AIX专家俱乐部

939 文章 12万 总阅读

查看TA的文章>

最全的知识图谱技术综述 | 收藏

2017-10-09 07:25 技术 / 百度 / 操作系统

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。以下内容涵盖了基本定义与架构、代表性知识图谱库、构建技术、开源库和典型应用。

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,其中，是知识库中的实体集合，共包含|E|种不同实体； 是知识库中的关系集合，共包含|R|种不同关系；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-

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。

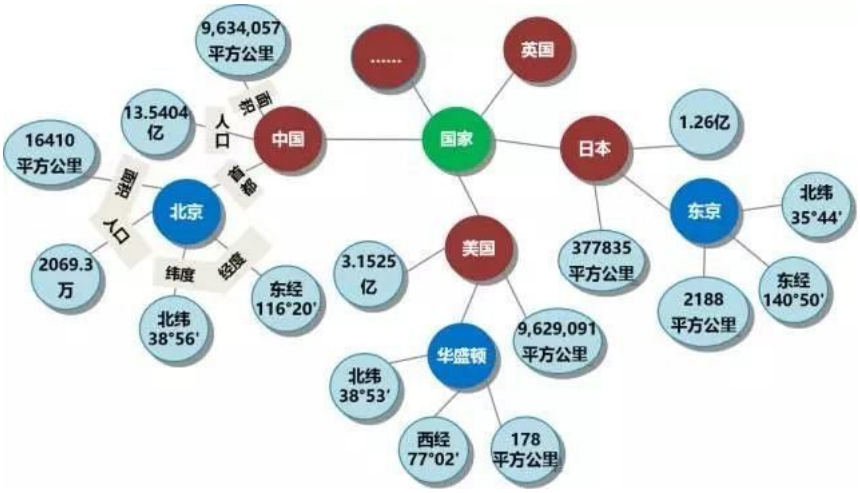

图1 知识图谱示例

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

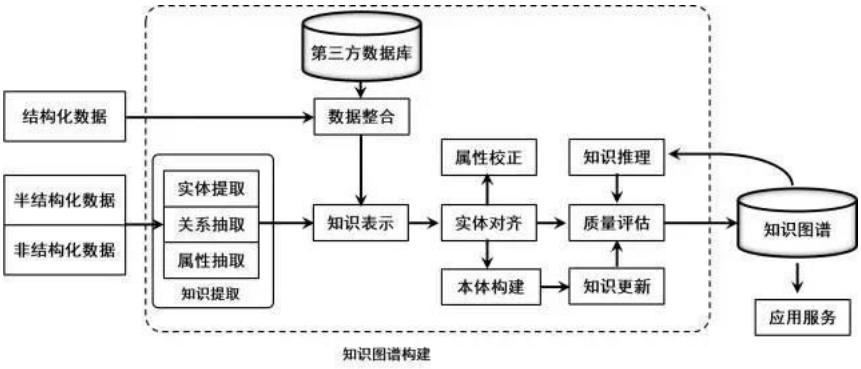

图2 知识图谱的技术架构

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

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

（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。

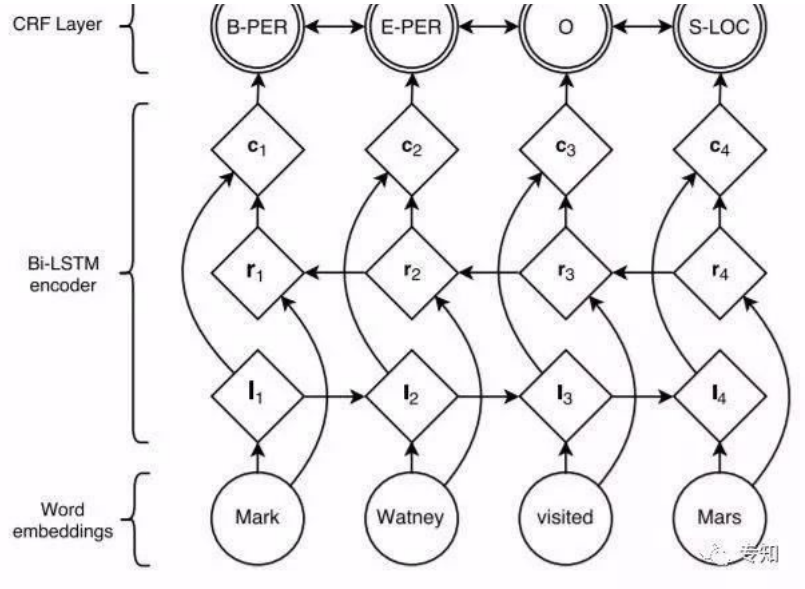

图4 基于BI-LSTM和CRF的架构

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

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

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：

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。

模式匹配。

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

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 Deion 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， ut的T次幂∈R的k次幂为关系 r 的向量化表示；g()为tanh函数； Mr,1×Mr,2∈R的k次幂是通过关系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

型为知识库中每个三元组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,Mr∈R的d×d次幂是通过关系r 定义的双线性变换矩阵；

lh×lt∈R的d次幂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。

4）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：

式中， ut的T次幂∈R的k次幂为关系 r 的向量化表示；g()为tanh函数； Mr∈d×k×k是一个三阶张量；Mr,1×Mr,2∈R的k次幂是通过关系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。

5）矩阵分解模型

通过矩阵分解的方式可得到低维的向量表示，故不少研究者提出可采用该方式进行知识表示学习，其中的典型代表是文献[46]提出的RESACL模型。在RESCAL模型中，知识库中的三元组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lh的T次幂×Mr×lt并使|Xhrt-lh的T次幂×Mr×l|尽量小。

6）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lr可以看作是从头实体向量到尾实体向量lt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|lh+lt≈lt|：，其损失函数为：fr(h,t)=‖h+lr-lt‖L1/L2, 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

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lr与其正交的法向量wr选取某一个超平面F， 然后将头实体向量lh和尾实体向量lt法向量wr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||h+lt≈lt||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lh-lt进行聚类，可将r分为若干个子关系rc 。

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。（本文来自网络，著作权归原作者所有） 返回搜狐，查看更多

还没有评论，快来抢沙发吧！

推荐
微信
魅族
HTC
快手
ICO
谷歌
乐视
OPPO
骁龙
华米
安卓
CES

推荐阅读

珠海银隆被爆拖欠货款超10亿，新能源造车难道全在“裸奔”？

投资界 · 今天 00:49 2

阿里云否认提供“挖矿平台”和虚拟货币，BAT对区块链表态谨慎

全天候科技 · 今天 08:30

切断AT&T与华为联系，反对中国移动进入，美国在顾虑什么？

Trendforce集邦 · 昨天 19:20 3

一加海外官网疑遭入侵，用户支付信息泄漏导致信用卡欺诈事件发生

黑客与极客 · 今天 08:30

手机业寒冬已至 厂商、渠道商举步维艰

21世纪经济报道 · 今天 05:39

怎样最廉价地体验外星生活？钻进滚筒洗衣机

不存在日报 · 昨天 21:22

大家都在闷声搞「全面屏」，中兴却发布了一台折叠手机 Axon M

爱范儿 · 昨天 22:02 7

36氪独家 | 滴滴不仅反对了ofo拿阿里融资，还自己造了20万台单车

36氪 · 今天 07:11

退休大爷，用闲钱炒股，跟着微信提示买，2周后乐坏了！

广告 · 今天 9:30

“数说”2018第一风口 谁是直播答题大“撒币”不能只看砸钱

懂懂笔记 · 今天 02:05

比特币“挖矿”太耗电？俄罗斯投资人干脆买下两座发电站

华尔街见闻 · 昨天 17:50 1

民航局承认开放机上PED使用条件基本成熟 最终实施仍需等待

金羊网 · 昨天 23:30

阿里云声明不会提供挖矿平台和虚拟货币；一加开始调查泄露用户信用卡信息…

极客公园 · 今天 07:40 1

经典老茶，陈皮普洱198元/桶，老板们都在喝

广告 · 今天 9:30

美国三大股指悉数收跌，迅雷收跌6.43%

IT之家 · 今天 08:19

90后00后们最爱用的十大APP，你的手机里有几款？

21世纪商业评论 · 昨天 23:32

启示录：2017年裁员最多的25家美国机构

资本实验室 · 今天 08:25

趣店罗敏终现身：过去是我错了，接下来我要做另一个千亿的生意

36氪 · 昨天 12:13

把握不好买卖点的散户注意了，这个微信群里每天都会提示！【免费入群】

广告 · 今天 9:30

深度长文：面对数字化颠覆，成功企业应当坚守 10 个原则

36氪 · 今天 07:21

Google Maps 将恢复中国完整体验，你的手机很快也能用了

动点科技 · 昨天 09:24

带上王健林，董明珠砸下30亿的这家公司，被供应商拉横幅讨债！

每日经济新闻 · 今天 08:02 2

厅能否迎来大爆发？

创业邦 · 今天 08:27

智者，大成。梅赛德斯-奔驰长轴距E级车

广告 · 今天 9:30

加载更多