LDA 算法，同时对多个社交媒体上的媒体流，例如 Twitter 和 Yahoo，进行事件检测，其思路是利用 LDA 分别在多个数据流上检测主题，再利用主题联系各个数据流[Hong et.al., 2011]。Wang 等将不同新闻媒体流定义成协同文本流 (Coordinated Text Streams)，建模的过程中考虑流间的相互增益，通过估计给定时间点话题的出现概率来检测突发话题，最后以不同流内共同的突发话题为关键点将新闻流在时间线上对齐[Wang et.al., 2007]。Wang 等将异步文本流的对齐和话题抽取放到同一框架中，模型通过引入一个自增强过程，有效地利用异步文本的语义关联性对齐和时间信息的关联性，将时序对齐和话题建模整合为统一目标函数[Wang et.al., 2009]。除了新闻和用户生成内容的文本信息，Lin 等进一步考虑了用户间关系 (即社区信息)。他们将新闻的传播以及社区的形成 (即用户关系的建立) 形式化为一个联合推理问题，分别使用混合话题模型和高斯马尔可夫随机场 (Gaussian Markov Random Field) 刻画用户生成内容的产生和用户间影响力的变化对问题进行建模求解[Lin et.al., 2011]。针对新闻和社交媒体间跨数据流的相互依赖，Hou 等提出特定事件内新闻和用户生成内容相互影响分析问题，并分别利用基于话题距离[Hou et.al., 2015b]和格兰杰因果测试的影响发现方法[Hou et.al., 2016]结合新闻传播学、语言学和社会认知学对新闻和用户生成内容间的相互影响进行量化分析。

## 参考文献

- [Ahmed and Xing, 2010] Ahmed A, Xing E P. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA: AUAI Press, 2010. 20–29.
- [Ahn, 2006] Ahn D. The stages of event extraction. [J]. In Proceedings of the Workshop on Annotating and Reasoning About Time and Events, 2006: 1–8.
- [Allan, 2012] Allan J. Topic detection and tracking: event-based information organization[J]. Springer Science and Business Media, 2012, 12.
- [Allan et.al., 1998a] Allan J, Carbonell J G, Doddington G, et al. Topic detection and tracking pilot study final report[M]. 1998.
- [Allan et.al., 1998b] Allan J, Papka R, Lavrenko V. On-line new event detection and tracking. [J]. In Proceedings of the 21st Annual International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval, 1998: 37–45.
- [Benhardus and Kalita, 2013] J. Benhardus and J. Kalita, “Streaming trend detection in twitter,” IJWBC, vol. 9, no. 1, pp. 122–139, 2013.
- [Blei and Lafferty, 2006] D.M. Blei, J.D. Lafferty. Dynamic Topic Model [C]. Proceedings of the International Conference on Machine Learning, 2006:113-120
- [Brants et.al., 2003] Brants T, Chen F, Farahat A. A system for new event detection[C]/Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. ACM, 2003: 330-337
- [Chai and Biermann, 1998] Chai J Y, Biermann. AW. Learning and generalization in the creation of information extraction systems. [J]. Citeseer, 1998.
- [Chemero, 2000] Chemero A. What events are. [J]. Ecological Psychology, 2000, 12(1): 37–42.
- [Chen and Ji, 2009] Chen Z, Ji H. Can one language bootstrap the other: a case study on event extraction. [J]. In Proceedings of the North American Chapter of the Association for Computational Linguistics 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, 2009: 66–74.
- [Chen and Ji, 2009] Chen Z, Ji H. Language specific issue and feature exploration in Chinese event extraction[J]. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009: 209–212.
- [Chen and Ng, 2012] Chen C, Ng V. Joint modeling for Chinese event extraction with rich linguistic features[J]. In Proceedings of International Conference on Computational Linguistics, 2012: 529–544.

- [Chen et.al., 2015] Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In ACL-IJCNLP.
- [Chieu and Ng, 2002] Chieu H L, Ng H T. A maximum entropy approach to information extraction from semi-structured and free text. [J]. In Eighteenth National Conference on Artificial Intelligence, 2002:786–791.
- [Chinchor and Marsh, 1998] Chinchor N, Marsh E. Muc-7 information extraction task definition[J]. In Proceeding of the seventh message understanding conference, Appendices, 1998: 359–367.
- [Cui et.al., 2011] Cui W, Liu S, Tan L, et al. Textflow: Towards better understanding of evolving topics in text. Visualization and Computer Graphics, IEEE Transactions on, 2011, 17(12):2412–2421.
- [Ding et.al., 2013] Ding X, Qin B, Liu T. Building Chinese event type paradigm based on trigger clustering. [J]. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, 2013: 311–319.
- [Doddington et.al., 2004] 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.
- [Dong et.al., 2010] Dong Z, Dong Q, Hao C. Hownet and its computation of meaning[C]//Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. Association for Computational Linguistics, 2010: 53-56.
- [Feng et.al., 2016] Feng X, Huang L, Tang D, et al. A language-independent neural network for event detection[J]. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 66.
- [Fernando, 2007] Fernando T. Observing events and situations in time. [J]. Linguistics and Philosophy, 2007, 30(5): 527–550.
- [Gao et.al., 2011] Gao Z J, Song Y, Liu S, et al. Tracking and connecting topics via incremental hierarchical dirichlet processes. Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, BC, Canada, 2011. IEEE. 1056–1061.
- [Ge et.al., 2016a] Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Ming Zhou: Event Detection with Burst Information Networks. In COLING 2016.
- [Ge et.al., 2016b] Tao Ge, Lei Cui, Baobao Chang, Sujian Li, Ming Zhou, Zhifang Sui: News Stream Summarization using Burst Information Networks. In EMNLP 2016.
- [Glasbey, 2004] Glasbey S. Event structure, punctuality, and when. [J]. Natural language semantics, 2004, 12(2): 191–211.
- [Griffiths and Steyvers, 2004] T.L. Griffiths, M Steyvers. Finding scientific topics [C] Proceedings of the National Academic Science USA, 2004, 101(1):5228-5235

- [Guille et.al., 2014] Guille A, Favre C. Mention-anomaly-based event detection and tracking in twitter[C]//Advances in Social Networks Analysis and Mining, 2014 IEEE/ACM International Conference on. IEEE, 2014: 375-382.
- [Hall et.al., 2008] D. Hall, D Jurafsky, C.D. Manning. Studying the History of Ideas Using Topic Models [C]. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008:363-371
- [Hong et.al., 2007] Hong Y, Zhang Y, Liu T, et al. Topic detection and tracking review[J]. Journal of Chinese information processing, 2007, 6(21): 77–79.
- [Hong et.al., 2011] Hong L, Dom B, Gurumurthy S, et al. A time-dependent topic model for multiple text streams. Proceedings of the 17th ACM International Conference on Knowledge Discovery in Data Mining, California, USA, 2011. 832–840.
- [Hong et.al., 2011] Hong Y, Zhang J, Ma B, et al. Using cross-entity inference to improve event extraction[J]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, 1: 1127–1136.
- [Hou et.al., 2015a] Hou L, Li J, Wang Z, et al. Newsminer: multifaceted news analysis for event search[J]. Knowledge-Based Systems, 2015, 76: 17-29.
- [Hou et.al., 2015b] Hou L, Li J, Li X L, et al. Measuring the influence from user-generated content to news via cross-dependence topic modeling[C]//International Conference on Database Systems for Advanced Applications. Springer International Publishing, 2015: 125-141.
- [Hou et.al., 2016] Hou L, Li J, Li X L, et al. Detecting Public Influence on News Using Topic-Aware Dynamic Granger Test[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, 2016: 331-346.
- [Hoxha et.al., 2016] Hoxha K, Baxhaku A, Ninka I. Bootstrapping an Online News Knowledge Base. International Conference on Web Engineering. Springer International Publishing, 2016: 501-506.
- [Hsi et.al., 2016] Hsi A, Yang Y, Carbonell J, et al. Leveraging multilingual training for limited resource event extraction. [J]. In Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, 2016: 1201–1210.
- [Hu et.al., 2011] Hu P, Huang M, Xu P, et al. Generating breakpoint-based timeline overview for news topic retrospection. Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, BC, Canada: IEEE, 2011. 260–269.
- [Hu et.al., 2015a] Linmei Hu, Juanzi Li, Xiaoli Li, Chao Shao, Xuzhong Wang. TSDPMM: Incorporating Prior Topic Knowledge into Dirichlet Process Mixture Models for Text Clustering. EMNLP 2015: 787-792.
- [Hu et.al., 2015b] Linmei Hu, Chao Shao, Juanzi Li, Heng Ji. Incremental learning from news events. Knowledge-Based System. 89: 618-626 (2015).

- [Hu et.al., 2015c] Linmei Hu, Xuzhong Wang, Mengdi Zhang, Juan-Zi Li, Xiaoli Li, Chao Shao, Jie Tang, Yongbin Liu. Learning Topic Hierarchies for Wikipedia Categories. ACL (2) 2015: 346–351.
- [Ji, 2009] Ji H. Cross-lingual predicate cluster acquisition to improve bilingual event extraction by inductive learning[J]. In Proceedings of the Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics, 2009: 27–35.
- [Ji and Grishman, 2008] Ji H, Grishman R. Refining event extraction through unsupervised Cross-document inference[C]. Proceedings of ACL-08 . HLT Columbus, USA. HLT, 2008: 254-262.
- [Jiang, 2005] Jifa J. An event ie pattern acquisition method. [J]. Computer Engineering, 2005, 31(15): 96–98.
- [Kim and Moldovan, 1995] Kim J, Moldovan D I. Acquisition of linguistic patterns for knowledge-based information extraction. [J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(5): 713–724.
- [Kumaran and Allan, 2004] Kumaran G, and Allan J. Text Classification and Named Entities for New Event Detection. In Proceedings of the 27th Annual International ACM SIGIR Conference, New York, NY, USA. ACM Press, 297–304. 2004.
- [Kuzey et.al., 2014] E. Kuzey, J. Vreeken, G. Weikum, a fresh look on knowledge bases: Distilling named events from news, In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, 2014, pp. 1689–1698.
- [Liang and Wu, 2006] Liang H, Wu P. Information extraction system based on event frame. [J]. Journal of Chinese Information Processing, 2006, 20(2): 40–46.
- [Liao and Grishman, 2010] Liao S, Grishman R. Using document level cross-event inference to improve event extraction. [J]. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010: 789–797.
- [Liao and Grishman, 2011a] Liao S, Grishman R. Can document selection help semi-supervised learning? a case study on event extraction[J]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, 2: 260–265.
- [Liao and Grishman, 2011b] Liao S, Grishman R. Using prediction from sentential scope to build a pseudo co-testing learner for event extraction. [J]. In Proceedings of 5th International Joint Conference on Natural Language Processing, 2011: 714–722.
- [Li et.al., 2012a] Li P, Zhu Q, Diao H, et al. Joint modeling of trigger identification and event type determination in Chinese event extraction. [J]. In Proceedings of the International Conference on Computational Linguistics, 2012: 1635–1652.
- [Li et.al., 2012b] Li P, Zhou G, Zhu Q, et al. Employing compositional semantics and discourse consistency in Chinese event extraction[J]. In Proceedings of the 2012

- Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 1006–1016.
- [Li et.al., 2013a] Li P, Zhu Q, Zhou G. Joint modeling of argument identification and role determination in Chinese event extraction with discourse-level information. [J]. In Proceedings of International Joint Conference on Artificial Intelligence, 2013: 612–622.
- [Li et.al., 2013b] Li P, Zhu Q, Zhou G. Argument inference from relevant event mentions in Chinese argument extraction[J]. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013: 1477–1487.
- [Li et.al., 2014] Li Q, Ji H, Hong Y, et al. Constructing information networks using one single model[J]. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014: 1846–1851.
- [Lin et.al., 2011] Lin C X, Mei Q, Han J, et al. The joint inference of topic diffusion and evolution in social communities. Proceedings of the 11th IEEE International Conference on Data Mining, Vancouver, BC, Canada, 2011. IEEE. 378–387.
- [Liu et.al., 2016a] Shulin Liu, Kang Liu and Jun Zhao, A Probabilistic Soft Logic Based Approach to Exploit Latent and Global Information in Event Classification, AAAI 2016
- [Liu et.al., 2016b] Liu S, Chen Y, He S, et al. Leveraging framenet to improve automatic event detection[C]/Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016, 1: 2134-2143.
- [Ludwig, 2001] Ludwig Wittgenstein: Tractatus logico-philosophicus[M]. Oldenbourg Verlag, 2001.
- [Mei and Zhai, 2005] Mei Q, Zhai C. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. Proceedings of the 11th ACM International Conference on Knowledge Discovery in Data Mining, Chicago, Illinois, USA, 2005. ACM. 198–207.
- [Meng, 2015] Meng J. Research on event extraction technology in the field of unexpected events. [J]. Master's thesis, Shanghai University, 2015.
- [Nguyen and Grishman, 2015] Nguyen H T, Grishman. R. Event detection and domain adaptation with convolutional neural networks[J]. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2015: 365–371.
- [Nguyen et.al., 2016] Nguyen T H, Cho K, Grishman R. Joint event extraction via recurrent neural networks. [J]. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 300–309.
- [Piskorski et.al., 2001] Piskorski J, Tanev H, Atkinson M, et al. Online news event extraction for global crisis surveillance [J]. In Transactions on computational collective intelligence V, 2001: 182–212.

- [Pustejovsky et.al., 2003] Pustejovsky J, Hanks P, Sauri R, et al. The timebank corpus. [J]. In Proceedings of the Corpus linguistics, 2003: 40–49.
- [Quine, 1985] Quine W V O. Events and reification. [J]. Actions and events: Perspectives on the philosophy of Donald Davidson, 1985: 162–171.
- [Riloff, 1993] Riloff E. Automatically constructing a dictionary for information extraction tasks[C]//AAAI. 1993: 811-816.
- [Riloff and Shoen, 1995] Riloff E, Shoen J. Automatically acquiring conceptual patterns without an annotated corpus. [J]. In Proceedings of the Third Workshop on Very Large Corpora, 1995, 3.
- [Roger, 1978] Wilensky R. Understanding goal-based stories. [J]. Technical report, DTIC Document, 1978.
- [Rospocher et.al., 2016] Rospocher M, van Erp M, Vossen P, et al. Building event-centric knowledge graphs from news[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2016, 37: 132-151.
- [Rouces et.al., 2015] Rouces J, de Melo G, Hose K. Framebase: Representing n-ary relations using semantic frames[C]. European Semantic Web Conference. Springer International Publishing, 2015: 505-521.
- [Rupnik et.al., 2016] Rupnik J, Muhic A, Leban G, et al. News across languages-cross-lingual document similarity and event tracking[J]. Journal of Artificial Intelligence Research, 2016, 55: 283-316.
- [Shamma et.al., 2011] D.A. Shamma, L. Kennedy, and E.F. Churchill, “Peaks and persistence: modeling the shape of microblog conversations,” in CSCW, 2011, pp. 355–358.
- [Tanev et.al., 2008] Tanev H, Piskorski J, Atkinson M. Real-time news event extraction for global crisis monitoring. [J]. In Proceedings of the International Conference on Application of Natural Language to Information Systems, 2008: 207–218.
- [Tao et.al., 2013] Tao F, Lei K H, Han J, et al. Eventcube: multi-dimensional search and mining of structured and text data. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013: 1494-1497.
- [Trabasso, 1985] Trabasso T, Broek P V D. Causal thinking and the representation of narrative events. [J]. Journal of memory and language, 1985, 24(5): 612–630.
- [Wang and Zhao, 2012] 王伟, 赵东岩. 中文新闻事件本体建模与自动扩充[J]. 计算机工程与科学, 2012, 34(4): 171-176.
- [Wang et.al., 2007] Wang X, Zhai C, Hu X, et al. Mining correlated bursty topic patterns from coordinated text streams. Proceedings of the 13th ACM International Conference on Knowledge Discovery in Data Mining, San Jose, California, USA, 2007. ACM. 784–793.
- [Wang et.al., 2009] Wang X, Zhang K, Jin X, et al. Mining common topics from multiple asynchronous text streams. Proceedings of the 2nd ACM International

- Conference on Web Search and Data Mining, Barcelona, Spain, 2009. ACM. 192–201.
- [Xu et.al., 2016] 徐佳俊, 杨 飚, 姚天昉, 付中阳: 基于 LDA 模型的论坛热点话题识别和追踪. In Journal of Chinese information proceeding Vol. 30. No 1. 2016
- [Yangarber et.al., 2000] Yangarber R, Grishman R, Tapanainen P, et al. Automatic acquisition of domain knowledge for information extraction[C]//Proceedings of the 18th conference on Computational Linguistics-Volume 2. Association for Computational Linguistics, 2000: 940-946.
- [Yang et.al., 1998] Yang Y, Pierce T, Carbonell J. A study of retrospective and on-line event detection. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 1998. 28–36.
- [Yang et.al., 1999] Yang Y, Carbonell J G, Brown R D, et al. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems and Their Applications, 1999, 14(4):32–43.
- [Yang et.al., 2002] Yang Y, Zhang J, Carbonell J, et al. Topic-conditioned novelty detection[C]//Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2002: 688-693.
- [Zacks, 2001] Zacks J M, Tversky B. Event structure in perception and conception. [J]. Psychological bulletin, 2001, 127(1): 3.
- [Zhang, 2006] 张阔. 新闻挖掘关键技术研究. 博士学位论文, 2006
- [Zhou, 2003] Zhou J. Research on financial event extraction technology-based on automatic rule acquisition [J]. Master's thesis, Tsinghua University, 2003.
- [Zhu et.al., 2014] Zhu Z, Li S, Zhou G, et al. Bilingual event extraction: a case study on trigger type determination. [J]. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 842–847.
- [Zwaan, 1999] Zwaan R. Five dimensions of narrative comprehension: The event-indexing model. [J]. Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso, 1999, 5(1): 93–110.

## 第六章 知识存储与查询

### 1. 任务定义、目标和研究意义

简单地说，“知识图谱”就是以图（Graph）的方式来展现“实体”、实体“属性”，以及实体之间的“关系”。目前知识图谱普遍采用了语义网框架中 RDF[W3C, 2014]（Resource Description Framework，资源描述框架）模型来表示数据。语义网是万维网之父蒂姆·伯纳斯-李（Tim Berners-Lee）在 1998 年提出的概念 [Wikipedia, 2018]，其核心是构建以数据为中心的网络，即 Web of Data；这是相对于我们目前的万维网是 Web of Pages 而提出的。语义网的核心是让计算机能够理解文档中的数据，以及数据和数据之间的语义关联关系，从而使得机器可以更加智能化地处理这些信息。因此我们可以把语义网想象成是一个全球性的数据库系统，也就是我们通常所提到的 Web of Data。本报告将从数据管理的角度去介绍在知识存储和查询方面的研究和应用问题。

RDF 是用于描述现实中资源的 W3C 标准。它被设计为提供一种描述信息的通用方法，这样就可以被计算机应用程序读取并理解。现实中任何实体都可以表示成 RDF 模型中的资源，比如图书的标题、作者、修改日期、内容以及版权信息。资源以唯一的 URI（统一资源标识——Uniform Resource Identifiers，通常使用的 URL 是它的一个子集）来表示，不同的资源拥有不同的 URI。这些资源可以用来作为知识图谱中对客观世界的概念、实体和事件的抽象。

图 1 给出了一个 RDF 数据实体示例，用来表示现实中一个著名欧洲哲学家亚里士多德（Aristotle）。在 RDF 数据模型中，亚里士多德就能通过亚里士多德头像上方所示的 URI 来进行唯一标识。客观世界中的概念、实体和事件很多都是有属性。图 1 中亚里士多德头像下方给出的属性和属性值描述了亚里士多德这个资源所对应的人的名字是“亚里士多德”。此外，客观世界中不同概念、实体和事件相互之间可能会有各种关系，所以 RDF 模型中不同资源之间也是会存在关系。比如，图 1 给出了亚里士多德和另一个表示希腊城市卡尔基斯（Chalcis）所对应的资源通过一个 placeOfDeath 关系连接了起来，描述了亚里士多德死于卡尔基斯这个事实。

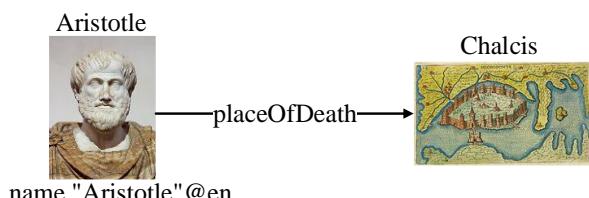


图 1. 示例 RDF 资源

利用这些属性和关系，很多资源就被连接起来形成了 RDF 数据集。每个资源的一个属性及属性值，或者它与其他资源的一条关系，都被称为一条知识。上述属性以及关系就能表示成三元组。每一条三元组又可被称为一条陈述。一条陈述包含三个部分，通常称之为“主体”、“谓词”和“宾语”。其中“主体”一定是一个被描述的资源。“谓词”可以表示“主体”的属性，或者表示“主体”和“宾语”之间某种关系。当“谓词”表示属性时，“宾语”就是属性值，通常是一个字面值；否则，“宾语”是另外一个资源。

图 2 的展示了一个著名 RDF 数据集 DBpedia[Lehmann J, et al., 2015] 的片段。这个片段中包括 15 条陈述，描述了欧洲哲学家 Aristotle(亚里士多德)和 Boethius(波伊提乌)所对应的资源及其相关陈述。

主语	谓词	宾语
Aristotle	influencedBy	Plato
Aristotle	mainInterest	Ethics
Aristotle	name	"Aristotle"
Aristotle	placeOfDeath	Chalcis
Aristotle	mainInterest	Physics
Boethius	influencedBy	Aristotle
Boethius	mainInterest	Religion
Boethius	name	"Boethius"
Boethius	placeOfDeath	Pavia
Plato	name	"Plato"
Chalcis	imageSkyline	Chalkida.JPG
Chalcis	country	Greece
Chalcis	postalCode	341 00
Pavia	country	Italy
Pavia	postalCode	27100

图 2. 示例 RDF 三元组

面向 RDF 数据集，W3C 提出了一种结构化查询语言 SPARQL[W3C, 2013]；它类似于面向关系数据库的查询语言 SQL。和 SQL 一样，SPARQL 也是一种描述性的结构化查询语言，即用户只需要按照 SPARQL 定义的语法规则去描述其想查询的信息即可，不需要明确指定如何进行查询的计算机的实现步骤。2008 年 1 月，SPARQL 成为 W3C 的正式标准。对于一个 SELECT 语句中，SELECT 子句指定查询应当返回的内容，FROM 子句指定将要使用的数据集，WHERE 子句由一组三元模式组成，以指定所返回的 RDF 知识图谱数据片段需要满足的模式。

图 3(a)给出了一个针对哲学家的 SPARQL 查询，目标在于查询出所有“受过亚里士多德影响的伦理学相关的哲学家”。这个查询在图 2 所示 RDF 数据集上所对应的匹配如图 3(b)所示，即“受过亚里士多德影响的伦理学相关的哲学家”有波伊提乌(Boethius)。

```
SELECT ?x ?n WHERE {  
?x mainInterest Ethics.  
?x influencedBy Aristotle.  
?x name ?n.  
}
```



?x → Boethius  
?n → "Boethius"

图 3. 示例基本图模式查询

我们也可以将 RDF 数据分别表示成图的形式。在这个图中，每个 RDF 资源或者 RDF 数据集中出现过的字符串可以被视为图上的点，每个三元组可以视为连接主体及客体的有向边，而三元组中的谓词就可以视为有向边上的标签。从语义角度上看，RDF 数据本质上就是通过预先定义的语义构成的一个或多个连通图。V.Bönström 等人提出[Bonstrom V, et al., 2003]，相比于将 RDF 数据视为 XML 格式数据或三元组的集合，RDF 的图模型包含了 RDF 数据中涵盖的语义信息。

图 4 展示了图 2 所示 RDF 知识图谱数据集所对应的 RDF 数据图。图 4 中所有的资源都是椭圆，而文本点都是矩形点。

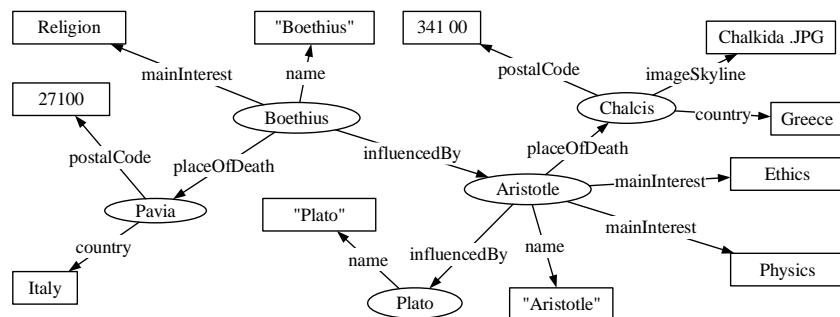


图 4. 示例 RDF 数据图

与 RDF 数据的图形式表示类似，一个 SPARQL 查询可以表示为一个查询图。查询中每个变量或者常量对应一个查询图上的点，每个 WHERE 子句中的三元模式对应一条边。图 5 给出了一个图 3(a)所示基本图模式查询所对应的查询图，用以查询 RDF 数据图上所有“受过亚里士多德影响的伦理学相关的哲学家”。

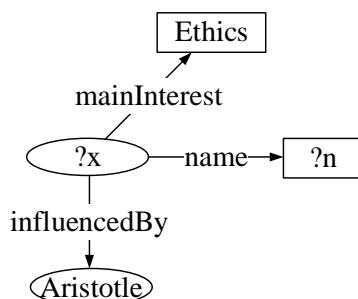


图 5. 示例 SPARQL 查询图

现有 RDF 数据存储与查询的基本问题就是：给定一个 RDF 数据集  $G$  和 SPARQL 查询  $Q$ ，找出  $Q$  在  $G$  上的匹配。当 RDF 数据和 SPARQL 查询都转化成图的形式，SPARQL 查询语句的查询结果就是其所对应的查询图在 RDF 数据图上的子图匹配[Zou L, et al., 2011, Zou L, et al., 2014]。

## 2. 研究内容和挑战问题

由于 RDF 结构的灵活性，现在 RDF 数据的应用范围日益广阔。越来越多的数据开始表示成 RDF 格式。比如，德国的莱比锡大学和柏林自由大学合作从维基百科上抽取结构化数据形成的知识库 DBpedia[4]已将近有 24 亿三元组；另外，德国 Max Planck 实验室从 Wikipedia 上抽取出的信息并结合 wordnet 中类型信息形成的 RDF 知识库 YAGO[Suchanek F, et al., 2008, Hoffart J, et al., 2013, Mahdisoltani F, et al., 2015] 也有将近 2 亿三元组；而由用户编辑和维基百科中信息抽取共同形成的 RDF 知识库 Freebase [Google, 2016] 也有 19 亿三元组。

因此 RDF 数据管理的一个核心问题是如何有效地存储和查询上述大规模的 RDF 数据集。在查询处理过程中，我们需要将 SPARQL 查询图中变量与 RDF 数据图上点进行绑定以得到所有 SPARQL 查询图在 RDF 数据图上的子图匹配。学术界和工业界当前已经构建了不少高效的 RDF 数据管理系统来进行 SPARQL 查询处理。

## 3. 技术方法和研究现状

知识图谱数据管理的一个核心问题是如何有效地存储和查询 RDF 数据集。总的来说，有两套完全不同的思路。其一是我们可以利用已有的关系数据模型来存储和管理知识图谱数据，将面向 RDF 数据的 SPARQL 查询转换为面向关系数据库的 SQL 查询，利用已有的关系数据库产品或者相关技术来回答查询。这里面最核心的研究问题是如何构建关系表来存储 RDF 数据，并且使得转换的 SQL 查询语句查询性能更高；其二是直接开发面向 RDF 知识图谱数据的 Native 的知识图谱数据存储和查询系统（Native RDF 图数据库系统），考虑到 RDF 数据管理的特性，从数据库系统的底层进行优化。针对以上两个方面的思路，我们分别加以介绍。

### 3.1 基于关系数据模型的 RDF 数据存储和查询

在数据管理方面，关系数据模型自提出以来取得了巨大成功。市面上已经产生了大量成熟的关系数据库。而 RDF 数据的三元组模型可以很容易完成对于关系模型的映射。因此，不少研究者都尝试使用关系数据模型来设计 RDF 存储方案。下面介绍几种经典的方法。

### 3.1.1. 简单三列表

现在已经有不少比较成熟的系统来利用关系数据库进行数据管理，包括 Jena[Wilkinson K, et al., 2003, Wilkinson K, et al., 2006]、Oracle[Inseok E, et al., 2005]、Sesame[Broekstra J, et al., 2003, Broekstra J, et al., 2002]、3store[Harris S, et al., 2003]以及 SOR[Lu J, et al., 2007]。这些系统通过维护一张巨大的三元组表来管理 RDF 数据。这张三元组表包含三列。这三列分别对应存储主体、谓词和客体（或者主体、属性和属性值）。当系统接收到用户输入的 SPARQL 查询时，这些系统将 SPARQL 查询转化为 SQL 查询。然后，根据所得 SQL 查询，这些系统通过对三元组表执行多次自连接操作以得到最终解。

虽然这种方法具有很好的通用性，但最大的问题是查询性能差。首先这张三列表的规模可能非常庞大。而且这种方法可能会产生大量的自连接操作，而在关系数据库系统中自连接操作非常耗时，特别是对于那些数据规模很大的表而言。所以这些方法都有很大的局限。

### 3.1.2 水平存储

所谓的水平方法（Horizontal Schema）[Pan Z, et al., 2003]是将知识图谱中的每一个 RDF 主体（subject）表示为数据库表中的一行。表中的列包括该 RDF 数据集合中所有的属性。这种的策略的好处在于设计简单，同时很容易回答面向某单个主体的属性值的查询，即星状查询，如图 6 所示。

Subject	name	mainInterest	mainInterest	placeOfDeath	influencedBy	postalCode	country	imageSkyline
Aristotle	"Aristotle"	Ethics	Physics	Chalcis	Plato			
Boethius	"Boethius"	Religion		Pavia	Aristotle			
Chalcis						341 00	Greece	Chalkida.JPG
Pavia						27100	Italy	
Plato	"Plato"							

图 6. 水平存储

根据图 6 表的结构，为了回答图 3 中的 SPARQL 查询，可以转换为下面的 SQL 语句。显然下面的 SQL 语句没有耗时的连接操作，因此其查询效率比较高。

```
SELECT name from T WHERE
    mainInterest = "Ethics" and
    influenceBy = "Aristotle".
```

图 7. 水平存储上的 SQL 查询

然而这种水平存储方法的缺点也是很明显的：其一，表中存在大量的列。一般来讲属性数目会比主体和属性值的个数少很多，但是还是有可能超过当前数据库能够承受的数量。其二，表的稀疏性问题。通常一个主体并不在所有的属性上有值。相反，主体仅仅在极少量的属性上有值。然而由于一个主体存成一行，那么表中将存在大量空值。空值不仅增加了存储负载，而且带来了其他的问题，比

如增大了索引大小，影响查询效率。其三，水平存储存在多值性的问题。一个表中列的数量是固定的，这就使得主体在一个属性上只能有一个值。而真实数据往往并不符合这个限制条件。其四，数据的变化可能带来很大的更新成本。在实际应用中，数据的更新可能导致增加属性或删除属性等改变，但是这就涉及到整个表结构的变化，水平结构很难处理类似的问题。

### 3.1.3 属性表

为了降低自连接操作次数，Jena[Wilkinson K, et al., 2003, Wilkinson K, et al., 2006] 和 Oracle[Inseok E, et al., 2005]在单张大三元组表之外还支持利用属性表进行 RDF 数据管理。具体而言，Jena 通过聚类的方式将一些类似的三元组聚类到一起，然后将每一个聚类的三元组统一到一张属性表中进行管理，这种方式下的属性表也被称之为聚类属性表；而 Oracle 利用 RDF 资源的类型信息将三元组进行分类，相同类的三元组放到同一张表中，这种方式下的属性表也被称之为分类属性表。对于上述两种情况，由于 RDF 数据表示的灵活性，会存在部分三元组无法放入任何一个属性表示。此时，Jena 和 Oracle 将这个三元组另起一张表来进行管理。同时，并不是属于某个属性表的每个资源在各个属性上都有值，所以属性表中可能存在若干位空的位置。图 8 显示了图 2 所示大表下的 RDF 数据在聚类属性表和分类属性表下不同的情况。

Subject	name	mainInterest	mainInterest	placeOfDeath	influencedBy
Aristotle	"Aristotle"	Ethics	Physics	Chalcis	Plato
Boethius	"Boethius"	Religion		Pavia	Aristotle

Subject	postalCode	country	imageSkyline
Chalcis	341 00	Greece	Chalkida.JPG
Pavia	27100	Italy	

Subject	name
Plato	"Plato"

聚类属性表

Subject	name	mainInterest	mainInterest	placeOfDeath	influencedBy
Aristotle	"Aristotle"	Ethics	Physics	Chalcis	Plato
Boethius	"Boethius"	Religion		Pavia	Aristotle
Plato	"Plato"				

Subject	postalCode	country	imageSkyline
Chalcis	341 00	Greece	Chalkida.JPG
Pavia	27100	Italy	

分类属性表

图 8. 属性表示例

属性表也有着一些先天性的缺陷。其一，虽然属性表对于某些查询能够提高查询性能，但是大部分的查询都会涉及多个表的连接或合并操作。对聚类属性表

而言,如果查询中属性作为变量出现,则会涉及多个属性表;对属性分类表而言,如果查询并未确定属性类别,则查询会涉及多个属性表。在这种情况下,属性表的优点就较不明显了。其二,RDF 数据由于来源庞杂,其结构性可能较差,从而属性和主体间的关联性可能并不强,类似的主体可能并不包含相同的属性。这时,空值的问题就出现了。数据的结构性越差,空值的问题就越发明显。其三,在现实中,一个主体在一个属性上可能存在多值。这时,用 RDBMS 管理这些数据时就带来麻烦。其中,前两个问题是相互影响的。当一个表的列数目减小时,对结构性要求较低,空值问题得到缓解,但查询会涉及更多的表;而当表的列数加大时,如果数据结构性不强,就会出现更多空值的问题。

### 3.1.4 垂直划分策略

针对属性表的问题,SW-Store[Abadi D, et al., 2009]提出了对 RDF 数据按照谓词(或属性)分割成若干表的方法。具体而言,SW-Store 将 RDF 三元组按照谓词(或属性)的不同分成不同的表,每张表能保存在谓词(或属性)上相同的三元组。SW-Store 称这种方法为垂直分割。这种方法的优势在于能避免大量的自连接操作,而变成不同表之间的连接。因为在现有的关系数据库中不同表之间的连接操作要快于自连接操作,所以 SW-Store 能一定程度提高效率。但是,垂直分割缺点在于无法很好地支持 SPARQL 查询中某个三元组模式在谓词(或属性)上是变量的情况。

### 3.1.5 全索引策略

如前所述,简单的三列表存储的缺点在于自连接次数较多。为了提高简单三列表存储的查询效率,目前一种普遍被认可的方法是“全索引(exhaustive indexing)”策略。如 Hexastore [Weiss C, et al., 2008] 和 RDF-3x[Neumann T, et al., 2008, Neumann T, et al., 2010a, Neumann T, et al., 2010b]。

为了加速 RDF 三元组在 SPARQL 查询处理过程中的连接操作速度,Hexastore [Weiss C, et al., 2008] 和 RDF-3x[Neumann T, et al., 2008, Neumann T, et al., 2010a, Neumann T, et al., 2010b] 都将三元组在主体、属性、客体之间各种排列下能形成各种形态构建都枚举出来,然后为它们构建索引。这样建立的索引恰好是六重索引。比如图 2 中三元组<Aristotle, placeOfDeath, Chalcis>,在这 Hexastore 和 RDF-3x 中都保存了六份,分别是<Aristotle, placeOfDeath, Chalcis>、<Aristotle, Chalcis, placeOfDeath>、<Chalcis, placeOfDeath, Aristotle>、<Chalcis, Aristotle, placeOfDeath>、<placeOfDeath, Aristotle, Chalcis> 和<placeOfDeath, Chalcis, Aristotle>。这些索引内容正好对应 SPARQL 查询中带变量三元组模式的各种可能,于是就能很好支持 SPARQL 查询。比如,<Chalcis, placeOfDeath, Aristotle>就可以很好地支持 SPARQL 中变量在主体位置的三元组。

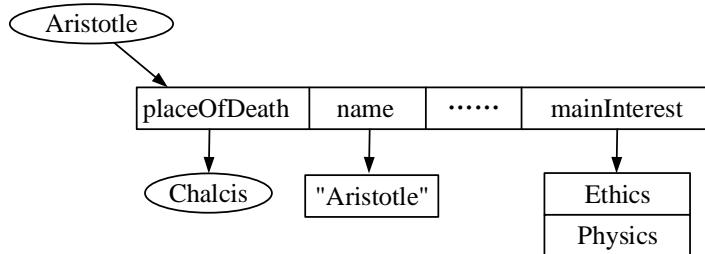


图 9. Hexastore 数据组织的示例

虽然用全索引策略可以弥补一些简单垂直存储的缺点,但三元存储方式难以解决的问题还有很多。其一,不同的三元组其主体/属性/属性值可能重复,这样的重复出现会浪费存储空间。其二,复杂的查询需要进行大量表连接操作,即使精心设计的索引可以将连接操作都转化为合并连接,当 SPARQL 查询复杂时,其连接操作的查询代价依然不可忽略。其三,随着数据量增长,表的规模会不断膨胀,系统的性能下降严重;而且目前此类系统都无法支持分布式的存储和查询,这限制了其系统的可扩展性。其四,由于数据类型多样,无法根据特定数据类型进行存储的优化,可能会造成存储空间的浪费(例如,客体的值可能多种多样,如 URI、一般字符串或数值。客体一栏的存储空间必须满足所有的取值,而无法进行存储优化)。为了解决这个问题,目前的全索引方法都是利用字典方式将所有的字符串和数值映射成一个独立的整数 ID。但是这种字典映射的方法很难支持带有数值范围约束和字符串中的子串约束的 SPARQL 查询。

### 3.2 基于图模型的 RDF 数据存储和查询

如前文所述,通过将 RDF 三元组看作带标签的边,RDF 数据自然的符合图模型结构。因此,很多研究者从 RDF 图模型结构的角度看待 RDF 数据。RDF 数据的图模型可以最大限度的保持 RDF 数据的语义信息,也有利于对语义信息的查询。在这种情况下,SPARQL 查询就可以视为在 RDF 数据图上进行子图匹配运算。子图匹配运算是图数据库中一个比较经典的问题。其问题定义在于给定一个数据图和一个查询图,找出数据上所有与查询图子图同态的位置。这个问题已被证明是一个 NP 难问题。

针对 RDF 数据的 SPARQL 查询已经有一些基于图模型的查询处理系统,如 gStore[Zou L, et al., 2011, Zou L, et al., 2014]、dipLODocus[RDF][Wylot M, et al., 2011]和 TurboHOM++[Kim J, et al., 2015]。它们都是利用 RDF 数据图的特点来构建索引。

gStore[Zou L, et al., 2011, Zou L, et al., 2014]是由北京大学计算机科学技术研究所数据管理实验室实现并维护的一个基于图的 RDF 知识图谱数据管理系统<sup>4</sup>。gStore 根据每个资源的所有属性和属性值将其映射到一个二进制位串上。图 10

<sup>4</sup> gStore 项目主页 <http://www.gstore-pku.com/>

显示 gStore 对一个 RDF 数据图进行二进制编码的示例。然后，gStore 将所有位串按照 RDF 背后对应的图结构组织成一棵签章树——VS\*-tree。VS\*-tree 被分为若干层，每一层都是整张 RDF 数据图的摘要。基于 VS\*-tree，gStore 可以完成高效的数据存储、更新与查询操作。当 SPARQL 查询进入时，将每个查询点在这个 VSTree 上进行检索，找到相应候选解，然后再将这些候选解通过连接操作拼接起来。

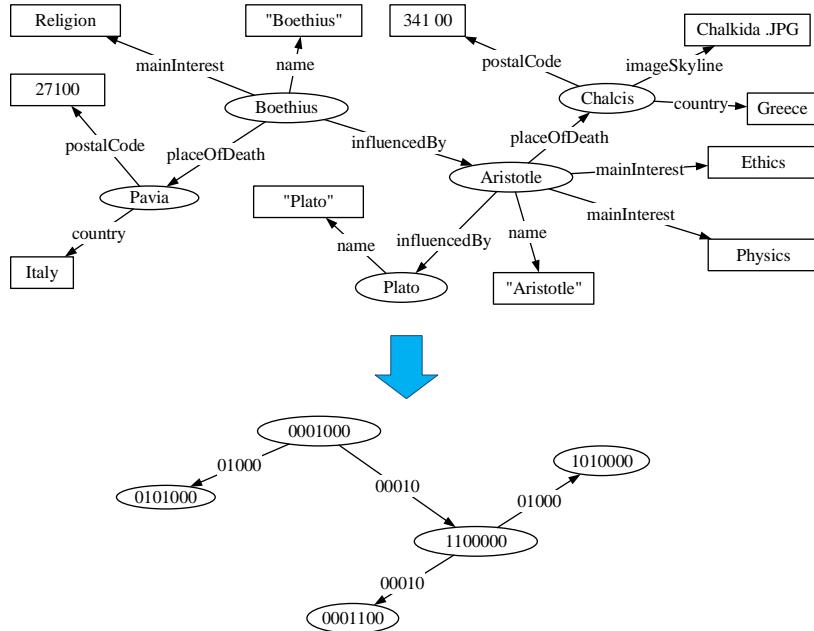


图 10. gStore 进行二进制编码的示例

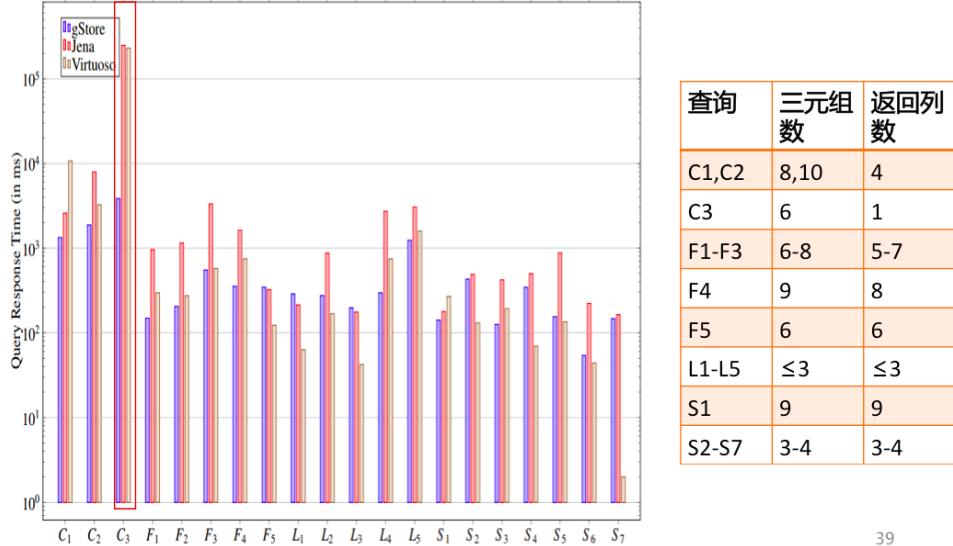
图 11 展示了在 3-5 亿规模的三元组的国际标准测试集（LUBM 和 WatDiv）上，gStore 系统和目前使用最为广泛的 RDF 知识图谱存储查询系统 Virtuoso 和 Apache Jena 之间的查询性能对比情况。由于基于图结构方法的索引可以考虑到查询图整体信息，因此总的来说查询图越复杂（例如查询图的边越多），gStore 相对于对比系统的性能会更好，有的可以达到一个数量级以上的性能优势。gStore 的分布式版本的在 10 台机器组成的 Cluster 上可以进行 50-100 亿规模的 RDF 知识图谱管理的任务。

dipLODocus[RDF] [Wylot M, et al., 2011]提出一个同时利用 RDF 图结构与考虑数据分析需求的混合存储模式。所谓利用 RDF 数据图结构，就是挖掘出 RDF 图中的若干存储模式，然后将 RDF 数据图中满足这些存储模式的结构存在一起。所谓考虑数据分析需求，就是利用列存储技术存储数值型数据，即将满足某个存储模式的所有结构中特定位置的数值按列存储组织在一起以方便聚集性查询处理。

TurboHOM++[ Kim J, et al., 2015]将子图匹配的技术应用到了 SPARQL 查询处理上。具体而言，TurboHOM++首先将 RDF 数据图基于每个资源的类信息转化为一般的普通数据图。然后，TurboHOM++在 SPARQL 查询图上从一个选定查

询点出发做宽度优先搜索，得到一颗树宽度优先搜索树。同时，TurboHOM++在数据图上从选定查询点候选出发结合宽度优先搜索树深度做深度优先搜索，得到选定查询点候选的候选区域，并在这个候选区域中结合一定匹配顺序找到最终 SPARQL 查询的解。

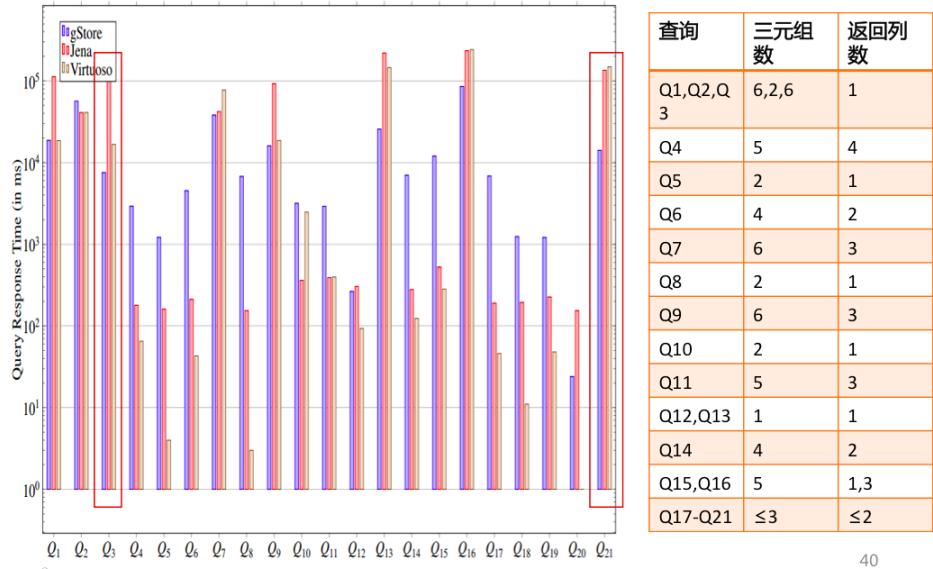
Query Performance over WatDiv 300M



39

(a) 在 WatDiv 3 亿三元组规模数据上的评测结果

Query Performance over LUBM 500M



40

(b) 在 LUBM 5 亿三元组规模数据上的评测结果

图 11. 在国际通用 RDF 测评数据集上的比较结果

#### 4. 技术展望与发展趋势

因为 RDF 模型的灵活性，越来越多的知识图谱数据提供方将自身的知识图

谱数据表示成 RDF 格式并发布到互联网上。这些发布在互联网上的的 RDF 数据之间通过 URI 相互链接起来，共同构成了一个庞大的覆盖整个互联网的知识图谱。这个庞大的覆盖整个互联网的知识图谱描述了整个互联网上的知识。这样，互联网就由一个文档的网络转化成一个数据的网络，而且是一个计算机可以理解的数据网络。为了让这个数据的网络更加的丰富和完善，W3C 在积极推进 LOD（Linked Open Data）项目[Heath T, 2018]。这个项目目的就是将网络上的 RDF 数据集相互链接起来以增强数据可用性。当前，LOD 已成功令数百个 RDF 数据集相互链接在一起。图 12 展示了 LOD 项目中相互连接的数据集。

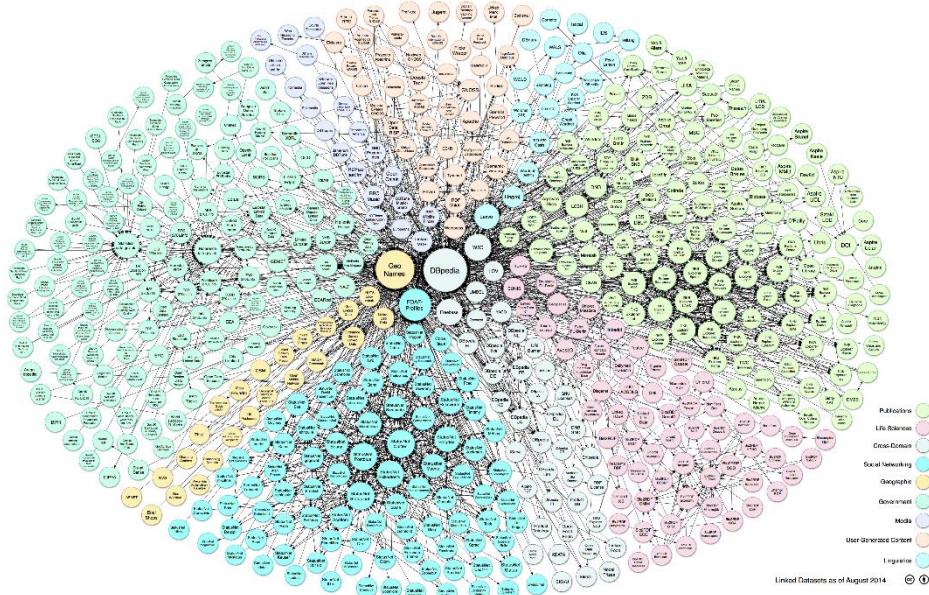


图 12. 关联数据

随着互联网上 RDF 知识图谱数据集的数量与规模日益增长，互联网上的 RDF 知识图谱数据集已经远超出了现有单机系统能力。于是，利用分布式数据库系统相关技术来进行 RDF 数据上的查询处理成为了未来研究的趋势。

现阶段，部分研究人员也已经有设计并实现了一些针对 RDF 数据的分布式查询处理方法。这些方法可以被分为三类：一类是基于已有云平台的分布式查询处理方法；一类是基于数据划分的分布式查询处理方法；还有一类是联邦型分布式 RDF 数据查询处理方法。

#### 4.1 基于已有云平台的分布式 RDF 数据查询处理方法

所谓基于已有云平台的分布式 RDF 数据查询处理方法都是利用已有云平台存储管理系统进行关联数据的存储，并利用这些已有云平台上成熟的任务处理模式进行查询处理。现有被用来进行查询处理的云平台系统包括 Hadoop[Shvachko K, et al., 2010]、Trinity[Shao B, et al., 2013]等。

因为 Hadoop[Shvachko K, et al., 2010]是目前最受欢迎的云计算平台，所以很多研究人员在研究如何利用 Hadoop 进行 RDF 数据上的查询处理。在进行数据

预处理的时候，现有基于 Hadoop 的 RDF 数据上的分布式查询处理方法将 RDF 数据转化为平面文件存储在 HDFS 上。在进行查询处理的时候，这些方法将查询分解成若干子查询。每个子查询通过在 HDFS 上的扫描得到候选解，然后用 MapReduce 将候选解连接起来以得到最终解。不同方法之间主要区别就是 RDF 数据转化 HDFS 平面文件的方式不同。

SHARD[Rohloff K, et al., 2010]以 RDF 数据中的主体为核心进行数据划分。SHARD 把一个主体相关的所有三元组聚集在一起并存储成 HDFS 文件中的一行。HadoopRDF[Husain M, et al., 2011]和 P-Partition[Zhang X, et al., 2012]都是以 RDF 数据中的属性为核心进行存储。它们把有相同属性的所有三元组聚集一起并存储于一个 HDFS 文件中。HadoopRDF[Husain M, et al., 2011]在以属性为核心的存储模式之外，还利用客体的类型信息进一步划分 RDF 数据。EAGRE[Zhang X, et al., 2013]提出了一个基于实体类型的存储模型。EAGRE 将 RDF 数据中所有主体视为一个实体，进而将 RDF 数据图压缩成一个实体图；然后，将相似的实体聚类成一个实体类，进而形成一个压缩实体图。EAGRE 把这个压缩实体图存储在内存，并利用已有图划分方法 METIS[Karypis G, et al., 1998]对这个压缩实体图进行划分。根据 METIS 划分结果，EAGRE 把这些实体以及相应三元组放到不同机器上。同时，每个实体是按照实体类来排序并存在 HDFS 中。

除了基于 Hadoop 的方法之外，现阶段还有部分研究工作是基于其他的云平台系统的。比如基于 Trinity[Shao B, et al., 2013]系统的 Trinity.RDF[Zeng K, et al., 2013]、基于 Parquet[Apache, 2014]的 Sempala[Schätzle A, et al., 2014]。

Trinity.RDF[Zeng K, et al., 2013]提出了利用 Trinity[Shao B, et al., 2013]进行 RDF 数据管理的方法。Trinity 是微软研发的一个基于内存的分布式图数据管理系统。Trinity.RDF 将 RDF 数据图的邻接表载入 Trinity 的内存云中。当用户提交查询之后，Trinity.RDF 依次对查询中每个变量  $v$  的候选点  $u$  进行图探索直到得到解。所谓图探索就是检查  $v$  候选点  $u$  的邻居是否能满足  $v$  在查询图上邻居的条件，如果不满足，则结束扩展；否则，看  $u$  邻居的邻居能否满足  $v$  在查询图上邻居的邻居。这个过程迭代直到找到最终解。Sempala[Schätzle A, et al., 2014]利用一个基于列存储的云存储系统 Parquet[Apache, 2014]进行 RDF 数据管理。Sempala 将 RDF 数据转化成基于属性的关系数据表存储在 Parquet 上。在查询处理阶段，Sempala 将查询改写成能在 Parquet 上 SQL 语句以执行得到结果。

总的来说，因为基于已有云平台的查询处理方法利用了现有的云计算框架，所以这些方法都有很好的可扩展性与容错性。但是，由于之前云计算框架很多并未针对 RDF 数据管理进行特殊的优化，所以这些方法进行查询处理的效率不高。

## 4.2 基于数据划分的分布式 RDF 数据查询处理方法

基于数据划分的分布式 RDF 数据查询处理方法首先将 RDF 数据划分成若

干子数据集，然后将这些子数据集分配到不同计算节点上。各个计算节点安装单机的 RDF 数据管理系统以管理被分配来的子数据集。当查询输入这些系统中后，这些方法首先将查询也划分成若干子查询，然后这些方法将这些子查询分配到各个计算节点上执行得到部分解，最后这些方法收集所有部分解通过连接得到最终解。不同基于数据划分的查询处理方法的主要区别在于数据划分时采用的策略不一样。

Jiewen Huang 等人提出的方法[Huang J, et al., 2011]使用现有成熟工具 METIS[Karypis G, et al., 1998]来对 RDF 数据的划分。划分出来每个子图对应一个数据分片，进而对应一个系统中的一个工作节点。在每个工作节点内部，Jiewen Huang 等人提出的方法使用已有的单机 RDF 数据管理系统对数据分片进行管理。SemStore[Wu B, et al., 2014]则是提出了一种称作有根子图的特殊结构做为划分基本单元对 RDF 知识图谱数据进行划分。所谓 RDF 数据图上点  $v$  的有根子图就是从  $v$  出发做遍历得到的所有点构成的子图。SemStore 首先找出能覆盖整个 RDF 数据图的一个有根子图集合，然后将这些有根子图聚成若干类。每一个类里面所有的作为有根子图一个分块被分配到一个对应的机器。

华中科技大学袁平鹏老师研究组还提出了一种基于 RDF 数据图上路径的划分方法[Wu B, et al., 2015]，该方法首先在 RDF 数据图上定义出“源点”和“沉入点”。所谓 RDF 数据图上的源点就是 RDF 数据图上没有入度的点；而所谓 RDF 数据图上的沉入点就是 RDF 数据图上没有出度的点。然后在源点和沉入点基础上定义出“末端到末端路径”，即从源点或者图上环中没有进入环的边的点到沉入点或者末端到末端路径已经路过点的路径。袁平鹏老师研究组的方法首先找出覆盖全图的末端到末端路径集合，然后袁平鹏老师研究组提出算法将覆盖全图的末端到末端路径集合分成  $k$  份，每份作为一个分块存储到一台机器上。

总的来说，基于数据划分的 RDF 数据上的分布式查询处理方法要求按照自身的算法设计进行 RDF 数据的划分与分配，以减少查询处理阶段的通信代价。但是，这些方法的系统性能受到每台机器上的单机 RDF 数据管理系统性能的限制。

### 4.3 联邦型分布式 RDF 数据查询处理方法

随着关联数据的发展，现在越来越多的数据发布者都愿意将数据表示成 RDF 数据格式并链接入关联数据上。其中很多数据发布者在将数据表示成 RDF 数据格式之外还提供 SPARQL 查询接口来让别人使用它的数据。这些 SPARQL 查询接口都属于“自治”的系统，即能各自独立地接受 SPARQL 查询并计算出匹配。每一个包含一定 RDF 数据和 SPARQL 查询接口的机器被称为一个 RDF 数据源。这些“自治”的 RDF 数据源被集成到一个系统平台下就形成了所谓的联邦型分布式 RDF 数据管理系统。

针对联邦型分布式 RDF 数据管理系统，现阶段也有一些研究在讨论来其上的查询处理技术。在联邦型分布式 RDF 数据管理系统中，因为各个 RDF 数据源之间相互独立地自治，所以系统在查询处理阶段无法中断各个 RDF 数据源的处理进程。因此，在联邦型分布式 RDF 数据管理系统中，系统需要提前将 SPARQL 查询分解成若干子查询并传送到它们对应的 RDF 数据源，以让这些对应的 RDF 数据源对子查询独立地进行处理并得到部分解。之后，系统将这些部分解收集起来并通过连接操作得到最终解。在这个过程中，不同方法之间的主要区别在于如何进行查询分解并确定每个子查询对应的 RDF 数据源。

DARQ [Quilitz B, et al., 2008] 是最早地讨论如何在联邦型分布式 RDF 数据管理系统上的进行 SPARQL 查询处理。当 SPARQL 查询输入以后，DARQ 根据服务描述的索引进行查询分解并确定出相关的 RDF 数据源。所谓服务描述，其中包含若干所谓的性能值。每个性能值对应一个数据源，其中包含若干元组  $t = (p, r)$ ，其中  $p$  表示该数据源有  $p$  这个属性， $r$  对应于当属性为  $p$  时主体或者客体若干限制。此外，在查询处理过程中，DARQ 还讨论了两个子查询结果连接方式：一是嵌套循环连接，就是一般的自然连接；二是绑定式连接，就是一个子查询先找出解，然后将解传输到另一个子查询那里，然后将解绑定到第二个子查询那里进行过滤。

在 DARQ 的服务描述之外，还有 SPLENDID[Görlitz O, et al., 2011]、HiBISCuS[Saleem M, et al., 2014]等方法。其中，SPLENDID 根据每个数据源的 VOID 信息建立一个倒排索引。这个索引将每个属性和类型信息映射到一个数据对  $(d, c)$ ，其中  $d$  表示属性或类型信息所在的 RDF 数据源， $c$  表示在  $d$  这个数据源上属性或类型信息的数量。HiBISCuS[Saleem M, et al., 2014]也构建了与 DARQ 类似的索引。只是，在确定各个子查询的相关 RDF 数据源阶段，HiBISCuS 将查询图建模成一个有向带标签的超图，并利用这个有向带标签的超图进一步减少每个子查询的候选 RDF 数据源。

不同于上述利用索引来确定相关 RDF 数据源的方法，FedX[Schwarze A, et al., 2011]则可以在查询处理阶段实时确定相关数据源。当查询输入以后，FedX[Schwarze A, et al., 2011]首先将查询中每个三元组模式都传到所有 RDF 数据源上并通过 SPARQL 查询语义中的 ASK 来确定相关数据源。之后，以三元组模式为单位进行查询优化，进而将若干三元组模式聚集在一起并得到连接操作顺序。FedX 所使用的连接方式也是和 DARQ 相似的绑定式连接，但是 FedX 在传输中间结果的时候不再是一个一个传，而是若干个中间结果合在一起传。

此外，本研究团队之前针对关联数据上被预先划分好的 RDF 数据还曾提出过一个基于“局部计算-再合并”的分布式 RDF 数据管理方法[Peng P, et al., 2016]。这种方法也是不干预 RDF 数据图预先定义的划分，即假设数据已经被划分并分布在不同的计算节点上。系统中每台机器根据自身上所存储的 RDF 数据计算出

整个 SPARQL 查询的局部匹配。所找出的局部匹配被定义为本地局部匹配。然后，所有被找出的本地局部匹配被归并起来并通过连接操作合并成最终匹配。

## 参考文献

- [Abadi D, et al., 2009] Daniel J. Abadi, Adam Marcus, Samuel Madden, Kate Hollenbach. SW-Store: a vertically partitioned DBMS for Semantic Web data management. VLDB J. 18(2): 385-406 (2009)
- [Apache, 2014]. Apache. Apache Parquet. 2014. <http://parquet.apache.org/>.
- [Bonstrom V, et al., 2003] Valerie Bonstrom, Annika Hinze, Heinz Scheppe. Storing RDF as a Graph. In Proceedings of LA-WEB'2003. pp.27-36
- [Broekstra J, et al., 2002] Jeen Broekstra, Arjohn Kampman, Frank van Harmelen. Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema. International Semantic Web Conference 2002: 54-68
- [Broekstra J, et al., 2003] Jeen Broekstra, Arjohn Kampman, Frank van Harmelen. Sesame: An Architecture for Storing and Querying RDF Data and Schema Information, in Spinning the Semantic Web (2003), pp. 197–222.
- [Google, 2016]. Google. Freebase. 2016. <https://developers.google.com/freebase/>.
- [Görlitz O, et al., 2011] Olaf Görlitz, Steffen Staab. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. COLD 2011.
- [Harris S, et al., 2003] Stephen Harris, Nicholas Gibbins. 3store: Efficient Bulk RDF Storage. PSSS 2003.
- [Heath T, 2018]. Tom Heath. Linked Data - Connect Distributed Data across the Web. 2018. <http://linkeddata.org/home>.
- [Hoffart J, et al., 2013] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Gerhard Weikum. YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. Artif. Intell. 194, 28–61 (2013).
- [Huang J, et al., 2011] Jiewen Huang, Daniel J. Abadi, Kun Ren. Scalable SPARQL Querying of Large RDF Graphs. PVLDB 4(11): 1123-1134 (2011).
- [Husain M, et al., 2011] Mohammad Farhan Husain, James P. McGlothlin, Mohammad M. Masud, Latifur R. Khan, Bhavani M. Thuraisingham. Heuristics-Based Query Processing for Large RDF Graphs Using Cloud Computing. IEEE Trans. Knowl. Data Eng. 23(9): 1312-1327 (2011)
- [Inseok E, et al., 2005] Eugene Inseok Chong, Souripriya Das, George Eadon, Jagannathan Srinivasan. An Efficient SQL-based RDF Querying Scheme. VLDB 2005: 1216-1227
- [Karypis G, et al., 1998] George Karypis, Vipin Kumar. A Fast and High Quality

- Multilevel Scheme for Partitioning Irregular Graphs. SIAM J. Scientific Computing 20(1): 359-392 (1998).
- [Kim J, et al., 2015] Jinha Kim, Hyungyu Shin, Wook-Shin Han, Sungpack Hong, Hassan Chafi. Taming Subgraph Isomorphism for RDF Query Processing. PVLDB 8(11): 1238-1249 (2015).
- [Lehmann J, et al., 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, Christian Bizer. DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web 6 (2), 167–195 (2015).
- [Lu J, et al., 2007] Jing Lu, Li Ma, Lei Zhang, Jean-Sébastien Brunner, Chen Wang, Yue Pan, Yong Yu. SOR: A Practical System for Ontology Storage, Reasoning and Search. VLDB 2007: 1402-1405.
- [Mahdisoltani F, et al., 2015] Farzaneh Mahdisoltani, Joanna Biega, Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. in CIDR (2015).
- [Neumann T, et al., 2008] Thomas Neumann, Gerhard Weikum. RDF-3X: A RISC-style Engine for RDF. PVLDB 1(1): 647-659 (2008)
- [Neumann T, et al., 2010a] Thomas Neumann, Gerhard Weikum. The RDF-3X Engine for Scalable Management of RDF Data. VLDB J. 19(1): 91-113 (2010)
- [Neumann T, et al., 2010b] Thomas Neumann, Gerhard Weikum. x-RDF-3X: Fast Querying, High Update Rates, and Consistency for RDF Databases. PVLDB 3(1): 256-263 (2010)
- [Pan Z, et al., 2003] Zhengxiang Pan, Jeff Heflin. DLDB: Extending Relational Databases to Support Semantic Web Queries. In Proceedings of PSSS'2003.
- [Peng P, et al., 2016] Peng Peng, Lei Zou, M. Tamer Özsu, Lei Chen, Dongyan Zhao. Processing SPARQL queries over distributed RDF graphs. VLDB J. 25(2): 243-268 (2016).
- [Quilitz B, et al., 2008] Bastian Quilitz, Ulf Leser. Querying Distributed RDF Data Sources with SPARQL. ESWC 2008: 524-538.
- [Rohloff K, et al., 2010] Kurt Rohloff, Richard E. Schantz. High-performance, Massively Scalable Distributed Systems using the MapReduce Software Framework: the SHARD Triple-store. PSI EtA 2010: 4.
- [Saleem M, et al., 2014] Muhammad Saleem, Axel-Cyrille Ngonga Ngomo. HiBISCuS: Hypergraph-Based Source Selection for SPARQL Endpoint Federation. ESWC

- 2014: 176-191.
- [Schätzle A, et al., 2014] Alexander Schätzle, Martin Przyjaciel-Zablocki, Antony Neu, Georg Lausen. Sempala: Interactive SPARQL Query Processing on Hadoop. Semantic Web Conference (1) 2014: 164-179.
- [Schwarze A, et al., 2011] Andreas Schwarze, Peter Haase, Katja Hose, Ralf Schenkel, Michael Schmidt. FedEx: Optimization Techniques for Federated Query Processing on Linked Data, International Semantic Web Conference 2011: 601-616.
- [Shao B, et al., 2013] Bin Shao, Haixun Wang, Yatao Li. Trinity: A Distributed Graph Engine on a Memory Cloud. SIGMOD Conference 2013: 505-516.
- [Shvachko K, et al., 2010] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler. The Hadoop Distributed File System. MSST 2010: 1-10.
- [Suchanek F, et al., 2008] Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. J. Web Sem. 6 (3), 203–217 (2008).
- [W3C, 2013]. W3C. SPARQL 1.1 Overview. 2013. <https://www.w3.org/TR/sparql11-overview/>.
- [W3C, 2014]. W3C. RDF - Semantic Web Standards. 2014. <http://www.w3.org/RDF/>.
- [Weiss C, et al., 2008] Cathrin Weiss, Panagiotis Karras, Abraham Bernstein. Hexastore: sextuple indexing for semantic web data management. PVLDB 1(1): 1008-1019 (2008)
- [Wikipedia, 2018]. Wikipedia. Semantic Web – Wikipedia. 2018. [https://en.wikipedia.org/wiki/Semantic\\_Web](https://en.wikipedia.org/wiki/Semantic_Web).
- [Wilkinson K, et al., 2003] Kevin Wilkinson, Craig Sayers, Harumi A. Kuno, Dave Reynolds. Efficient RDF Storage and Retrieval in Jena2. SWDB 2003: 131-150
- [Wilkinson K, et al., 2006] Kevin Wilkinson. Jena Property Table Implementation. in SSWS, Athens, Georgia, USA (2006), pp. 35–46.
- [Wu B, et al., 2014] Buwen Wu, Yongluan Zhou, Pingpeng Yuan, Hai Jin, Ling Liu. SemStore: A Semantic-Preserving Distributed RDF Triple Store. CIKM 2014: 509-518
- [Wu B, et al., 2015] Buwen Wu, Yongluan Zhou, Pingpeng Yuan, Ling Liu, Hai Jin. Scalable SPARQL Querying using Path Partitioning. ICDE 2015: 795-806.
- [Wylot M, et al., 2011] Marcin Wylot, Jigé Pont, Mariusz Wisniewski, Philippe Cudré-Mauroux. dipLODocus[RDF] - Short and Long-Tail RDF Analytics for Massive Webs of Data. International Semantic Web Conference (1) 2011: 778-793.
- [Zeng K, et al., 2013] Kai Zeng, Jiacheng Yang, Haixun Wang, Bin Shao, Zhongyuan

- Wang. A Distributed Graph Engine for Web Scale RDF Data. PVLDB 6(4): 265-276 (2013)
- [Zhang X, et al., 2012] Xiaofei Zhang, Lei Chen, Min Wang. Towards Efficient Join Processing over Large RDF Graph Using MapReduce. SSDBM 2012: 250-259.
- [Zhang X, et al., 2013] Xiaofei Zhang, Lei Chen, Yongxin Tong, Min Wang. EAGRE: Towards Scalable I/O Efficient SPARQL Query Evaluation on the Cloud. ICDE 2013: 565-576.
- [Zou L, et al., 2011] Lei Zou, Jinghui Mo, Lei Chen, M. Tamer Özsü, Dongyan Zhao. gStore: Answering SPARQL Queries via Subgraph Matching. PVLDB 4(8): 482-493 (2011)
- [Zou L, et al., 2014] Lei Zou, M. Tamer Özsü, Lei Chen, Xuchuan Shen, Ruizhe Huang, Dongyan Zhao. gStore: A Graph-based SPARQL Query Engine. VLDB J. 23(4): 565-590 (2014)

## 第七章 知识推理

### 1. 任务定义、目标和研究意义

随着知识图谱研究的深入，研究人员发现知识图谱在各种应用中存在以下质量问题：

第一个问题是知识图谱的不完备性，即知识图谱中的关系缺失或者属性缺失，例如人物的职业信息缺失。这个问题可能是因为构建知识图谱的数据本身就是不完备的，也可能是信息抽取算法无法识别到一些关系或者抽取到属性值。

第二个问题是知识图谱中存在错误的关系，如人物知识图谱中可能包含错误的人物关系。这个问题可能是因为构建知识图谱的数据有错误，也可能是因为知识图谱构建时采用了统计方法，而统计方法很难保证学习的知识是绝对正确的。

这两个问题对于智能问答等应用来说有较大影响，对于问答来说，前者会导致提出的问题没有答案，而后者会导致系统给出的答案是错误的。为了解决这两个问题，就需要对知识图谱的推理进行研究。知识图谱的推理指的是从给定的知识图谱推导出新的实体跟实体之间的关系，可以粗略地分为基于符号的推理和基于统计的推理。在人工智能的研究中，基于符号的推理一般是基于经典逻辑（一阶谓词逻辑或者命题逻辑）或者经典逻辑的变异（比如说缺省逻辑）。基于符号的推理一般是通过规则或者本体从一个已有的知识图谱推理出新的实体间关系，从而有助于解决第一个问题；而且基于符号的推理可以对知识图谱进行逻辑的冲突检测，从而有助于解决第二个问题。基于统计的方法一般指关系机器学习方法，即通过统计规律从知识图谱中学习到新的实体间关系，从而处理第一个问题；并且对新学到的关系进行评分，去掉那些可能错误的关系，从而处理第二个问题。

知识图谱之所以被认为是实现人工智能的一个重要研究方向，是因为知识图谱上的推理使之能够支撑人工智能的很多应用，而这也是知识图谱区别于传统关系数据模型的关键所在。如上所述，基于符号的推理包括基于本体的推理和基于规则的推理两种，前者包括概念的定义和分类，以及概念中实例的推断等推理，后者考虑的是将规则应用于图谱，实现图谱上新的关系推断以及基于图谱的决策支持。基于符号的推理被广泛用于生物医学中术语定义和概念分类、电商数据的一致检测和查询重写以及智能问答中的知识扩充等。基于统计的推理包括模式归纳和实体关系学习，前者考虑的是从知识图谱中挖掘概念的关系，后者考虑的是通过统计方法推断出两个实体之间的关系。模式归纳用于构建知识图谱的模式知识，提供概念之间的上下文关系和关系的定义域与值域，模式知识可以用于符号逻辑推理，也可以用于知识图谱的构建。实体关系学习对于知识图谱的补全有很大作用，可以用于智能问答等知识图谱的应用。

## 2. 研究内容和关键科学问题

知识图谱的推理首先需要考虑的是知识如何表达的问题，即知识图谱的知识表示，这里有基于图结构的表示以及相应的逻辑基础，还有基于张量的表示。其次需要考虑的是逻辑推理算法以及优化方法，实现高效的逻辑推理机。再次，需要考虑基于统计的知识图谱推理算法，重点介绍基于表示学习的方法和基于图特征的方法。最后介绍如何从知识图谱中通过统计方法来学习本体的方法。

### 2.1 知识图谱的表示

知识图谱表示 (representation) 指的是用什么数据结构来表示一个知识图谱。顾名思义，知识图谱是以图的方式来展示知识，但是这并不代表知识图谱必须采用图的表示。从图的角度看，知识图谱是一个语义网络，即一种用互联的节点和边来表示知识的一个结构 [Sowa, 1991]。语义网络中的节点可以代表概念 (concept)、属性 (attribute)、事件 (event) 或者实体 (entity)，而边则用来表示节点之间的关系，边的标签指明了关系的类型。语义网络中的语义主要体现在图中边的含义。为了赋予这些边语义，研究人员提出了术语语言 (terminological language)，并最终提出了描述逻辑 (description logic)，描述逻辑是一阶谓词逻辑的一个子集，推理复杂度是可判定的 (decidable)。W3C 采用了以描述逻辑为逻辑基础的本体语言 OWL (Ontology Web Language) 作为定义 Web 术语的标准语言，还推出了另外一种用于表示 Web 本体的语言 RDF Schema (简称 RDFS)。虽然描述逻辑以及 RDFS 的理论已经成熟，但是这些理论还没有很好的应用于知识图谱，目前还缺乏针对知识图谱的一个逻辑表示语言。最近，基于向量 (vector) 的知识表示开始流行，这类表示将知识图谱三元组中的主谓宾表示成数值向量，通过向量的知识表示，可以采用统计或者神经网络的方法来进行推理，对知识图谱中的实体间的关系进行预测。知识图谱的向量表示主要考虑事实性 (factual) 知识图谱的表示，如何对模式 (schematic) 知识图谱进行数值表示是一个难点。

### 2.2 基于符号的并行知识推理

基于符号的知识图谱推理一般是应用推理规则到知识图谱上，通过触发规则的前件来推导出新的实体关系，这里的推理规则可能是知识表示语言所有的，也可能是人工设定或者通过机器学习技术获取的。

基于符号的推理虽然有各种优化方法来提高推理效率，但是还是跟不上数据增长的速度，特别是当数据规模大到目前基于内存的服务器无法处理的情况下。为了应对这一挑战，研究人员开始考虑将描述逻辑和 RDFS 的推理并行来提升推理的效率和可扩展性，并且取得了很多成果。并行推理工作所借助的并行技术分为以下两类：

1. 单机环境下的多核、多处理器技术，比如多线程、GPU 技术等；

2. 多机环境下基于网络通信的分布式技术，比如 MapReduce 计算框架、Peer-To-Peer 网络框架等。

现有的并行推理方法主要集中在前向链推理，即应用推理规则到知识图谱生成新的三元组，所以对于动态知识图谱的推理还处理不好。另外，前向链推理会导致知识图谱存储大量冗余知识，也不利于高效的知识检索和查询。

### 2.3 实体关系学习方法

实体关系学习的目的是通过统计方法或者神经网络方法学习知识图谱中实体之间的关系。这方面的工作非常多，也是最近几年知识图谱的一个热门研究方向。相关研究工作大体可以分为两类：基于表示学习的方法和基于图特征的方法。

基于表示学习的方法将知识图谱中的实体与关系统一映射至低维连续向量空间，以此来刻画它们的潜在语义特征。通过比较、匹配实体与关系的分布式表示，可以得到知识图谱中潜在成立的实体间的关系。此类方法灵活自由，通常具有较高的计算效率，然而可解释性较差，同时对于困难的推理问题往往精度不足。如何提升这类方法的推理精度仍然是当前研究的热点与难点。

基于图特征的方法利用从知识图谱中观察到的图特征来预测一条可能存在的边。代表性工作包括归纳逻辑程序设计（ILP）[Quinlan, 1990]、关联规则挖掘（ARM）[Galarraga et al., 2013]、路径排序算法（PRA）[Lao et al., 2010]等。此类方法在推理的同时能从知识图谱中自动挖掘推理规则，具备明确的推理机理。然而图特征的提取效率较低，尤其对于时下超大规模的知识图谱更是如此。如何提高效率是这类方法亟待突破的壁垒。

### 2.4 模式归纳方法

模式归纳方法是从知识图谱中学习本体的模式层信息或丰富已有本体，包括概念层次、属性层次、不交公理、属性的值域与定义域和一些属性或概念的约束等公理的学习。

知识图谱的迅猛增长，为人们提供日益丰富的相互关联的数据可用。但是，这些数据大都处于实例层，描述个体及个体之间的关系，缺少用于约束个体的模式层信息，例如概念或属性的层次关系、不交公理、属性或概念的一些约束公理等的缺失。模式层信息的缺失，给知识图谱的整合、查询和维护等关键任务带来了重重困难。因此，研究人员针对这些问题提出不少模式归纳的方法[Subhashree et al., 2018]，基于知识图谱进行各种各样模式层公理的学习。这方面的主要研究大致分为以下三类：

1. 基于归纳逻辑编程（ILP）进行模式归纳的研究。这类方法结合了机器学习和逻辑编程技术，从实例和背景知识中获得逻辑结论，构建本体。

2. 基于关联规则挖掘进行模式归纳的研究。这类方法常首先从知识图谱中收集所需信息，然后将之用事务表表示出来，再利用传统的关联规则挖掘方法找出规则，而这些规则往往可直接转换成本体中的公理。
3. 基于机器学习进行模式归纳的研究。这类方法使用一些机器学习的方法，例如贝叶斯网络和聚类，将本体学习转换成一个机器学习的问题，将知识图谱用采纳的学习模型进行表示、建模和推理，获得新的公理。

知识图谱采用的是开放世界假设，而传统的关联规则挖掘或者机器学习则是采用封闭世界假设，如何应对这个不同世界假设的问题是研究者们不断努力的方向。另外，鉴于知识图谱规模的巨大，开发出高效的、扩展性强的模式归纳算法也是一大难题。

### 3. 技术方法和研究现状

#### 3.1 知识图谱的表示

虽然语义网络作为一种自然语言的知识表示方法受到了很多关注，Woods 在 1975 年的一篇论文指出语义网络中的弧的语义存在问题[Woods, 1975]。由于一阶谓词逻辑具有形式化语义，所以很多研究人员开始建立一阶谓词逻辑和语义网络之间的对应关系，从而赋予语义网络形式化语义（如文献[Allen et al., 1982]）。但是，由于一阶谓词逻辑的推理复杂度是不可判定的，所以一般业界很少关注这种对应。为了使得语义网络同时具备形式化语义和高效推理，一些研究人员提出了易处理（tractable）概念语言，如 KL-ONE[Brachman, 1978] 和 CLASSIC [Brachman et al. 1999]，并且开发了一些商用化的语义网络系统。这些系统的提出，使得针对概念描述的一系列逻辑语言，统称描述逻辑（description logic），得到了学术界和业界广泛关注。但是这些系统的推理效率难以满足日益增长的数据的需求，最终没能得到广泛应用。这一困局被利物浦大学的 Ian Horrocks 教授打破，他开发的 FaCT 系统可以处理一个比较大的医疗术语本体 GALEN，而且性能比其他类似的推理机要好得多 [Horrocks, 1998]。描述逻辑最终成为了 W3C 推荐的 Web 本体语言 OWL 的逻辑基础 [Horrocks et al. 2003]。描述逻辑为知识图谱提供了一些逻辑基础，但是并没有被很好的应用。在文献[Krötzsch et al. 2016] 和[Krötzsch et al. 2017a]中，Krötzsch 等人给出了如何扩展描述逻辑或者采用存在规则来表示知识图谱的本体，并提高逻辑推理服务，但是这些工作并没有实用化的系统。[Krötzsch, 2017b]给出了除了描述逻辑以外的一些支持逻辑推理的知识图谱表示语言，比如说，除了描述逻辑以外，还可以采用属性逻辑（Attributed Logics）来表示知识图谱的本体，这里的属性逻辑比较适合表示具有时空维度的知识图谱。RDF 的表示是基于三元组的，但是现实生活中经常需要处理的知识不

是三元组可以描述的，比如说奥巴马 2008 年担任美国总统，就无法用三元组来描述，因为三元组只能描述（奥巴马 担任 总统），没法继续描述时间。为了解决这一问题，Udrea 等人在[Udrea et al. 2010]中提出了带标签的 RDF，即给每个 RDF 三元组一个标签，用来描述该三元组的元信息。著名的知识图谱 Yago 采用了一种新的方法来表示三元组的时空维度的信息[Rebele et al. 2016]，即给每个三元组一个标识符，用来代表该三元组，然后对该标识符赋予时空维度的信息。

## 3.2 基于符号的并行知识推理

### 3.2.1 基于多核、多处理器技术的大规模推理

单机环境下的并行技术以共享内存模型为特点，侧重于提升本体推理的时间效率。对于实时性要求较高的应用场景，这种方法成为首选。对于表达能力较低的语言，比如 RDFS、OWLEL，单机环境下的并行技术显著地提升了本体推理效率。Goodman 等人在文献 [Goodman et al. 2011] 中利用高性能计算平台 Cray XMT 实现了大规模的 RDFS 本体推理，利用平台计算资源的优势限制所有推理任务在内存完成。然而对于计算资源有限的平台，内存使用率的优化成为了不可避免的问题。Motik 等人在文献 [Motik et al. 2014] 工作中将 RDFS，以及表达能力更高的 OWL RL 等价地转换为 Datalog 程序，然后利用 Datalog 中的并行优化技术来解决内存的使用率问题。在文献 [Urbani et al. 2015] 中，作者尝试利用并行与串行的混合方法来提升 OWL RL 的推理效率。Kazakov 等人在文献 [Kazakov et al. 2015] 中提出了利用多线程技术实现 OWLEL 分类 (classification) 的方法，并实现推理机 ELK。Zhou 等人在[Zhou et al. 2016] 中提出了一套并行推理的理论和算法，可以用来实现并行的 OWL 推理。

### 3.2.2 基于分布式技术的大规模推理

尽管单机环境的推理技术可以满足高推理性能的需求，但是由于计算资源有限（比如内存、存储容量），推理方法的可伸缩性 (scalability) 受到不同程度的限制。因此，很多工作利用分布式技术突破大规模数据的处理界限。这种方法利用多机搭建集群来实现本体推理。Oren 等人[Oren et al. 2009]首次尝试利用 Peer-To-Peer 的分布式框架实现 RDF 数据推理，表明利用分布式技术可以完成很多在单机环境下无法完成的大数据量推理任务。很多工作基于 MapReduce 的开源实现（如 Hadoop、Spark 等）设计提出了大规模本体的推理方法，其中较为成功的一个尝试是 Urbani 等人在 2010 年公布的推理系统 WebPIE[Urbani et al. 2010]，在大集群上可以完成上百亿的 RDF 三元组的推理。他们又在此基础上研究提出了基于 MapReduce 的 OWL RL 查询算法[Urbani et al. 2011]。文献[Zhou et al. 2013]利用 MapReduce 来实现 OWLEL 本体的推理算法，证明 MapReduce 技术同样可以解决大规模的 OWL EL 本体推理，并在后续工作中进一步扩展，

使得推理可以在多个并行计算平台完成[Zhou et al. 2017]。

### 3.3 实体关系学习方法

#### 3.3.1 基于表示学习的方法

知识图谱表示学习旨在于将知识图谱中的实体与关系统一映射至低维连续向量空间，以刻画它们的潜在语义特征。通过比较实体与关系在该向量空间中的分布式表示，可以推断出实体和实体之间潜在的关系。近年来，随着深度神经网络的蓬勃发展，知识图谱表示学习技术异军突起，获得了学界广泛关注。早期的研究通过设计简单的向量空间操作来建模实体间的关系。例如 TransE 模型[Bordes et al., 2013]将关系视作头尾实体之间的位移操作(translation)，认为头实体向量经过关系的位移后应尽可能接近尾实体向量。RESCAL 模型[Nickel et al., 2011]将关系表示为方阵以刻画潜在特征间的两两关联，通过头尾实体和关系的双线性匹配来判断关系成立的可能性。后期的工作一方面致力于设计更加合理的实体间关系的建模方式，如 TransE 模型的拓展[Wang et al., 2014a] [Lin et al., 2015a] [Ji et al., 2015] [Xiao et al., 2016] [He et al., 2015]、RESCAL 模型的拓展[Yang et al., 2015] [Nickel et al., 2016a] [Trouillon et al., 2016] [Liu et al., 2017]，以及一系列新兴的基于神经网络架构的表示学习模型[Socher et al., 2013] [Bordes et al., 2014] [Shi et al., 2017] [Dettmers et al., 2018]等；另一方面也尝试在实体间关系的基础上，进一步融入其他形式的信息以辅助表示学习，如实体类型[Guo et al., 2015] [Xie et al., 2016a]、关系路径[Neelakantan et al., 2015] [Lin et al., 2015b] [Guu et al., 2015]、实体描述文本[Wang et al., 2014b] [Xie et al., 2016b] [Wang et al., 2016]，甚至是逻辑规则[Rocktäschel et al., 2015] [Guo et al., 2016] [Demeester et al., 2016] [Guo et al., 2018]等。文献[Wang et al., 2017]对当下主流的知识图谱表示学习技术进行了系统的综述。

#### 3.3.2 基于图特征的方法

基于图特征的方法借助从知识图谱中抽取出的图特征来预测两个实体间可能存在的不同类型的边(关系)。例如，根据两个实体“姚明”和“叶莉”在知识图谱中的联通路径“姚明  $\xrightarrow{\text{女儿}}$  姚沁蕾  $\xleftarrow{\text{女儿}}$  叶莉”，可以预测出他们之间大概率具备“配偶”关系。此类方法与基于符号的推理有着很深的渊源，长久以来都是人工智能、机器学习等领域关注的热点。早在上世纪九十年代初期，Quinlan 就提出了著名的 FOIL 算法[Quinlan, 1990]，采用序贯覆盖框架自顶向下地从关系数据库中自动归纳一阶规则(first-order rules)，并将这些规则应用到数据库上，推出新的关系实例。FOIL 算法遵循封闭世界假设(closed world assumption)，认为但凡未声明是正例的关系样例一律是反例。然而，现实世界中，通过众包或信

息自动抽取方式构建的大规模知识图谱通常是高度不完备的[Dong et al., 2014], 并不符合该假设。鉴于此, Galárraga 等人提出了部分完整性假设 (partial completeness assumption) 并设计了 AMIE 算法[Galarraga et al., 2013], 以实现不完备知识图谱中的关联规则挖掘。随后, AMIE+算法[Teflioudi et al., 2015]在规则的搜索与评价方面做出了改进, 显著提高了规则学习的效率。由于知识图谱中包含的仅仅是实体之间的二元关系, 因此规则与知识图谱中的关系路径存在对应关系。路径排序算法(PRA)[Lao et al., 2010]就是以两实体间的联通路径作为特征, 来学习目标关系的分类器, 据此判断这两个实体是否属于目标关系。文献[Gardner et al., 2015]进一步探索了路径排序算法中不同的特征抽取和特征值计算策略对整体效率以及性能的影响。文献[Nickel et al., 2016b]对基于图特征的推理方法进行了简要概述。

### 3.4 模式归纳方法

#### 3.4.1 基于 ILP 的模式归纳方法

基于 ILP 的方法进行本体学习的早期工作在文献[Amato et al., 2010]与[Fanizzi et al., 2010]给出了很好的综述。Jens Lehmann 等在[Lehmann et al., 2007]中提出用向下精化算子学习 ALC 的概念定义公理的方法, 并在后续工作[Hellmann et al., 2009]中将原有方法扩展到处理大规模知识库上。相关的算法都在本体学习工具 DL-Learner 中得到实现[Lehmann et al., 2009], 并且在[Bühmann et al., 2016]工作中得到进一步扩展, 涉及到框架的设计和可扩展性的提升等方面。

#### 3.4.2 基于关联规则挖掘的模式归纳方法

在基于关联规则挖掘的方法中, 文献[Völker et al., 2011]将公理学习转换成关联规则挖掘的过程, 可学到原子概念间的包含关系等, 在后续工作中使用负关联规则挖掘技术学习不交公理[Fleischhacker et al., 2011], 并在文献[Völker et al., 2015]中给出了丰富的结果。文献[Galarraga et al., 2013]针对语义数据的开放世界假设的特点提出新的置信度定义, 用来学习属性的包含与等价关系, 并在后期工作[Teflioudi et al., 2015]中针对可扩展性做了进一步提升。文献[Ell et al., 2016]利用频繁项集挖掘进行属性定义域和值域的学习。[Irny et al., 2017] 中提出谓词偏好因子这个度量方法, 并提出考虑谓词语义相似度的机制, 用于学习相反和对称公理。鉴于已有方法往往只利用实例层信息而忽略模式层的存在, 文献[Barati et al., 2017]提出利用模式层信息给规则的挖掘提供更多的语义。文献[Gao et al., 2018]对传统关联规则挖掘技术进行了改进, 事务表中用 0 到 1 之间的一个实数代替原来的 0 或者 1, 使得提出的方法更符合语义数据开放的特点。

#### 3.4.3 基于机器学习的模式归纳方法

在基于统计的方法中, 文献[Zhu et al., 2015]提出将本体学习转换为在扩展的

贝叶斯描述逻辑网络 (BelNet+) 下的推理，即将用于学习的本体用 BelNet+ 表示出来，然后通过推理获得新的公理。文献[Zhang et al., 2015] 使用无监督数据驱动方法计算属性之间的等价关系，包含了相似度阈值的自动设置和等价属性群的发现。文献[Bühmann et al., 2012] 针对可以通过 SPARQL 查询终端获取所需信息，再利用 DL-Learner 中的机器学习算法计算公理的置信度，该工作在[Bühmann et al., 2013] 中保持其原有高效性的同时，使之可处理更多频繁的公理类型。文献 [Barchi et al., 2015] 利用聚类的算法学习关系的定义域和值域。文献[Munoz et al., 2017] 应用统计的方法过滤属性的使用，并找出准确、健壮的模式，用于学习属性的数量约束公理。

## 4. 技术展望与发展趋势

本文介绍了知识图谱的两类推理方法，即基于符号的推理方法和基于统计的推理方法。两种推理方法具有各自的应用场景，并且具有互补性。基于符号的推理方法更多考虑确定性知识的推理，一般通过给定的规则对知识图谱进行推理。为了使得推理具有高效性和可扩展性，现有工作采用了多种并行框架实现并行推理。基于统计的推理是一种不确定性推理，通过统计规律对知识图谱中的缺失的边进行补全。通过采取并行框架可以对推理进行加速，从而实现海量知识图谱的关系补全。虽然知识图谱上的推理技术已经取得了很多进展，特别是逻辑推理方面取得了很多理论和系统上的进展，但是总体来说大部分推理技术还离实际应用还有一段距离，存在以下问题需要解决。

### 4.1 基于符号推理的发展趋势

首先，知识图谱的表示还缺乏一个统一的方法。基于语义网络的方法缺乏逻辑基础；描述逻辑虽然是一种具有逻辑基础的语义网络，但是由于其提出是面向语义网的本体标准，不能完全适合知识图谱的表示；基于张量的表示方法则是对图的一种变种，并不能作为知识图谱的表示基础。对于知识图谱的表示，需要兼顾实用性和理论性。实用性方面，需要考虑到知识图谱构建并不一定会遵循某个标准，所以知识图谱的表达方式必然是多样化的，如何建立一套围绕知识图谱的表示方法，可以对各种应用的知识图谱进行表示值得研究的问题。这里重点可以围绕以下几个具体问题进行研究：1. 时空知识图谱如何进行表示，这里会涉及到知识图谱的动态性，即动态知识图谱的表示；2. 事件图谱如何进行表示，事件图谱的表示需要考虑到事件之间的因果关系和时序关系等。

其次，知识图谱的推理方法还缺乏实用的工具。本体推理机有很多原型系统，但是这些原型系统大都没有在实践中检验，在系统的稳定性和效率方面还存在诸多问题，需要进一步工程化。知识图谱的推理比较实用的还是基于规则引擎的推理，典型的工具是 RDFox[Motik et al. 2014]，但是很多规则引擎还局限于线下推

理，而且缺乏规则管理的功能，从而限制了其实用性。这里重点可以围绕以下几个具体问题进行研究：1. 目前针对易处理（Tractable）描述逻辑本体的并行推理已经取得了不少突破，但是对于更复杂本体的推理还需要研究通过近似推理等技术来优化；2. 本体推理机主要还是考虑前向链推理，而实际应用的时候需要考虑查询重写，如何融合前向链推理和查询重写是值得研究的课题；3. 无论是本体推理还是规则推理，都需要考虑本体或者规则的管理，比如说本体的融合和更新，本体的不一致性处理，规则的冲突检测和规则的权重设置等，目前还缺乏全面考虑这些知识管理等推理引擎。

## 4.2 基于统计推理的发展趋势

知识图谱表示学习方面虽然提出了很多技术，而且有一些开源工具，但是这些工具大都只能应对某些基准数据，未能在实际数据上取得很好的应用效果。如何进一步提升基于表示学习的推理方法的推理精度是亟待解决的问题。当前的一大发展趋势是利用深度神经网络架构建模实体与实体之间的关系，期待通过复杂的网络结构和非线性变换捕捉实体之间复杂的语义关联，从而提升推理的精度。如何将表示学习模型与更加丰富信息形态（尤其是图模式和符号规则）有效结合，充分发挥后者在表示学习中的效用，也是未来值得研究的问题。基于图特征的方法因其良好的可解释性，近年来也受到了广泛的关注。然而大规模知识图谱上的图特征抽取，尤其是复杂图特征抽取，效率低下，耗时极长。如何提高图特征抽取的效率，使其真正意义上适用于大规模知识图谱是当前一大技术挑战。另外，知识图谱的稀疏性也会导致大量对推理有用的图特征无法被有效抽取。如何突破知识图谱联通性壁垒，抽取出更加丰富的图特征也是未来重要的研究方向。

最后，基于知识图谱进行模式归纳的方法在资源的利用和实用工具的开发上还存在较大的提高空间。现有的模式归纳方法大都是基于实例层信息进行模式层公理的获取，即便在测试的实例上已经存在一些本体，已有方法很少利用这些有价值的资源，因此如何更好地利用这些模式层信息辅助基于实例的本体学习的工作，还存在较大的提升空间。另外，知识图谱是不断发展变化的，新的事实不断产生，已有的事实也在不停地发生改变，这些动态信息资源，对于已有本体会产生什么影响、或者针对这些动态数据怎样进行模式归纳都具有很大挑战。根据已有工作我们可以发现，模式归纳的工具还是比较欠缺的，大多模式归纳算法的实现没有提供用户界面，提供用户界面的工具缺少可扩展性，即其他种类公理的学习算法不方便整合进去，因此如何开发一个易扩展、界面友好的模式归纳工具是个亟待解决的问题。

## 参考文献

- [Allen et al., 1982] James F. Allen and Alan M. Frisch: What's in a Semantic Network.  
Proceedings of the 20th annual meeting on Association for Computational

- Linguistics (ACL 1982), 19-27, 1982.
- [Amato et al., 2010] Claudia d'Amato, Nicola Fanizzi, Floriana Esposito, Inductive Learning for the Semantic Web: what does it buy? Semantic Web (2010) 53-59
- [Barati et al., 2017] Barati M, Bai Q, Liu Q. Mining Semantic Association Rules from RDF Data[J]. Knowledge-Based Systems, 2017.
- [Barchi et al., 2015] Barchi P H, Hruschka E R. Never-ending ontology extension through machine reading. International Conference on Hybrid Intelligent Systems. IEEE, 2015:266-272.
- [Bordes et al., 2013] Bordes, A., Usunier, N., García-Durán, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in Neural Information Processing Systems (NIPS), pages 2787-2795.
- [Bordes et al., 2014] Bordes, A., Glorot, X., Weston, J., and Bengio, Y. (2014). A semantic matching energy function for learning with multi-relational data. Machine Learning, 94(2): 233-259.
- [Brachman, 1978] Ron J. Brachman. A Structural Paradigm for Representing Knowledge, Ph.D. thesis. Harvard University, May 1977 Also. BBN Report No.3605, Bolt Beranek and Newman Inc., May 1978.
- [Brachman et al. 1999] Ronald J. Brachman, Deborah L. McGuinness, Peter F. Patel-Schneider, Alexander Borgida: "Reducing" CLASSIC to Practice: Knowledge Representation Theory Meets Reality. Artificial Intelligence, 114(1-2): 203-237, 1999.
- [Bühmann et al., 2012] Lorenz Bühmann, Jens Lehmann: Universal OWL Axiom Enrichment for Large Knowledge Bases. EKAW 2012: 57-71
- [Bühmann et al., 2013] Lorenz Bühmann, Jens Lehmann: Pattern Based Knowledge Base Enrichment. In Proceedings of the 12th International Semantic Web Conference (1) 2013: 33-48
- [Bühmann et al., 2016] Lorenz Bühmann, Jens Lehmann, Patrick Westphal: DL-Learner - A framework for inductive learning on the Semantic Web. J. Web Sem. 39: 15-24 (2016)
- [Demeester et al., 2016] Demeester, T., Rocktäschel, T., and Riedel, S. (2016). Lifted rule injection for relation embeddings. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1389-1399.
- [Dettmers et al., 2018] Dettmers, T., Pasquale, M., Pontus, S., and Riedel, S. (2018). Convolutional 2D knowledge graph embeddings. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI).
- [Dong et al., 2014] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pages 601-610.

- [Ell et al., 2016] Basil Ell, Sherzod Hakimov, Philipp Cimiano: Statistical Induction of Coupled Domain/Range Restrictions from RDF Knowledge Bases. KEKI/NLP& DBpedia @ ISWC 2016: 27-40
- [Fanizzi et al., 2010] Nicola Fanizzi, Claudia D'Amato, Floriana Esposito, Machine Learning Methods for Ontology Mining, 2010, pp. 131 – 153.
- [Fleischhacker et al., 2011] Daniel Fleischhacker, Johanna Völker: Inductive Learning of Disjointness Axioms. OTM Conferences (2) 2011: 680-697
- [Fleischhacker et al., 2012] Fleischhacker, D., Voelker, J., Stuckenschmidt, H.: Mining RDF data for property axioms. In: Meersman, R., et al. (eds.) OTM 2012. LNCS, vol. 7566, pp. 718 – 735. Springer, Heidelberg (2012).
- [Galarraga et al., 2013] Galárraga, L.A., Teflioudi, C., Hose, K., Suchanek, F.M.: AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In: WWW (2013)
- [Gao et al., 2018] Huan Gao, Guilin Qi, Qiu Ji. Schema Induction from Incomplete Semantic Data. Journal of Intelligent Data Analysis. 2018 (to appear)
- [Gardner et al., 2015] Gardner, M. and Mitchell, T. (2015). Efficient and expressive knowledge base completion using subgraph feature extraction. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1488-1498.
- [Goodman et al. 2011] Eric L. Goodman, Jimenez Edward, Mizell David, al Saffar Sinan, Adolf Bob, Haglin David: High-performance Computing Applied to Semantic Databases. Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011), 31-45, 2011.
- [Guo et al., 2015] Guo, S., Wang, Q., Wang, B., Wang, L., and Guo, L. (2015). Semantically smooth knowledge graph embedding. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 84-94.
- [Guo et al., 2016] Guo, S., Wang, Q., Wang, L., Wang, B., and Guo, L. (2016). Jointly embedding knowledge graphs and logical rules. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 192-202.
- [Guo et al., 2018] Guo, S., Wang, Q., Wang, L., Wang, B., and Guo, L. (2018). Knowledge graph embedding with iterative guidance from soft rules. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI).
- [Guu et al., 2015] Guu, K., Miller, J., and Liang, P. (2015). Traversing knowledge graphs in vector space. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 318-327.
- [He et al., 2015] He, S., Liu, K., Ji, G., and Zhao, J. (2015). Learning to represent knowledge graphs with Gaussian embedding. In Proceedings of the 24th ACM

- International on Conference on Information and Knowledge Management (CIKM), pages 623-632.
- [Hellmann et al., 2009] Sebastian Hellmann, Jens Lehmann, Sören Auer. Learning of OWL class descriptions on very large knowledge bases. International Journal on Semantic Web and Information Systems 5(2), 25 – 48 (2009)
- [Horrocks, 1998] Ian Horrocks: Using an Expressive Description Logic: FaCT or Fiction? Proceedings of the Sixth International Conference on Principles of Knowledge Representation and Reasoning (KR 1998), 636-649, 1998.
- [Horrocks et al. 2003] Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen: From SHIQ and RDF to OWL: The Making of a Web Ontology Language. Journal of Web Semantics, 1(1):7-26, 2003.
- [Irny et al., 2017] Irny R, Kumar P S. Mining Inverse and Symmetric Axioms in Linked Data. Joint International Semantic Technology Conference. Springer, Cham, 2017:215-231.
- [Ji et al., 2015] Ji, G., He, S., Xu, L., Liu, K., and Zhao, J. (2015). Knowledge graph embedding via dynamic mapping matrix. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 687-696.
- [Kazakov et al. 2015] Yevgeny Kazakov, Pavel Klinov: Advancing ELK: Not Only Performance Matters. Description Logics, 233-248, 2015.
- [Krötzsch et al. 2016] Markus Krötzsch, Veronika Thost: Ontologies for Knowledge Graphs: Breaking the Rules. International Semantic Web Conference (1) 2016: 376-392
- [Krötzsch, 2017a] Markus Krötzsch: Ontologies for Knowledge Graphs? Description Logics 2017.
- [Krötzsch et al. 2017b] Markus Krötzsch, Maximilian Marx, Ana Ozaki, Veronika Thost: Attributed Description Logics: Ontologies for Knowledge Graphs. International Semantic Web Conference (1) 2017: 418-435
- [Lao et al., 2010] Lao, N. and Cohen, W. W. (2010). Relational retrieval using a combination of path-constrained random walks. Machine Learning, 81(1): 53-67.
- [Lehmann et al., 2007] Jens Lehmann, Pascal Hitzler: A Refinement Operator Based Learning Algorithm for the ALC Description Logic. ILP 2007: 147-160
- [Lehmann et al., 2009] Jens Lehmann: DL-Learner: Learning Concepts in Description Logics. Journal of Machine Learning Research 10: 2639-2642 (2009)
- [Lin et al., 2015a] Lin, Y., Liu, Z., Sun, M., Liu, Y., and Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI), pages 2181-2187.
- [Lin et al., 2015b] Lin, Y., Liu, Z., Luan, H., Sun, M., Rao, S., and Liu, S. (2015). Modeling relation paths for representation learning of knowledge bases. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language

- Processing (EMNLP), pages 705-714.
- [Liu et al., 2017] Liu, H., Wu, Y., and Yang Y. (2017). Analogical inference for multi-relational embeddings. In Proceedings of the 34th International Conference on Machine Learning (ICML), pages 2168-2178.
- [Motik et al. 2014] Boris Motik, Yavor Nenov, Robert Piro, Ian Horrocks: Parallel Materialisation of Datalog Programs in Main-Memory RDF Databases. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI 2014), 2014.
- [Munoz et al., 2017] Muñoz E, Nickles M. Mining Cardinalities from Knowledge Bases. International Conference on Database and Expert Systems Applications. Springer, Cham, 2017:447-462.
- [Neelakantan et al., 2015] Neelakantan, A., Roth, B., and McCallum, A. (2015). Compositional vector space models for knowledge base completion. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP), pages 156-166.
- [Nickel et al., 2011] Nickel, M., Tresp, V., and Kriegel, H.-P. (2011). A three-way model for collective learning on multi-relational data. In Proceedings of the 28th International Conference on Machine Learning (ICML), pages 809-816.
- [Nickel et al., 2016a] Nickel, M., Rosasco, L., and Poggio, T. (2016). Holographic embeddings of knowledge graphs. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI), pages 1955-1961.
- [Nickel et al., 2016b] Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1): 11-33.
- [Oren et al. 2009] Eyal Oren, Spyros Kotoulas, George Anadiotis, Ronny Siebes, Annette ten Teije, Frank van Harmelen: Marvin: Distributed reasoning over large-scale Semantic Web data. Journal of Web Semantics, 305-316, 2009.
- [Quinlan, 1990] Quinlan, J. R. (1990). Learning logical definitions from relations. Machine Learning, 5(3): 239-226.
- [Rebele et al. 2016] Thomas Rebele, Fabian M. Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, Gerhard Weikum: YAGO: A Multilingual Knowledge Base from Wikipedia, Wordnet, and Geonames. International Semantic Web Conference (2) 2016: 177-185
- [Rocktäschel et al., 2015] Rocktäschel, T., Singh, S., and Riedel, S. (2015). Injecting logical background knowledge into embeddings for relation extraction. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), pages 1119-1129.
- [Shi et al., 2017] Shi, B. and Weninger, T. (2017). ProjE: Embedding projection for

- knowledge graph completion. In Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI), pages 1236-1242.
- [Socher et al., 2013] Socher, R., Chen, D., Manning, C. D., and Ng, A. Y. (2013). Reasoning with neural tensor networks for knowledge base completion. In Advances in Neural Information Processing Systems (NIPS), pages 926-934.
- [Sowa, 1991] John F. Sowa: Principles of Semantic Networks: Exploration in the Representation of Knowledge, Morgan Kaufmann Publishers, INC. San Mateo, California, 1991.
- [Subhashree et al., 2018] Subhashree S, Irny R, Kumar P S. Review of Approaches for Linked Data Ontology Enrichmen. International Conference on Distributed Computing and Internet Technology. Springer, Cham, 2018:27-49.
- [Teflioudi et al., 2015] Teflioudi C, Hose K, Suchanek F M. Fast rule mining in ontological knowledge bases with AMIE+. VLDB Journal, 2015, 24(6):707-730.
- [Trouillon et al., 2016] Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In Proceedings of the 33rd International Conference on Machine Learning (ICML), pages 2071-2080.
- [Udrea et al. 2010] Octavian Udrea, Diego Reforgiato Recupero, V. S. Subrahmanian: Annotated RDF. ACM Trans. Comput. Log. 11(2): 10:1-10:41 (2010)
- [Urbani et al. 2010] Jacopo Urbani, Spyros Kotoulas, Jason Maassen, Frank van Harmelen, Henri E. Bal: OWL Reasoning with WebPIE: Calculating the Closure of 100 Billion Triples. Proceedings of the Extended Semantic Web Conference (ESWC 2010), 180-195, 2010.
- [Urbani et al. 2011] Jacopo Urbani, Frank van Harmelen, Stefan Schlobach, Henri E. Bal: QueryPIE: Backward Reasoning for OWL Horst over Very Large Knowledge Bases. Proceedings of the International Semantic Web Conference (ISWC 2011), 730-745, 2011.
- [Urbani et al. 2015] Jacopo Urbani, Ceriel J. H. Jacobs: RDF-SQ: Mixing Parallel and Sequential Computation for Top-Down OWL RL Inference. GKR 2015: 125-138,2015.
- [Völker et al., 2011] Johanna Völker, Mathias Niepert. Statistical Schema Induction. Proceedings of 8th Extended Semantic Web Conference (ESWC) (1), 2011: 124-138
- [Völker et al., 2015] Johanna Völker, Daniel Fleischhacker, Heiner Stuckenschmidt. Automatic Acquisition of Class Disjointness. Journal of Web Semantics. 35: 124-139 (2015)
- [Wang et al., 2014a] Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge graph embedding by translating on hyperplanes. In Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI), pages 1112-1119.
- [Wang et al., 2014b] Wang, Z., Zhang, J., Feng, J., and Chen, Z. (2014). Knowledge graph and text jointly embedding. In Proceedings of the 2014 Conference on

- Empirical Methods in Natural Language Processing (EMNLP), pages 1591-1601.
- [Wang et al., 2016] Wang, Z. and Li, J. (2016). Text-enhanced representation learning for knowledge graph. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pages 1293-1299.
- [Wang et al., 2017] Wang, Q., Mao, Z., Wang, B., and Guo, L. (2017). Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12): 2724-2743.
- [Woods, 1975] William A. Woods, What's in a link: Foundations for Semantic Networks. Representation and Understanding, D. Bobrow and A. Collins (eds.), Academic Press, 1975.
- [Xiao et al., 2016] Xiao, H., Huang, M., and Zhu, X. (2016). From one point to a manifold: Knowledge graph embedding for precise link prediction. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pages 1315-1321.
- [Xie et al., 2016a] Xie, R., Liu, Z., and Sun, M. (2016). Representation learning of knowledge graphs with hierarchical types. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), pages 2965-2971.
- [Xie et al., 2016b] Xie, R., Liu, Z., Jia, J., Luan, H., and Sun, M. (2016). Representation learning of knowledge graphs with entity descriptions. In Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI), pages 2659-2665.
- [Yang et al., 2015] Yang, B., Yih, W.-t., He, X., Gao, J., and Deng, L. (2015). Embedding entities and relations for learning and inference in knowledge bases. In Proceedings of the International Conference on Learning Representations (ICLR).
- [Zhang et al., 2015] Zhang, Z., Gentile, A.L., Blomqvist, E., Augenstein, I., Ciravegna, F.: An unsupervised data-driven method to discover equivalent relations in large linked datasets. *Semantic Web* 8(2), 1-27 (2015)
- [Zhou et al. 2013] Zhangquan Zhou, Guilin Qi, Chang Liu, Pascal Hitzler, Raghava Mutharaju: Scale reasoning with fuzzy-EL+ ontologies based on MapReduce. *Proceedings of the IJCAI-2013 Workshop on Weighted Logics for Artificial Intelligence*, 87-93, 2013.
- [Zhou et al. 2016] Zhangquan Zhou, Guilin Qi, Birte Glimm: Exploring Parallel Tractability of Ontology Materialization. *Proceedings of the 22nd European Conference on Artificial Intelligence (ECAI 2016)*, 73-81, 2016.
- [Zhou et al. 2017] Zhangquan Zhou, Guilin Qi: GEL: A Platform-Independent Reasoner for Parallel Classification with OWL EL Ontologies Using Graph Representation. *International Journal on Artificial Intelligence Tools* 26(1): 1-33 (2017)
- [Zhu et al., 2015] Man Zhu, Zhiqiang Gao, Jeff Z. Pan, Yuting Zhao, Ying Xu, Zhibin Quan: TBox Learning from Incomplete Data by Inference in BelNet+. *Knowledge- Based Systems*. 75: 30-40 (2015)

## 第八章 通用和领域知识图谱

### 1. 任务定义、目标和研究意义

知识图谱本质是语义网络，一种基于图的数据结构，由“节点-边-节点”组成。其中节点代表“概念”或“实体”，边则代表两个节点之间的关系，用以描述现实世界中的概念、实体以及他们之间丰富的关联关系。2012年，Google 提出知识图谱用以改善搜索体验，提高搜索质量，引起了社会各界纷纷关注。一方面各大企业开始陆续着手构建自己的知识图谱，如百度“知心”，搜狗“知立方”等，另一方面学术界对知识图谱构建技术与应用的研究也在不断地加深。

随着研究的深入，作为一种知识表示的新方法和知识管理的新思路，知识图谱不再局限于搜索引擎应用，而在其它系统如 IBM Watson 中也开始崭露头角，扮演着越来越重要的角色。但知识图谱的构建是一项庞大而复杂的工程，现阶段的知识图谱系统还远远不能满足人们的应用需求，构建一个完善的知识图谱仍然面临着诸多挑战，知识图谱在各行业中的应用场景也并不十分明确，有待探索。基于现实世界中复杂而庞大的数据量、多源异构的数据模式，如何使用知识图谱技术将其进行融合与关联，并探索更多的应用形态，这一任务被称为知识图谱的构建与应用。

本章围绕通用和领域知识分别论述各自的构建方法及应用形态。通用知识图谱可以形象地看成一个面向通用领域的“结构化的百科知识库”，其中包含了大量的现实世界中的常识性知识，覆盖面极广。领域知识图谱又叫行业知识图谱或垂直知识图谱，通常面向某一特定领域，可看成是一个“基于语义技术的行业知识库”。领域知识图谱基于行业数据构建，通常有着严格而丰富的数据模式，对该领域知识的深度、知识准确性有着更高的要求。

### 2. 研究内容和关键科学问题

由于现实世界的知识丰富多样且极其庞杂，通用知识图谱主要强调知识的广度，通常运用百科数据进行自底向上（Top-Down）的方法进行构建。而领域知识图谱面向不同的领域，其数据模式不同，应用需求也各不相同，因此没有一套通用的标准和规范来指导构建，而需要基于特定行业通过工程师与业务专家的不断交互与定制来实现。虽然如此，通用与领域知识图谱的构建与应用也并非完全没有互通之处，如下图所示，其从无到有的构建过程可分为六个阶段，被称为知识图谱的生命周期<sup>5</sup>。本节以生命周期为视角来阐述知识图谱构建过程中的研究内容及关键科学问题。

---

<sup>5</sup> 引自 CCKS2017《行业知识图谱构建与应用》



图1 领域知识图谱生命周期

## 2.1 知识建模

知识建模是建立知识图谱的概念模式的过程，相当于关系型数据库的表结构定义。为了对知识进行合理组织，更好地描述知识本身与知识之间的关联，需要对知识图谱的模式进行良好的定义。一般来说，相同的数据可以有若干种模式定义的方法，设计良好的模式可以减少数据的冗余，提高应用效率，因此在进行知识建模时，需要结合数据特点与应用特点来完成模式的定义。

知识建模通常采用两种方式：一种是自顶向下（Top-Down）的方法，即首先为知识图谱定义数据模式，数据模式从最顶层概念构建，逐步向下细化，形成结构良好的分类学层次，然后再进行将实体添加进概念。

另一种则是自底向上（Bottom-Up）的方法，即首先对实体进行归纳组织，形成底层概念，然后逐步往上抽象，形成上层概念。该方法可基于行业现有标准转换生成数据模式，也可基于高质量行业数据源映射生成。

为了保证知识图谱质量，通常在建模时需要考虑以下几点关键问题：1) 概念划分的合理性，如何描述知识体系及知识点之间的关联关系[Guarino,1995]；2) 属性定义方式，如何在冗余程度最低的条件下满足应用和可视化展现；3) 事件、时序等复杂知识表示，通过匿名节点的方法还是边属性的方法来进行描述，各自的优缺点是什么[Raisig et al,2009]；4) 后续的知识扩展难度，能否支持概念体系的变更以及属性的调整。

## 2.2 知识获取

知识获取是指从不同来源、不同数据中进行知识提取，形成知识存入到知识图谱的过程。由于真实世界中的数据类型及介质多种多样，所以如何高效、稳定地从不同的数据源进行数据接入至关重要，其会直接影响到知识图谱中数据的规模、实时性及有效性。

现有的数据源中，数据大致可分为三类：一类是结构化的数据，这类数据包

括以关系型数据库（Mysql, Oracle 等）为介质的关系型数据，以及之前提到的开放链接数据，如 Yago, Freebase 等；第二类为半结构化数据，如百科数据（Wikipedia, 百度百科等），或是垂直网站中的数据，如 IMDB, 丁香园等；第三类是以文本为代表的非结构化数据。

结构化数据中会存在一些复杂关系，针对这类关系的抽取是此类研究的重点，主要方法包括直接映射（Direct mapping）或者映射规则定义（R2RML）等；半结构化数据通常采用包装器（Wrapper）的方式对网站进行解析，包装器是一个针对目标数据源中的数据制定了抽取规则的计算机程序。包装器的定义、自动生成以及如何对包装器进行更新及维护以应对网站的变更，是当前获获取需要考虑的问题；非结构化数据抽取难度最高，如何保证抽取的准确率和覆盖率则是这类数据上进行知识获取需要考虑的科学问题。

## 2.3 知识融合

知识融合指将不同来源的知识进行对齐、合并的工作，形成全局统一的知识标识和关联。知识融合是知识图谱构建中不可缺少的一环，知识融合体现了开放链接数据中互联的思想。良好的融合方法能有效地避免信息孤岛，使得知识的连接更加稠密，提升知识应用价值，因此知识融合是构建知识图谱过程中的核心工作与研究重点问题。

知识图谱中的知识融合包含两个方面，即数据模式层的融合和数据层的融合。数据模式层的融合包含概念合并、概念上下位关系合并以及概念的属性定义合并，通常依靠专家人工构建或从可靠的结构化数据中映射生成，在映射的过程中，一般会通过设置融合规则确保数据的统一。数据层的融合包括实体合并、实体属性融合以及冲突检测与解决。

进行知识融合时需要考虑使用什么方式实现不同来源、不同形态知识的融合；如何对海量知识进行高效融合[Dong et al,2014]；如何对新增知识进行实时融合以及如何进行多语言融合等问题[Bryl et al, 2014]。

## 2.4 知识存储

知识存储，顾名思义为针对构建完成的知识图谱设计底层存储方式，完成各类知识的存储，包括基本属性知识、关联知识、事件知识、时序知识、资源类知识等。知识存储方案的优劣会直接影响到查询的效率，同时也需要结合知识应用场景进行良好设计。

目前主流的知识存储解决方案包括单一式存储和混合式存储两种。在单一存储中，可以通过三元组，属性表或者垂直分割等方式进行知识的存储[Özsu et al,2016]。其中，三元组的存储方式较为直观，但在进行连接查询时开销巨大[Harris et al,2003]。属性表指基于主语的类型划分数据表，其缺点是不利于缺失

属性的查询[Wilkinson et al,2006]。垂直分割指基于谓词进行数据的划分，其缺点是数据表过多，且写操作的代价比较大[Bobrov et al,2017]。

对于知识存储介质的选择，可以分为原生（neo4j, allegrograph 等）的和基于现有数据库（Mysql, Mongo 等）两类。原生存储的优点是其本身已经提供了较为完善的图查询语言或算法的支持，但不支持定制，灵活程度不高，对于复杂节点等极端数据情况的表现非常差。因此有了基于现有数据库的自定义方案，这样做好处是自由程度高，可以根据数据特点进行知识的划分、索引的构建等，但增加了开发和维护成本。

从上述介绍中可以得知，目前尚没有一个统一的可以实现所有类型知识存储的方式。因此，如何根据自身知识的特点选择知识存储方案，或者进行存储方案的结合，以满足针对知识的应用需要，是知识存储过程中需要解决的关键问题。

## 2.5 知识计算

知识计算是领域知识图谱能力输出的主要方式，通过知识图谱本身能力为传统的应用形态赋能，提升服务质量效率。其中图挖掘计算和知识推理是最具代表性的两种能力，如何将这两种能力与传统应用相结合是需要解决的一个关键问题。

图挖掘计算指基于图论的相关算法，实现对图谱的探索与挖掘。图计算能力可辅助传统的推荐、搜索类应用。知识图谱中的图算法一般包括图遍历、最短路径、权威节点分析、族群发现最大流算法、相似节点等，大规模图上的算法效率是图算法设计与实现的主要问题。

知识推理一般运用于知识发现，冲突与异常检测，是知识精细化工作和决策分析的主要实现方式。知识推理又可以分为基于本体的推理和基于规则的推理。一般需要依据行业应用的业务特征进行规则的定义，并基于本体结构与所定义的规则，执行推理过程，给出推理结果。知识推理的关键问题包括：大数据量下的快速推理，记忆对于增量知识和规则的快速加载[Wang et al, 2018]。

## 2.6 知识应用

知识应用指将知识图谱特有的应用形态与领域数据与业务场景相结合，助力领域业务转型。知识图谱的典型应用包括语义搜索、智能问答以及可视化决策支持三种。如何针对业务需求设计实现知识图谱应用，并基于数据特点进行优化调整，是知识图谱应用的关键研究内容。

其中，语义搜索指基于知识图谱中的知识，解决传统搜索中遇到的关键字语义多样性及语义消歧的难题，通过实体链接实现知识与文档的混合检索。语义检索需要考虑如何解决自然语言输入带来的表达多样性问题，同时需要解决语言中实体的歧义性问题。

而智能问答指针对用户输入的自然语言进行理解，从知识图谱中或目标数据

中给出用户问题的答案，其关键技术及难点包括：1) 准确的语义解析，如何正确理解用户的真实意图；2) 对于返回的答案，如何评分以确定优先级顺序。

可视化决策支持则是指通过提供统一的图形接口，结合可视化、推理、检索等，为用户提供信息获取的入口，需要考虑的关键问题包括：1) 如何通过可视化方式辅助用户快速发现业务模式；2) 如何提升可视化组件的交互友好程度，比如高效地缩放和导航；3) 大规模图环境下底层算法的效率。

### 3.技术方案和研究现状

#### 3.1 知识图谱构建方案研究

##### 3.1.1 自底向上的构建方法

通用知识图谱的构建采用自底向上的方法，主要依赖开放连接数据集和百科，从这些结构化的知识中进行自动学习，主要分为实体与概念的学习、上下位关系的学习、数据模式的学习。

开放数据链接和百科中拥有丰富的实体和概念信息，且开放链接和百科中的数据通常以一定的结构组织生成，因此，从这类数据源中抽取概念和实体较为容易。由于百科的分类体系都是经过了百科管理或是高级编辑人员的校验，其分类系统中的数据可靠性非常高，因此从百科中抽取概念和实体，通常将标题作为实体的候选，而将百科中的分类系统直接作为概念的候选。对于概念的学习，[关键,2010]提出了一种基于语言学和基于统计学的多策略概念抽取方法，该方法提高了领域内概念抽取的效果。

实体对齐的目标是将从不同百科中学习到的描述同一目标的实体或概念进行合并，再将合并后的实体集与开放链接数据中抽取的实体进行合并。实体对齐过程主要分为六步：1) 从开放链接数据集中抽取同义关系；2) 基于结构化的数据对百科中的实体进行实体对齐；3) 采用自监督的实体对齐方法对百科的文章进行对齐；4) 将百科中的实体与链接数据中的实体进行对齐；5) 基于语言学模式的方法抽取同义关系；6) 实体基于 CRF 的开放同义关系抽取方法学习同义词关系。[黄峻福,2016]提出了一种基于实体属性信息及上下文主题特征相结合进行实体对齐的方法。[万静等人,2018]提出了一种独立于模式的基于属性语义特征的实体对齐方法。

对于上下位关系，开放链接数据集中拥有明确的描述机制，针对不同的数据集编写相应的规则直接解析即可获取。百科中描述了两种上下位关系，一种是类别之间的上下位关系，对于概念的层次关系；另一种则是类别与文章之间的上下位关系，对应实体与概念之间的从属关系。实体对齐可从开放链接数据集以及百科中抽取上下位关系。[Wang,et al,2018]等人引入了弱监督学习框架来提取来自

用户生成的类别关系，并提出了一种基于模式的关系选择方法，来解决学习过程中“语义漂移”。

数据模式的学习又称为概念的属性学习，一个属性的定义包含三个部分：属性名、属性的定义域、属性的值域。但概念的属性被定义好，属于该属性的实体则默认具备此属性，填充属性的值即可。概念属性的变更会直接影响到它的实体，及其子概念以及这些概念下的实体，因此概念的属性定义十分重要，通常大部分知识库中的概念属性都是采用人工定义等方式生成，通用知识图谱则可以从开放数据集中获取概念的属性，然后从在线百科中学习实体的属性，并对实体属性进行往上规约从而生成概念的属性。在进行属性往上规约的过程中，需要通过一定的机制保证概念属性的准确性，对于那些无法自动保证准确性的属性，需要进行人工校验。[\[Su et al,2016\]](#)提出了一种新的半监督方法自动从维基百科页面自动提取属性。[\[Logan et al,2017\]](#)我们提出了多模态属性提取的任务，用来提取实体的基础属性。

### 3.1.2 自顶向下的构建方法

领域知识图谱通常采用自顶向下的方法进行构建，针对特定的行业，由该行业专家定义数据模式，进行知识建模。国内外现有可借助的建模工具以 Protégé、PlantData 为代表。Protégé<sup>6</sup>是一套基于 RDF(S)、OWL 等语义网规范的开源本体编辑器，拥有图形化界面，适用于原型构建场景。Protégé 同时提供在线版本的 WebProtégé，方便在线进行知识图谱语义本体的自动构建。PlantData<sup>7</sup>知识建模工具是一款商用知识图谱智能平台软件。该软件提供了本体概念类，关系，属性和实例的定义和编辑，屏蔽了具体的本体描述语言，用户只需在概念层次上进行领域本体模型的构建，使得建模更加便捷。

为保证可靠性，数据模式的构建基本都经过了人工校验，因此知识融合的关键任务是数据层的融合。工业界在进行知识融合时，通常在知识抽取环节中就对数据进行控制，以减少融合过程中的难度及保证数据的质量。在这些方面，工业界均做了不同角度的尝试，如 DBpedia Mapping<sup>8</sup>采用属性映射的方式进行知识融合。zhishi.me 采用离线融合的方式识别实体间的 sameAs 关系完成知识融合[\[Tianxing ,2014\]](#)，并通过双语主题模型针对中英文下知识体系进行跨语言融合[\[Wu et al,2016\]](#)。

接着需要根据数据源的不同进行知识获取，其方法主要分为三种：第一种是使用 D2R 工具，该方法主要针对结构化数据，通过 D2R 工具将关系型数据映射

---

<sup>6</sup> <http://protege.stanford.edu/>

<sup>7</sup> <http://www.plantdata.ai/>

<sup>8</sup> <http://mappings.dbpedia.org/>

为 RDF 数据。常用的开源 D2R 工具有 D2RQ<sup>9</sup>、D2R Server<sup>10</sup>、db2triples<sup>11</sup>等。D2RQ 通过 D2RQ Mapping Language 将关系型数据转化成 RDF 数据，同时支持基于该语言在关系数据上直接提供 RDF 形式的数据访问 api; D2R Server 提供对 RDF 数据的查询访问接口，以供上层的 RDF 浏览器、SPARQL 查询客户端以及传统的 HTML 浏览器调用；db2triples 支持基于 W3C 的 R2RML 和 DM 的标准将数据映射成 RDF 形式。

第二种是使用包装器，该方法主要针对半结构化数据，通过使用构建面向站点的包装器解析特定网页、标记语言文本。包装器通常需要根据目标数据源编写特定的程序，因此学者们的研究主要集中于包装器的自动生成。[Muslea et al,2001] 等人基于层次化信息抽取的思想，提出了一个包装器自动生成算法“STALKER”；[Pan et al,2002]等人开发了一个名为“Wargo”的半自动生成包装器的工具。

第三种是借助信息抽取的方法，该方法主要针对非结构化的文本。文本抽取按照抽取范围的不同可分为 OpenIE 和 CloseIE 两种。OpenIE 面向开放领域抽取信息，是一种基于语言学模式的抽取，无法实现获知待抽取知识的关系类型，通常抽取规模大，精度较低。典型工具有 ReVerb<sup>12</sup>、TextRunner<sup>13</sup>等。CloseIE 面向特定领域抽取信息，因其基于领域专业知识进行抽取，可以预先定义好抽取的关系类型，且通常规模小，精度较高。DeepDive 是 CloseIE 场景中的典型工具，其基于联合推理的算法让用户只需要关心特征本身，让开发者更多的思考特征而不是算法。

知识图谱的存储一般通过两种方式实现，分别是 RDF 存储和图数据库（Graph Database）。RDF 是语义网技术栈中的资源描述框架，常见的 RDF 存储有 AllegroGraph、Virtuoso、Jena 等。基于 RDF 的存储设计一直是热门的研究方向，[Harris et al,2003]设计了三元组表式存储 3store，[Wilkinson et al,2006]等则通过属性表的方式来完成三元组的存储，[Wang et al,2010]等提出了列式存储在 RDF 上表现要优于行式存储，并建立了六重索引来提升查询效率。图数据库是 NoSQL 中的重要代表，较为知名的图数据库有 Neo4j，Titan 等。Neo4j<sup>14</sup>是一个高度可扩展的图形数据库，将数据中的实体和关系作为顶层类型，支持节点及边上的属性操作。Titan [Chang et al,1997]是一个分布式的图数据库，支持横向扩展，支持事务，并且可以支撑上千并发用户和计算复杂图形遍历。

---

<sup>9</sup> D2RQ [EB/OL].[2014-02-26]. <http://d2rq.org/>.

<sup>10</sup> Bizer C, Cyganiak R. D2r server-publishing relational databases on the semantic web[C]//Poster at the 5th international semantic web conference. 2006, 175.

<sup>11</sup> Db2triples [EB/OL].[2014-02-26]. <http://www.antidot.net/en/ecosystem/db2triples/>.

<sup>12</sup> <http://reverb.cs.washington.edu/>

<sup>13</sup> [https://www.researchgate.net/publication/220816876\\_TextRunner\\_Open\\_Information\\_Extraction\\_on\\_the\\_Web](https://www.researchgate.net/publication/220816876_TextRunner_Open_Information_Extraction_on_the_Web)

<sup>14</sup> Developers N J. Neo4j[J]. Graph NoSQL Database [online], 2012.

## 3.2 知识图谱研究现状

### 3.2.1 通用知识图谱案例

国外的 DBpedia [Auer et al,2007]使用固定的模式从维基百科中抽取信息实体，当前拥有 127 种语言的超过两千八百万实体以及数亿 RDF 三元组；Yago [Suchanek et al,2007]则整合维基百科与 WordNet 的大规模本体，拥有 10 种语言约 459 万个实体，2400 万个事实；Babelnet [Navigli et al,2012]则采用将 WordNet 词典与 Wikipedia 百科集成的方法，构建了一个目前最大规模的多语言词典知识库，包含 271 种语言 1400 万同义词组、36.4 万词语关系和 3.8 亿链接关系。

国内的 Zhishi.me 从开放的百科数据中抽取结构化数据，当前已融合了包括百度百科、互动百科、中文维基三大百科的数据，拥有 1000 万个实体数据、一亿两千万个 RDF 三元组；以通用百科为主线，结合垂直领域的 CN-DBpedia [Xu et al,2017]，则从百科类网站的纯文本页面中提取信息，经过滤、融合、推断等操作后形成高质量的结构化数据；XLore[Wang et al,2013]则是基于中文维基百科、英文维基百科、百度百科、互动百科构建的大规模中英文知识平衡知识图谱。

### 3.2.2 领域知识图谱案例

领域知识图谱常常用来辅助各种复杂的分析应用或决策支持，如下图所示，在多个领域均有应用，不同领域的构建方案与应用形式则有所不同，本文将以电商、企业商业、图情、创投四个领域为例，从图谱构建与知识应用两个方面介绍领域知识图谱的技术构建应用与研究现状。



图2 行业知识图谱应用一览<sup>15</sup>

<sup>15</sup> 引自 2017CCKS《行业知识图谱的构建与应用》

### 3.2.2.1 电商知识图谱的构建与应用

当下，电商的交易规模巨大，对我们每个人的生活都有影响。因而电商知识图谱这个垂直图谱变得非常重要。相对于通用知识图谱，它有很多不同。首先，电商平台是围绕着商品，买卖双方在线上进行交易的平台。故而电商知识图谱的核心是商品。整个商业活动中有品牌商、平台运营、消费者、国家机构、物流商等多角色参与，相对于网页来说，数据的产生、加工、使用、反馈控制得更加严格，约束性更强。从而，如果电商数据以知识图谱的方法组织，可以从数据的生产端开始，就遵循顶层设计。电商数据的结构化程度相对于通用域来说做的更好。此外，面向不同的消费者和细分市场，不同角色，不同市场，不同平台对商品描述的侧重都不同，使得对同一个实体描述时会有不同的定义。知识融合就变得非常重要。最后，与通用知识图谱比较而言，电商知识图谱有大量的国家标准、行业规则，法律法规对商品描述进行着约束。存在大量的人的经验来描述商品做到跟消费者需求的匹配。从而，知识推理显得更为重要。以下以阿里巴巴知识图谱为例分块介绍电商知识图谱的相应技术模块和应用。

#### ● 知识建模

电商知识图谱以商品为核心，以人、货、场为主要框架。目前共涉及 9 大类一级本体和 27 大类二级本体。一级本体分别为：人、货、场、百科知识、行业竞对、品质、类目、资质和舆情。人货场构成了商品信息流通的闭环，其他本体主要给予商品更丰富的信息描述。下图描述了商品知识图谱的数据模型，数据来源包含国内-国外数据，商业-国家数据，线上-线下等多源数据。目前有百亿级的节点和百亿级的关系边。

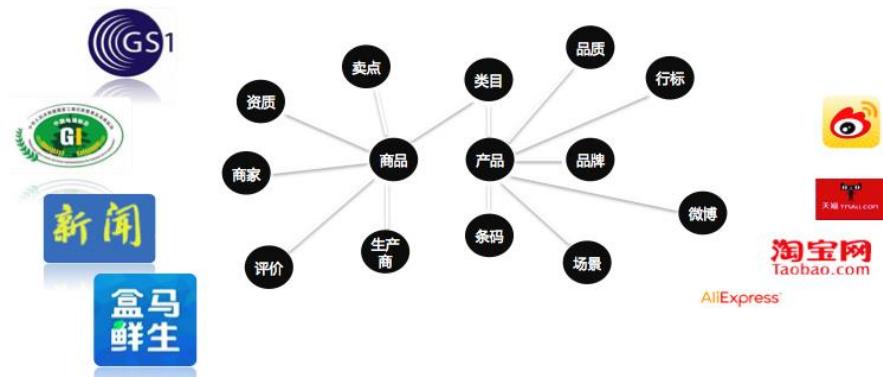


图3 电商图谱 Schema

#### ● 知识获取

电商知识图谱主要的获取来源为知识众包，这其中的关键就是知识图谱本体设计，设计上要考虑商品本身，又要考虑消费者需求和平台运营的管理方便。另一个核心工作是要开发面向电商各种角色的数据采集工具，例如面向卖家的商品发布端。

此外，电商知识的另一个来源是文本数据例如商品标题、图片、详情、评价、舆情中的品牌、型号、卖点、场景等信息。这就要求命名识别系统具有跨域大规模实体类型的识别能力，能够支持电商域数据、人机语言交互自然语言问题以及更广泛的微博、新闻等舆情域数据的识别，并且把识别出的实体与知识图谱链接，主要包括如下，特别是商品属性和属性值涉及上千类别的实体类型。

商品域：类目、产品词、品牌、商品属性、属性值、标准产品。

LBS 域：小区、超市、商场、写字楼、公司。

通用域：人物、数字、时间。

最后，对知识图谱实体描述，除了基础的属性与属性值外，很多是通过实体标签来实现的。标签相对来说变化快，易扩展。这类知识很大一部分是通过推理获得的。例如食品的标签生成中，知识推理通过食品的配料表数据和国家行业标准，如：

无糖：碳水化合物 $\leq 0.5 \text{ g}/100 \text{ g}$ （固体）或 $100 \text{ mL}$ （液体）

无盐：钠 $\leq 5 \text{ mg}/100 \text{ g}$  或 $100 \text{ mL}$

做逻辑推理。可以把配料表数据转化为“无糖”“无盐”等知识点。从而真正地把数据变成了知识标签。改善了消费者购物体验。

### ● 知识融合

电商知识图谱的知识融合，主要涉及商品和产品这两个核心节点知识融合。涉及大规模聚类、大规模实体链指、大规模层次分类等技术。主要依据商品或产品的图片、文本、属性结构化等数据。图片涉及相似图计算、OCR 等技术。

大规模层次分类需要把目标商品/产品归到上千个商品 1 级和 2 级类目中去。这里面的难度在于类目的细分和混淆度，以及大规模训练数据的生成和去噪。

大规模聚类目的是把统一数据源的信息先做一次融合。大规模实体链指核心是把知识图谱的候选实体排序，把新的实体跟知识图谱目标识别做关联。从而把新知识融入知识图谱。新知识融入工程中，涉及不同数据源属性名称和属性值的映射和标准化。这就需要大规模电商词林的建设和挖掘。

### ● 知识存储

知识图谱的存储方案的选择要考虑很多因素例如：支持的查询方式、支持的图查询路径长度，响应时间，机器成本等。通常来说，电商知识图谱的实体量比通用知识图谱的实体量要大很多。所以存储也采取多种存储方式混合的方案。

考虑的成本因素，全量的图谱数据通过离线关系数据库存储，三种表类型：实体表、关系表、类目表。为了更好的支持在线图查询和逻辑查询，与在线业务相关的知识图谱子图采用在线图数据库来存储。离线关系数据库支持向在线图数据库导入。考虑图数据的查询性能与节点路径长度关系很大。为保证毫秒级的在线响应，部分数据采用在线关系数据库来支持查询。

## ● 知识应用

电商知识图谱，这个商品‘大脑’的一个应用场景就是导购。而所谓导购，就是让消费者更容易找到他想要的东西，比如说买家输入“我需要一件漂亮的真丝丝巾”，“商品大脑”会通过语法词法分析来提取语义要点“一”、“漂亮”、“真丝”、“丝巾”这些关键词，从而帮买家搜索到合适的商品。在导购中为让发现更简单，“商品大脑”还学习了大量的行业规范与国家标准，比如说全棉、低糖、低嘌呤等。此外，它还有与时俱进的优点。“商品大脑”可以从公共媒体、专业社区的信息中识别出近期热词，跟踪热点词的变化，由运营确认是否成为热点词，这也是为什么买家在输入斩男色、禁忌之吻、流苏风等热词后，出现了自己想要的商品。最后，智能的“商品大脑”还能通过实时学习构建出场景。比如输入“海边玩买什么”，结果就会出现泳衣、游泳圈、防晒霜、沙滩裙等商品。

再者，电商平台管控从过去的“巡检”模式升级为发布端实时逐一检查。在海量的商品发布量的挑战下，最大可能地借助大数据、人工智能阻止坏人、问题商品进入电商生态。为了最大限度地保护知识产权，保护消费者权益，电商知识图谱推理引擎技术满足了智能化、自学习、毫秒级响应、可解释等更高地技术要求。实现了良好的社会效益。例如：上下位和等价推理，检索父类时，通过上下位推理把子类的对象召回，同时利用等价推理（实体的同义词、变异词、同款模型等），扩大召回。以为保护消费者我们需要拦截“产地为某核污染区域的食品”为例，推理引擎翻译为“找到产地为该区域，且属性项与“产地”同义，属性值是该区域下位实体的食品，以及与命中的食品是同款的食品”。

### 3.2.2.2 图情知识图谱的构建与应用

图情知识图谱是指聚焦某一特定细分行业，以整合行业内资源为目标的知识图谱。提供知识搜索、知识标引、决策支持等形态的知识应用，服务于行业内的从业人员，科研机构及行业决策者。

图情领域与知识图谱的结合由来已久。英国的大英博物馆通过结合语义技术对馆藏品各类数据资源进行语义组织，通过语义细化、多媒体资源标注等方式提供多样化的知识服务形式<sup>16</sup>；英国广播公司 BBC [Kobilarov et al,2009]在其音乐、体育野生动物等板块定义了知识本体，将新闻转化为机器可读的信息源（RDF / XML, JSON 和 XML）进行内容管理与报道自动生成。国内图情领域也越来越重视对知识图谱技术的利用。上海图书馆<sup>17</sup>借鉴美国国会书目框架 BibFrame [Kroeger et al,2013]对家谱、名人、手稿等资源构建知识体系，打造家谱服务平台为研究者们提供古籍循证服务；中国农科院<sup>18</sup>则聚焦于水稻细分领域，整合论文、

<sup>16</sup> 王昊奋.知识图谱概览[R].小象学院公开课,2017-10-24

<sup>17</sup> 翠娟,刘炜,陈涛,张磊.家谱关联数据服务平台的开发实践[J].中国图书馆学报,2016,42(03):27-38.

<sup>18</sup> 国家水稻数据中心. <http://www.ricedata.cn>

专利、新闻等行业资源，构建水稻知识图谱，为科研工作者提供了行业专业知识服务平台。

为了使读者对图情知识图谱有更清晰的认识，以下章节笔者将介绍图谱知识图谱一般的构建过程，并展示图情知识图谱中典型的应用场景。

### ● 知识建模

图情知识图谱的构建一般采用自顶向下的方式进行知识建模，通常从资源类型数据入手，整理出资源的发表者（人物），发表机构（机构），关键词（知识点），发表载体（刊物）等等类型的实体及各自之间的关系，同时通过人物、机构的主页进行实体属性的扩充。下图是一张典型的图情知识图谱 schema 模型，展示了概念与概念间的关系以及部分属性。

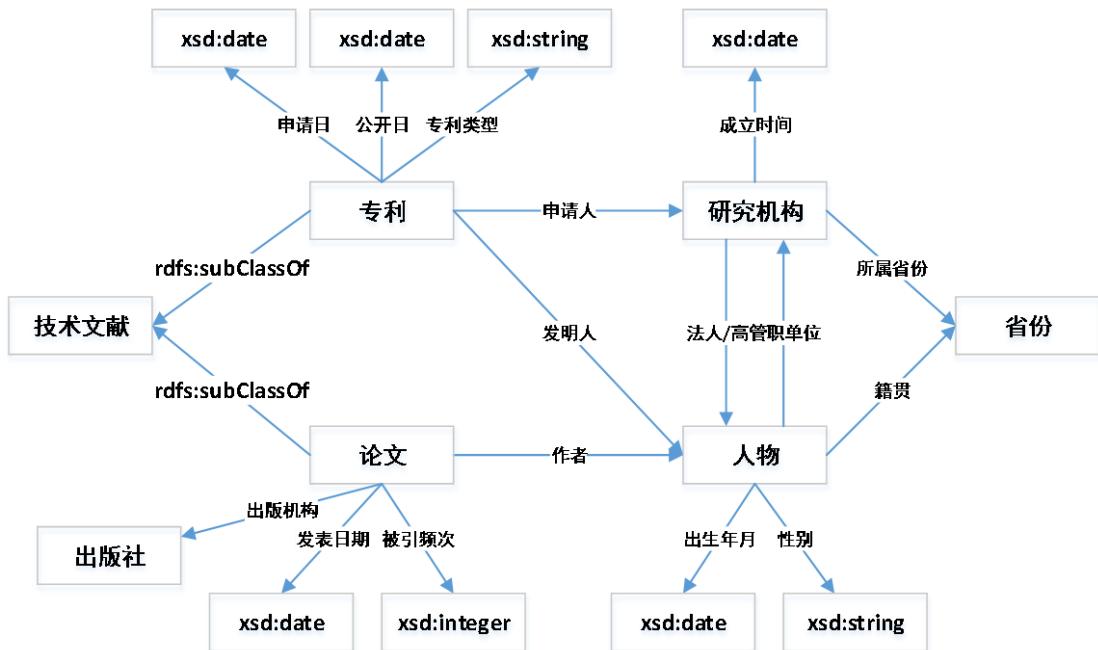


图4 图情行业典型 Schema 模型

### ● 知识获取

图情领域的数据源主要包括四类。第一类是知网、专利局等的文献类网站，第二类是开放通用数据，包括百科类网站以及 DBpedia 等的开放链接数据集，第三类是行业垂直的新闻门户，第四类是行业内企业和科研机构内部积累的既有数据。知识获取的方法视数据类型而异，具体可参考本章 2.2 节的介绍。

### ● 知识融合

图情领域的知识融合需要考虑实体层面的融合以及知识体系的融合。对于实体融合，主要解决不同来源实体的属性缺失、冲突等问题，一般采用多数投票的方式来进行实体属性的对齐。对于多知识体系的融合，通常确定置信度最高的体系作为基准，如专利的 IPC 分类，继而将其他来源的知识点进行对齐。由于知识体系的质量影响到了整个知识图谱的知识描述能力与准确性，所以一般允许较多的人工介入来进行体系的融合梳理。

## ● 知识存储

图情知识图谱的存储设计时需要兼顾实体、概念等图谱数据与论文、新闻等资源类型数据。对于图谱数据，推荐使用基于 RDF 的存储，如 AllegroGraph, Jena 等，对数据中的语义描述有着天然的支持，能更快的实现语义搜索等应用。对于资源数据，则可以使用面向搜索设计的数据库，如 Elasticsearch, Solr 等，以获得更好的搜索支持。

## ● 知识计算

图情领域中的知识计算主要包括图论算法、知识统计以及知识推理。通过实现基本图论算法来辅助进行各类业务分析。如：通过图遍历算法进行机构合作的谱系分析；基于社区发现算法寻找学术研究热点；图排序算法帮助进行权威分析等。通过统计学方法能帮助进行宏观层面的分析，如行业发展趋势，机构研究分布等。通过知识推理完成新知识的补充，如专家合作关系，公司上下游关系等。

## ● 知识应用

图情知识图谱的典型应用包括知识搜索、知识标引、决策支持等，下面逐一进行介绍。

知识搜索是图情领域的基础性服务，而知识图谱技术可以从准确性和形态上为其赋能。图谱中的实体识别技术能够提高搜索的命中率，同时允许用户通过自然语言的方式进行知识的语义搜索。而通过知识卡片、知识推荐等结果的返回也可以提升用户的交互体验。

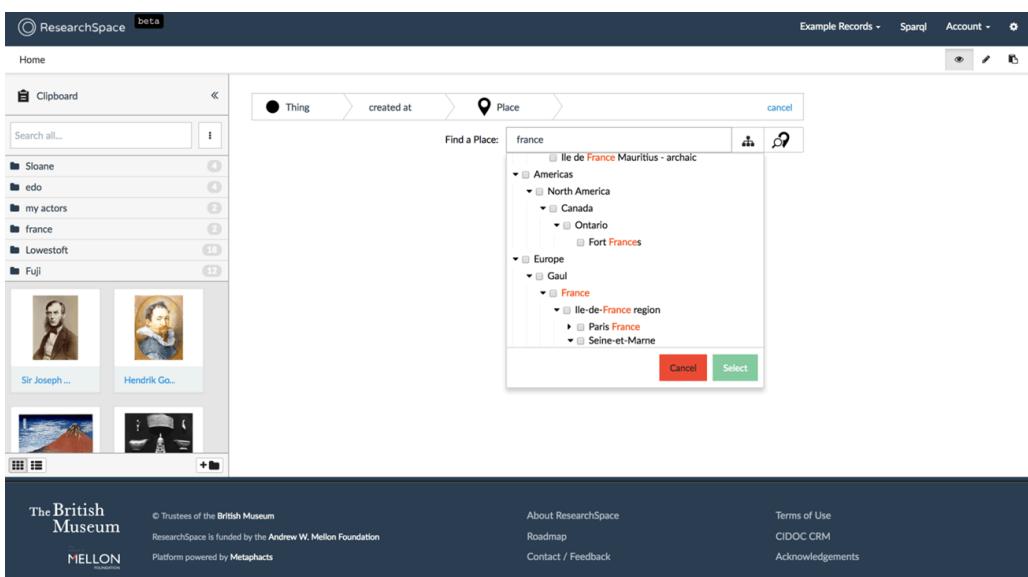


图5 大英博物院语义搜索

## ● 知识标引

知识标引指的是根据构建完成的图情知识图谱，对新闻、文献等文本的内容进行知识标注的过程。知识标引既是图谱构建过程中的重要工作，同时是图谱应用的一种形态，可以依托标引技术打造在线的阅读工具，或者集成 office、pdf

reader 等文档类应用，提供知识卡片、知识推荐等服务辅助终端用户阅读。

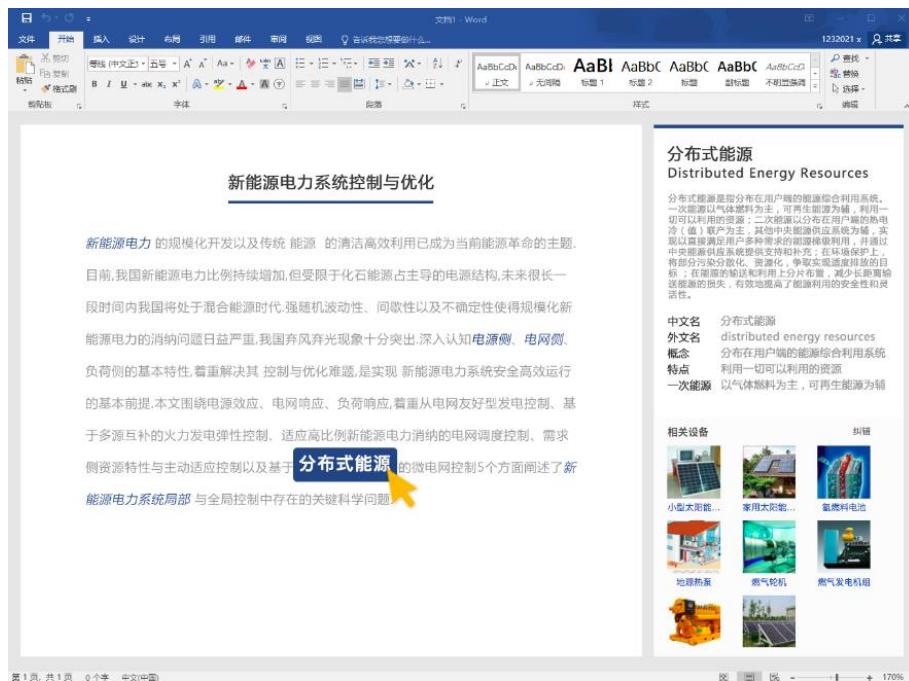


图6 基于知识标引的辅助阅读

决策支持基于路径分析、关联分析、节点聚类等图算法进行辅助分析，并通过图谱可视化的方式展示知识间的关联。可以对关联参数，如步长，过滤条件等，以及可视化的形态、如节点颜色、大小、距离等进行定制，从而为可视化决策支持赋予不同的业务含义。以下是几个典型的可视化决策支持场景。

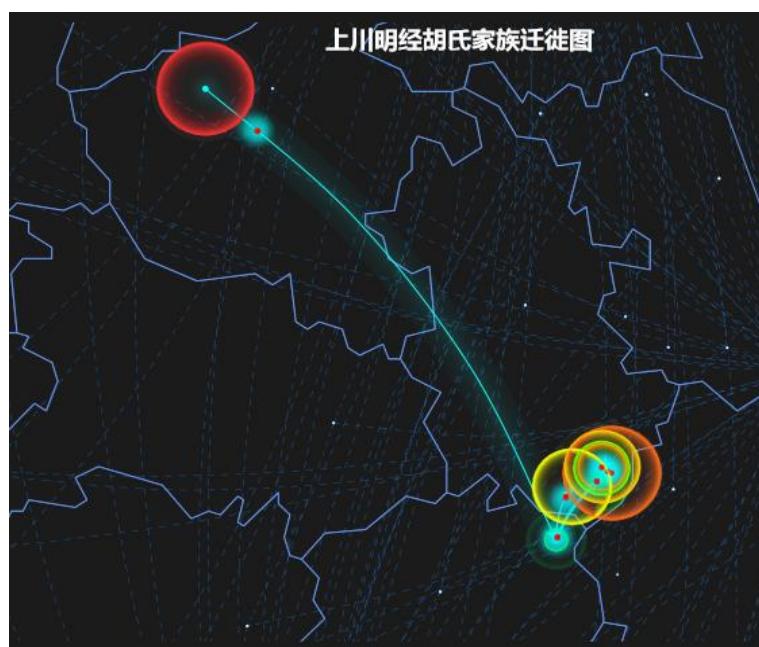


图7 上川明胡氏家族迁徙图

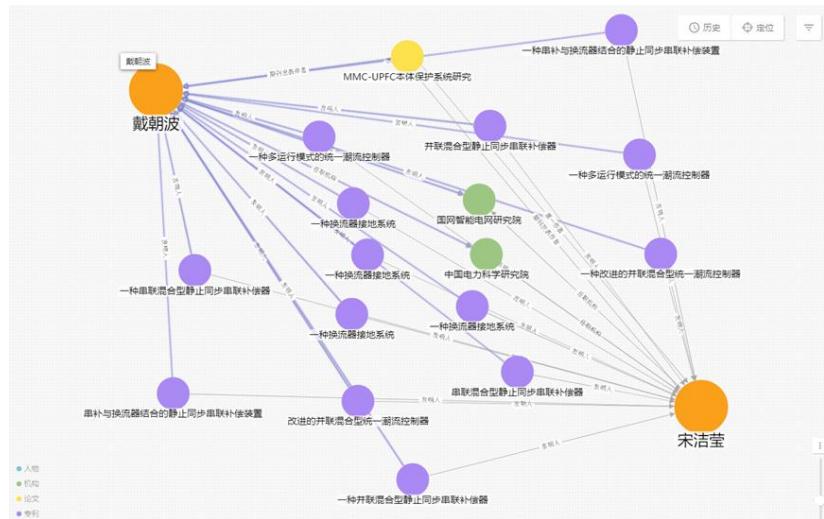


图8 专家合作分析

### 3.2.2.3 企业商业知识图谱的构建与应用

丰富多维度的企业信息在基本面分析中十分重要，当前全国企业总量超过三千万家，数量十分庞大，数据多源，需要构建统一的企业商业知识图谱，企业商业知识图谱企业、人物、专利等信息构成，关注企业与人物之间任职及股权关系、专利与企业人物是所属权关系，以完善企业及个人画像，助力企业潜在客户获取、客户背景调查、多层次研究报告、风险管控；辅助发现不良资产、企业风险、非法集资等。

量子魔镜<sup>19</sup>以全国全量企业的全景数据资源为研究基础打造企业信用风险洞察平台，天眼查<sup>20</sup>、启信宝<sup>21</sup>则专注服务于个人与企业信息查询工具，为用户提供企业、工商、信用等相关信息的查询。企查查<sup>22</sup>立足于企业征信，通过深度学习、特征抽取以及知识图谱技术对相关信息进行整合，并向用户提供数据信息。中信建投将全国企业知识图谱整合进客户关系管理系统中，构建全面清晰的客户视图，以实现高效客户关系管理。笔者将企业商业知识图谱的构建方式梳理如下，以供读者借鉴与参考。

#### ● 知识建模

构建企业商业知识图谱，通常关注企业与人物的任职关系、投资关系、分支机构关系。

从相应网站中抽取企业信息、人物形象、诉讼信息以及信用信息，再添加上市公司、股票等概念和相应属性。企业招投标信息、上市公司的股票信息可从相关网站进行采集。企业的竞争关系、并购事件则从百科站点中进行抽取。这些信息存在于信息框、列表、表格等半结构化数据，以及无结构的纯文本中。其 schema

<sup>19</sup> <https://www.datathea.com/>

<sup>20</sup> <https://www.tianyancha.com/>

<sup>21</sup> <http://www.qixin.com/>

<sup>22</sup> <http://www.qichacha.com/>

模式如下图所示。

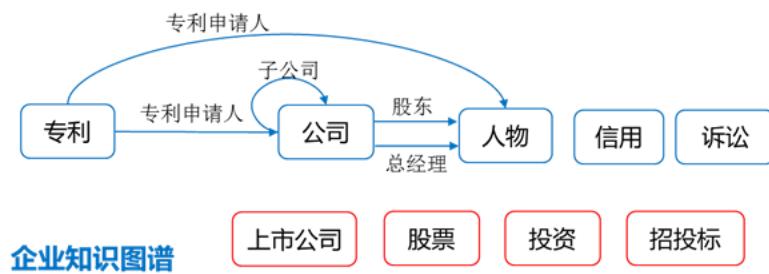


图9 企业商业知识图谱

#### ● 知识获取

企业商业知识图谱数据源主要包含两大类：1) 半结构化的网页数据，其中包括全国企业信用信息公示系统、中国裁判文书网、中国执行信息公开网、国家知识产权局、商标局、版权局等。2) 文本数据，如招投标信息公告、法律文书、新闻、企业年报等。通过 D2R 工具、包装器、文本信息抽取等方式对以上数据分别进行抽取。

#### ● 知识融合

企业知识图谱的数据来源多种多样，因此需要在数据层对数据进行融合。数据层的集成主要是对公司和人物两类实体进行融合，目标是解决由企业全称与简称产生的描述不一致的问题，以及人物重名问题。公司的融合推荐基于公司名的全称进行链接，人物实例的融合则推荐使用基于启发式规则进行集成。

#### ● 知识存储

全国企业商业知识图谱包含全国上千万家企业信息，十亿级别的三元组，形成知识图谱庞大而复杂，因此对存储方式提出了挑战，要求能够对海量的图数据进行存储，且具有良好的可伸缩性和灵活性，对此我们推荐采用图数据库的方式进行存储，并可以扩展分布式存储方案以提高服务可用性与稳定性。

#### ● 知识计算

企业商业知识图谱中的图计算主要集中在知识推理的计算，从而应用于金融反欺诈、辅助信贷审核的功能。当前知识图谱中的推理主要是基于规则进行推理计算。如：在金融反欺诈中，多个借款人联系方式属性相同，但地址属性不同，则可通过不一致性验证的方式来判断借款人是否有欺诈风险。

#### ● 知识应用

全国企业知识图谱通过异常关联挖掘、企业风险评估、关联探索、最终控制人和战略发展等方式为行业客户提供智能服务和风险管理。

异常关联挖掘是通过路径分析、关联探索等操作，挖掘目标企业谱系中的异常关联。基于企业商业知识图谱从多维度构建数据模型，进行全方位的企业风险评估，有效规避潜在的经营风险与资金风险。

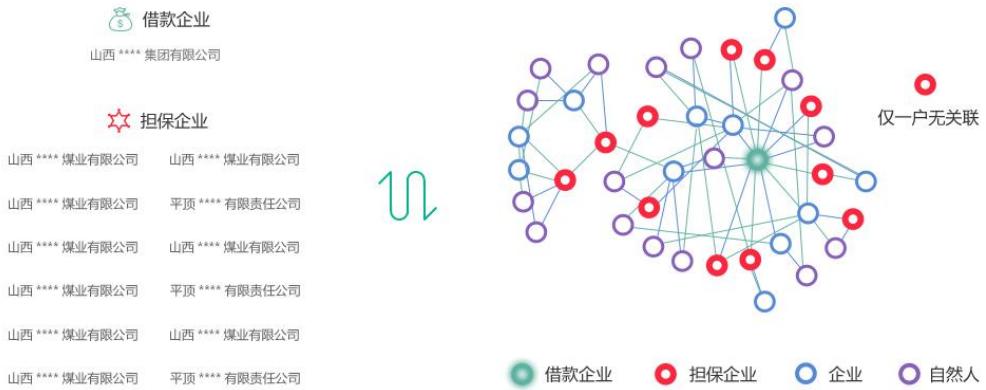


图10 异常关联挖掘

最终控制人是基于股权投资关系寻找持股比例最大的股东，最终追溯至自然人或国有资产管理部门。

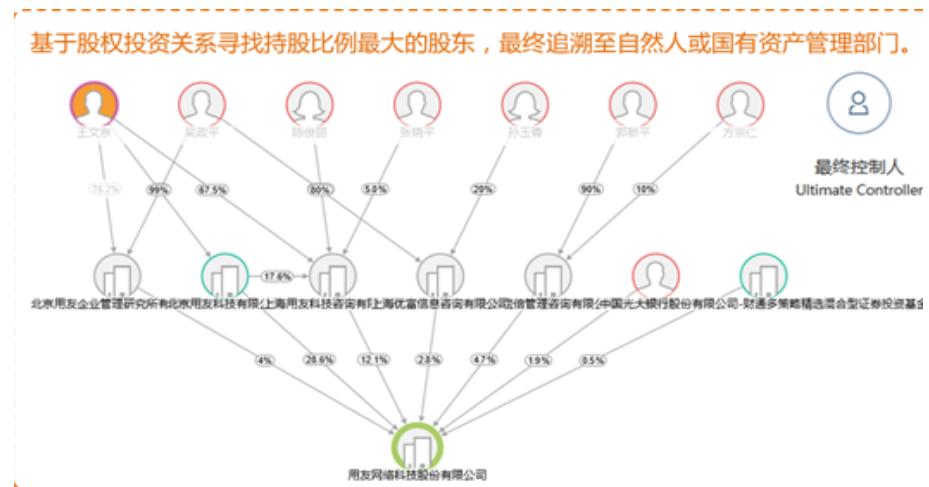


图11 最终控制人分析

战略发展则以“信任圈”的形式将目标企业的对外投资企业从股权上加以区分，探寻其全资、控股、合营、参股的股权结构及发展战略，从而理解竞争对手和行业企业的真实战略，发现投资行业结构、区域结构、风险结构、年龄结构等。

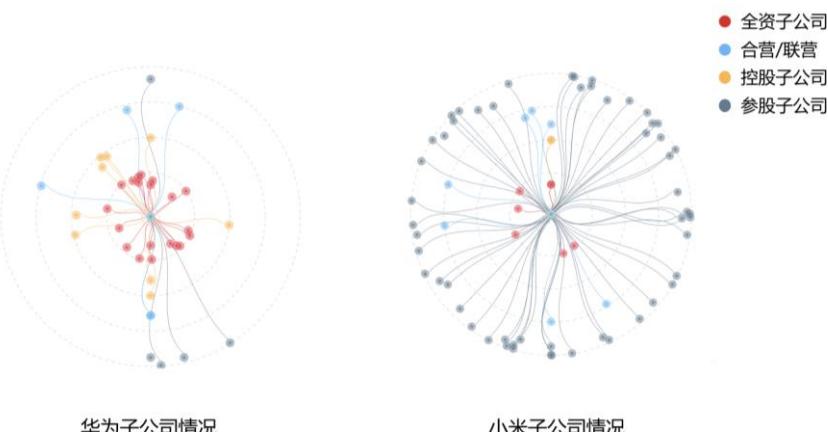


图12 企业社交图谱

### 3.2.2.4 创投知识图谱的构建与应用

创投是创业投资的简称。创投知识图谱聚焦于工商知识图谱的一部分数据内容，旨在展现企业、投融资事件、投资机构之间的关系。据 IT 桔子的不完全统计，截止 2018 年 2 月，全国拥有初创公司超过 9 万家，投资机构超过 6 千家，9 万多名的创业者，投资事件超过 4 万起。

作为公司发展过程中的重要阶段，创投领域的发展正得到越来越多数据与技术公司的关注。2007 年在美国旧金山创立的 Crunchbase<sup>23</sup>，其核心业务是围绕初创公司及投资机构的生态为企业提供数据服务。国内企业中 TechNode 于 2017 年发布了数据棱镜平台<sup>24</sup>，构建创投知识图谱，为专业人员提供创业投资数据分析工具；因果树<sup>25</sup>是一家人工智能股权投资服务平台，依托大数据和人工智能技术，提升一级市场效率，推动一级市场量化。

依照惯例，笔者将创投知识图谱的构建方法与典型应用形态进行了总结，在后续章节中为大家一一呈现。

#### ● 知识建模

要定义创投知识图谱 schema，首先要理解创投领域的相关概念跟关系。创投领域 Schema 中涉及的概念主要包括初创公司、投资机构、投资人、公司高管、行业以及投融资事件等。融资事件是创投领域的核心，不同于实体节点，融资事件描述的是一个事实，具有抽象性。典型的创投 schema 如下图所示：

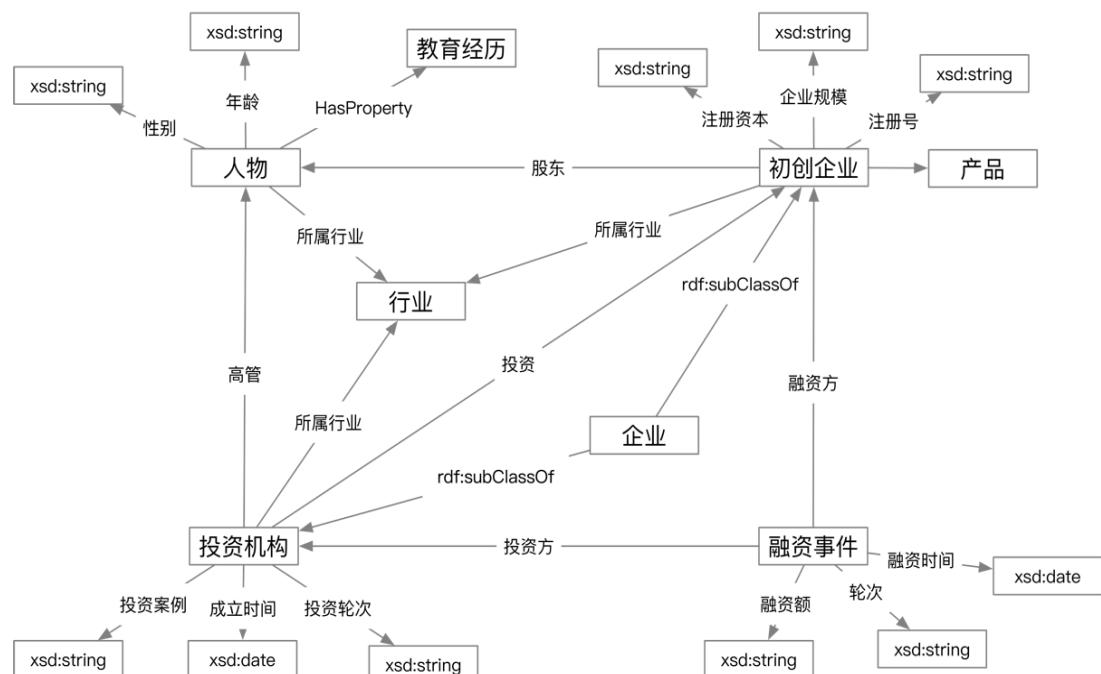


图13 创投 schema

<sup>23</sup> <https://www.crunchbase.com/>

<sup>24</sup> <http://lengjing.io/>

<sup>25</sup> <https://www.innotree.cn/>

### ● 知识获取

创投数据主要来源于虎嗅、IT 桔子、36Kr 等科技型媒体网站。IT 桔子是以融资事件为核心的的关注 IT 互联网行业的结构化的公司数据库和商业信息服务提供商，其中包含了各类结构化的投资机构库和融资讯息。虎嗅和 36Kr 则主要是以商业科技资讯为主的新闻数据来源网站。

### ● 知识融合

构建创投知识图谱时同样需要考虑数据融合的问题，典型问题包括：1) 数值属性表示不一致，例如金额的阿拉伯数字与中文写法区别；2) 实体同义，例如企业的全称与简称；3) 不同数据源中数据冲突。一般采用先实体对齐后属性对齐的方法来进行融合操作。

### ● 知识存储

创投知识图谱的存储主要考虑融资事件的存储设计，通常采用两种方式对此类信息进行存储。第一种是在传统三元组的基础上加入其它描述字段进行时间、轮次等信息存储；第二种方式是，通过匿名节点对事件进行存储，把时间、地点等相关信息作为事件节点的属性。对于融资事件来说，虽然它不是客观世界中一个具体的事物，但它包含了丰富的属性信息，如融资时间、融资轮次、融资额等。因此比较适合单独引入一类节点来进行存储和表示。

### ● 知识计算

创投知识图谱的知识计算主要集中于使用社区发现、基于图的排序、最短路径等图算法对合作分析、时序、相似公司等应用进行能力输出。例如，通过最短路径算法辅助合作分析；基于社区发现算法寻找行业研究热点；图排序算法帮助进行权威分析等；通过分析展现公司的发展情况。

### ● 知识应用

创投领域知识图谱主要应用形态包括知识检索以及可视化决策支持。

知识检索依托创投知识图谱可以在原有知识全文搜索的基础上实现语义搜索与智能问答的应用形态。其中，语义搜索提供自然语言式的搜索方式，由机器完成用户搜索意图识别。而作为知识搜索的终极形态，智能问答允许用户通过对话的方式对领域内知识进行问答交互，同时通过配置问题模板实现复杂业务问题的回答。



图14 语义搜索

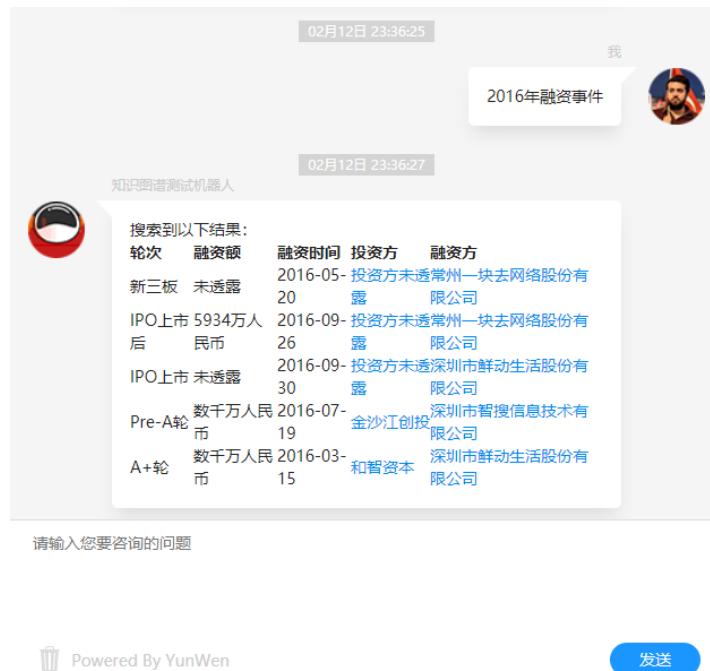


图15 智能问答

决策支持通过图谱可视化技术可对创投图谱中的初创公司发展情况，投资机构投资偏好等进行解读，通过节点探索、路径发现、关联探寻等可视化分析技术展示公司的全方位信息，通过知识地图、时序图谱等形态对地理分布、发展趋势等进行解读，为投融资决策提供支持。



图16 融资事件地图

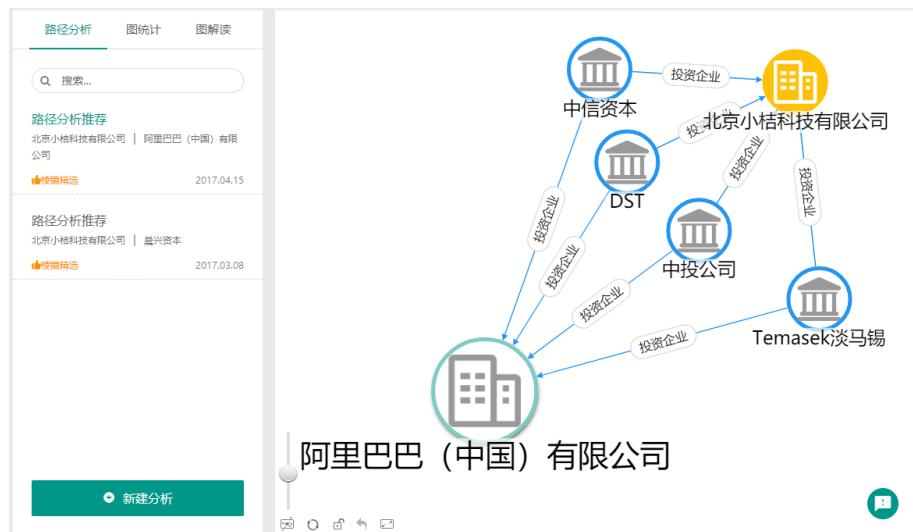


图17 路径分析

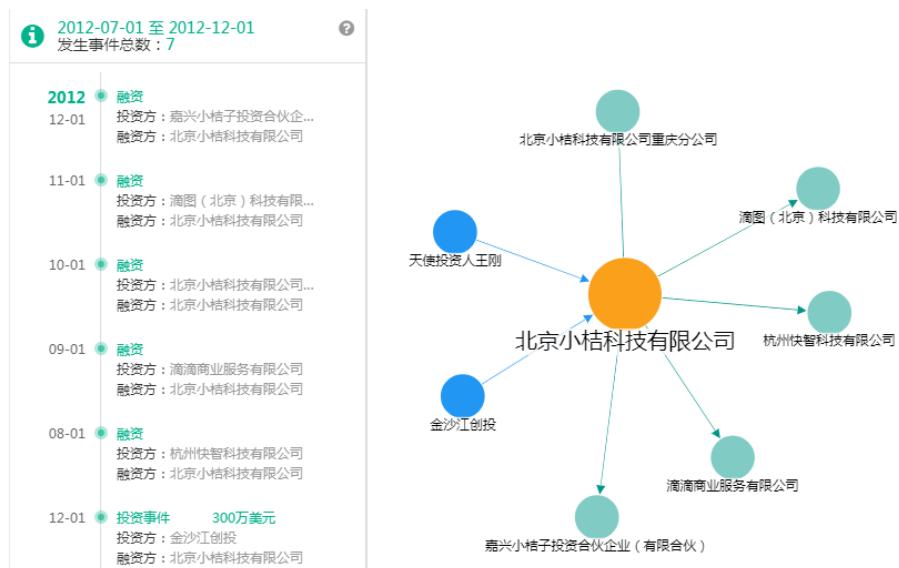


图18 时序分析

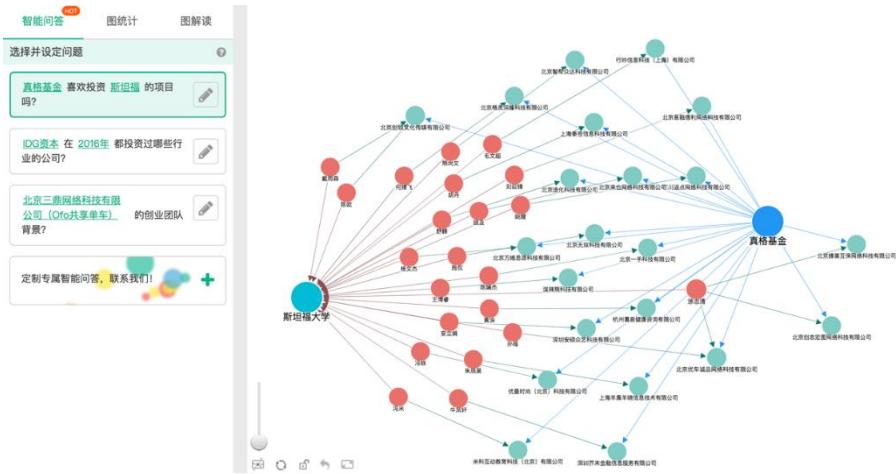


图19 自然语言 BI

## 4. 技术展望和发展趋势

结合知识图谱研究发展态势，并结合当前知识图谱的构建与应用未来现状，对知识图谱未来技术发展及趋势发展做一个展望。

### ● 知识图谱构建

现阶段，基于本体工程的知识描述和表示仍是知识图谱建模的主流方法，而且仅仅用到了一些 RDFS 及 OWL 中定义的基础元属性来完成知识图谱模式层构建，知识图谱所关注的重点也仍然是数据中的概念、实体、属性等。随着人们对知识的认知层次的提升，势必会对现有的知识表示方法进行扩展，逐步扩充对于时序知识、空间知识[Zhuang et al, 2017]、事件知识[Hernes et al, 2018]等的表示。而知识图谱本身也会逐步将关注重点转移到时序、位置、事件等动态知识中去，来更有效地描述事物发展的变化，为预测类的应用形态提供支持。

其次，对于知识图谱构建任务来说，最困难最无法标准化实现的一个环节就是对于文本数据的信息抽取。知识图谱面向开放领域的信息抽取普遍存在着召回率低、算法准确性低、限制条件多、拓展性差等问题。随着计算机计算能力的日益提高与深度学习技术不断研究发展，NLP 领域发生了翻天覆地的变化，CNN、RNN 等经典神经网络结构已经被应用于 NLP 中，尝试完成机器翻译、命名实体识别任务。未来，深度学习的思想和方法会越来越多的应用于文本信息抽取中，优化的抽取方式，提高知识的覆盖率与准确率[Xu et al, 2017] [Londhe et al, 2017]。其他如跨语言知识融合[Wu et al, 2017]，知识嵌入[Wang et al, 2017]等方向也会在深度学习技术的加持下迸发新的研究浪潮。

### ● 知识图谱应用

知识图谱应用方面，未来将会出现更多应用形态，如基于知识图谱的智能文本编制，通过知识图谱将行业中的业务知识与文档相结合，在文档编制过程中，进行实时的智能提示、知识校验、知识生产等，辅助文档编制。又如基于知识图

谱的自然语言理解与自然语言生成，通过知识图谱对知识的建模能力，结合深度学习对知识的学习与抽象能力，实现以自然语言形式进行输入和输出的下一代问答系统。随着知识表示技术和推理技术的发展，结合一些新型的可视化方法，我们还可以展望一些预测分析类的应用形态，如疾病预测、行情预测、政治意识形态检测[Chen et al,2017]、城市人流动线分析 [Zhuang et al,2017]。除此之外，知识图谱在辅助多媒体数据处理方面也是一个有待深入研究的方向，如物体检测[Fang et al,2017]、图像理解[Ruimao et al,2017]等。

总之，知识图谱作为人工智能技术中的知识容器和孵化器，会对未来 AI 领域的发展起到关键性的作用。无论是通用知识图谱还是领域知识图谱，其构建技术的发展和对应用场景的探索仍然会不断的持续下去。知识图谱技术不单指某一项具体的技术，而是从知识表示、抽取、存储、计算、应用等一系列技术的集合。随着这些相关技术的发展，我们有理由相信，知识图谱构建技术会朝着越来越自动化方向前进，同时知识图谱也会在越来越多的领域找到能够真正落地的应用场景，在各行各业中解放生产力，助力业务转型。

## 参考文献

- [Auer et al,2007] Auer S, Bizer C, Kobilarov G, et al. DBpedia: ANucleus for a Web of Open Data. [C]// The SemanticWeb, International Semantic Web Conference, AsianSemantic Web Conference, ISWC 2007 + Aswc 2007, Busan, Korea, November. DBLP, 2007:722-735.
- [Bobrov et al,2017] Bobrov N, Chernishev G, Novikov B. Workload-independent data-driven vertical partitioning[C]//Advances in Databases and Information Systems. Springer, Cham, 2017: 275-284.
- [Bryl et al, 2014] Bryl V, Bizer C, Isele R, et al. Interlinking and knowledge fusion[M]//Linked Open Data--Creating Knowledge Out of Interlinked Data. Springer, Cham, 2014: 70-89.
- [Chen et al,2017] Chen W, Zhang X, Wang T, et al. Opinion-aware Knowledge Graph for Political Ideology Detection[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 2017: 3647-3653.
- [Cui et al,2009] G. Cui, Q. Lu, W. Li, and Y. Chen. Mining Concepts from Wikipedia for Ontology Construction. Web Intelligence and Intelligent Agent Technologies. 2009(3): 287 - 290.
- [Dong et al,2014] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion[J]. Proceedings of the VLDB Endowment, 2014, 7(10): 881-892.

- [Fang et al,2017] Fang Y, Kuan K, Lin J, et al. Object detection meets knowledge graphs[J]. 2017.
- [Guarino,1995] Guarino N. Formal ontology, conceptual analysis and knowledge representation[J]. International journal of human-computer studies, 1995, 43(5-6): 625-640.
- [Harris et al,2003] Harris S, Gibbins N. 3store: Efficient bulk RDF storage[J]. 2003.
- [Hernes et al, 2018] Hernes M, Bytniewski A. Knowledge Representation of Cognitive Agents Processing the Economy Events[C]//Asian Conference on Intelligent Information and Database Systems. Springer, Cham, 2018: 392-401.
- implementation[J]. 2006.
- [ Kobilarov et al,2009] Kobilarov G. et al. (2009) Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In: Aroyo L. et al. (eds) The Semantic Web: Research and Applications. ESWC 2009. Lecture Notes in Computer Science, vol 5554. Springer, Berlin, Heidelberg
- [Kroeger et al,2013] Kroeger, Angela. "The road to BIBFRAME: the evolution of the idea of bibliographic transition into a post-MARC Future." Cataloging & classification quarterly 51.8 (2013): 873-890.
- [Logan et al,2017] Logan I V, Robert L, Humeau S, et al. Multimodal Attribute Extraction[J]. arXiv preprint arXiv:1711.11118, 2017.
- [Londhe et al,2017] Londhe S N, Shah S. A novel approach for knowledge extraction from Artificial Neural Networks[J]. ISH Journal of Hydraulic Engineering, 2017: 1-13.
- [Muslea et al,2001] I. Muslea, S. Minton, and C. Knoblock. Hierarchical Wrapper Induction for Semistructured Information Sources. J. Autonomous Agents and Multi-Agent systems. 2001, 4: 93-114.
- [Navigli et al,2012] Navigli, Roberto, and Simone Paolo Ponzetto. "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." Artificial Intelligence 193 (2012): 217-250.
- [Özsu et al,2016] Özsu M T. A survey of RDF data management systems[J]. Frontiers of Computer Science, 2016, 10(3): 418-432.
- [Pan et al,2002] A. Pan, J. Raposo, M. Álvarez, J. Hidalgo, and Á. Viña. Semi-Automatic Wrapper Generation for Commercial Web Sources. J. Engineering Information Systems in the Internet Context. 2002, 103: 265-283.

- [Raisig et al,2009] Raisig S, Welke T, Hagendorf H, et al. Insights into knowledge representation: The influence of amodal and perceptual variables on event knowledge retrieval from memory[J]. *Cognitive science*, 2009, 33(7): 1252-1266.
- [Ruimao et al,2017] Ruimao Z, Jiefeng P, Yang W, et al. The Semantic Knowledge Embedded Deep Representation Learning and Its Applications on Visual Understanding[J]. *Journal of Computer Research and Development*, 2017, 6: 010.
- [Suchanek et al,2007] Suchanek F M, Kasneci G, Weikum G. Yago: A Core of Semantic Knowledge[C]// International Conference on World Wide Web. 2007:697-706.
- [Su F, et al,2017] Su F, Rong C, Huang Q, et al. Attribute extracting from Wikipedia pages in domain automatically[M]//Information Technology and Intelligent Transportation Systems. Springer, Cham, 2017: 433-440.
- [Tianxing et al ,2014] Tianxing Wu, Guilin Qi, Haofen Wang. Zhishi.schema Explorer: A Plaborm for Exploring Chinese Linked Open Schema. *Semantic Web and Web Science* 2014: 174-181
- [Titan et al,1997] Chang C, Moon B, Acharya A, et al. Titan: a high-performance remote-sensing database[C]//Data Engineering, 1997. Proceedings. 13th International Conference on. IEEE, 1997: 375-384.
- [Wang et al,2010] Wang X, Wang S, Du P, et al. Storing and indexing RDF data in a column-oriented DBMS[C]//Database Technology and Applications (DBTA), 2010 2nd International Workshop on. IEEE, 2010: 1-4.
- [Wang et al,2013] Wang Z, Li J, Wang Z, et al. XLoRe: A Large-scale English-Chinese Bilingual Knowledge Graph[C]//International semantic web conference (Posters & Demos). 2013, 1035: 121-124.
- [Wang et al, 2017] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: A survey of approaches and applications[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(12): 2724-2743.
- [Wang, et al,2018] Wang C, Fan Y, He X, et al. Predicting hypernym–hyponym relations for Chinese taxonomy learning[J]. *Knowledge and Information Systems*, 2018: 1-26.
- [Wang et al, 2018] Wang S, Wan J, Li D, et al. Knowledge Reasoning with Semantic Data for Real-Time Data Processing in Smart Factory[J]. *Sensors*, 2018, 18(2): 471.
- [Wilkinson et al,2006] Wilkinson K, Wilkinson K. Jena property table implementation[J]. 2006.

- [Wu et al,2016] Wu T, Qi G, Wang H, et al. Cross-Lingual Taxonomy Alignment with Bilingual Biterm Topic Model[C]//AAAI. 2016: 287-293.
- [Wu et al,2017] Wu T, Zhang D, Zhang L, et al. Cross-Lingual Taxonomy Alignment with Bilingual Knowledge Graph Embeddings[C]//Joint International Semantic Technology Conference. Springer, Cham, 2017: 251-258.
- [Xu et al, 2017] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.
- [Xu et al,2017] Xu B, Xu Y, Liang J, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[C]//International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 2017: 428-438.
- [Zhuang et al, 2017] Zhuang C, Yuan N J, Song R, et al. Understanding people lifestyles: construction of urban movement knowledge graph from GPS trajectory[C]//Proceedings of the 26th International Joint Conference on Artificial Intelligence. AAAI Press, 2017: 3616-3623.
- [Zirn et al,2008] C. Zirn, V. Nastase, and M. Strube. Distinguishing between instances and classes in the wikipedia taxonomy. In: Proceedings of the 5th European semantic web conference on the semantic web, 2008: 376-387.
- [万静等,2018]万静, 李琳, 严欢春, 等. 基于 VS-Adaboost 的实体对齐方法[J]. 北京化工大学学报 (自然科学版), 2018, 45(1): 72-77.
- [关键,2010]关键.面向中文文本本体学习概念抽取的研究[J].吉林:吉林大学,2010
- [黄峻福,2016]黄峻福.中文 RDF 知识库构建问题研究与应用[D].西南交通大学, 2016.

# 第九章 语义集成

## 1. 任务定义、目标和研究意义

语义网是下一代万维网的发展方向。目前，万维网创始人 Tim Berners-Lee 先生倡导的链接数据理念以及相应的链接开放数据项目赢得了世界范围的广泛支持。2012 年 5 月 Google 正式提出的知识图谱概念也受到了各界的热烈响应，出现了一批以 DBpedia、Freebase、Wikidata、YAGO 等为代表的大型知识库，汇集了数百亿条 RDF 三元组，涵盖了包括社交网络、生物医学、地理、政务、电影、音乐等众多方向，其中蕴含的海量知识被广泛用于语义搜索、智能问答、个性化推荐等领域。

知识图谱可以由任何机构和个人自由构建，其背后的数据来源广泛、质量参差不齐，导致它们之间存在多样性和异构性。例如，对于相交领域（甚至是相同领域），通常会存在多个不同的实体指称真实世界中的相同事物。语义集成的目标就是将不同知识图谱融合为一个统一、一致、简洁的形式，为使用不同知识图谱的应用程序间的交互建立互操作性。常用技术方法包括本体匹配（也称为本体映射）、实例匹配（也称为实体对齐、对象共指消解）以及知识融合等。

语义集成是知识图谱研究中的一个核心问题，对于链接数据和知识融合至关重要。语义集成研究有助于提升基于知识图谱的信息服务水平和智能化程度，推动语义网以及人工智能、数据库、自然语言处理等相关领域的研究发展，具有重要的理论价值和广泛的应用前景，可以创造巨大的社会和经济效益。

## 2. 研究内容和关键科学问题

图 1 展示了一个语义集成的常见流程，主要包括：输入、预处理、匹配、知识融合和输出 5 个环节。一般来说，系统或框架型的研究工作通常涵盖了图中的大部分内容，而方法或算法型的研究工作更侧重图中的某个局部。下面将根据这 5 个环节分别详细介绍主要内容和关键科学问题。

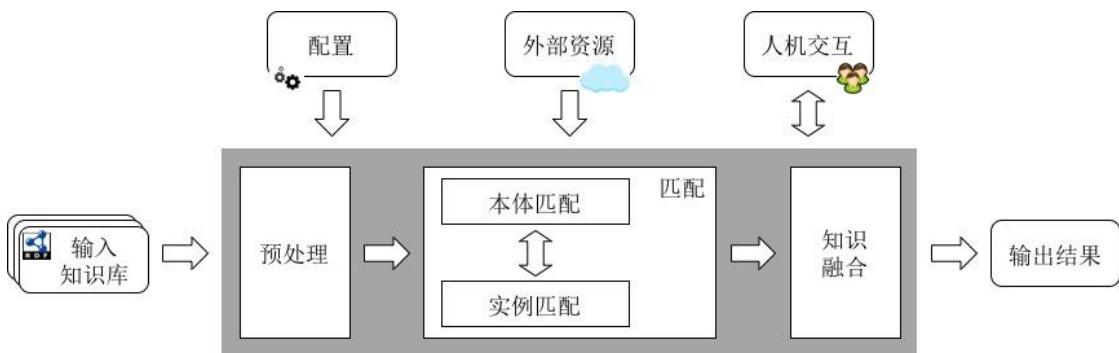


图 1. 语义集成的常见流程

## 2.1 语义集成的输入

语义集成的输入包括待集成的若干个知识库以及配置、外部资源等。待集成的知识库常见为两个，但也有一些工作支持输入更多的知识库，这些工作通过基于全局的优化方法来获得更好的结果。待集成的知识库格式一般为 RDF/OWL 数据文件或 SPARQL 端点（endpoint）。虽然 SPARQL 端点更加灵活且易于扩展，但是远程访问可能带来性能和数据规模上的限制。输入的配置通常包括需要预设的参数、阈值、规则等。除通过手动配置外，一些工作也使用了自动配置以减轻用户手动配置的难度。

外部资源可以被认为是语义集成过程中使用到的背景知识，例如字/辞典背景知识（例如 WordNet）、常识背景知识（例如 Cyc）、实时背景知识（例如搜索引擎）等。另一方面，语义集成过程中也可能涉及人机交互，例如雇佣人来对部分数据或结果进行标注。这些标注数据对语义集成的多个环节中均有作用，例如帮助训练匹配器或者提升融合的效果。众包（crowdsourcing）和主动学习是经常使用的技术，使用它们可以更为高效地利用昂贵的人力资源。

## 2.2 预处理

预处理主要包括预先对输入知识库进行清洗和后续步骤的准备。清洗主要是为了解决输入质量问题，与自由文本不同，知识库通常基于 RDF/OWL 语言构建，质量较好，并且现有工具已经可以较为全面地解析、调试知识库。因此，在大多数语义集成工作中清洗过程都较为简单、通用。

后续步骤的准备分为配置和数据两方面。针对配置的准备，一些工作使用遗传算法自动生成适合输入知识库的集成规则，还有一些工作通过分析输入知识库自适应或者使用无监督学习计算出合适的模型参数。

针对数据的准备通常使用索引技术以提高后续环节的处理速度和规模。分块（blocking）是一项被广泛使用的技术，通过对索引的设计，可以避免例如匹配环节达到知识库规模平方级的复杂度。就传统数据库而言，索引是对数据预先排序得到的一种存储结构，使用索引可以快速访问数据库中的记录。知识库分块则是通过索引键值将知识库中的元素划分成一组不相交或相交的区块，后续仅考虑对应区块间的匹配和融合。这种方法剪枝过滤掉完全无关的区块对，从而提高集成效率。这里的一个关键科学问题是对于区块大小和数量的权衡，在尽量不丢失可能结果的情况下使分块尽可能的小。

## 2.3 匹配

根据匹配对象的不同，匹配一般分为本体匹配和实例匹配两方面。本体匹配侧重发现（模式层）等价或相似的类、属性或关系，而实例匹配侧重发现指称真实世界相同对象的不同实例。本体匹配和实例匹配间也可相互影响，例如基于实

例匹配的本体匹配。如何从语义上消解知识库间的异构性是匹配环节待解决的关键科学问题。

文本相似性度量是发现匹配的最基础方法。无论是本体匹配还是实例匹配，文本相似性度量方法均较为相似，大致分为四种类型：基于字符的（例如 Levenshtein 编辑距离）、基于单词的（例如 Jaccard 系数）、混合型（例如 soft TF-IDF）和基于语义的（例如 WordNet）。

由于知识库通常可以表示为一个节点和边均带标签的有向图结构，因此可以基于图结构进行匹配。从是否利用图结构上下文信息这一点来看，匹配方法可以分为两种：成对匹配和集体（collective）匹配。成对匹配又叫基于元素的匹配，这类方法中匹配之间不会相互影响。集体匹配又叫基于结构的匹配，这类方法会根据匹配之间的影响调整匹配的相似度，也是目前的研究焦点。一般而言，由于本体的图规模较小，并且结构更丰富，因此一些复杂的集体匹配方法多用在本体匹配上；而实例匹配更多使用一些简单或者局部的集体匹配方法以保证运行时间在一个可以接受的范围内。第 3 节将具体介绍一些代表性方法。

## 2.4 知识融合

在匹配的基础上，知识融合一般通过冲突检测、真值发现等技术消解知识集成过程中的冲突，再对知识进行关联与合并，最终形成一个一致的结果。如何处理冲突是知识融合环节的主要研究问题。

目前常见的冲突处理策略分为以下 3 类：冲突忽略、冲突避免和冲突消解。冲突忽略选择忽略知识冲突，将冲突交给用户解决。冲突避免承认冲突的存在，但是不解决冲突，对于所有情形使用统一的规则，例如不同知识来源具有不同的优先级。冲突消解聚焦于如何根据知识本身和元数据的特征来消解冲突，这也是目前的主流研究方向。冲突消解方法主要可以分为 3 类：第一类是基于投票的方法，例如采用简单的多数投票（majority voting）策略。第二类是基于质量的方法，这种方法在投票过程中考虑知识来源的可信度，从而推出高质量的结果。常见的方法主要是基于概率模型的，第 3 节将介绍一些最新研究进展。第三类基于关系的方法，这种方法在基于质量的方法上考虑不同知识来源之间的关系，关系会影响投票或可信度评估的结果。

## 2.5 语义集成的输出

语义集成的输出是一个统一的、一致的、简洁的知识库。注意，这个知识库可以是虚构形式。除此以外，输出还可以包括结果和过程的图形化展示。

# 3. 技术方法和研究现状

语义集成领域已经经历过了一段时间的蓬勃发展，随着大数据时代的来临，

大数据背景下的语义集成已成为当下的研究热点。同时，表示学习（representation learning）技术在图像、视频、语音、自然语言处理等领域的成功也引起了语义集成领域研究人员的注意，如何将表示学习技术运用于语义集成过程成为新的热点。由于篇幅有限，下面仅具体介绍语义集成方向的近期研究动向和一些代表性技术方法。

### 3.1 本体匹配

#### 3.1.1 多本体全体匹配

伴随链接数据的蓬勃发展，本体的数量越来越多。现有大多数本体匹配方法处理的是成对的本体，但是成对匹配方法在同时匹配多个本体时会产生一些问题，最主要的问题是它们得到的结果从全局看可能存在冲突。因此，针对多个本体的全体匹配（holistic matching）研究变得越发关键。

LPHOM [Megdiche et al., 2016] 是一种多本体全体匹配方法，该方法在匹配多个本体的同时还能保证其结果是全局最优解。LPHOM 将全体匹配问题建模成基于最大权图匹配的线性规划问题，在这之上增加了 4 种针对本体匹配问题的一般性约束，这些约束用于保证匹配结果的一致性。约束具体包括：每个类和属性仅能参与最多一个匹配（1:1 规则）；类和属性的匹配要满足本体中定义的不相交关系；对象属性的匹配与其定义域类、值域类的匹配结果相互制约；数据属性的匹配与其定义域类的匹配结果相互制约。

#### 3.1.2 跨语言本体匹配

随着多语言知识库的发展，跨语言本体匹配方法的重要性已经凸显。由于语言不同，跨语言本体匹配相较一般本体匹配更为困难，特别是影响文本相似性度量的准确性。也正因为如此，一些模型与方法的优势在跨语言场景得以体现。近几年来，较有代表性的工作包括：

EAFG [Zhang et al., 2017] 是一个用于解决跨语言属性匹配问题的因子图模型，该模型同时考虑了属性对自身的特征和属性对之间的相关性。图 2 展示了如何将属性匹配问题转化到 EAFG 模型。左侧是关系图，表示不同语言维基百科 K1 和 K2 间的多种关系。其中，属性层包含了属性和模板间关系，文章层包含了文章和目录间关系。两层间的虚线表示文章和属性之间的使用关系，红色虚线表示已存在的跨语言链接。右侧是因子图，其中白色节点表示变量，黑色节点表示因子。变量分为两种：观察变量  $x_i$ ，对应每个候选对；隐变量  $y_i$ ，表示观察变量  $x_i$  的一个布尔标签（是否匹配）。因子共有 3 种，各自对应一种特征函数，用于将关系转为可计算的特征：局部特征、模板特征和同义词特征分别对应因子  $f$ 、 $g$  和  $h$ 。根据这 3 种特征定义联合概率分布函数，通过最大化联合概率确定匹配。

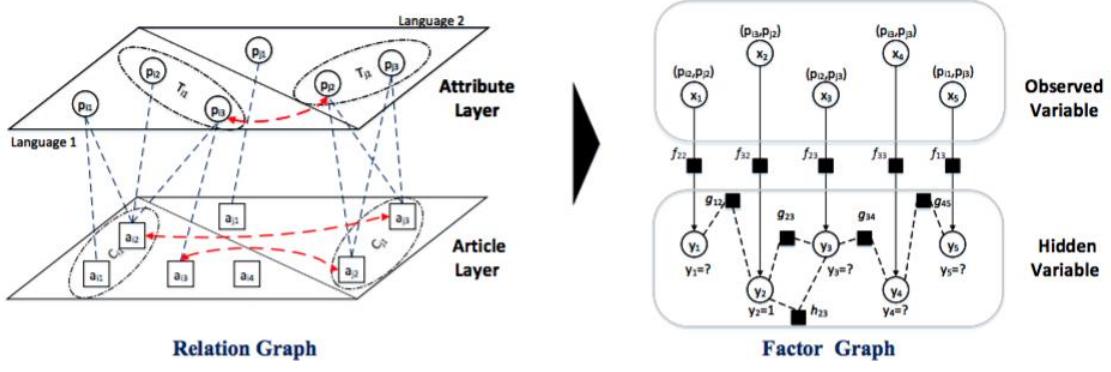


图 2. 跨语言属性匹配模型 EAFG

双语主题模型也被用于解决跨语言本体匹配问题。在匹配的过程中，首先使用常规方法获得候选匹配对，之后使用双语主题模型从匹配对象的文本上下文中获得其主题分布，从而在相同主题空间内表示不同语言的匹配对象。主题向量的余弦相似度被作为得分用于确定最终的匹配。BiBTM [Wu et al., 2016] 是一个近期工作，其建模对象是双语文档对中出现的无序单词对，称为 bitem。在双语文档对中，任何两个不同的单词均会组成一个 bitem。一个 bitem 的生成过程被定义为：首先，从所有 bitem 的全局主题分布中选择一个主题；然后，根据对应语言的主题-单词分布选择一个单词；最后，推导上下文的具体主题则是通过计算上下文中生成 bitem 的主题概率的期望值实现。其改进版本 CC-BiBTM [Wu et al., 2017] 进一步加入了类之间的单词共现关系和层次结构关系。

### 3.2 实例匹配

#### 3.2.1 基于人机协作的实例匹配

众包和主动学习等人机协作方法是目前实例匹配的研究热点。这些方法雇佣普通用户，通过付出较小的人工代价来获得丰富的先验数据，从而提高匹配模型的性能。

Hike [Zhuang et al., 2017] 是一个解决大规模知识库间实例匹配的众包方法，其框架流程如图 3 所示。该方法为实例对之间定义偏序关系，根据构建的偏序结构和已知问题答案可以推断未知问题答案。之后基于假设定义实例对和问题集的推断期望，根据推断期望选出最佳问题分发给众包参与者。

与通过众包直接解决大规模实例匹配不同，链接发现工具 Silk [Isele & Bizer, 2013] 和 LIMES [Ngonga Ngomo & Lyko, 2012] 均通过结合主动学习和遗传算法来生成链接规约 (link specification)。链接规约由以下两种操作组合得到：求值操作和相似度操作。求值操作根据输入的实例输出一组值，例如取出一组特定属性或者对特定属性、数据做小写、分词等变换处理；相似度操作则是针对输入的一对实例求得或者聚合相似度。Silk 和 LIMES 将学习到的链接规约组织成树的结构。在向用户提问的策略上，LIMES 选择能够最大化投票熵的候选，而 Silk 则

选择能够最小化信息增益的候选。

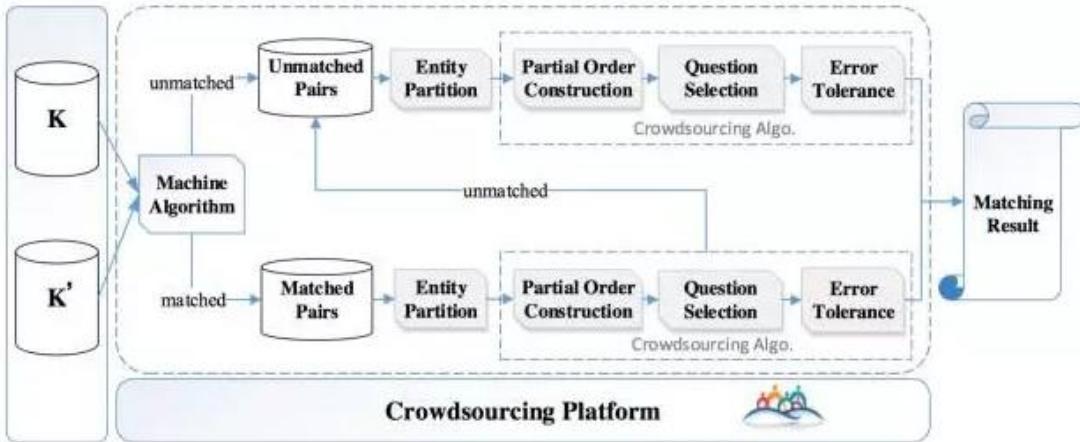


图 3. 基于众包的实例匹配方法 Hike

### 3.2.2 基于表示学习的实例匹配

随着表示学习技术在诸如图像、视频、语音、自然语言处理等领域的成功，一些研究人员开始着手研究面向知识图谱的表示学习技术，将实体、关系等转换成一个低维空间中的实值向量（即分布式语义表示），并在知识图谱补全、知识库问答等应用中取得了不错的效果。

MTransE [Chen et al., 2017] 通过基于转移的方法解决跨语言知识图谱的表示学习和匹配问题。它首先使用 TransE [Bordes et al., 2013] 对单个知识图谱进行表示学习，然后学习不同知识表示空间的线性变换来进行实例匹配。MTransE 包含了 3 种不同的转移方法：轴标定法，转移向量法和线性变换法。通过使用不同的损失函数，MTransE 一共设计了 5 种不同变种。

IPTransE [Zhu et al., 2017] 和 JAPE [Sun et al., 2017] 基于先验实例匹配，使用联合表示学习技术直接将不同知识图谱中的实体和关系嵌入到统一的向量空间中，将不同知识图谱中实例间的匹配过程转换为计算它们的向量表示间距离的过程。IPTransE 使用迭代的方式不断更新实例匹配，而 JAPE 则利用了属性及文字描述信息来增强实例的表示学习。

### 3.2.3 基于强化学习的实例匹配

近年来，强化学习取得了一系列进展，如何在语义集成中运用强化学习逐渐成为新的动向。ALEX [El-Roby & Aboulnaga, 2015] 是一个通过利用用户提供的查询答案反馈来提高实例匹配质量的系统，其整体结构如图 4 所示。ALEX 从一组候选链接开始，搜索与用户在先前查询反馈中批准的匹配相似的匹配，以发现新的匹配。具体而言，ALEX 使用蒙特卡罗强化学习方法来学习如何在某个匹配周边进行探索。其中，将每个匹配视作一个状态，用户反馈被转换为行为奖励，通过最大化收集到的行为奖励改善策略。

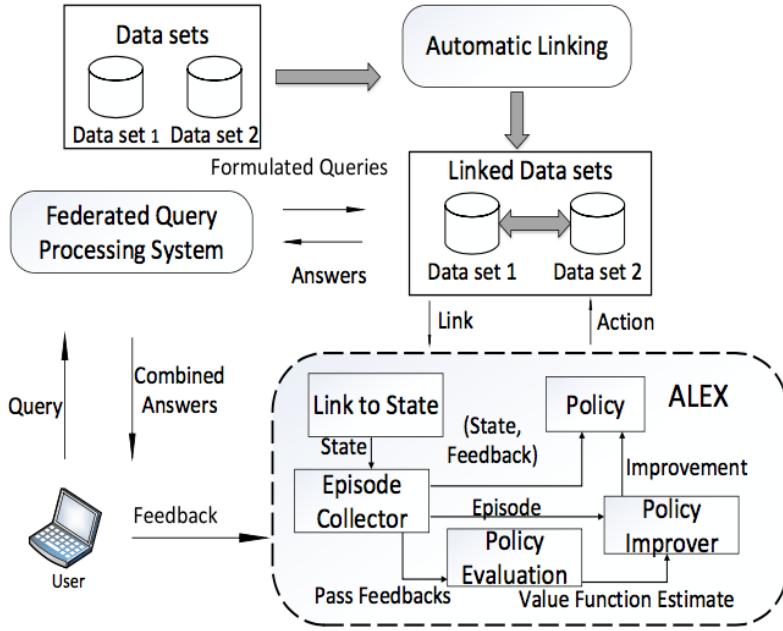


图 4. 基于强化学习的实例匹配方法 ALEX

### 3.3 知识融合

早期的知识融合主要是借鉴传统数据融合的方法，近几年比较流行的冲突消解方法是基于图模型的方法，这类方法通常为每条知识分配一个概率，将冲突消解问题看作图中节点的概率预测问题，通过知识之间已有的关系或现有的外部知识来预测不同来源知识可能的真值。

文献[Dong et al., 2015a]从传统数据融合方法中挑选出了一些最新方法，并针对知识融合的数据特征对这些方法做出改进，将其用于知识融合。主要改进包括将数据源与相应的知识抽取工具配对，共同作为数据融合任务中的一个数据源，数据融合结果的真假二值也被拓展为输出概率。文献[Wang et al., 2015]提出了一种将不同搜索引擎得到的知识卡片进行融合的方法，与文献[Dong et al., 2015a]类似，该方法也是将知识融合问题进行转化之后，再应用数据融合技术进行求解。不同的是，该方法提出了一种新的消歧概率评分算法，并设计了一种基于学习的方法来匹配代表相同对象的知识卡片的属性。与上述借鉴传统数据融合的方法不同，文献[Dong et al., 2015b]针对不同知识抽取工具的结果正确性存在差异这一问题，提出了使用多层概率模型对数据源准确性和知识正确性进行联合推断。

长尾 (long tail) 实体是指那些拥有三元组数量很少的实体。虽然现有很多知识库中已经存在数以千万计的三元组，然而实验表明其中长尾实体仍占了相当大的比重。由于长尾实体本身拥有的信息很少，例如没有足够的文本描述或语义关系，传统的基于属性相似度或者结构相似性的方法可能效果较差。FACTY [Li et al., 2017] 框架是一个用于在知识融合过程中进行知识验证的框架。为了解决长

尾实体缺少信息的问题，FACTY 从已有知识库、互联网和搜索引擎查询日志中搜集能够支持输入知识的证据，并且分析证据识别正确的知识。

### 3.4 语义集成评测

标准的评测数据集对于语义集成的研究十分重要，这些数据集提供了一个横向比较各种方法优劣的平台。随着语义集成研究的蓬勃发展，用于语义集成评测的数据集也有了一些变化。OAEI<sup>26</sup> (*ontology alignment evaluation initiative*) 是一个评测和比较本体匹配的比赛，如今也包含了除本体匹配外的其他评测任务，近期的一些变化如下：

- Ontology alignment for query answering (oa4qa)是 2014 年、2015 年增加的新任务。该任务不使用经典的本体匹配评估方法。参与的系统需要生成本体匹配来回答一组基于本体的问题，该任务根据其回答问题的能力计算精度和召回率。
- Disease and phenotype 是 2016 年增加的新任务，该任务关注疾病本体、表型本体的匹配。这些子任务可以帮助理解疾病、表型与基因的关系。
- Process model matching 是 2016 年增加的新任务。该任务源自 PMMC (*process model matching contest*)，关注流程模型的匹配。流程模型最早由 BPML (*business process modeling language*)语言表示，在 OAEI 中被转换为本体模型，由此产生的任务是实例匹配的一种特例。
- HOBBIT link discovery 是 2017 年增加的新任务。该任务旨在测试基于字符串和拓扑方法的链接发现工具的性能。该任务针对结果质量和时间性能来评估不同的框架，其中时间性能是之前 OAEI 很少考虑的。
- 另外，OAEI 也取消了 2 项评测任务：Library（2015 年起取消）和 Benchmark（2017 年起取消）。Library 任务针对经济学词库 STW 和社会学词库 TheSoz 进行匹配，这些词库常被图书馆用于检索。该任务旨在评估匹配系统是否能够处理这些包含大量概念和附加描述的轻量级本体。Benchmark 是最早的本体匹配任务之一，其目标是提供一个覆盖广泛的、不断改进的、稳定的标准测试集。

## 4. 技术展望与发展趋势

语义集成在过去几年里得到了广泛而深入的研究，未来可能的研究方向包括：

首先，表示学习是近年来的研究热点，未来仍会延续这一发展趋势。当前，针对语义集成的表示学习研究均是以 TransE 系列模型为基础构建的，这些模型大多以知识图谱补全为目标，目前缺少在模型层面针对语义集成问题设计的表示

---

<sup>26</sup> <http://oaei.ontologymatching.org/>

学习方法，导致现有基于表示学习的语义集成方法精度不高。未来研究面临许多挑战，例如如何有效利用复杂的属性及文字描述信息以及各种外部知识。

其次，人机协作可以有效提高语义集成的效果，目前已经得到较多关注。人机协作方法目前的研究主要集中于如何提出标注问题和利用已有标注。未来除了延续这一趋势以外可能还会关注其他实际问题，例如，如何将标注问题合适地展示给用户，以提高其标注质量；如何区分用户对不同领域的熟悉程度，评估其针对不同领域的标注质量。

最后，目前针对语义集成的评测数据集主要分为两类：一类是人造数据集，例如 OAEI 提供的多个数据集；另一类是真实世界中的数据集，例如 DBpedia、Freebase 等。这两类数据集均存在一定的问题：OAEI 数据集的优点是质量较高，但是其数据规模偏小，无法较好地评测大规模语义集成方法；真实世界数据集虽然数据规模很大，但针对其建立高质量的参考标准却十分困难。基于以上原因，迫切需要建立一个通用的大规模评测数据集。

## 参考文献

- [Bordes et al., 2013] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko. TransE: Translating embeddings for modeling multi-relational data. In: Proceedings of NIPS, 2787-2795, 2013
- [Chen et al., 2017] M. Chen, Y. Tian, M. Yang, C. Zaniolo. MTransE: Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In: Proceedings of IJCAI, 1511-1517, 2017
- [Dong et al., 2015a] X.L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, W. Zhang. From data fusion to knowledge fusion. Proceedings of the VLDB Endowment, 7(10):881-892, 2015
- [Dong et al., 2015b] X.L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaressi, S. Sun, W. Zhang. Knowledge-based trust: estimating the trustworthiness of web sources. Proceedings of the VLDB Endowment, 8(9), 938-949, 2015
- [El-Roby & Aboulnaga, 2015] A. El-Roby, A. Aboulnaga. ALEX: Automatic link exploration in linked data. In: Proceedings of SIGMOD, 1839-1853, 2015
- [Isele & Bizer, 2013] R. Isele, C. Bizer. Active learning of expressive linkage rules using genetic programming. Journal of Web Semantics, 23:2-15, 2013
- [Li et al., 2017] F. Li, X.L. Dong, A. Langen A, Y. Li. Knowledge verification for long-tail verticals. Proceedings of the VLDB Endowment, 10(11):1370-1381, 2017
- [Megdiche et al., 2016] I. Megdiche, O. Teste, C. Trojahn. An extensible linear

- approach for holistic ontology matching. In: Proceedings of ISWC, 393-410, 2016
- [Ngonga Ngomo & Lyko, 2012] A.-C. Ngonga Ngomo, K. Lyko. Eagle: Efficient active learning of link specifications using genetic programming. In: Proceedings of ESWC, 149-163, 2012
- [Sun et al., 2017] Z. Sun, W. Hu, C. Li. Cross-lingual entity alignment via joint attribute-preserving embedding. In: Proceedings of ISWC, 628-644, 2017
- [Wang et al., 2015] H. Wang, Z. Fang, L. Zhang, J.Z. Pan, T. Ruan. Effective online knowledge graph fusion. In: Proceedings of ISWC, 286-302, 2015
- [Wu et al., 2016] T. Wu, G. Qi, H. Wang, K. Xu, X. Cui. Cross-lingual taxonomy alignment with bilingual biterm topic model. In: Proceedings of AAAI, 287-293, 2016
- [Wu et al., 2017] T. Wu, L. Zhang, G. Qi, X. Cui, K. Xu. Encoding category correlations into bilingual topic modeling for cross-lingual taxonomy alignment. In: Proceedings of ISWC, 728-744, 2017
- [Zhang et al., 2017] Y. Zhang, T. Paradis, L. Hou, J. Li, J. Zhang, H. Zheng. Cross-lingual infobox alignment in wikipedia using entity-attribute factor graph. In: Proceedings of ISWC, 745-760, 2017
- [Zhu et al., 2017] H. Zhu, R. Xie, Z. Liu, M. Sun. IPTransE: Iterative entity alignment via joint knowledge embeddings. In: Proceedings of IJCAI, 4258-4264, 2017
- [Zhuang et al., 2017] Y. Zhuang, G. Li, Z. Zhong, J. Feng. Hike: A hybrid human-machine method for entity Alignment in large-scale knowledge bases. In: Proceedings of CIKM, 1917-1926, 2017

# 第十章 语义搜索

## 1. 任务定义、目标和研究意义

随着 Internet 的爆炸性增长，万维网已经发展成为包含多种信息资源、站点遍布全球的巨大动态信息服务网络，为用户提供了一个极具价值的信息源。然而，传统搜索技术仍以关键词匹配、倒排索引和网页的链接结构为搜索依据，其查全率和查准率均无法满足用户日益提高的标准 [Arvind, et al., 2001] [Guo, et al., 2003] [Zhang, et al., 2007]。与传统搜索技术不同，语义搜索是指搜索引擎的工作不再拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句背后的真实意图，并依此来进行搜索，从而更准确地向用户返回最符合其需求的搜索结果。

语义搜索和传统的搜索系统有很大不同。以 Google 为例，Google 的传统搜索主要依据网站中是否存在关键词、有多少其他网站链向这个网站、用户的点击率等其他各种因素来决定呈现什么结果。Google 其实并不知道搜索词的含义。比如当你在 Google 中搜索“中国最大的城市”，Google 给你呈现的是一页包含这些关键词的链接，Google 并不知道这个问题的真正含义。相反，借助于语义网和知识图谱，语义搜索能直接给出这个问题的答案，而不是一页相关的链接。可以肯定的说，未来的搜索将会超越搜索词本身，进入由各种实体 (entities)、实体的属性和实体的相互关系所组成的世界。语义搜索的目的即是借助于对实体的理解，它们之间的交互行为，用户对这些实体的理解获取准确的答案而不是一条条链接，通过利用语义技术，将推理结合到检索过程中，可以极大的提高当前的搜索效果，在语义 Web 环境下，可以更高效地发现信息资源。

## 2. 研究内容和关键科学问题

事实上，传统搜索技术提升的困难并不是因为目前的搜索技术本身不够先进，其根源存在于 Web 上现有的信息表达和组织方式过于简单。Web 上的网页使用的 HTML 语言或其改进版本，通过 DOM 树描述了网页的结构和格式，并引入包括图片、声音以及视频等多种媒体格式，使得信息的显示更加生动、形象。此外，相关文档之间可以采用超链接互相定向。然而，这种信息的表达和组织方法主要是为人们阅读服务的，对于计算机而言，缺失了 Web 页面所承载的语义信息。比如，某个 Web 页面中说明“小米 Note3 6GB 手机的价格是 2500 元人民币”。HTML 语言难以使得计算机理解：小米是一个移动设备的制造公司，Note3 是其生产的一款手机型号，6GB 是手机的内存容量，2650 的含义是手机的销售价格，其单位是人民币。

目前有不少针对自然语言理解的研究，希望通过自动的方式将自然语言的语义转换为计算机可以理解的表达方式，但是当前的研究进展无论是处理的效率还是结果的精确度都不能达到令人满意的程度。因此，现有的信息表达机制限制了计算机帮助人们处理、综合和分析信息的能力。为此，万维网发明人 Tim Berners-Lee 在 20 世纪九十年代末提出了语义网的构想。他指出，“语义网是现有万维网的扩展，在其中信息被赋予明确的、完善的语义，以使得计算机和人能够更好地进行协作” [Tim, et al., 2001] [Nigel, et al., 2006] 。为了实现在 Web 上表达语义的需求，包括万维网联盟(W3C) 和因特网工程技术组(IETF)在内的研究机构制定和开发了一系列技术规范。它们是在 Web 上进行语义表达和处理的技术基础，构成了一个层次化的技术框架。语义网是对万维网中信息表达方式的一次革新，它给出了一套技术框架使得 Web 上的信息可以方便地被计算机进行处理和理解。语义搜索是架构在语义网上的搜索引擎，将语义 Web 技术引入搜索引擎，为用户提供精准的检索结果。近两年来国外学者采用不同的方法和技术对该问题进行了深入的研究，并得出了不少有益的结论，也建立了相关的原型系统。但是一方面，由于语义 Web 处于发展阶段，另一方面也由于技术条件的限制，目前并不存在一个“通用”的解决方案，现有的语义搜索引擎系统也都处于起步研究阶段，离实用的商业化水平还相距甚远。

总的来说，语义网背景下的语义搜索主要面临的问题有以下三点：

(1) 与传统的 Web 文档相比，语义网文档的本质是 RDF Graph。给定一个 RDF Graph，可以采取多种语法格式对其进行序列化，如，RDF / XML, Natation3 等。采用不同的语法进行序列化，生成的语义网文档之间可能具有显著的差别，然而它们表达的语义却是一致的。有时，即使采用相同的语法，也会导致不同的结果文档，比如采用不同的 name space 前缀。因此，对于语义网文档的搜索而言，如何针对 RDF 数据模型的特点进行文档分析、索引建立和查询匹配即变得极为重要。

(2) 理解一个 URI 所指称的实体对于判断语义网上的实体共指问题非常重要。实体共指是指客观世界的同一个对象，在语义网上（通常是被不同的信息发布者）使用不同的 URI 来指称。这种共指现象给语义网数据的整合和建立在其上的搜索均带来了困难。自动的共指消解技术能够帮助人们快速地找到可能的共指 URI 列表。要更好地解决实体共指问题，当前还是以人工参与为主。因此，提供一种快速、高效的办法理解一个 URI 所指称的实体，将能够很好地帮助人们做出共指判断，进而帮助人们理解所获取的信息的真实含义。

(3) 在现有缺乏必要的手段形成语义网的背景下，如何利用语义网技术改进传统的 Web 信息检索系统对用户来说极为重要。传统 Web 是基于自然语言的方式进行组织的，而语义网提供的一系列的技术规范，包括语义的明确表达和语

义网数据查询，能够以一个特定领域的搜索系统为切入点，利用语义网技术帮助获取传统 Web 上的信息。

### 3. 技术方法和研究现状

语义搜索的研究涉及到多个领域，包括搜索引擎、语义 Web、数据挖掘和知识推理等。运用的主要方法可归纳为：（1）图理论；在语义网的技术框架中，RDF(Resource Description Framework)是一个非常基础、且又非常重要的数据模型。通过 RDF 数据模型可将语义网中的本体组织为图结构，图中的弧和由结点和弧组成的路径中都包含着信息，因此在语义搜索中应用到了不同形式的图遍历方法，如实例扩展及查询的形式化方法等；（2）匹配算法，在语义搜索中需进行概念与关键字或者实例与关键字的匹配，关键字提供了一种快速定位信息的入口，而关键字和概念的匹配方法是语义搜索中重要的一环；（3）逻辑特别是描述逻辑、模糊逻辑等。逻辑和推理已经被整合到未来的语义 Web 框架中。描述逻辑是知识的一种形式化表示方法[Baader, et al., 2003]，作为本体语言的基础为人们所熟知[Horrocks, et al., 2003]，如 OIL, DAML+OIL, OWL。语义搜索的目的是为了准确地理解用户的输入，因此必须要使计算机具有逻辑推理能力，即如果输入为“小米 Note3 是 Note2 的升级版吗？价格是多少？”计算机要确切理解“小米”、“Note2”、“Note3”代表的含义，并且理解“Note2”和“Note3”之间的关系。

#### 3.1. 主流语义网搜索引擎

在新一代的语义搜索引擎中较为典型的有两个，且都是基于本体的语义搜索引擎，分别为：Swoogle 和 TUCUXI。其中，Swoogle 从搜索返回结果的 Web 文档中提取出本体，然后依据本体间的语义关联性确定出文档间的语义关系；TUCUXI 则通过所获得的本体在 Web 上以特定规则爬行，并通过语义处理找出最符合要求的网页。目前已开发出许多建立于本体上的语义搜索引擎，如，Congnition、Hakia、DeepDyve、Factbites、Kngine 等。

Swoogle 是由马里兰大学计算机科学和电气工程系于美国国家科学基金会(NSF)和美国国防部下署高级研究计划署(DARPA)的资助下所建立的。与那些传统意义上的语义网搜索引擎不同，Swoogle 在资源获取方面拥有一系列突出的解决方案，可自动发现语义网中 RDF 格式的文档，通过 Link-Following 和 Meta-Search 的方式识别出语义网文档(SWDs)，通过语义分析不断发现新的语义网文档，并可对其中元数据建立相关索引提供高效率的查询服务，利用 Rational Random Surfing 模型提供高质量的排序结果[Ding, et al., 2004] [Ding and Finin, 2006]。Swoogle 的核心功能有：

- 提取语义网中的实例数据；

- 支持对语义网的浏览，提供语义网中文档的元数据；
- 搜寻语义网中的术语，譬如通过属性与类定义的 URIs 等；
- 搜索提取语义网中的本体，并使用独有的算法提供高质量的排序结果；
- 可存储各种类型的语义网文档。

**Swoogle** 与通常的本体存储器或本体标注系统相比，其最大的与众不同之处在于能够鉴别出异源本体，此外还具有语义网文档自动发现化制。

**Cognition**: 目前可提供三个 Demo，Cognition Q&A，Medline Semantic Search 及 Wikipedia Semantic Search，涉及法律、医学与消费者信息等深度内容，且是首个真正实现人机对话界面的语义搜索引擎。

**Hakia**: 由 Xerox 公司推出的 Hakia 搜索引擎通过理解用户查询，并利用本体进行查询扩展，将各种基于主题的相关信息汇总。其利用的技术包括：词形变换、同义词扩展、概念具体化、自然语言理解等，可为用户提供语义搜索范围内解决方案，能够满足用户对于低成本、高效率的搜索需求。其搜索范围包括新闻、网页、博客、维基词条、Pubmed 等，返回结果的呈现方式有深度语义（Galleries、Pubmed、可信站点）、表面语义（新闻、博客、网页）、常规搜索（Twitter 与图像）加结果页面链接。

**Factbites**: 可依据事实进行回答，与结果链接相比，其更专注于内容分析，并可使搜索结果更有意义，到目前也只有简单搜索方式。其搜索结果呈现方式是从网页中所抽取出来的有意义的、完整的语句清单加 URL。

**DeepDyve**: 是深网或者隐形网络搜索引擎，可提供深度网络学术资源租赁服务与全文预览服务。其搜索范围可包括来自 Nature、IEEE、Elsevier、Wiley-Blackwell、Springer 等一流出版社的有关健康科学、生命科学、人文社会科学、物理科学与工程学等领域的权威评审期刊与专利等等深度网络学术资源，并同时可搜索 Wikipedia，现正慢慢扩展至更多的领域。其搜索结果主要为 PDF 文档，而搜索结果呈现方式是结果过滤项（主题、类型（可租用、仅供预览、免费）、时间、作者、期刊）加结果页链接。

**Kngine**: 其可对任何主题进行搜索，能够支持移动端搜索，其语种包括英语、德语、西班牙语、阿拉伯语。以选项卡形式展现搜索结果，在选项卡下方可选择显示与每项相关的术语和网页，其搜索方式包括语音搜索和简单搜索。

## 3.2 技术要点及研究现状

随着计算机的普及和万维网技术的发展，万维网已经成为人类历史上最大的信息系统，也成为人们获取信息的重要来源。传统的搜索引擎引入的“关键词匹配导致难以理解用户意图”和“缺乏有效方法分析数据间关系”的问题无法保证返回用户满意的结果。基于此，结合语义信息的搜索引擎-语义搜索越发被

学者和工业界重视起来。事实上，语义搜索是传统搜索的进化，传统的搜索技术对于结合检索与推理的语义搜索有许多可借鉴的经验。因此，可在传统搜索引擎技术的基础上对语义搜索进行更深入的研究，建立实用性更强的语义搜索系统，改善当前的搜索效果，以期在更广泛的语义 Web 环境中发挥更大的作用。国外学者在近几年已经运用不同的方法对语义搜索领域进行了深入的学习与研究，并成功设计及实现了多个系统原型。但是一方面受制于语义网仍处于起步阶段，另一方面也由于目前技术水平的限制，至今还不存在一个既精准又高效的通用解决方案[Anuar, et al., 2016]。事实上，语义搜索研究目前仍处于探索阶段，现有的有关语义搜索的研究点主要有：

### 3.2.1 引入推理和关联关系的语义搜索

在语义 Web 设计中，Web 中的资源用 URI 统一标示，并利用 RDF / OWL 标识资源的语义信息，由于数据间的语义明确便于计算机理解，基于此结构良好的数据的搜索克服了关键词查询的歧义性，同时在这些数据上还可以通过推理实现知识发现，推理出新的知识。随着语义 Web 研究的深入和应用的更新，Web 上的 RDF 资源对越来越多，基于推理的知识型语义搜索越来越得到关注，或许将成为未来语义搜索的主要方式。

Stanford 大学研制的 Triple 系统是一个基于逻辑程序设计的 RDF 查询系统，逻辑子句的问题求解能力使它能够解答较为复杂的问题，类似于“迈克尔杰克逊的《This Is It》专辑中有哪些歌曲？”这类推理型问题[Sintek and Decker, 2002]。马里兰大学设计的 HOWLIR 系统是基于 DAML 描述框架的语义 web 信息检索系统，它采用 DAML-JESSDB(一个基于 DAML 的推理系统)作为推理引擎。该系统自动产生并提取网页中的语义标签，同时也实现推理以产生更多关于网页的语义信息[Shah, et al., 2002]。搜索请求可以是针对语义信息的形式化查询，也可以是针对文本信息的关键字查询。文献[Dzbor and Motta, 2006]提出的 Swangler 系统则将语义标注转化为一般的文本查询关键字。清华大学提出了一种细粒度语义网检索模型，可对用户提供基于关键字的查询接口，检索系统以 RDF 图构建搜索策略，以 URI 资源为检索单位，查询结果是包含关键字在内的三元组集合[吴刚,等., 2005]。

资源间由关联关系引入的链接路径在某些特定领域比资源本身更具价值，比如在国家安全领域通常需要搜索资源之间的链接关系，这些关系可能意味着某些潜在的安全威胁。关联搜索中的主要问题在于如何定义链接的兴趣尺度，且这种定义方法不仅能够消除用户不感兴趣的关联关系，而且可以搜索到数据之间复杂的、隐藏的关联关系。文献[Anyanwu and Sheth, 2003]提出了一种大众化且简单的形式化计算方法，尝试发现资源间有价值的相关关系。语义搜索不仅要能够探索到资源之间的关联关系还需要获得合理的排序结果 [Aleman-Meza, et al., 2003]

[Anyanwu, et al., 2005]。知识库中实体之间关系的个数往往会超出实体本身，语义关联就是指实体之间的复杂关系。传统搜索引擎采用的排序方法只能对检索得到链接文本进行排序，无法对结构信息排序。为对语义搜索中获取的结构信息（多为 RDF 三元组）排序，目前多是将传统的结果排序算法做出改变以应用于语义搜索结果排序。文献[Bamba and Mukherjea, 2004]即利用语义 Web 资源的重要性对结果集进行排序。文献[Bai, et al., 2009]试图发现元数据上复杂的关系，提出了一种预测用户需求的排序方法来识别语义关联。

### 3.2.2 语义搜索中的查询扩展

传统的搜索引擎经常会因为词语含义的多样性而产生无意义的检索结果。产生词语多样性问题的根本原因在于，人们在现实生活中描述同样的对象或事件的用词存在着多样性。例如，单车和脚踏车都是对自行车这一概念的称谓。为解决这个问题，人们提出了基于概念的语义查询扩展（Semantic based QE），用概念来描述查询主旨，找到与查询语义相关的概念对查询进行扩展，因为概念是专门用来描述现实世界对象的。基于概念，可以消除现实世界中人们对同一真实对象的不同表达方式的理解差异。语义网的构建目标即是将网络中的概念构成网状结构，利用概念间的联系形成拓扑网络，而本体（语义网中的结点）则视为概念的具体表现形式。因此这种基于概念的查询扩展一直是语义搜索领域的研究重点。

目前语义搜索的研究侧重点围绕于查询语句或是文档中的语义发掘，注重发现目标资源间的关联，通过深度的查询理解而获得更高的查准率。文献[Jothilakshmi, et al., 2013]在领域模型的基础上提出了一种语义查询扩展方法，即结合概念级别（基于领域知识）、语法级别（基于 WordNet 的术语词汇）和随机模型 ME-HMM2（隐马尔可夫模型与最大熵模型相结合），取得了较好的效果。Pal 等在文献[Pal, et al., 2014]中提出了基于组合的概念映射查询扩展方法，考虑到每个候选扩展术语实用性的三个方面：其在相关文献和目标语料库中的分布、与查询术语的统计关联及术语在 WordNet 中的定义及其与查询术语间的语义关系，这种不同信息来源的组合能够对测试集合产生较好的效果。文献[Ngo and Cao, 2010]中提出了一种基于本体的广义向量空间模型进行文本的语义搜索，利用命名实体及其潜在的相关命名实体的本体特性获得文档和查询词的语义，并在此基础上构建了一个框架通过结合不同的本体，利用它们之间的互补优势进行语义标注和搜索。文献[Chauhan, et al., 2012]中提出了基于语义查询扩展的信息组织与检索系统，所提出的语义查询扩展方法包括一个基于领域本体的数学模型来计算概念之间的语义相似性和查询扩展算法，利用查询的概念及这些概念的同义词来执行查询扩展。Zhao 等在文献[Zhao, et al., 2015]中提出了一个基于物联网环境的、带有主题发现与语义感知功能的索引构建方案和检索系统 Acrost，通过以多主题为中心的搜集组合获得感兴趣信息的初始内容，基于聚合正则表达式和条件

随机域方法提取元数据，通过分析查询和对相关性内容排序进行语义感知检索。Bashar 等在文献[Bashar and Myaeng, 2014]中利用维基百科页面中的语义标注提出一种新颖的语义查询扩展方法用于对初始查询词消除歧义并丰富语义，然后将该方法应用于专利搜索、专利分类。

### 3.2.3 语义搜索中的索引构建

创建合理和有效的索引是保证搜索顺利进行的保证。传统的基于关键字的搜索引擎不能很好的解决一词多义,多词一义的问题。用户将花费很大的代价从搜索引擎返回的结果中寻找所需的结果或者换关键字重新查询。建立语义索引则是为搜索引擎解决以上问题提供了新方向。利用语义网中的本体去分析文档和查询语句的语义信息，从而为海量的无结构网页数据建立语义索引,查询时通过匹配用户意图和文档中以本体标识的概念的相关性给出结果。这种方法避免了基于关键词搜索的一词多义和多词一义问题。如文献[Ma, et al., 2007]提出了一种利用本体获取词间的语义关系,消除自然语言的多义性以标识文档的方法。文献[Mihalcea and Moldovan, 2000]介绍了基于 WordNet，用布尔模型添加词语语义信息到传统索引上从而建立了一个结合了语义和传统关键字的新型索引。文献[Buscaldi and Zargayouna, 2013]介绍了一种标注文档中概念的语义检索系统YaSemIR，并且这个系统可配置以和不同的本体、不同的文档一起工作。文献[Setchi, et al., 2011]将文档中最有意义、最有代表性的词语找出,并且获取其语义以组成一个语义核心，以这些语义核心建立索引，查询匹配的时候仅就语义核心和查询语句进行匹配。文献[Roger, 2008]考虑了词语对的关联性，并依据关联性强度快速地构建了一个潜在语义索引分析系统。文献[Kokiopoulou and Saad, 2004]在改进的 K-近邻算法基础上，消除了传统潜在语义索引时间复杂度高的特点，并应用文档索引结果做基于反馈的文本过滤。

## 4. 技术展望与发展趋势

国内外主流的搜索引擎厂商对于语义搜索的前景极为看好，普遍认为其是机遇与挑战并存的新领域。为了改进搜索效果和提升竞争力，处于主流地位的传统搜索引擎巨头也开始尝试语义搜索技术。2008 年微软收购了语义搜索引擎 Powerset<sup>27</sup>，希望以此提高 Bing 的语义功能和认知度以缩小与 Google 在搜索质量上的差距。百度则在 2009 年即开始涉足语义搜索领域，与哈尔滨工业大学建立合作研究实验室，专门对语义搜索中的关键技术-自然语言处理进行研发，并推出一款基于语义的“框计算”应用，专门用于对中文中的生僻字进行查询[Yang, et al., 2000]。Google 也于 2012 年 5 月推出知识图谱，将其应用于搜索引擎中以增强搜索结果，标志着大规模知识图谱在互联网语义搜索中的成功应用，视其为

---

<sup>27</sup> <https://www.cnet.com/news/report-microsoft-to-buy-powerset/>

下一代语义搜索的第一步。

虽然各大互联网公司都试图在语义搜索上有所突破，但目前国内科研机构对语义搜索的研究还处于初步探索阶段，并未形成一种通用的框架和方案。虽然提出了多种系统，但受限于语义网技术的尚未普及，并未有一套实用的语义搜索系统。已提出的系统有的只是对传统的信息检索功能进行补充和完善，有的只能提供形式化的查询，有的仅仅是有限利用了本体中的结构数据，并不存在能紧密结合两者功能的系统，实现的推理功能尚处于初步尝试过程中，目前也不存在较为成熟的基于语义的结果排序方法。未来的语义搜索研究方向可沿以下几点展开：

- ① 语义搜索概念模型。语义模型能改善当前搜索引擎的搜索效果，未来可扩展成为构建在语义Web上的新一代搜索引擎。
- ② 语义搜索本体知识库的构建、维护与进化。研究垂直领域的本体知识库构建方法、本体知识库设计方法和本体知识库查询方法，构建完备的领域本体知识库，探索本体知识库的维护方案，随着领域本体知识库的丰富还要研究并解决多领域异构的本体知识库的融合问题，提供本体相容性冲突检测方案。
- ③ 语义搜索的推理机制。结合领域本体，研究语义搜索中基于描述逻辑及模糊逻辑的推理问题，提高基于描述逻辑的本体推理技术的推理效率，扩大其推理算法的适用范围，结合文本信息获取用户的查询语义，提高处理用户查询需求的准确度。
- ④ 语义搜索的结果排序。传统搜索引擎采用的排序方法只能对文本信息进行排序，不能对实体之间的复杂关系排序，无法实现语义搜索结果的排序，因此需研究基于语义的结果排序方法，实现本体知识库中实体及实体之间关系的排序，提高返回结果的相关性。
- ⑤ 语义搜索的原型系统实现。基于以上研究，实现语义搜索引擎系统原型，在应用环境中进行测试并实现性能优化。

## 参考文献

- [Arvind, et al., 2001] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, Andreas Paepcke, and Sriram Raghavan. Searching the web. ACM Transactions on Internet Technology, 2001, 1(1): 2-43.
- [Aleman-Meza, et al., 2003] Boanerges Aleman-Meza, Chris Halaschek, I. Budak Arpinar, and Amit Sheth. Context-aware semantic association ranking. In Proceedings of the First International Conference on Semantic Web and Databases, 2003, 24-41.
- [Anyanwu and Sheth, 2003] Kemafor Anyanwu and Amit Sheth. P-Queries: Enabling querying for semantic associations on the semantic web. In Proceedings of the

- 12th International Conference on World Wide Web, 2003, 690-699.
- [Anyanwu, et al., 2005] Kemafor Anyanwu, Angela Maduko, and Amit Sheth. SemRank: Ranking complex relationship search results on the semantic web. In Proceedings of the 14th International Conference on World Wide Web, 2005, 117-127.
- [Anuar, et al., 2016] Fatahiyah Mohd Anuar, Rossitza Setchi, and Yu-Kun Lai. Semantic retrieval of trademarks based on conceptual similarity. IEEE Transactions on Systems, Man, and Cybernetics, 2016, 46(2): 220-233.
- [Baader, et al., 2003] Franz Baader, Diego Calvanese, Deborah L McGuinness, Daniele Nardi, and Peter F. Patel-Schneider. The description logic handbook: Theory, implementation and applications. Cambridge University Press. 2007.
- [Bamba and Mukherjea, 2004] Bhuvan Bamba and Sougata Mukherjea. Utilizing resource importance for ranking semantic web query results. In Proceedings of the Second International Conference on Semantic Web and Databases, 2004, 185-198.
- [Bai, et al., 2009] Bing Bai, Jason Weston, David Grangier, Ronan Collobert, Kunihiko Sadamasa, Yanjun Qi, Olivier Chapelle, and Kilian Weinberger. Supervised semantic indexing. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, 2009, 187-196.
- [Buscaldi and Zargayouna, 2013] Davide Buscaldi and Ha fa Zargayouna. YaSemIR: Yet another semantic information retrieval system. In Proceedings of the 6th International Workshop on Exploiting Semantic Annotations in Information Retrieval, 2013, 13-16.
- [Bashar and Myaeng, 2014] A. S. Bashar and S. H. Myaeng. Wikipedia-based query phrase expansion in patent class search. Information Retrieval, 2014, 17(5): 430-451.
- [Chauhan, et al., 2012] R. Chauhan, R. Goudar, R. Rathore, P. Singh, and S. Rao. Ontology based automatic query expansion for semantic information retrieval in sports domain. In Proceedings of the International Conference on Eco-friendly Computing and Communication systems, 2012, 422-433.
- [Ding, et al, 2004] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal Doshi, and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In Proceedings of the 13th ACM International Conference on Information and Knowledge Management, 2004, 652-659.
- [Ding and Finin, 2006] Li Ding and Tim Finin. Characterizing the semantic web on the

- web. In Proceedings of the 5th International Semantic Web Conference, 2006: 242-257.
- [Dzbor and Motta, 2006] Martin Dzbor and Enrico Motta. Study on integrating semantic applications with magpie. In Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, 2006, 66-76.
- [Guo, et al., 2003] Lin Guo, Feng Shao, Chavdar Botev, and Jayavel Shanmugasundaram. XRANK: Ranked keyword search over XML documents. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data, 2003, 16-27.
- [Horrocks, et al., 2003] Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. From SHIQ and RDF to OWL: The making of a Web Ontology Language. Web Semantics: Science, Services and Agents on the World Wide Web, 2003, 1(1): 7-26.
- [Jothilakshmi, et al., 2013] R. Jothilakshmi, N. Shanthi, and R. Babisraraswthi. An approach for semantic query expansion based on maximum entropy-hidden markov model. In Proceedings of the 4th International Conference on Computing, Communication and Networking Technologies, 2013, 1-5.
- [Kokiopoulou and Saad, 2004] E. Kokiopoulou and Y. Saad. Polynomial filtering in latent semantic indexing for information retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004, 104-111.
- [Mihalcea and Moldovan, 2000] Rada Mihalcea and Dan Moldovan. Semantic indexing using WordNet senses. In Proceedings of the ACL-2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval: Held in Conjunction with the 38th Annual Meeting of the Association for Computational, 2000, 35-45.
- [Ma, et al., 2007] Wenhui Ma, Wenbin Fang, Gang Wang, and Jing Liu. Concept index for document retrieval with peer-to-peer network. In Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/ Distributed Computing, 2007, 1119-1123.
- [Nigel, et al., 2006] Shadbolt Nigel, Hall Wendy, and Berners-Lee Tim. The semantic web revisited. IEEE Intelligent Systems, 2006, 21(3): 96-101.
- [Ngo and Cao, 2010] V. M. Ngo, and T. H. Cao. Ontology-based query expansion with latently related named entities for semantic text search. Advances in Intelligent

- Information and Database Systems, 2010, 283: 41-52.
- [Pal, et al., 2014] D. Pal, M. Mitra, K. Datta. Improving query expansion with latently related named entities for semantic text search. In Proceedings of the 2nd Asian Conference on Intelligent Information and Database Systems, 2010, 41-45.
- [Roger, 2008] Bradford B. Roger. An empirical study of required dimensionality for large-scale latent semantic indexing applications. In Proceedings of the 17th ACM Conference on Information and Knowledge Management, 2008, 153-162.
- [Shah, et al., 2002] Urvi Shah, Tim Finin, Anupam Joshi, R. Scott Cost, and James Matfield. Information retrieval on the semantic web. In Proceedings of the 11th International Conference on Information and Knowledge Management, 2002, 461-468.
- [Sintek and Decker, 2002] Michael Sintek and Stefan Decker. TRIPLE-A query, inference, and transformation language for the semantic web. In Proceedings of the First International Semantic Web Conference on the Semantic Web, 2002, 364-378.
- [Setchi, et al., 2011] Rossi Setchi, Qiao Tang, and Ivan Stankov. Semantic-based information retrieval in support of concept design. Advanced Engineering Informatics, 2011, 25(2): 131-146.
- [Tim, et al., 2001] Berners-Lee Tim, Hendler James, and Lassila Ora. The semantic web. Scientific American: Feature Article, 2001.
- [吴刚, 等., 2005] 吴刚, 唐杰, 李涓子, 王克宏. 细粒度语义网检索. 清华大学学报(自然科学版), 2005, 45(1): 139-146.
- [Yang, el al., 2000] Qiang Yang, Hai-Feng Wang, Ji-Rong Wen, and H. M. Zhang. Towards a next-generation search engine. In Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence, 2000, 1-12.
- [Zhang, et al., 2007] Lei Zhang, Qiaoling Liu, Jie Zhang, Haofen Wang, Yue Pan, and Yong Yu. Semplore: An IR approach to scalable hybrid query of semantic[C]. In Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, 2007, 652-665.
- [Zhao, et al., 2015] F. Zhao, Z. Sun, and H. Jin. Topic-centric and semantic-aware retrieval system for internet of things. Information Fusion, 2015, (23): 33-42.

# 第十一章 基于知识的问答

## 1. 任务定义、目标和研究意义

问答系统 (Question Answering, QA) 是指让计算机自动回答用户所提出的问题，是信息服务的一种高级形式。不同于现有的搜索引擎，问答系统返回用户的不再是基于关键词匹配的相关文档排序，而是精准的自然语言形式的答案。华盛顿大学图灵中心主任 Etzioni 教授 2011 年曾在 Nature 上发表文章《Search Needs a Shake-Up》，其中明确指出：“以直接而准确的方式回答用户自然语言提问的自动问答系统将构成下一代搜索引擎的基本形态”[Etzioni O., 2011]。因此，问答系统被看做是未来信息服务的颠覆性技术之一，被认为是机器具备语言理解能力的主要验证手段之一。因此，对其开展研究具有非常重要的学术和实际意义。特别是近些年，随着人工智能热潮到来，无论是学术界还是产业界，都给予其极大关注和投入。

纵观问答系统的技术演进，其一直伴随的人工智能技术的发展而发展。近些年，问答系统更是取得一系列倍受关注的成果。2011 年，IBM Watson 自动问答机器人在美国智力竞赛节目 Jeopardy 中战胜人类选手，在业内引起了巨大的轰动。随着人工智能技术的突飞猛进，各大 IT 巨头更是相继推出以问答系统为核心技术的产品和服务，如移动生活助手 (Siri、Google Now、Cortana、小冰等)、智能音箱 (HomePod、Alexa、叮咚音箱等、公子小白等) 等，这似乎让人们看到了黎明前的阳光，甚至认为现有的问答技术已经十分成熟。

尽管 IBM Watson 系统在 Jeopardy 中战胜了人类选手，但是其核心技术并没有突破传统基于“检索+抽取”的问答模式，缺乏对于文本语义深层次的分析和处理，难以实现知识的深层逻辑推理，无法达到人工智能的高级目标。Watson 的成功也已经被证明仅仅局限于限定领域、特定类型的问题，离语义的深度理解以及智能问答还有很大的距离，其他问答系统，如 Siri 等，也存在同样的问题。因此，面对已有问答模式的不足，为了提升信息服务的准确性与智能性，研究者近些年逐步把目光投向知识图谱 (Knowledge Graph)。其意图是通过信息抽取、关联、融合等手段，将互联网文本转化为结构化的知识，利用实体以及实体间语义关系对于整个互联网文本内容进行描述和表示，从数据源头对于信息进行深度的挖掘和理解。同时，互联网中已经有一些可以获取的大规模知识图谱，例如 DBpedia[Lehmann et al., 2014]、Freebase[Bollacker, 2008]、YAGO[Suchanek et al., 2007] 等。这些知识图谱多是以实体、关系为基本单元所组成的图结构。

基于这样的结构化的知识，分析用户自然语言问题的语义，进而构建的结构化知识图谱中通过检索、匹配或推理等手段，获取正确答案，这一任务称之为

为知识库问答 (Question Answering over Knowledge Base, KBQA)。这一问答范式由于已经在数据层面通过知识图谱的构建对于文本内容进行了深度挖掘与理解，能够有效地提升问答的准确性。

## 2. 研究内容和关键科学问题

知识库问答系统在回答用户问题时，需要正确理解用户所提的自然语言问题，抽取其中的关键语义信息，然后在已有单个或多个知识库中通过检索、推理等手段获取答案并返回给用户。其中所涉及的关键技术包括：词法分析、句法分析、语义分析、信息检索、逻辑推理、语言生成等。传统知识库问答系统多集中在限定领域，针对有限类型的问题进行回答。然而伴随大数据的飞速发展，已有知识图谱的规模在不断增大，所涉及的领域不断增多。现有研究趋向于开放域、面向大规模、开放域、多源异构知识库问答系统构建。总体来讲，主要面临如下三个关键科学问题。

### 2.1 问句语义解析

知识库问答要回答用户的问题，首先就要正确理解用户所提问题的语义内容。面对结构化知识库，需要将用户问题转化为结构化的查询语句，进而在知识图谱进行进行查询、推理等操作，获取正确答案。因此，对于用户问题的语义解析是知识库问答研究所面临的首要科学问题。具体过程需要分析用户问题中的语义单元与知识图谱中的实体、概念进行链接，并分析问句中这些语义单元之间的语义关系，将用户问题解析成为知识图谱中所定义的实体、概念、关系所组成的结构化语义表示形式。其中涉及词法分析、句法分析、语义分析等多项关键技术，需要自底向上从文本的多个维度理解其中包含的语义内容。在词语层面，需要在开放域环境下，研究实体(Entity)和术语 (Terminology)的识别、答案类型词(Lexical Answer Type)识别、实体消歧 (Entity Disambiguation) 等关键技术。在句法层面，需要解析句子中词与词之间、短语与短语之间的句法关系，分析出句子句法结构。在语义层面，需要根据词语层面、句法层面的分析结果，将自然语言问句解析成可计算的结构化的逻辑表达形式(如一阶谓词逻辑表达式)。传统知识库问答方法面对单一领域有限规模知识图谱，多涉及的实体、概念、关系规模较小，通常采用模板、或者小规模机器学习算法进行语义解析。但是当面对大规模、多领域知识库时，随着实体、概念、关系规模增大，语义解析算法的复杂度也指数增加，如何获取实体提及，如何进行开放域关系抽取等问题仍然是学术界需要面对的难点问题。目前，已有一些工作利用深度神经网络将用户问题解析成为隐式表达的分布式数值向量的形式，其中蕴含的用户问句的关键语义，但是如何在分布式表示过程中与知识图谱相关联，反映其中所蕴含的实体、关系等关键语义也

是另一个科学问题。

## 2.2 大规模知识推理

在问答过程中，并不是所有的答案都能通过在知识图谱中进行检索或查询就可以获取答案。主要原因是已有知识库本身的覆盖度有限。需要在已有的知识体系中，通过知识推理的手段获取这些隐含的答案。例如，知识库中包括了一个人的“出生地”信息，但是没包括这个人的“国籍”信息，虽然知识库中对于人物对应了“国籍”属性，但是由于没有直接给出该属性的值，因此还是不能回答诸如“某某人是哪国人？”这样的问题。但是实际上我们都知道，一般情况下，一个人的“出生地”所属的国家就是他（她）的“国籍”。这些隐含知识天然存在于人的常识知识体系中，但在已有知识库中，并未被编码进去。面对知识库问答，就需要通过推理的方式学习到这样的模式。传统推理方法基于符号的知识表示形式，通过人工构建的推理规则推理出答案。但是面对大规模、开放域的问答场景，如何自动进行规则学习，如何解决规则冲突仍然是亟待解决的难点问题。目前，伴随深度学习的飞速发展，基于分布式表示的知识表示学习方法能够讲实体、概念以及它们之间的语义关系表示为低维空间中的对象（向量、矩阵等），通过在低维空间中的数值计算完成知识推理任务。虽然就目前来说，这类推理的效果离实用还有段距离，但是我们认为这是值得探寻的方法，特别是如何将已有的基于符号表示的逻辑推理与基于分布式表示的数值推理相结合，研究融合符号逻辑和表示学习的知识推理技术，是知识推理任务中的关键难点问题。

## 2.3 异构知识关联

由于用户问题的复杂性和多样性，问题的答案往往不能够在单一知识库中找到，需要综合多个知识库（多种语言、多种领域、多种模态）内的知识才能给出答案[Bizer et al., 2009]。例如

“谁出演了变形金刚并且和《Monkey Business》的演唱者结婚了？”

“谁出演了变形金刚”的信息需要在电影知识库中搜寻答案；而有关“结婚”的信息通常位于人物知识库中；“《Monkey Business》的演唱者”信息则位于音乐知识库中。因此，回答这个问句，需要综合电影、人物以及音乐三个不同知识库的信息，才能推出最终的答案：“乔什·杜哈明”。由于多源知识库之间存在结构差异、内容差异、语言差异、模态差异，要完成这一任务并不简单。（1）在面向多源异构知识库问答过程中，相对于面向单一知识库的问答，问句文本歧义更加严重。同一短语，在不同知识库中会映射为更多的概念（实体、关系）候选，这使得问句的语义解析更加困难。（2）问句中不同的子问题需要在不同的知识库中进行求解，这需要问答系统对于子问题进行精准的划分，同时确定子问题求解范围。（3）

不同源异构知识库之间存在冗余关联，不同知识库中的不同实体、关系间具有同指关系。多知识库问答需要利用这种同指关系对于多个知识库中的知识进行综合，从而回答用户问题。然而多源异构知识库间的同指关系通常并没有显式给出，是一种隐含关系。因此，系统需要挖掘知识库间的同指关系，完成异构知识库的关联与对齐，这对于构建多源异构知识库的问答系统有着重要的作用。

### 3. 技术方法和研究现状

根据技术路线的不同，已有知识库问答技术大致可以分为两类：1) 基于语义解析（Semantic Parsing）的知识库问答方法；2) 基于检索排序的知识库问答方法。下面将分别简要介绍技术现状。

#### 3.1 基于语义解析的知识库问答方法

在结构化数据形式的知识图谱上进行查询、匹配、推理等操作，最有效的方式是利用结构化的查询语句，例如：SQL、SPARQL 语句等。然而，这些语句通常是由专家编写，对于普通用户来说，自然语言仍然是最直接的交互方式。因此，如何把用户的自然语言问句转化为结构化的知识库查询语句便是进行问答的核心所在，其关键是对于自然语言问句进行解析（如图 1 所示）。目前，主流方法是通过语义解析，将用户的自然语言问句转化成结构化的语义表示，例如 $\lambda$ 范式 [Kwiatkowski, et al., 2011] 和 DCS-Tree[Liang, et al., 2011]。相对应的也提出了很多语义解析语法或方法，例如组合范畴语法（Category Compositional Grammar, CCG）[Kwiatkowski, et al., 2011] 以及依存组合语法（Dependency-based Compositional Semantics, DCS）[Liang, et al., 2011] 等。

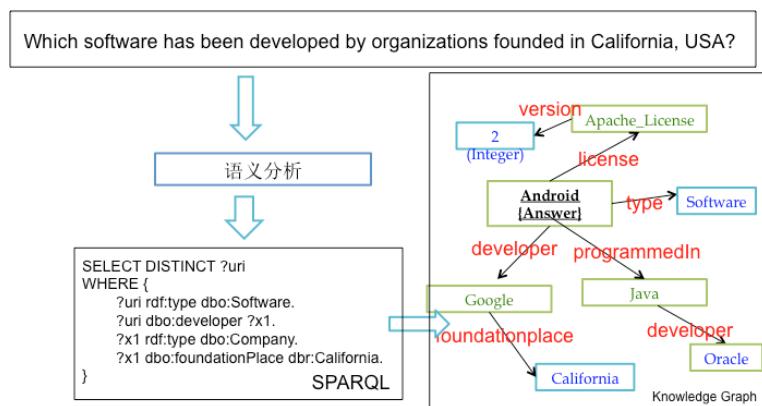


图 1. 基于语义解析的知识库问答过程

尽管很多语义解析方法在限定领域内能达到很好的效果，在这些工作中，很多重要组成部分（比如 CCG 中的词汇表和规则集）都是人工编写的。但是，当面对大规模多源异构知识库，传统的语义分析方法存在以下几个缺陷：(1) 资源（例如词汇表、规则集）标注费时费力，传统方法在有限的训练数据下性能大打

折扣；(2) 大规模知识库的开放域特性使得文本歧义问题更加严重，传统语义分析方法难以处理这一问题；(3) 在很多场景下，回答一个问题需要多个知识库的综合运用。然而，不同来源知识库是异构的，它们在结构和内容都具有异质性 [Lopez, et al., 2011]，如何处理多知识库间的冗余和差异性，是面向多知识库的问答系统面临的主要挑战。因此，很多研究者针对上述问题进行研究，取得一系列进展。在面对训练数据标注困难这一挑战时，目前主要采用半监督学习[Clarke et al., 2010]、外部资源[Cai, et al., 2013]、弱监督学习[Liang, et al., 2011] [Berant, et al., 2013]、同义对应[Fader, et al., 2013]等手段。在面对歧义更加严重的挑战时，主要采用联合模型[Lu, 2014]、图搜索[Tran, et al., 2009]等手段。在面对多源异构知识库时，主要采用子问题划分[Fader, et al., 2014] [Lopez, et al., 2012]、整数线性规划[Zhang, et al., 2016]等手段。

### 3.2 基于检索排序的知识库问答方法

但是，基于语义解析的知识库问答系统的处理范式通常仍然是基于符号逻辑的，缺乏灵活性。在分析问句语义过程中，易受到符号间语义鸿沟影响。同时从自然语言问句到结构化语义表达需要多步操作，多步间的误差传递对于问答的准确度也有很大的影响。近年来，深度学习技术以及相关研究飞速发展，越来越多的研究者开始研究深度学习技术在自然语言处理问题中的应用，例如情感分析、机器翻译、句法分析等等。知识库问答系统也不例外，已有相关的工作包括[Bordes, et al., 2014; Bordes, et al., 2014; Dong, et al., 2015; Hao, et al., 2017]。与传统基于符号的知识库问答方法相比，基于表示学习的知识库问答方法更具鲁棒性，其在效果上已经逐步超过传统方法，如图 2 所示。这些方法的基本假设是把知识库问答看做是一个语义匹配的过程。通过表示学习，我们能够用户的自然语言问题转换为一个低维空间中的数值向量(分布式语义表示)，同时知识库中的实体、概念、类别以及关系也能够表示成为同一语义空间的数值向量。那么传统知识库问答任务就可以看成问句语义向量与知识库中实体、边的语义向量相似度计算的过程。

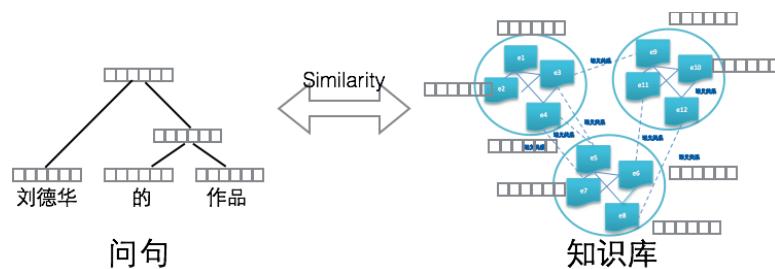


图 2、基于表示学习的知识库问答示意图

### 3.3 技术现状

已有的评测主要针对于一些限定领域的知识库进行问答。已有方法也取得了不错的结果。比如，在 Geoquery<sup>28</sup>(美国地理知识查询)数据集上(600 个训练样本，280 个测试样本)上，使用 CCG 和本体匹配的方法 F 值能达到 89.0%，使用 DCS 的方法 F 值能达到 91.1%；在求职(JOBS)数据集上(500 个训练样本，140 个测试样本)，使用 CCG 的方法 F 值能达到 79.3%，使用 DCS 的方法 F 值能达到 95%。在这一方面，QALD (Question Answering over Linked Data) 评测的举办更是推动了这方面的研究。QALD 每年举办一届，目前已经举办了第六届。每一次评测，组织者都会给出一些问题，要求参加评测系统在给定知识库的基础上，将所给问题转化为结构化的 SPARQL 查询语句，并在给定知识库上查询答案。但是，目前的研究趋势是从限定领域的知识库向大规模开放域甚至是多领域知识库进行扩展，例如 Freebase。与限定领域知识库相比，大规模开放知识库包含的资源和关系数量要大得多，比如 Geoquery 中只包含 8 个关系谓词，而 Freebase 包含上万个关系。因此开放知识库上的语义解析效果有明显的下降。例如利用 Freebase 知识库，开放查询测试的最好的效果只有 39.9% [Berant, 2014]；而在 QALD 评测中，在 DBpedia 上、开放查询中，表现最好的问答系统的正确率只有 40% [He, et al., 2014]。下图给出在面对开放域知识库 Freebase 时，在公开问题库 WebQuestion 上，已有系统能够达到的精度。

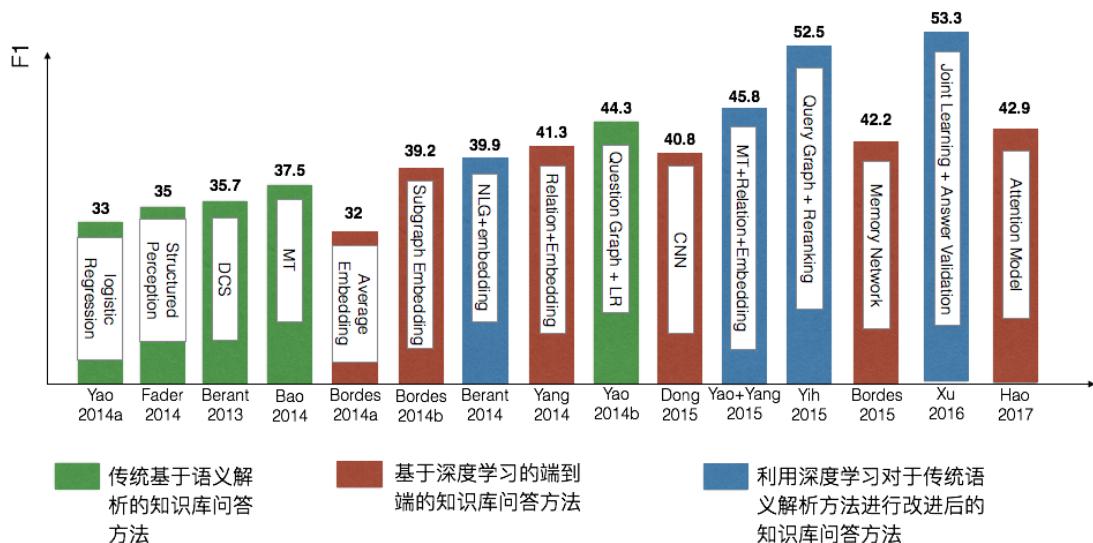


图 3. 已有知识库问答方法在 WebQuestion 问题集上的性能

### 4. 技术展望与发展趋势

纵观知识库问答研究发展的态势和技术现状，以下研究方向或问题将可能成

<sup>28</sup> <http://www.cs.utexas.edu/users/ml/nldata/geoquery.html>

为未来整个领域和行业重点关注的方向：

- 面向复杂问句的深度学习知识库问答方法

在实际问答场景下，用户的问题往往是复杂问句，其中所包含的语义关系复杂多样，子问题嵌套、文本歧义现象尤为突出。然而，已有基于深度学习的知识库问答方法目前尚只能解决简单类型问题（包含单一关系的问题类型）。在面对复杂问题时，例如有限制条件的问题（what did obama do before he was elected president?）、聚合问题（when's the last time the steelers won the superbowl?）等，已有方法处理手段单一，常忽略知识库与文本语义的关联与约束，缺乏在已有知识资源约束下的文本语义表示学习手段。因此，如何利用深度学习的方法解决复杂问题值得继续关注。

- 分布式表示与符号表示相结合的知识库问答

目前，基于深度学习的知识库问答方法试图通过高质量已标注的问题-答案建立联合学习模型，同时学习知识库库和问题的语义表示及他们之间的语义映射关系，试图通过分布式表示（向量）间的简单数值运算对于复杂的问答过程进行建模。这类方法的优势在于把传统的问答语义解析的复杂步骤转变为一个可学习的过程，虽然取得了一定的效果，但是训练过程容易受到训练数据质量的影响，缺乏已有知识的约束。同时，问答过程也缺乏可解释性。从目前自然语言处理很多任务来看，将统计与知识相结合是未来技术的重点突破方向。知识库问答也不例外，目前已经有将基于分布式表示的方法（深度学习）与传统基于符号表示的语义解析方法相结合的初步工作，例如神经图灵机[Liang, et al., 2017]。但是工作还很初步。如何将深度学习与传统语义方法进行深度融合，使这两种技术路线相互融合、相互约束，提升知识库问答的效果，是一个很值得深入研究的方向。

- 面向问答的深度推理

尽管已有知识图谱规模已经十分巨大，能够覆盖多个领域，但仍旧面临信息缺失的现象，这对于知识库问答带来巨大的挑战。这就需要面向问答的深度推理。传统基于符号逻辑的逻辑推理方法基于严格的符号匹配，过分依赖于推理规则的生成，因此具有领域适应性差、无法进行大规模推理的缺点。而深度学习基于分布式语义表示，利用语义空间中的数值模糊计算替代传统问答过程中的符号严格匹配，为解决上述问题供一种途径，但也存在推理结果准确度低、可解释性差的问题。因此，如果利用深度学习大规模、可学习的特点，在深度神经网络框架下，融入传统的逻辑推理规则，构建精准的大规模知识推理引擎是自动问答迫切需要解决的难点问题。

- 对话中的自然语言形式回复

传统的自动问答都是采用一问一答的形式。然而在很多场景下，需要提问者和系统进行多轮对话交互，实现问答过程。这时，需要系统返回用户的答案不再

只是单一实体、概念、关系的形式，而是需要是以自然语言的形式返回答案。这就需要自动生成自然语言的回复。现有方法多利用 sequence-to-sequence 模型进行自然语言生成，在这一过程中，如何与知识库相结合，将知识库问答的答案加入自然语言回复中，仍是亟待解决的问题。

总之，自动问答作为人工智能技术的有效评价手段，已经研究 60 余年了。整体上，知识库问答技术的发展趋势是从限定领域向开放领域、从单个数据源向多个数据源、从浅层语义分析向深度推理不断推进。我们有理由相信，随着自然语言处理、深度学习、知识工程等相关技术的飞速发展，知识库问答技术在未来有可能得到相当程度的突破。伴随着更多 AI 应用的实际落地，我们期待看到这一技术将在不远的未来得到更大、更广的应用。

## 参考文献

- [Bizer C., et al., 2009] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1--22, 2009.
- [Berant J, et al., 2014] Berant J, Liang P. Semantic parsing via paraphrasing[C]. In *Proceedings of ACL*. 2014, 7(1): 92.
- [Bollacker, 2008] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.
- [Bordes, et al., 2014] Bordes A, Weston J, Usunier N. Open question answering with weakly supervised embedding models, *Machine Learning and Knowledge Discovery in Databases*. Springer Berlin Heidelberg, 2014: 165-180.
- [Bordes, et al., 2014] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings In *Proceedings of EMNLP* 2014.
- [Cai Q, et al., 2013] Cai Q, Yates A. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension[C]. *ACL* (1). 2013: 423-433.
- [Dong, et al., 2015] Dong L, Wei F, Zhou M, et al. Question Answering over Freebase with Multi-Column Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. 2015, 1: 260-269
- [Etzioni O., 2011] O. Etzioni, Search needs a shake-up, *Nature*, vol. 476, no. 7358, pp. 25–26, 2011.
- [Fader A, et al., 2013] Fader A, Zettlemoyer L S, Etzioni O. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of ACL 2013*: 1608-1618.
- [Fader, et al., 2014] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings In *Proceedings of EMNLP* 2014
- [Hao, et al., 2014] Yanchao Hao, Yuanzhe Zhang, Shizhu He, Kang Liu and Jun Zhao, An End-to-End Model for Question Answering over Knowledge Base with Cross-

- Attention Combining Global Knowledge, in Proceedings of ACL 2017, Vancouver, Canada, July 30-August 4.
- [He, 2014] He S, Liu K, Zhang Y, Xu LH, Zhao J. Question Answering over Linked Data Using First-order Logic. In Proceedings of EMNLP. 2014
- [Kwiatkowski, et al., 2011] T. Kwiatkowski, L. Zettlemoyer, S. Goldwater, and M. Steedman, Lexical generalization in ccg grammar induction for semantic parsing, in Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, 2011, pp. 1512–1523. [11]  
[SEP]
- [Lehmann et al., 2014] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer et al., Dbpedia a large-scale, multilingual knowledge base extracted from wikipedia, Semantic Web, 2014.
- [Liang, et al., 2011] P. Liang, M. I. Jordan, and D. Klein, Learning dependency- based compositional semantics, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1, 2011, pp. 590–599.
- [Liang, et al., 2017] Chen Liang, Jonathan Berant, Quoc Le, Kenneth D. Forbus, Ni Lao, Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision , In Proceedings of ACL 2017
- [Lopez V, et al., 2012] Lopez V, Fernández M, Motta E, et al. Powerqua: Supporting users in querying and exploring the semantic web[J]. Semantic Web, 2012, 3(3): 249-265.
- [Lu, 2014] Lu W. Semantic Parsing with Relaxed Hybrid Trees[C]. In Proceedings of EMNLP. 2014.
- [Suchanek et al., 2007] F. M. Suchanek, G. Kasneci, and G. Weikum, Yago: a core of semantic knowledge, in Proceedings of the 16th international conference on World Wide Web, 2007. [11]  
[SEP]
- [Tran T, et al., 2009] Tran T, Wang H, Rudolph S, et al. Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data[C]. Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on. IEEE, 2009: 405-416.
- [Zhang, et al., 2016] Yuanzhe Zhang, Shizhu He, Kang Liu and Jun Zhao, A Joint Model for Question Answering over Multiple Knowledge Bases, in Proceedings of AAAI 2016, Phoenix, USA, February, 12-17.