【干货】最全知识图谱综述#1: 概念以及构建技术

原创：Xu/Shi/Quan et. 专知 2017-09-28

【导读】知识图谱技术是人工智能技术的组成部分，其强大的语义处理和互联组织能力，为智能化信息应用提供了基础。我们专知的技术基石之一正是知识图谱-构建AI知识体系-专知主题知识树简介。下面我们特别整理了关于知识图谱的技术全面综述，涵盖基本定义与架构、代表性知识图谱库、构建技术、开源库和典型应用。主要基于的参考文献来自[22]和[40], 本人(Quan)做了部分修整。

引言

随着互联网的发展，网络数据内容呈现爆炸式增长的态势。由于互联网内容的大规模、异质多元、组织结构松散的特点，给人们有效获取信息和知识提出了挑战。知识图谱（Knowledge Graph)以其强大的语义处理能力和开放组织能力，为互联网时代的知识化组织和智能应用奠定了基础。最近，大规模知识图谱库的研究和应用在学术界和工业界引起了足够的注意力[1-5]。一个知识图谱旨在描述现实世界中存在的实体以及实体之间的关系。知识图谱于2012年5月17日由[Google]正式提出[6]，其初衷是为了提高搜索引擎的能力，改善用户的搜索质量以及搜索体验。随着人工智能的技术发展和应用，知识图谱作为关键技术之一，已被广泛应用于智能搜索、智能问答、个性化推荐、内容分发等领域。

知识图谱的定义

在维基百科的官方词条中：知识图谱是Google用于增强其搜索引擎功能的知识库。本质上, 知识图谱旨在描述真实世界中存在的各种实体或概念及其关系,其构成一张巨大的语义网络图，节点表示实体或概念，边则由属性或关系构成。现在的知识图谱已被用来泛指各种大规模的知识库。 在具体介绍知识图谱的定义，我们先来看下知识类型的定义：

知识图谱中包含三种节点：

- **实体**: 指的是具有可区别性且独立存在的某种事物。如某一个人、某一个城市、某一种植物等、某一种商品等等。世界万物有具体事物组成，此指实体。如图1的“中国”、“美国”、“日本”等。 ，实体是知识图谱中的最基本元素，不同的实体间存在不同的关系。
- **语义类（概念）**：具有同种特性的实体构成的集合，如国家、民族、书籍、电脑等。 概念主要指集合、类别、对象类型、事物的种类，例如人物、地理等。
- **内容**: 通常作为实体和语义类的名字、描述、解释等，可以由文本、图像、音视频等来表达。
- **属性(值)**: 从一个实体指向它的属性值。不同的属性类型对应于不同类型属性的边。属性值主要指对象指定属性的值。如图1所示的“面积”、“人口”、“首都”是几种不同的属性。属性值主要指对象指定属性的值，例如960万平方公里等。

- **关系**: 形式化为一个函数，它把kk个点映射到一个布尔值。在知识图谱上，关系则是一个把kk个图节点(实体、语义类、属性值)映射到布尔值的函数。

基于上述定义。基于三元组是知识图谱的一种通用表示方式，即$G=(E,R,S)$,其中$\mathrm{E}=\{\mathrm{e}_1,e_2,\ldots,e_{|E|}\}$，是知识库中的实体集合，共包含|E|种不同实体；$\mathrm{R}=\{\mathrm{r}_1,r_2,\ldots,r_{|R|}\}$是知识库中的关系集合，共包含|R|种不同关系；$\mathrm{S}\subseteq E\times R\times E$代表知识库中的三元组集合。三元组的基本形式主要包括(实体1-关系-实体2)和(实体-属性-属性值)等。每个实体(概念的外延)可用一个全局唯一确定的ID来标识，每个属性-属性值对(attribute-value pair，AVP)可用来刻画实体的内在特性，而关系可用来连接两个实体，刻画它们之间的关联。如下图1的知识图谱例子所示，中国是一个实体，北京是一个实体，中国-首都-北京 是一个（实体-关系-实体）的三元组样例北京是一个实体 ，人口是一种属性2069.3万是属性值。北京-人口-2069.3万构成一个（实体-属性-属性值）的三元组样例。

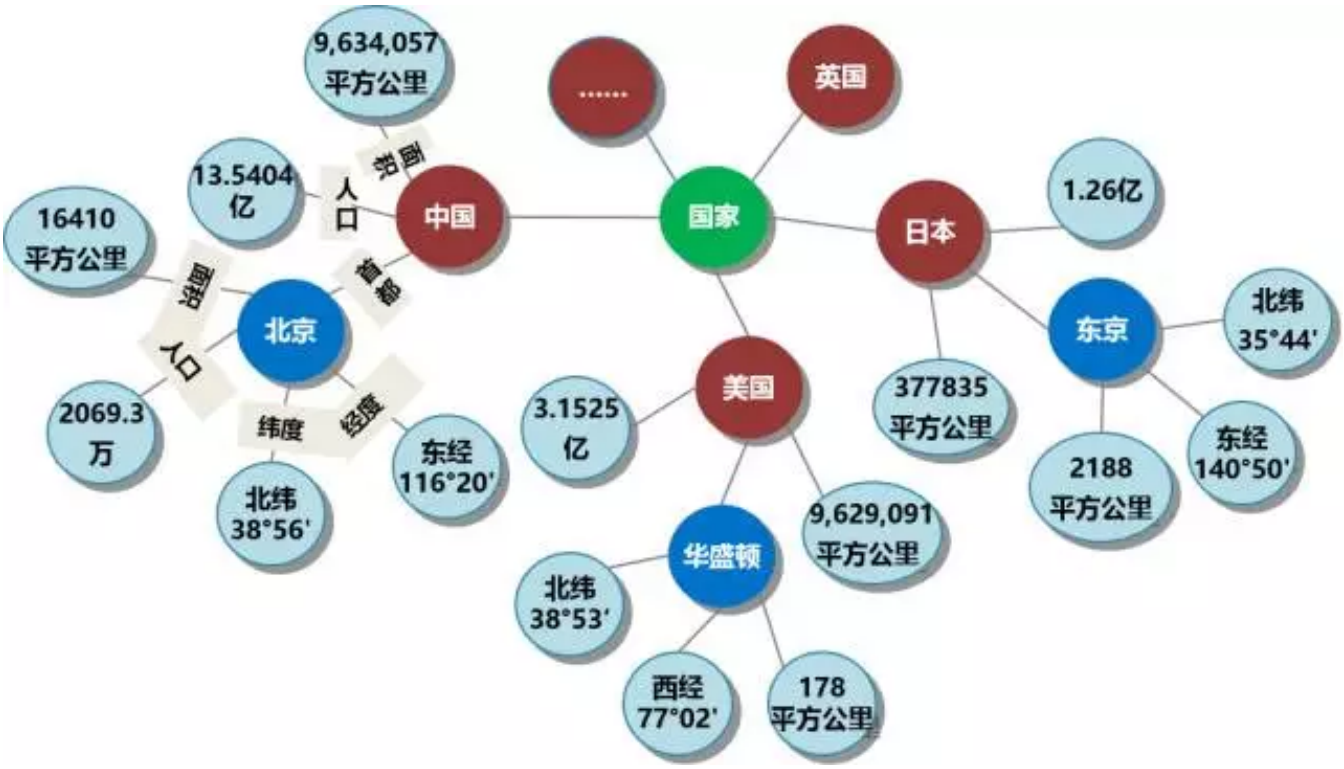

图1 知识图谱示例

知识图谱的架构

知识图谱的架构包括自身的逻辑结构以及构建知识图谱所采用的技术（体系）架构。

1） 知识图谱的逻辑结构

知识图谱在逻辑上可分为模式层与数据层两个层次，数据层主要是由一系列的事实组成，而知识将以事实为单位进行存储。如果用(实体1，关系，实体2)、(实体、属性，属性值)这样的三元组来表

达事实，可选择图数据库作为存储介质，例如开源的Neo4j[7]、Twitter的FlockDB[8]、sones的GraphDB[9]等。模式层构建在数据层之上，是知识图谱的核心，通常采用本体库来管理知识图谱的模式层。本体是结构化知识库的概念模板，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小。

2）知识图谱的体系架构

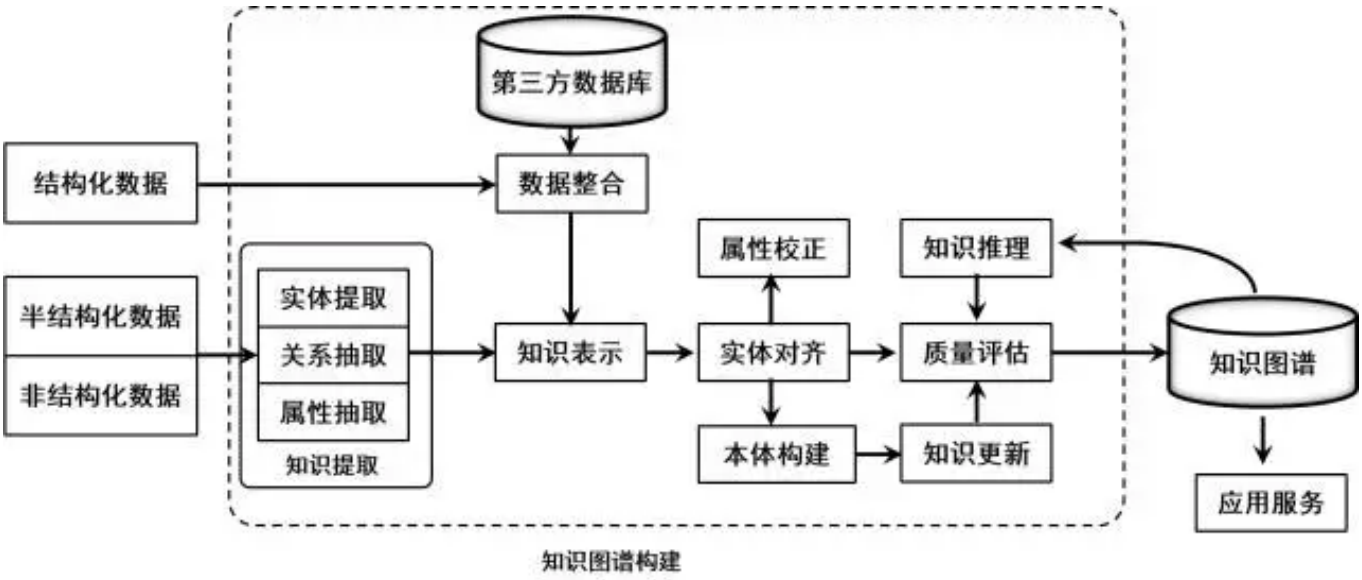

图2 知识图谱的技术架构

知识图谱的体系架构是其指构建模式结构，如图2所示。其中虚线框内的部分为知识图谱的构建过程，也包含知识图谱的更新过程。知识图谱构建从最原始的数据（包括结构化、半结构化、非结构化数据）出发，采用一系列自动或者半自动的技术手段，从原始数据库和第三方数据库中提取知识事实，并将其存入知识库的数据层和模式层，这一过程包含：信息抽取、知识表示、知识融合、知识推理四个过程，每一次更新迭代均包含这四个阶段。知识图谱主要有自顶向下(top-down)与自底向上(bottom-up)两种构建方式。自顶向下指的是先为知识图谱定义好本体与数据模式，再将实体加入到知识库。该构建方式需要利用一些现有的结构化知识库作为其基础知识库，例如Freebase项目就是采用这种方式，它的绝大部分数据是从维基百科中得到的。自底向上指的是从一些开放链接数据中提取出实体，选择其中置信度较高的加入到知识库，再构建顶层的本体模式[10]。目前，大多数知识图谱都采用自底向上的方式进行构建，其中最典型就是Google的Knowledge Vault[11]和微软的Satori知识库。现在也符合互联网数据内容知识产生的特点。

代表性知识图谱库

根据覆盖范围而言，知识图谱也可分为开放域通用知识图谱和垂直行业知识图谱[12]。开放通用知识图谱注重广度，强调融合更多的实体，较垂直行业知识图谱而言，其准确度不够高，并且受概念范围的影响，很难借助本体库对公理、规则以及约束条件的支持能力规范其实体、属性、实体间的关系等。通用知识图谱主要应用于智能搜索等领域。行业知识图谱通常需要依靠特定行业的数据来

构建，具有特定的行业意义。行业知识图谱中，实体的属性与数据模式往往比较丰富，需要考虑到不同的业务场景与使用人员。下图展示了现在知名度较高的大规模知识库。

知识图谱库名称	机构	特点、构建手段	应用产品
FreeBase	MetaWeb(2010年被谷歌收购)	• 实体、语义类、属性、关系； • 自动+人工：部分数据从维基百科等数据源抽取而得到；另一部分数据来自人工协同编辑 • https://developers.google.com/freebase/	Google Search Engine，Google Now
Knowledge Vault（谷歌知识图谱）	Google	• 实体、语义类、属性、关系； • 超大规模数据库；源自维基百科、Freebase、《世界各国纪实年鉴》 • https://research.google.com/pubs/pub45634	Google Search Engine，Google Now
DBpedia	莱比锡大学、柏林自由大学、 OpenLink Software	• 实体、语义类、属性、关系 • 从维基百科抽取 •	DBPedia
维基数据(Wikidata)	维基媒体基金会(Wikimedia Foundation)	• 实体、语义类、属性、关系,与维基百科紧密结合 • 人工（协同编辑）	Wikipedia
Wolfram Alpha	沃尔夫勒姆公司(Wolfram Research)	• 实体、语义类、属性、关系, 知识计算 • 部分知识来自于Mathematica；其它知识来自于各个垂直网站	Apple Siri
Bing Satori	Microsoft	• 实体、语义类、属性、关系, 知识计算 • 自动+人工	Bing Search Engine, Microsoft Cortana
YAGO	马克斯·普朗克研究所	• 自动：从维基百科、WordNet和GeoNames提取信息	YAGO
Facebook Social Graph	Facebook	• Facebook 社交网络数据	Social Graph Search
百度知识图谱	百度	• 搜索结构化数据	百度搜索
搜狗知立方	搜狗	• 搜索结构化数据	搜狗搜索
ImageNet	斯坦福大学	• 搜索引擎 • 亚马逊 AMT	计算机视觉相关应用

图3 代表性知识图谱库概览

知识图谱构建的关键技术

大规模知识库的构建与应用需要多种技术的支持。通过知识提取技术，可以从一些公开的半结构化、非结构化和第三方结构化数据库的数据中提取出实体、关系、属性等知识要素。知识表示则通过一定有效手段对知识要素表示，便于进一步处理使用。然后通过知识融合，可消除实体、关系、属性等指称项与事实对象之间的歧义，形成高质量的知识库。知识推理则是在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库。分布式的知识表示形成的综合向量对知识库的构建、推理、融合以及应用均具有重要的意义。接下来，本文将以知识抽取、知识表示、知识融合以及知识推理技术为重点，选取代表性的方法，说明其中的相关研究进展和实用技术手段 。

1 知识提取

知识抽取主要是面向开放的链接数据，通常典型的输入是自然语言文本或者多媒体内容文档（图像或者视频）等。然后通过自动化或者半自动化的技术抽取出可用的知识单元，知识单元主要包括实体(概念的外延)、关系以及属性3个知识要素，并以此为基础，形成一系列高质量的事实表达，为上层模式层的构建奠定基础。

1.1 实体抽取

实体抽取也称为命名实体学习(named entity learning)或命名实体识别（named entity recognition)，指的是从原始数据语料中自动识别出命名实体。由于实体是知识图谱中的最基本元素，其抽取的完整性、准确率、召回率等将直接影响到知识图谱构建的质量。因此，实体抽取是知识抽取中最为基础与关键的一步。参照文献[13]，我们可以将实体抽取的方法分为4种：基于百科站点或垂直站点提取、基于规则与词典的方法、基于统计机器学习的方法以及面向开放域的抽取方法。基于百科站点或垂直站点提取则是一种很常规基本的提取方法；基于规则的方法通常需要为目标实体编写模板，然后在原始语料中进行匹配；基于统计机器学习的方法主要是通过机器学习的方法对原始语料进行训练，然后再利用训练好的模型去识别实体；面向开放域的抽取将是面向海量的Web语料[14]。

1) 基于百科或垂直站点提取

基于百科站点或垂直站点提取这种方法是从百科类站点（如维基百科、百度百科、互动百科等）的标题和链接中提取实体名。这种方法的优点是可以得到开放互联网中最常见的实体名，其缺点是对于中低频的覆盖率低。与一般性通用的网站相比，垂直类站点的实体提取可以获取特定领域的实体。例如从豆瓣各频道(音乐、读书、电影等)获取各种实体列表。这种方法主要是基于爬取技术来实现和获取。基于百科类站点或垂直站点是一种最常规和基本的方法。

2) 基于规则与词典的实体提取方法

早期的实体抽取是在限定文本领域、限定语义单元类型的条件下进行的，主要采用的是基于规则与词典的方法，例如使用已定义的规则，抽取出文本中的人名、地名、组织机构名、特定时间等实体[15]。文献[16]首次实现了一套能够抽取公司名称的实体抽取系统，其中主要用到了启发式算法与规则模板相结合的方法。然而，基于规则模板的方法不仅需要依靠大量的专家来编写规则或模板，覆盖的领域范围有限，而且很难适应数据变化的新需求。

3) 基于统计机器学习的实体抽取方法

鉴于基于规则与词典实体的局限性，为具更有可扩展性，相关研究人员将机器学习中的监督学习算法用于命名实体的抽取问题上。例如文献[17]利用KNN算法与条件随机场模型，实现了对Twitter

文本数据中实体的识别。单纯的监督学习算法在性能上不仅受到训练集合的限制，并且算法的准确率与召回率都不够理想。相关研究者认识到监督学习算法的制约性后，尝试将监督学习算法与规则相互结合，取得了一定的成果。例如文献[18]基于字典，使用最大熵算法在Medline论文摘要的GENIA数据集上进行了实体抽取实验，实验的准确率与召回率都在70%以上。近年来随着深度学习的兴起应用，基于深度学习的命名实体识别得到广泛应用。在文献[19]，介绍了一种基于双向LSTM深度神经网络和条件随机场的识别方法，在测试数据上取得的最好的表现结果。

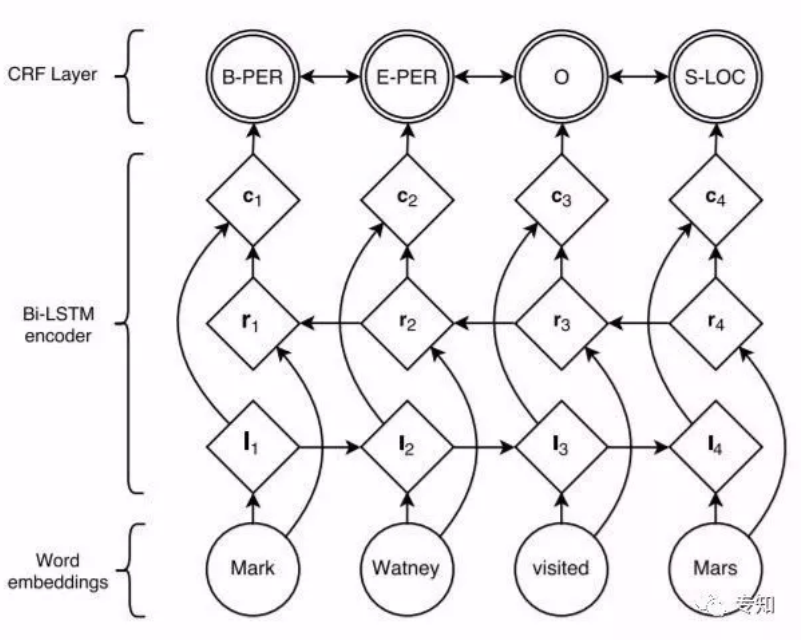

图4 基于BI-LSTM和CRF的架构

4) 面向开放域的实体抽取方法

针对如何从少量实体实例中自动发现具有区分力的模式，进而扩展到海量文本去给实体做分类与聚类的问题，文献[20]提出了一种通过迭代方式扩展实体语料库的解决方案，其基本思想是通过少量的实体实例建立特征模型，再通过该模型应用于新的数据集得到新的命名实体。文献[21]提出了一种基于无监督学习的开放域聚类算法，其基本思想是基于已知实体的语义特征去搜索日志中识别出命名的实体，然后进行聚类。

1.2 语义类抽取

语义类抽取是指从文本中自动抽取信息来构造语义类并建立实体和语义类的关联, 作为实体层面上的规整和抽象。以下介绍一种行之有效的语义类抽取方法，包含三个模块：并列度相似计算、上下位关系提取以及语义类生成 [22]。

1) 并列相似度计算

并列相似度计算其结果是词和词之间的相似性信息，例如三元组（苹果，梨，s1）表示苹果和梨的相似度是s1。两个词有较高的并列相似度的条件是它们具有并列关系（即同属于一个语义类），并

且有较大的关联度。按照这样的标准，北京和上海具有较高的并列相似度，而北京和汽车的并列相似度很低（因为它们不属于同一个语义类）。对于海淀、朝阳、闵行三个市辖区来说，海淀和朝阳的并列相似度大于海淀和闵行的并列相似度（因为前两者的关联度更高）。

当前主流的并列相似度计算方法有分布相似度法（distributional similarity）和模式匹配法（pattern Matching）。分布相似度方法[23-24]基于哈里斯（Harris）的分布假设（distributional hypothesis）[25]，即经常出现在类似的上下文环境中的两个词具有语义上的相似性。分布相似度方法的实现分三个步骤：第一步，定义上下文；第二步，把每个词表示成一个特征向量，向量每一维代表一个不同的上下文，向量的值表示本词相对于上下文的权重；第三步，计算两个特征向量之间的相似度，将其作为它们所代表的词之间的相似度。 模式匹配法的基本思路是把一些模式作用于源数据，得到一些词和词之间共同出现的信息，然后把这些信息聚集起来生成单词之间的相似度。模式可以是手工定义的，也可以是根据一些种子数据而自动生成的。分布相似度法和模式匹配法都可以用来在数以百亿计的句子中或者数以十亿计的网页中抽取词的相似性信息。有关分布相似度法和模式匹配法所生成的相似度信息的质量比较参见文献。

2) 上下位关系提取

该该模块从文档中抽取词的上下位关系信息，生成（下义词，上义词）数据对，例如（狗，动物）、（悉尼，城市）。提取上下位关系最简单的方法是解析百科类站点的分类信息（如维基百科的“分类”和百度百科的“开放分类”）。这种方法的主要缺点包括：并不是所有的分类词条都代表上位词，例如百度百科中“狗”的开放分类“养殖”就不是其上位词；生成的关系图中没有权重信息，因此不能区分同一个实体所对应的不同上位词的重要性；覆盖率偏低，即很多上下位关系并没有包含在百科站点的分类信息中。

在英文数据上用Hearst 模式和IsA 模式进行模式匹配被认为是比较有效的上下位关系抽取方法。下面是这些模式的中文版本（其中NPC 表示上位词，NP 表示下位词）：

NPC { 包括| 包含| 有} {NP、}* [等| 等等]

NPC { 如| 比如| 像| 象} {NP、}*

{NP、}* [{ 以及| 和| 与} NP] 等 NPC

{NP、}* { 以及| 和| 与} { 其它| 其他} NPC

NP 是 { 一个| 一种| 一类} NPC

此外，一些网页表格中包含有上下位关系信息，例如在带有表头的表格中，表头行的文本是其它行的上位词。

3) 语义类生成

该模块包括聚类和语义类标定两个子模块。聚类的结果决定了要生成哪些语义类以及每个语义类包含哪些实体，而语义类标定的任务是给一个语义类附加一个或者多个上位词作为其成员的公共上位词。此模块依赖于并列相似性和上下位关系信息来进行聚类和标定。有些研究工作只根据上下位关系图来生成语义类，但经验表明并列相似性信息对于提高最终生成的语义类的精度和覆盖率都至关重要。

1.3 属性和属性值抽取

属性提取的任务是为每个本体语义类构造属性列表（如城市的属性包括面积、人口、所在国家、地理位置等），而属性值提取则为一个语义类的实体附加属性值。属性和属性值的抽取能够形成完整的实体概念的知识图谱维度。常见的属性和属性值抽取方法包括从百科类站点中提取，从垂直网站中进行包装器归纳，从网页表格中提取，以及利用手工定义或自动生成的模式从句子和查询日志中提取。

常见的语义类/ 实体的常见属性/ 属性值可以通过解析百科类站点中的半结构化信息（如维基百科的信息盒和百度百科的属性表格）而获得。尽管通过这种简单手段能够得到高质量的属性，但同时需要采用其它方法来增加覆盖率（即为语义类增加更多属性以及为更多的实体添加属性值）。

图5 爱因斯坦信息页

由于垂直网站（如电子产品网站、图书网站、电影网站、音乐网站）包含有大量实体的属性信息。例如上图的网页中包含了图书的作者、出版社、出版时间、评分等信息。通过基于一定规则模板建立，便可以从垂直站点中生成包装器（或称为模版），并根据包装器来提取属性信息。从包装器生成的自动化程度来看，这些方法可以分为手工法（即手工编写包装器）、监督方法、半监督法以及无监督法。考虑到需要从大量不同的网站中提取信息，并且网站模版可能会更新等因素，无监督包装器归纳方法显得更加重要和现实。无监督包装器归纳的基本思路是利用对同一个网站下面多个网页的超文本标签树的对比来生成模版。简单来看，不同网页的公共部分往往对应于模版或者属性名，不同的部分则可能是属性值，而同一个网页中重复的标签块则预示着重复的记录。

属性抽取的另一个信息源是网页表格。表格的内容对于人来说一目了然，而对于机器而言，情况则要复杂得多。由于表格类型千差万别，很多表格制作得不规则，加上机器缺乏人所具有的背景知识等原因，从网页表格中提取高质量的属性信息成为挑战。
上述三种方法的共同点是通过挖掘原始数据中的半结构化信息来获取属性和属性值。与通过“阅读”句子来进行信息抽取的方法相比，这些方法绕开了自然语言理解这样一个“硬骨头”而试图达到以柔克刚的效果。在现阶段，计算机知识库中的大多数属性值确实是通过上述方法获得的。但现实情况是只有一部分的人类知识是以半结构化形式体现的，而更多的知识则隐藏在自然语言句子中，因此直接从句子中抽取信息成为进一步提高知识库覆盖率的关键。当前从句子和查询日志中提取属性和属性值的基本手段是模式匹配和对自然语言的浅层处理。图6 描绘了为语义类抽取属性名的主框架（同样的过程也适用于为实体抽取属性值）。图中虚线左边的部分是输入，它包括一些手工定义的模式和一个作为种子的（词，属性）列表。模式的例子参见表3，（词，属性）的例子如（北京，面积）。在只有语义类无关的模式作为输入的情况下，整个方法是一个在句子中进行模式匹配而生成（语义类，属性）关系图的无监督的知识提取过程。此过程分两个步骤，第一个步骤通过将输入的模式作用到句子上而生成一些（词，属性）元组，这些数据元组在第二个步骤中根据语义类进行合并而生成（语义类，属性）关系图。在输入中包含种子列表或者语义类相关模式的情况下，整个方法是一个半监督的自举过程，分三个步骤：

1. 模式生成：在句子中匹配种子列表中的词和属性从而生成模式。模式通常由词和属性的环境信息而生成。
2. 模式匹配。
3. 模式评价与选择：通过生成的（语义类，属性）关系图对自动生成的模式的质量进行自动评价并选择高分值的模式作为下一轮匹配的输入。

1.3 关系抽取

关系抽取的目标是解决实体语义链接的问题。关系的基本信息包括参数类型、满足此关系的元组模式等。例如关系BeCapitalOf（表示一个国家的首都）的基本信息如下：
参数类型：（Capital， Country）
模式：

$$\begin{cases} \{0\}\text{be the capital of }\{1\} \\ \{0\}\text{be the capital in }\{1\} \\ \cdots, \end{cases}$$

元组：（北京，中国）；（华盛顿，美国）；Capital 和 Country表示首都和国家两个语义类。

早期的关系抽取主要是通过人工构造语义规则以及模板的方法识别实体关系。随后，实体间的关系模型逐渐替代了人工预定义的语法与规则。但是仍需要提前定义实体间的关系类型。 文献[26]提出了面向开放域的信息抽取框架 (open information extraction,OIE)，这是抽取模式上的一个巨大进步。但OIE方法在对实体的隐含关系抽取方面性能低下，因此部分研究者提出了基于马尔可夫逻辑网、基于本体推理的深层隐含关系抽取方法[27]。

开放式实体关系抽取

开放式实体关系抽取可分为二元开放式关系抽取和n元开放式关系抽取。在二元开放式关系抽取中，早期的研究有KnowItAll[28]与TextRunner[27]系统，在准确率与召回率上表现一般。文献[29]提出了一种基于Wikipedia的OIE方法WOE，经自监督学习得到抽取器，准确率较TextRunner有明显的提高。针对WOE的缺点，文献[30]提出了第二代OIE ReVerb系统，以动词关系抽取为主。文献[31]提出了第三代OIE系统OLLIE(open language learning for information extraction)，尝试弥补并扩展OIE的模型及相应的系统，抽取结果的准确度得到了增强。
然而，基于语义角色标注的OIE分析显示：英文语句中40%的实体关系是n元的[32]，如处理不当，可能会影响整体抽取的完整性。文献[33]提出了一种可抽取任意英文语句中n元实体关系的方法KPAKEN，弥补了ReVerb的不足。但是由于算法对语句深层语法特征的提取导致其效率显著下降，并不适用于大规模开放域语料的情况。

基于联合推理的实体关系抽取

联合推理的关系抽取中的典型方法是马尔可夫逻辑网MLN(Markov logic network)[34]，它是一种将马尔可夫网络与一阶逻辑相结合的统计关系学习框架，同时也是在OIE中融入推理的一种重要实体关系抽取模型。基于该模型，文献[35]提出了一种无监督学习模型StatSnowball，不同于传统的OIE，该方法可自动产生或选择模板生成抽取器。在StatSnowball的基础上，文献[27,36]提出了一种实体识别与关系抽取相结合的模型EntSum，主要由扩展的CRF命名实体识别模块与基于StatSnowball的关系抽取模块组成，在保证准确率的同时也提高了召回率。文献[27,37]提出了一种简易的Markov逻辑TML(tractable Markov logic)，TML将领域知识分解为若干部分，各部分主要来源于事物类的层次化结构，并依据此结构，将各大部分进一步分解为若干个子部分，以此类推。TML具有较强的表示能力，能够较为简洁地表示概念以及关系的本体结构。

2 知识表示

传统的知识表示方法主要是以RDF(Resource Description Framework资源描述框架)的三元组SPO(subject,property,object)来符号性描述实体之间的关系。这种表示方法通用简单，受到广泛认可，但是其在计算效率、数据稀疏性等方面面临诸多问题。近年来，以深度学习为代表的以深度学习为代表的表示学习技术取得了重要的进展，可以将实体的语义信息表示为稠密低维实值向

量，进而在低维空间中高效计算实体、关系及其之间的复杂语义关联，对知识库的构建、推理、融合以及应用均具有重要的意义[38-40]。

2.1 代表模型

知识表示学习的代表模型有距离模型、单层神经网络模型、双线性模型、神经张量模型、矩阵分解模型、翻译模型等。详细可参见清华大学刘知远的知识表示学习研究进展。相关实现也可参见 [39]。

1）距离模型

距离模型在文献[41] 提出了知识库中实体以及关系的结构化表示方法(structured embedding，SE)，其基本思想是：首先将实体用向量进行表示，然后通过关系矩阵将实体投影到与实体关系对的向量空间中，最后通过计算投影向量之间的距离来判断实体间已存在的关系的置信度。由于距离模型中的关系矩阵是两个不同的矩阵，使得协同性较差。

2）单层神经网络模型

文献[42]针对上述提到的距离模型中的缺陷，提出了采用单层神经网络的非线性模型(single layer model，SLM)，模型为知识库中每个三元组(h,r,t)定义了以下形式的评价函数：

$$f_r(h,t)=u_t^T g(M_{r,1}l_h+M_{r,2}l_t)$$

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

3）双线性模型

双线性模型又叫隐变量模型(latent factor model，LFM)，由文献[43-44]首先提出。模型为知识库中每个三元组(h,r,t)定义的评价函数具有如下形式：

$$f_r(h,t)=l_h^T M_r l_t$$

式中,$M_r\in\mathbb{R}^{d\times d}$是通过关系r 定义的双线性变换矩阵；$l_h l_t\in\mathbb{R}^d$是三元组中头实体与尾实体的向量化表示。双线性模型主要是通过基于实体间关系的双线性变换来刻画实体在关系下的语义相关性。模型不仅形式简单、易于计算，而且还能够有效刻画实体间的协同性。基于上述工作，文献[45]尝试将双线性变换矩阵r M 变换为对角矩阵， 提出了DISTMULT模型，不仅简化了计算的复杂度，并且实验效果得到了显著提升。

3）神经张量模型

文献[45]提出的神经张量模型，其基本思想是：在不同的维度下，将实体联系起来，表示实体间复杂的语义联系。模型为知识库中的每个三元组(h,r,t)定义了以下形式的评价函数：

$$f_r(h,t)=u_t^T g(l_h M_r l_t + M_{r,1} l_h + M_{r,2} l_t + b_r)$$

式中，$u_t^T \in \mathbb{R}^k$关系 r 的向量化表示；$g()$为tanh函数；$M_r \in d \times d \times k$是一个三阶张量；$M_{r,1} M_{r,2} \in \mathbb{R}^k$是通过关系r定义的两个矩阵。

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

4）矩阵分解模型

通过矩阵分解的方式可得到低维的向量表示，故不少研究者提出可采用该方式进行知识表示学习，其中的典型代表是文献[46]提出的RESACL模型。在RESCAL模型中，知识库中的三元组(h,r,t)集合被表示为一个三阶张量，如果该三元组存在，张量中对应位置的元素被置1，否则置为0。通过张量分解算法，可将张量中每个三元组(h,r,t)对应的张量值解为双线性模型中的知识表示形式$l_h^T M_r l_t$并使$|X_{hrt} - l_h^T M_r l_t|$尽量小。

5）翻译模型

文献[47]受到平移不变现象的启发，提出了TransE模型，即将知识库中实体之间的关系看成是从实体间的某种平移，并用向量表示。关系l_r可以看作是从头实体向量l_h 到尾实体向量l_t 的翻译。对于知识库中的每个三元组(h,r,t),TransE都希望满足以下关系：$|l_h + l_t \approx l_t|$，其损失函数为：$f_r(h,t) = |l_h + l_r - l_t|_{L_1/L_2}$, 该模型的参数较少，计算的复杂度显著降低。与此同时，TransE模型在大规模稀疏知识库上也同样具有较好的性能和可扩展性。

2.2 复杂关系模型

知识库中的实体关系类型也可分为1-to-1、1-to-N、N-to-1、N-to-N4种类型[47]，而复杂关系主要指的是1-to-N、N-to-1、N-to-N的3种关系类型。由于TransE模型不能用在处理复杂关系上[39]，一系列基于它的扩展模型纷纷被提出，下面将着重介绍其中的几项代表性工作。

1）TransH模型

文献[48]提出的TransH模型尝试通过不同的形式表示不同关系中的实体结构，对于同一个实体而言，它在不同的关系下也扮演着不同的角色。模型首先通过关系向量l_r与其正交的法向量w_r 选取某一个超平面F， 然后将头实体向量l_h 和尾实体向量l_t法向量w_r的方向投影到F, 最后计算损失函数。TransH使不同的实体在不同的关系下拥有了不同的表示形式，但由于实体向量被投影到了关系的语义空间中，故它们具有相同的维度。

2）TransR模型

由于实体、关系是不同的对象，不同的关系所关注的实体的属性也不尽相同，将它们映射到同一个语义空间，在一定程度上就限制了模型的表达能力。所以，文献[49]提出了TransR模型。模型首先将知识库中的每个三元组(h, r,t)的头实体与尾实体向关系空间中投影，然后希望满足$|l_h + l_t \approx l_t|$的关系，最后计算损失函数。

文献[49]提出的CTransR模型认为关系还可做更细致的划分，这将有利于提高实体与关系的语义联系。在CTransR模型中，通过对关系r 对应的头实体、尾实体向量的差值$l_h - l_t$进行聚类，可将r分为若干个子关系r_c 。

3）TransD模型

考虑到在知识库的三元组中，头实体和尾实体表示的含义、类型以及属性可能有较大差异，之前的TransR模型使它们被同一个投影矩阵进行映射，在一定程度上就限制了模型的表达能力。除此之外，将实体映射到关系空间体现的是从实体到关系的语 义联系，而TransR模型中提出的投影矩阵仅考虑了不同的关系类型，而忽视了实体与关系之间的交互。因此，文献[50]提出了TransD模型，模型分别定义了头实体与尾实体在关系空间上的投影矩阵。

4）TransG模型

文献[51]提出的TransG模型认为一种关系可能会对应多种语义，而每一种语义都可以用一个高斯分布表示。TransG模型考虑到了关系r 的不同语义，使用高斯混合模型来描述知识库中每个三元组(h,r,t)头实体与尾实体之间的关系，具有较高的实体区分度。

5）KG2E模型

考虑到知识库中的实体以及关系的不确定性，文献[52]提出了KG2E模型，其中同样是用高斯分布来刻画实体与关系。模型使用高斯分布的均值表示实体或关系在语义空间中的中心位置，协方差则表示实体或关系的不确定度。

知识库中，每个三元组(h,r,t)的头实体向量l_h与尾实体向量l_t间的

$$P_e = l_h - l_t \sim N(\mu_h - \mu_t, \Sigma_h + \Sigma_r)$$

关系r可表示为：

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

3 知识融合

通过知识提取，实现了从非结构化和半结构化数据中获取实体、关系以及实体属性信息的目标。但是由于知识来源广泛，存在知识质量良莠不齐、来自不同数据源的知识重复、层次结构缺失等问题，所以必须要进行知识的融合。知识融合是高层次的知识组织[53]，使来自不同知识源的知识在同一框架规范下进行异构数据整合、消歧、加工、推理验证、更新等步骤[54]，达到数据、信息、方法、经验以及人的思想的融合，形成高质量的知识库。

3.1 实体对齐

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

在大数据的环境下，受知识库规模的影响，在进行知识库实体对齐时，主要会面临以下3个方面的挑战[55]：1) 计算复杂度。匹配算法的计算复杂度会随知识库的规模呈二次增长，难以接受；2) 数据质量。由于不同知识库的构建目的与方式有所不同，可能存在知识质量良莠不齐、相似重复数据、孤立数据、数据时间粒度不一致等问题[56]；3) 先验训练数据。在大规模知识库中想要获得这种先验数据却非常困难。通常情况下，需要研究者手工构造先验训练数据。

基于上述，知识库实体对齐的主要流程将包括[55]：1) 将待对齐数据进行分区索引，以降低计算的复杂度；2) 利用相似度函数或相似性算法查找匹配实例；3) 使用实体对齐算法进行实例融合；4) 将步骤2)与步骤3)的结果结合起来，形成最终的对齐结果。对齐算法可分为成对实体对齐与集体实体对齐两大类，而集体实体对齐又可分为局部集体实体对齐与全局集体实体对齐。

1）成对实体对齐方法

① 基于传统概率模型的实体对齐方法

基于传统概率模型的实体对齐方法主要就是考虑两个实体各自属性的相似性，而并不考虑实体间的关系。文献[57]将基于属性相似度评分来判断实体是否匹配的问题转化为一个分类问题，建立了该问题的概率模型，缺点是没有体现重要属性对于实体相似度的影响。文献[58]基于概率实体链接模型，为每个匹配的属性对分配了不同的权重，匹配准确度有所提高。文献[59]还结合贝叶斯网络对属性的相关性进行建模，并使用最大似然估计方法对模型中的参数进行估计。

② 基于机器学习的实体对齐方法

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

通过属性比较向量来判断实体对匹配与否可称为成对实体对齐。这类方法中的典型代表有决策树[60]、支持向量机[61]、集成学习[62]等。文献[63]使用分类回归树、线性分析判别等方法完成了实体辨析。文献[64]基于二阶段实体链接分析模型，提出了一种新的SVM分类方法，匹配准确率远高于TAILOR中的混合算法。

基于聚类的实体对齐算法，其主要思想是将相似的实体尽量聚集到一起，再进行实体对齐。文献[65]提出了一种扩展性较强的自适应实体名称匹配与聚类算法，可通过训练样本生成一个自适应的距离函数。文献[66]采用类似的方法，在条件随机场实体对齐模型中使用监督学习的方法训练产生距离函数，然后调整权重，使特征函数与学习参数的积最大。

在主动学习中，可通过与人员的不断交互来解决很难获得足够的训练数据问题，文献[67]构建的ALIAS系统可通过人机交互的方式完成实体链接与去重的任务。文献[68]采用相似的方法构建了ActiveAtlas系统。

2）局部集体实体对齐方法

局部集体实体对齐方法为实体本身的属性以及与它有关联的实体的属性分别设置不同的权重，并通过加权求和计算总体的相似度，还可使用向量空间模型以及余弦相似性来判别大规模知识库中的实体的相似程度[69]，算法为每个实体建立了名称向量与虚拟文档向量，名称向量用于标识实体的属性，虚拟文档向量则用于表示实体的属性值以及其邻居节点的属性值的加权和值[55]。为了评价向量中每个分量的重要性，算法主要使用TF-IDF为每个分量设置权重，并为分量向量建立倒排索引，最后选择余弦相似性函数计算它们的相似程度[55]。该算法的召回率较高，执行速度快，但准确率不足。其根本原因在于没有真正从语义方面进行考虑。

3）全局集体实体对齐方法

① 基于相似性传播的集体实体对齐方法

基于相似性传播的方法是一种典型的集体实体对齐方法，匹配的两个实体与它们产生直接关联的其他实体也会具有较高的相似性，而这种相似性又会影响关联的其他实体[55]。

相似性传播集体实体对齐方法最早来源于文献[70-71]提出的集合关系聚类算法，该算法主要通过一种改进的层次凝聚算法迭代产生匹配对象。文献[72]在以上算法的基础上提出了适用于大规模知识库实体对齐的算法SiGMa，该算法将实体对齐问题看成是一个全局匹配评分目标函数的优化问题进行建模，属于二次分配问题，可通过贪婪优化算法求得其近似解。SiGMa方法[55]能够综合考虑实体对的属性与关系，通过集体实体的领域，不断迭代发现所有的匹配对。

② 基于概率模型的集体实体对齐方法基于概率模型的集体实体对齐方法主要采用统计关系学习进行计算与推理，常用的方法有LDA模型[73]、CRF模型[74]、Markov逻辑网[75]等。

文献[73]将LDA模型应用于实体的解析过程中，通过其中的隐含变量获取实体之间的关系。但在大规模的数据集上效果一般。文献[74]提出了一种基于图划分技术的CRF实体辨析模型，该模型以观察值为条件产生实体判别的决策，有利于处理属性间具有依赖关系的数据。文献[66]在CRF实体辨析模型的基础上提出了一种基于条件随机场模型的多关系的实体链接算法，引入了基于canopy的索引，提高了大规模知识库环境下的集体实体对齐效率。文献[75]提出了一种基于Markov逻辑网的实体解析方法。通过Markov逻辑网，可构建一个Markov网，将概率图模型中的最大可能性计

算问题转化为典型的最大化加权可满足性问题，但基于Markov网进行实体辨析时，需要定义一系列的等价谓词公理，通过它们完成知识库的集体实体对齐。

3.2 知识加工

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

1）本体构建

本体是同一领域内不同主体之间进行交流、连通的语义基础[76]，其主要呈现树状结构，相邻的层次节点或概念之间具有严格的“IsA”关系，有利于进行约束、推理等，却不利于表达概念的多样性。本体在知识图谱中的地位相当于知识库的模具，通过本体库而形成的知识库不仅层次结构较强，并且冗余程度较小[77]。

本体可通过人工编辑的方式手动构建，也可通过数据驱动自动构建，然后再经质量评估方法与人工审核相结合的方式加以修正与确认。在海量的实体数据面前，人工编辑构建的方式工作量极其巨大，故当前主流的本体库产品，都是面向特定领域，采用自动构建技术而逐步扩展形成的。例如Microsoft的Probase本体库就是采用数据驱动的方法，利用机器学习算法从网页文本中抽取概念间的“IsA”关系，然后合并形成概念层次结构。目前，Probase所包含的概念总数已达到千万级别，准确率高达92.8%，是目前为止包含概念数量最多，同时也是概念可信程度最高的知识库[78]。

数据驱动的本体自动构建过程主要可分为以下3个阶段[79]：① 纵向概念间的并列关系计算。通过计算任意2个实体间并列关系的相似度，可辨析它们在语义层面是否属于同一个概念。计算方法主要包括模式匹配与分布相似度两种[80]。② 实体上下位关系抽取。上下位关系抽取方法包括基于语法的抽取与基于语义的抽取两种方式，例如目前主流的信息抽取系统KnowItAll、TextRunner、NELL[81]等，都可以在语法层面抽取实体的上下位关系，而Probase则是采用基于语义的抽取模式[82]。③ 本体生成。对各层次得到的概念进行聚类，并为每一类的实体指定1个或多个公共上位词。文献[83]基于主题层次聚类的方法构建了本体结构。与此同时，为了解决主题模型不适用于短文本的问题，提出了基于单词共现网络的主题聚类与上下位词抽取模型。

2）质量评估

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

文献[84]基于LDIF框架，提出了一种新的知识质量评估方法，用户可根据业务需求来定义质量评估函数，或者通过对多种评估方法的综合考评来确定知识的最终质量评分。例如在对REVERRB系统的信息抽取质量进行评估时，文献[85]采用人工标注的方式对1 000个句子中的实体关系三元组进行了标注，并以此作为训练集，使用logistic回归模型计算抽取结果的置信度。例如Google的

KnowledgeVault项目则根据指定数据信息的抽取频率对信息的可信度进行评分，然后利用从可信知识库中得到的先验知识对可信度进行修正。实验结果表明：该方法可以有效地降低对数据信息正误判断的不确定性，提高知识的质量[85]。

3.2 知识更新

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

根据知识图谱的逻辑结构，其更新主要包括模式层的更新与数据层的更新。模式层的更新是指本体中元素的更新，包括概念的增加、修改、删除，概念属性的更新以及概念之间上下位关系的更新等。其中，概念属性的更新操作将直接影响到所有直接或间接属性的子概念和实体[87]。通常来说，模式层的增量更新方式消耗资源较少，但是多数情况下是在人工干预的情况下完成的，例如需要人工定义规则，人工处理冲突等。因此，实施起来并不容易[88]。数据层的更新指的是实体元素的更新，包括实体的增加、修改、删除，以及实体的基本信息和属性值。由于数据层的更新一般影响面较小，因此通常以自动的方式完成。

(未完待续，敬请明天继续关注)

参考文献

1. Fabian M. Suchanek, Gerhard Weikum: Knowledge harvesting from text and Web sources. ICDE 2013: 1250-1253.
2. Gerhard Weikum, Martin Theobald: From information to knowledge: harvesting entities and relationships from web sources. PODS 2010: 65-76.
3. A. Singhal. Introducing the Knowledge Graph: things, not strings. Official Google Blog, May, 2012.
4. Gallagher, Sean. (June 7, 2012). How Google and Microsoft taught search to understand the Web http://arstechnica.com/information-technology/2012/06/inside-the-architecture-of-googles- knowledge-graphand-microsofts-satori/.
5. Omkar Deshpande, Digvijay S. Lamba, Michel Tourn, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, AnHai Doan: Building, maintaining, and using knowledge bases: a report from the trenches. SIGMOD Conference 2013: 1209-1220.
6. AMIT S. Introducing the knowledge graph[R]. America:Official Blog of Google, 2012.

7. Shenshouer. Neo4j[EB/OL]. [2016-05-09]. http://neo4j.com/.
8. FlockDB Official. FlockDB[EB/OL]. [2016-05-09]. http://webscripts.softpedia.com/script/Database-Tools/FlockDB-66248.html.
9. Graphdb Official. Graphdb[EB/OL]. [2016-05-09]. http://www.graphdb.net/.
10. 刘峤, 李杨, 杨段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582-600.LIU Qiao, LI yang, YANG Duan-hong, et al. Knowledgegraph construction techniques[J]. Journal of Computer Research and Development, 2016, 53(3): 582-600.
11. DONG X, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a web-scale approach to probabilistic knowledgefusion[C]//Proc of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM, 2014.
12. BOLLACKER K, COOK R, TUFTS P. Freebase: a shared database of structured general human knowledge[C]//Proc of the 22nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2007: 1962-1963.
13. 孙镇, 王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术, 2010(6): 42-47.
14. 赵军, 刘康, 周光有, 等. 开放式文本信息抽取[J]. 中文信息学报, 2011, 25(6): 98-110.
15. CHINCHOR N, MARSH E. Muc-7 information extraction task definition[C]//Proc of the 7th Message Understanding Conf. Philadelphia: Linguistic Data Consortium, 1998:359-367.
16. RAU L F. Extracting company names from text[C]//Proc of the 7th IEEE Conf on Artificial Intelligence Applications.Piscataway, NJ: IEEE, 1991: 29-32.
17. LIU Xiao-hua, ZHANG Shao-dian, WEI Fu-ru, et al. Recognizing named entities in tweets[C]//Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 359-367.
18. LIN Yi-feng, TSAI T, CHOU Wen-chi, et al. A maximum entropy approach to biomedical named entity recognition[C]//Proc of the 4th ACM SIGKDD Workshop on Data Mining in Bioinformatics. New York: ACM, 2004.
19. Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer: Neural Architectures for Named Entity Recognition. HLT-NAACL 2016: 260-270
20. WHITELAW C, KEHLENBECK A, PETROVIC N, et al. Web-scale named entity recognition[C]//Proc of the 17th ACM Conf on Information and Knowledge Management. New York: ACM, 2008.
21. JAIN A, PENNACCHIOTTI M. Open entity extraction from web search query logs[C]//Proc of the 23rd Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2010 :510-518.
22. 史树明. "自动和半自动知识提取." 中国计算机学会通讯 9, no. 8 (2013): 65-73.

23. P. Pantel, E. Crestan, A. Borkovsky, A.-M. Popescu, and V. Vyas. Web-scale distributional similarity and entity set expansion. EMNLP'2009. Singapore
24. P. Pantel, D. Lin. Discovering word senses from text. SIGKDD'2002
25. Z. Harris. Distributional structure. The Philosophy of Linguistics. 1985
26. BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction for the Web[C]//Proc of the 20th Int Joint Conf on Artificial Intelligence. New York: ACM, 2007: 2670-2676.
27. 杨博, 蔡东风, 杨华. 开放式信息抽取研究进展[J]. 中文信息学报, 2014, 4: 1-11.
28. ETZIONI O, CAFARELLA M, DOWNEY D, et al. Unsupervised named-entity extraction from the Web: an experimental study[J]. Artificial Intelligence, 2005, 165(1): 91-134.
29. WU F, WELD D S. Open information extraction using Wikipedia[C]//Proceedings of Annual Meeting of the Association for Computational Linguistics. Sweden: ACL, 2010: 118-127.
30. FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing. Washington: EMNLP, 2015.
31. SCHMITZ M M, BART R, SODERLAND S, et al. Open language learning for information extraction[C]// Proceedings of Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning(EMNLP-CONLL). New York: ACM,2012.
32. JANARA C M, STEPHEN S, OREN E. An analysis of open information extraction based on semantic role labeling[C]//Proceedings of K-CAP. New York: ACM, 2011: 113-120.
33. AKBIK A, LOSER A. KRAKEN: N-ary facts in open information extraction[C]//Proceedings of AKBC-WEKEX at NAACL. New York: ACM, 2012: 52-56.
34. DOMINGOS P, LOWD D. Markov logic: an interface layer for artificial intelligence[M]. San Rafael, CA: Morgan & Claypool, 2009.
35. ZHU Jun, NIE Zai-qing, LIU Xiao-jiang, et al. StatSnowball: a statistical approach to extracting entity relationships[C]//Proceedings of the 18th International Conference on World Wide Web. 2009.
36. LIU Xiao-jiang, YU Neng-hai. People summarization by combining named entity recognition and relation extraction[J]. Journal of Convergence Information Technology, 2010, 5(10): 233-241.
37. DOMINGOS P, WEBB A. A tractable first-order probabilistic logic[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence. San Francisco, CA: AAAI, 2012.

38. BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1): 1-7.
39. 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 1-16.
40. 徐增林, 盛泳潘, 贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4):589-606.
41. BORDES A, WESTON J, COLLOBERT R, et al. Learnting structured embeddings for knowledgebases[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2011:301-306.
42. SOCHER R, CHEN D, MANNING C D, et al. Reasoning with neural tensor networks for knowledge base completion [C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 926-934.
43. JENATTON R, ROUX N L, BORDES A, et al. A latent factor model for highly multi-relational data[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2012: 3167-3175.
44. SUTSKEVER I, TENENBAUM J B, SALAKHUTDINOV R. Modelling relational data using Bayesian clustered tensor factorization[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2009: 1821-1828.
45. YANG B, YIH W, HE X, et al. Embedding enties and relations for learning and inference in knowledge bases[C]//Proc of Int Conf on Learning Representations (ICLR). France: ICLR Press, 2015.
46. NICKEL M, TRESP V, KRIEGEL H. A three-way model for collective learning on multi-relational data[C]//Proc of ICML. New York: ACM, 2011: 809-816.
47. BORDES A, USUNIER N, GARCIA-DURAN A, et al. Translating embeddings for modeling multi-relational data[C]//Proc of NIPS. Cambridge, MA: MIT Press, 2013: 2787-2795.
48. WANG Z, ZHANG J, FENG J, et al. Knowledge graph embedding by translating on hyperplanes[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2014: 1112-1119.
49. LIN Y, LIU Z, SUN M, et al. Learning entity and relation embedding for knowledge graph completion[C]//Proc of AAAI. Menlo Park, CA: AAAI, 2015.
50. JI G, HE S, XU L, et al. Knowledge graph embedding via dynamic mapping matrix[C]//Proc of ACL. Stroudsburg, PA: ACL, 2015: 687-696.
51. XIAO H, HUANG M, HAO Y, et al. TransG: a generative mixture model for knowledge graph embedding[J]. Arxiv Preprint ArXiv, 2015, 1509: 05488.
52. HE S, LIU K, JI J, et al. Learning to represent knowledge graphs with Gaussian embedding[C]//Proc of CIKM. New York: ACM, 2015: 623-632.
53. 徐绪堪, 房道伟, 蒋勋, 等. 知识组织中知识粒度化表示和规范化研究[J]. 图书情报知识, 2014(6): 101-106,90.
54. 张坤. 面向知识图谱的搜索技术(搜狗)[EB/OL].(2015-02-18). http://www.cipsc.org.cn/kg1/.ZHANG Kun.

55. 庄严, 李国良, 冯建华. 知识库实体对齐技术综述[J].计算机研究与发展, 2016, 01: 165-192.
56. 蒋勋, 徐绪堪. 面向知识服务的知识库逻辑结构模型[J].图书与情报, 2013(6): 23-31.
57. NEWCOMBE H B, KENNEDY J M, AXFORD S J, et al. Automatic linkage of vital records[J]. Science, 1959, 130(3381): 954-959.
58. HERZOG T N, SCHEUREN F J, WINKLER W E. Data quality and record linkage techniques[M]. Berlin: Springer, 2007.
59. WINKLER W E. Methods for record linkage and Bayesian networks, RRS2002/05[R]. Washington DC: US Bureau of the Census, 2001.
60. HAN J W, KAMBE M. Data mining: Concepts and techniques[M]. San Francisco, CA: Morgan Kaufmann, 2006.
61. VAPNIK V. The nature of statistical learning theory[M]. Berlin: Springer, 2000.
62. KANTARDZIC M. Data mining[M]. Hoboken, NJ: John Wiley & Sons, 2011.
63. COCHINWALA M, KURIEN V, LALK G, et al. Efficient data reconciliation[J]. Information Sciences, 2011, 137(14): 1-15.
64. CHRISTEN P. Automatic training example selection for scalable unsupervised record linkage[C]//LNAI 5012: Advances in Knowledge Discovery and Data Mining, 12th Pacific-Asia Conf. Berlin: Springer, 2008.
65. COHEN W W, RICHMAN J. Learning to match and cluster large high-dimensional data sets for data integration[C]//Proc of the 2002 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 475-480.
66. MCCALLUM A, WELLNER B. Conditional models of identity uncertainty with application to noun coreference[C]//Proc of Advances in Neural Information Processing System. Cambridge, MA: MIT Press, 2005: 905-912.
67. SARAWAGI S, BHAMIDIPATY A. Interactive deduplication using active learning[C]//Proc of the 2002 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 269-278.
68. TEJADA S, KNOBLOCK C A, MINTON S. Learning domain independent string transformation weights for high accuracy object identification[C]//Proc of the 2002 ACM
69. LI Juan-zi, WANG Zhi-chun, ZHANG Xiao, et al. Large scale instance matching via multiple indexes and candidate selection[J]. Knowledge-Based Systems, 2013, 50: 112-120.
70. DONG X. Reference reconciliation in complex information spaces[C]//Proc of the 2005 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2005: 85-96.
71. BHATTACHARYA I, GETOOR L. Collective entity resolution in relational data[J]. ACM Trans on Knowledge Discovery from Data, 2007, 1(2): 9-15.

72. LACOSTE-Julien S, PALLA K, DAVIES A, et al. SIGMA: Simple greedy matching for aligning large knowledge bases[C]//Proc of the 2013 ACM SIGKDD Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 572-580.
73. BHATTACHARYA I, GETOOR L. Alatent dirichlet allocation model for unsupervised entity resolution[C]// Proc of the 6th SIAM Int Conf on Data Mining. Philadelphia, PA: SIAM, 2006: 47-58.
74. DOMINGOS P. Multi-relational record linkage[C]//Proc of the KDD-2004 Workshop on Muti-Relational Data Mining. New York: ACM, 2004.
75. SINGLA P, DOMINGOS P. Entity resolution with Markov logic[C]//Proc of 2006 IEEE Int Conf on Data Mining(ICDM 2006). Piscataway, NJ: IEEE, 2006.
76. STUDER R, BENJAMINS V R, FENSEL D. Knowledge engineering: Principles and methods[J]. Data & Knowledge Engineering, 1998, 25(1): 161-197.
77. WONG W, LIU Wei, BENNAMOUN M. Ontology learning from text: a look back and into the future[J]. ACM Computing Surveys, 2012, 44(4): 18-24.
78. WU Wen-tao, LI Hong-song, WANG Hai-xun, et al. Probase: a probabilistic taxonomy for text understanding [C]//Proc of the 31st ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2012.
79. 史树明. 自动和半自动知识提取[J]. 中国计算机学会通讯, 2013, 9(8): 65-73.
80. HARRIS Z S. Distributional structure[J]. Word, 1954, 10(23): 146-162.
81. Carnegie Mellon University. NELL[EB/OL]. [2016-06-08]. http://rtw.ml.cmu.edu/rtw/.
82. ZENG Yi, WANG Dong-sheng, ZHANG Tie-lin, et al. CASIA-KB: a multi-source Chinese semantic knowledge base built from structured and unstructured Web data[C]//Semantic Technology. Berlin: Springer, 2014:
75-88.
83. WANG C, DANILEVSKY M, DESAI N, et al. A phrase mining framework for recursive construction of a topical Hierarchy[C]//Proc of the 19th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM,2013: 437-445.
84. FADER A, SODERLAND S, ETZIONI O. Identifying relations for open information extraction[C]//Proc of the Conf on Empirical Methods in Natual Language Processing. Stroudsburg, PA: ACL, 2011: 1535-1545.
85. MENDES P N, MUHLEISEN H, BIZER C. Sieve: Linked data quality assessment and fusion[C]//Proc of the 2nd Int Workshop on Linked Web Data Management at Extending Database Technology. New York: ACM, 2012: 116-123.
86. DONG Xin, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion[C]//Proc of the 20th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2014: 601-610.

87. TAN C H, AGICHTEIN E, IPEIROTIS P, et al. Trust, but verify: Predicting contribution quality for knowledge base construction and curation[C]//Proc of the 7th ACM Int Conf on Web Search and Data Mining. New York: ACM, 2014: 553-562.
88. 耿霞, 张继军, 李蔚妍. 知识图谱构建技术综述[J]. 计算机科学, 2014, 41(7): 148-152.

-END-

欢迎使用专知

专知，提供一个新的认知方式。目前主要聚焦在**人工智能、AI技术、算法**等内容，为科研工作者、人工智能领域从业者提供专业可信的知识服务。

访问使用方法>>点击文章下方“**阅读原文**”访问专知网站。

专·知

长按识别二维码，关注了解更多。

微信号：Quan_Zhuanzhi

一站式AI知识服务

基于知识图谱的内容分发

阅读原文