《知识图谱》读书报告

194568 郑秋硕

# 第一章 概述

## 1.1 什么是知识图谱

根据维基百科的描述，知识图谱是谷歌公司用来知识从语义角度组织网络数据，从而提供智能搜索服务的知识库。知识图谱使用结构化三元组的形式存储现实世界中的实体以及实体关系。

## 1.2 知识图谱发展历程

知识图谱的发展历程可以从两个领域追溯，一个是人工智能，另一个是语义网。

对于人工智能领域来说，知识图谱其实是伴随着人工智能技术发展而发展的。由于人工智能中对于在信息获取、分析、理解的需要，知识工程这一概念首先被提出。之后的研究者使用各种表示方法构建知识库，这就是知识图谱。

对于语义网来说，万维网现在采用的超文本标记语言HTML，是面向人的存储和共享信息的媒介，Web上的内容是提供给人而不是机器来理解和浏览的。由于Web内容没有采用形式化的表示方式，并且缺乏明确的语义信息，故而计算机“看到的”Web内容只是普通的二进制数据，对其内容无法进行识别。如果机器不能充分理解网页内容的含义，就无法实现Web内容的自动处理。

另外，万维网是根据网页的地址，而非内容的语义来定位信息资源的（缺少语义联接），网上所有信息都是由不同的网站发布的，相同主题的信息分散在全球众多不同的服务器上，又缺少有效工具能将不同来源的相关信息综合起来，因此形成了一个个信息孤岛(其中有大量重复)，查找自己所需的信息就像大海捞针一样困难。

语义网就是以Web数据的内容（即数据的语义）为核心，用机器能够理解和处理的方式链接起来的海量分布式数据库。语义网是对现有Web增加了语义支持，它是现有万维网的延伸与变革，其目标是帮助机器在一定程度上理解Web信息的含义，使得高效的信息共享和机器智能协同成为可能。语义网将会为用户提供动态、主动的服务，从而更便于机器和机器、人和机器之间的对话及协同工作。

## 1.3 知识图谱类型

根据知识的主客观性，分为下面两种：

1.事实性（客观性）知识：指那些确定性的、不随状态变化而改变的知识。

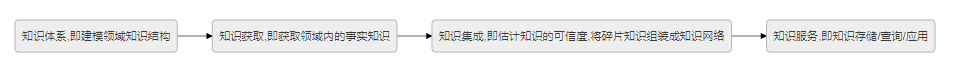
2.主观性知识：指某个人或者群体性感知的知识。

根据知识的变化性质，已有的知识可以分为静态知识和动态知识。另外，还可以分为领域知识、百科知识、场景知识、语言知识、常识知识等等。而目前已经构建的、比较有名的知识图谱有以下几个，其他的知识图谱还有很多，如影视知识图谱IMDB等。

|  |  |  |
| --- | --- | --- |
| **类别** | **名称** | **简介** |
| **人工构建** | **Cyc/OpenCyc** | **一个高质量英文电子词典和语言本体** |
| **WordNet** | **一个人工撰写的常识知识库** |
| **基于维基百科** | **DBPedia** | **一个部分结构化的Wikipedia，社区人工构建Taxonomy** |
| **YAGO** | **一个部分结构化的Wikipedia** |
| **WikiTaxonomy** |  |
| **Freebase** | **一个大规模协同构建知识库，整合Wikipedia，WordNet以及多个百科知识库** |
| **开放知识抽取** | **KnowItAll** |  |
| **NELL** | **一个基于文本信息抽取技术持续不断更新的知识库** |
| **Probase** |  |
| **企业知识图谱** | **MS sotor，Google KG** |  |
| **百度知心，搜狗知立方** |  |

## 1.4 知识图谱生命周期

知识图谱的生命体系是指知识图谱从无到有，从构建到应用的过程。其主要的过程如下所示：



### 1.4.1 知识体系构建

即知识建模，指用什么样的方式表达知识和形式构建一个本体对目标知识进行描述。目前流行的表示方法是使用RDF（资源描述框架）进行描述。其基本数据结构包含了三种类型：资源、谓词以及陈述。

### 1.4.2 知识获取

目前知识图谱中的数据主要来源有各种各样的结构化数据、半结构化数据以及非结构化文本数据。工业界主要是从结构化数据、半结构化数据中抽取知识来构建知识图谱，而学术界主要是通过非结构化文本数据中抽取三元组来构建知识图谱。

### 1.4.3 知识融合

即利用各种各样的知识和对已有的知识图谱进行补充。

### 1.4.4 知识存储

即研究如何对获取的知识进行存储。目前主要有两种形式，一种是RDF格式存储另一种是图格式存储。

### 1.4.5 知识推理

利用推理手段发现已有知识中的隐藏知识。

### 1.4.6 知识应用

目前常用的知识图谱的应用领域主要包括智能搜索、自动问答、智能推荐、决策支持等领域。

## 1.5 知识图谱与深度学习

目前，在知识图谱构建和应用的各个环节中，越来越多的任务采用深度神经网络建模，并取得了一定的效果。

# 知识表示

知识表示是基于知识的人工智能应用中的核心部分。知识表示有五大重要角色，

1. 知识表示是一种代理，基于对事物的表示，我们无须实践，而是通过思考和推理就可以得到有关外部世界的结论。

2. 知识表示是一组本体论约定的集合，说明我们以什么样的方式来思考世界。

3. 知识表示是智能推理的组成部分：推理需要对知识进行表示，但知识表示不是推理的全部。

4. 知识表示是高效计算的媒介，通过对知识进行有效组织，支持高效的推理。

5. 知识表示是人类表达的媒介，基于通用表示框架，方便人们表达和分享对世界的认知。

## 2.1 经典知识表示理论

### 2.1.1 逻辑

逻辑从简单到复杂分为：命题逻辑、一阶谓词逻辑、高阶逻辑。

1.命题逻辑：具有最简单的语法，定义了具有真假值的原子命题，通过与、或、非、蕴含、当且仅当等将多个原子命题组合成复合命题，而推理过程则根据逻辑连接词的真值表进行推导，适合进行简单的推理。这是一种非常简单但是非常实用的逻辑。

2. 一阶谓词逻辑（一阶逻辑）:在命题逻辑基础上引入了全称量词和存在量词，使得一阶逻辑可以量化实体和概念。它具有完备的逻辑推理算法。

3.高阶逻辑：量化谓词或集合、二阶量化集合、三阶量化集合的集合

优点：谓词逻辑是一种强大的知识表示语言，一阶谓词逻辑具有完备的逻辑推理算法，它可以保证知识的一致性，也能够确保推理结果的正确性。它可以表达对象集合的属性，而不用逐一列举所有对象。

缺点：难以表示过程性知识和不确定性知识。特别是在数据量较大时，工作效率非常低。

### 2.1.2 语义网络

语义网络是奎利恩等人提出的一种模型。语义网络是一个通过语义关系连接的概念网络，它将知识表示为相互连接的点和边的模式，其中节点表示实体、事件、值等，边表示对象之间的语义关系。语义单元的三元组形式表示为：<节点1，关系，节点2>。

语义网络的关系类型有：

实例关系（ISA）：

具体与抽象，一个事物是另一个事物的一个实例，如“小王是一个人”。

分类关系（AKO）：

一个事物是另一个事物的一种类型，如“篮球是一种球”

成员关系：

个体与集体，一个事物是另一个事物的成员，如“小于是一个公务员”

属性关系：

事物和其属性之间的关系，如人的身高

聚合关系：

组织或结构特征的部分与整体的关系，如“凳子是桌子的一部分”

时间关系：

不同事件发生时间的先后次序关系，常用在前，在后

位置关系：

不同事物在位置方面的关系，在上，在下，在内

相近关系：

不同事物在形状、内容等方面相似，如狮子老虎食物链位置相近

语义网络中的多元关系会被转化为多个二元关系来表示。

需要注意的是，语义网络和一阶谓词具有相同的表达能力，但他以一种简单的形式统一描述所有的知识，非常有利于计算机的存储和检索。其缺点也非常明显，就是无法进行推理。

### 2.1.3 框架

从认知学的角度，人类都是以一种类似于框架的结构存储在记忆中。在根据实际情况对框架进行填充，填充的部分叫做槽，框架以及槽的粒度则根据人类对事物的认知程度而定。

FrameNet是一个经典的基于框架表示的知识库。它针对词汇级的概念进行框架的建模，认为大部分词汇的语义能够通过语义框架的形式进行表示。FrameNet的数据是以层级结构进行表示和存储的。FrameNet中位于最上层的节点表示框架，框架之间的边表示框架之间的关系。框架之间有两种边，无向边和有向边，他们。也代表了两种关系，即无向关系和有向关系。他已经被证明对一系列的自然语言处理任务具有明显的效果。可以说在某些方面框架是一种非常有效的知识表示方法。

### 2.1.4 脚本

脚本是一种和框架类似的知识表示方法。它通过一系列的原子动作来表示事物的基本行为，按照时间叙述描述事物的发生，类似于电影剧本。脚本表示的知识有确定的时间和因果的顺序，必须是前一个动作完成后会触发下一个动作的开始。可以说脚本是用来描述动态过程而非静态知识的表示方法，它与框架有着不同的目的。

这种表示方法虽然有很强的时序关系表示能力，但在对于基本属性的描述上能力有限。也很难描述多变事件发展的可能方向。

## 2.2 语义网中的知识表示方法

### 2.2.1 语义网表示方法

语义网指的是对万维网的功能进行扩展，来提高其智能程度。语义网革命指的并不是在科学上有革命性的突破，而是工程上的挑战，其中标准化，规模化，系统开发与集成，用户交互等都是语义网技术面临的挑战。

目前的Web2.0使用的是Html文档对网页进行表示，但语义网使用的是Xml格式文档来表示网页。这样的表示方法，可以让计算机更好的理解网页中的信息。网页中的标签也不再仅仅是网页格式的标志，而是含有自身的语义。

### 2.2.2 语义网知识描述体系

目前，语义网。知识表示体系。主要有以下三个层次。一是XML，即全称可扩展标记语言。它是最早的语义网表示语言，是从网页标签是语言向语意表达语言的一次飞跃。二是RDF，即全称资源描述框架。他可以看成XML的扩展或简化。三是OWL，全称网络本体语言。他是本体的语义表示语言，建立在RDF 和RDF Schema的基础上。能够表达本体知识和刻画属性之间的关系。

Xml并不是我转为语义网而设计的。它的最初版本是在上世纪80年代初被提出来用来处理动态信息的显示问题，以及为了解决Html在数据表示和描述方面混乱的问题的技术标准。它的主要优点在于灵活。在一个使用了Xml的系统中。系统设计者完全可以灵活的设计所需要的元素和属性的标签。但是，在某些情况下，一旦没有提供这些标签的说明，那么。一些自定义、个性化的标签的语义便难以知晓。这对系统的使用和更新都会带来一定麻烦。

在链接开放数据项目出现后，开发于不同时期和不同组织的语义知识是需要相互兼容，因此需要统一且无歧义的语义定义方式。W3C又提出了资源描述框架RDF。RDF假定任何复杂的语义都可以通过若干的三元组的组合来表达，并定义这种三元组的形式为“对象-属性-值”和“主语-谓词-宾语”。在使用RDF知识表示的数据库中，为了存取相关知识，研究者们为它开发了一种类似于SQL语句中SELECT-FROM-WHERE的查询方式。标准的RDF和XML一样同样是领域无关的。这既是他的优点，也是他的缺点。这一特性虽然使其具有更大的自由度，但也使得同一领域中的不同知识难以交互和融合。RDF局限于二元谓词，RDFs则限制与子类和属性层次及其属性的定义域，值域。

为了解决RDF局限性。W3C又提出了网络本体语言作为语义网的领域本体表示工具。

## 2.3 知识图谱中的知识表示方法

### 2.3.1 表示框架

通常情况下，一个知识本体主要涵盖以下几个方面的内容，

1、事务：客观世界中的实体或者对象；

2、概念：具有相似本体特征的一类事物，也称类型；

3、属性：事务和概念具有的特征和特性等。；

4、关系：概念与实体之间的关联方式；

5、函数：事物或概念之间进行转化的形式表达。；

6、约束：某项断言成立的限制条件的形式化描述。；

7、规则：根据某项断言得到逻辑推论的因果关系知识的形式化描述；

8、公理：永远为真的断言；

目前大部分知识图谱主要是对前四部分内容进行建模，只有很少的知识图谱建模了简单的规则结构。这同时也反映了不同层次知识代表是上的复杂程度是不同的。

我们所描述的狭义知识图谱可以看成是知识库的图结构表示。除了谷歌知识图谱之外。FreeBase，Yago等具有图结构的三元组知识库，也是一种狭义的知识图谱。用统一的三元组形式来表示知识，不论是对人类操作的便捷性，还是对计算机计算的高效性，都具有非常大的优势。这一章描述的知识图谱局限在现存的具有图结构的三元组知识库上，泛指三元组r(h,t)知识组成的有向图结构。但是我们也应该注意到,知识图谱的知识表示绝不仅仅体现在以RDF为基础框架的三元组之上，还体现在实体、类别、属性、关系的多颗粒度，多层次语义单元的关联之中，他是一个知识系统。以一种统一的公式表示的知识定义和知识实力两个层次的知识。

另一方面。知识图谱也可以看成是语义网的工程实现，不太专注于对知识框架的定义，而专注于如何以工程方式从文本中自动抽取，或者依靠众包的方式获取并组建广泛的具有平铺结构的知识实例，最后要求使用它的方式具有容错、模糊匹配等机制。这种对内容的宽泛定义，可以看成是狭义的知识图谱和语义网的主要区别。

## 2.4 知识图谱中的数值化表示方法

### 2.4.1 符号的数值化表示

知识表示的一大重要目标就是进行语义计算。很多知识表示方法，用符号显示表示概念及其关系，概念的种类和关系的类型都是人们总结的结果，其中难免存在有遗漏的情况。

另一方面。目前的大多数及语义计算任务都采用基于数值计算的统计机器学习方法。而作为知识载体的数据表示是机器学习中的基础工作数据表示的好坏直接影响到整个机器学习系统的性能。因此，人们投入了大量精力去研究如何针对具体任务设计一种合适的数据表示方法，以提升机器学习系统的性能。这一环节也被称作特征工程。特征工程是人类对特定任务，智慧和灵感的体现。特征工程在传统技艺学习算法中有着不可替代的地位，但是由于需要大量的人力和专业知识，这也成为了积极学习系统性能提升的瓶颈。为了让机器学习算法有更好的扩展性，研究人员希望可以减少对特征工程的依赖。这样当靶机学习算法推广到新的领域中，十就可以省去大量增加，在新领域上分析和探索。加快应用的进程，使得系统更为智能。

### 2.4.2 文本的数值化表示

书中关注的知识表示以文本处理和自然语言处理为核心。在自然语言处理领域，深度学习技术并没有产生类似图像和语音领域那样的突破。其中一个主要原因是在图像和语音领域，最基本的数据是信号数据，我们可以通过一些距离度量判断信号是否相似，而文本是符号数据。两个词仅是字面不同，就难以刻画它们之间的联系。因此，在判断两个词是否相似时，还需要更多的背景知识才能做出回答。我们也希望计算机可以从大规模无标注的文本数据中自动学习得到，我们表示，这种表示，需要包含对应语言单元的语义信息，同时可以直接通过这种表示度量语言单元之间的语义相似度。

谁是知识表示的最基本单元？而传统不用符号表示这个词的方式不包含任何语义信息。如何将语义融入到词表示中，Harris在1954年提出的分布假设，为这一设想提供了理论基础：上下文相似的词，其语义也相似。Frith之后对其进行进一步的阐述和明确，词的语义由上下文决定。之后，随着统计方法，在自然语言处理中逐渐成为主流，人们发明了很多基于机器学习和深度学习的构造词表示的方法。

### 2.4.3 知识图谱的数值表示

知识图谱表示学习的方法，主要有张量分解模型和基于能量函数的模型等，他们都有各自的学习方式。这两部分涉及到机器学习和深度学习。

# 知识体系构建和知识融合

事实上，知识图谱不仅包含具体的实例知识数据，还包括了对知识数据的描述和定义，这部分对数据进行描述和定义的“元”数据被称为知识体系（Schema）或者本体（Ontology）。能够以一种统一的形式（三元组形式）表示实例型数据和描述型数据，是知识图谱得以广泛应用的重要特点。

知识融合通过框架匹配和实例对齐，把分散的知识资源联合起来，可以极大的增加知识图谱的覆盖领域和共享程度。

## 3.1 知识体系构建

知识体系个方面核心内容：

1.对概念的分类

2.概念属性的描述

3.概念之间相互关系

知识体系的基本形态的五个层次：

词汇

概念

分类关系

非分类关系

公理

### 3.1.1 人工构建方法

知识体系具有很高的抽象性以及盖过去。目前高质量的知识体系只能通过人工构建。体系的过程可以分为如下六个主要阶段。确定任务及领域；体系复用；罗列要素；确定分类体系；定义属性及关系；定义约束。以上六个阶段，在实践中并非严格的线性关系，有时需要回退到更早的阶段。

### 3.1.1.1 确定任务及领域

知识图谱作为人工智能应用的基础设施，其构建过程不能不了解具体的应用任务，也不能抛开领域建立一个高大全的、无法被广泛使用的产品。

事实上。知识体系与具体的领域密切相关。因此，在创建知识体系之前，首先应该确定知识图谱面向的领域，限定知识体系应该包含的知识范围，在领域内构建出各种各样的知识体系。

如果想要构建更为合适的体系，需要回答以下几个问题：

1、我们为什么要使用这个知识体系

2、这种知识体系呢个构帮助回答哪些类型的问题

3、谁会使用并维护这个知识体系

这些问题应该贯穿于知识体系构建的每一个阶段，并且随着体系构建的推荐，我们可能会对上述问题有更加深入的认知，原来的答案可能会被推翻。

### 3.1.1.2 体系复用

知识体系具有很强的抽象性和概括性，从零开始构建不仅成本高，而且质量难以保证。事实上，在具体实践时很多项目选择，先构建一个轻量级的知识体系，然后尽可能基于他们进行扩展。因此真正进行构建之前，应该广泛调研现有的第三方知识体系和与之相关的资源，尽可能多的参考前人已有的成果。这些资源主要包括以下几种：

1、领域词典：一些领域的专家会编撰领域内的词典。这些词典在构建限定领域的知识体系时，具有重要的参考价值。

2、语言学资源：在自然语言处理领域，有很多语言资源可以用于帮助知识体系的构建。

3、开源知识图谱：现有大规模开源知识图谱已经在第一张进行了介绍。这些知识图谱的知识体系都是由专家人工制定的，具有较高的质量，并且涵盖的领域非常广泛，对于定制新的知识体系具有较高的参考价值。

4、网络百科：网络百科是成千上万用户共同编辑走到的，其中包含的知识非常广泛，它的更新和知识添加都十分及时。但是由于歧视开放编剧的。其可信度可能会有所下降。

### 3.1.1.3 罗列要素

根据我们所确定的领域，罗列期望在知识图谱中出现的要素列表，主要包括概念，属性以及关系。这一步实际上是为后续步骤准备原材料，因此不需要对上述概念进行清晰的分类，只需要尽可能多地罗列出期望的元素即可。

### 3.1.1.4 确定分类体系

确定了相关要素之后，需要将其中表示概念的要素，组织生成层级结构的分类体系。构建过程中主要有两种方式。他们分别是自顶向下方法和自底向上方法。

### 3.1.1.5 定义属性及关系

定义了知识图谱的分类体系后，需要为其中的每一个类别定义属性以及关系。属性用于描述概念的内在特征，关系用于刻画不同概念之间的关系，属性的定义需要受到分类体系的约束，下层类别必须继承上所有上层类别的属性。

### 3.1.1.6 定义约束

不同属性和关系具有不同的值域和定义域。这一些约束主要是用来保证数据的一致性，避免异常值的出现。

### 3.1.2 自动构建方法

人工构建知识体系是一个耗时、昂贵、高度技巧化的任务，并且构建的过程繁琐而枯燥，很容易出错，因此自动地从数据中学习知识体系具有重要的意义。在不同数据源上学习知识体系需要不同的技术，根据数据员结构化程度的不同，知识体系的学习技术可以分为三大类：基于非结构化数据的知识体系学习、基于结构化数据的知识体系学习和基于半结构化数据的知识体系学习。其中，后两类研究工作较少，他们大部分采用与人工构建结合的方式工作。

### 3.1.2.1 基于非结构化数据的知识体系学习

非结构化数据通常指文本数据。例如新闻报道等。基于文本数据构建知识体系，也称为基于文本的本体学习，这类方法的基本思想是：首先利用自然语言处理工具对文本进行分词、句法分析、命名实体识别等预处理操作，然后利用模板匹配，统计学习等手段，从文本中抽取重要信息，主要包括领域概念，实例以及概念之间的关系。概括来说。基于非结构化文本的知识体系，学习方法，主要包括以下的三个主要步骤：

1.领域概念抽取；2.分类体系构建；3.概念属性及关系抽取。

1.领域概念抽取：这一主要步骤的目标是从文本数据中抽取出构建知识体系所需的关键元素，包括实体类型名，属性名，关系名等，这些关键元素称为该领域的术语。术语的抽取主要分为如下三步，第一，抽取候选术语；第二，术语过滤；第三，术语合并。

2.分类体系构建：构建分类体系，实际上是要获取不同概念之间的继承关系，语言学上称之为上下位关系，下位词世上为此概念的具体化。和同义词识别任务类似，基于词典的方法和基于统计的方法同样是解决上下位关系识别的主要方法。

3.概念属性及关系抽取：属性和关系也可以看作是一种概念，因此属性及关系的抽取过程和概念的抽取过程类似。

### 3.1.2.2 基于结构化数据的知识体系学习

结构化数据是指具有严格定义模式的数据，主要指存储于关系数据库中的数据。

关系数据库采用关系模型对现实世界中的信息进行建模，这种模型具有两个明显的优点：1、首先是关系模型结构简单，便于理解，所以我的对象在关系数据库中都通过二维表格进行存储及表示。2、关系模型具有很强的理论基础，关系代数强有力的支持了关系模型，使得关系数据库能够得到广泛的应用。

目前，基于结构化数据知识体系学习的主要任务是分析关系模型中蕴含的语义信息，并将其映射到知识体系的相应部分。

### 3.1.2.3 基于半结构化数据的知识体系学习

半结构化数据和非结构化数据相比具有一定的模式，但这种模式并不严格，典型的半结构化数据有Xml格式的数据，Html格式的网页数据，以及他们遵守的文档类型定义。由于这类数据是介于结构化数据和非结构化数据之间的一类数据，因此上述两类方法也能够应用于该类数据。另外，机器可读的知识词典也是一种特殊的半结构化数据。

### 3.1.3 经典知识体系

SUMO

Schema.org

freebase(已经被关闭)

protege

常被应用于基于框架的知识表示模型，一般是先定义类，在定义类中的属性，最后定义类和属性的约束。

## 3.2 知识融合

随着互联网上只是数量的不断增长，多个垂直领域都形成了专业的领域知识库，例如IMDb就是电影领域的知识库，MusicBrianz是音乐领域的知识库。这些专业领域的知识库中包含很多通用知识库中没有的专业知识。大多数时候只有将这些知识库联合起来应用，才能够满足互联网用户跨领域的信息需求。

另一方面，从融合的知识图谱类型来看，知识融合包括竖直方向的融合和水平方向的融合。竖直方向的融合是指融合高层通用本体与底层领域本体或实例数据。水平方向的融合是指融合相同层次的知识图谱，如BabelNet融合不同语言的已购知识图谱，实现跨语言的知识关联和共享。

不同机构，不同个人都可以自由地构建所需知识图谱，各个知识图谱的数据来源非常广泛，其质量也会层次不齐，关注领域也不尽相同，知识图谱之间存在多样性以及异构性。知识融合通过对多个相关知识图谱的对齐、关联和合并，使其成为一个整体。按融合对象的不同分为框架匹配和实体对齐。

框架匹配是指对概念、属性、关系等知识描述体系进行匹配和融合。实体对齐是指对齐合并相同的实体完成知识融合。通过框架匹配和实力，对其可以把不同的知识图谱关联在一起，但是，多个知识图谱中的实例知识有时有冲突，需要对其检测和消解。

### 3.2.1 框架匹配

这是体系能够在认知和语义层次上对领域知识进行建模和表达，确定领域内共同认可的词汇，通过概念之间的关系来描述概念的语义，提供对领域知识的共同理解。框架匹配主要解决知识体系之间的异构性，是知识融合的重要组成部分。框架匹配也称为本体对齐。

按照使用技术的不同，框架匹配可以分为元素级匹配和结构级匹配。元素级匹配独立判断两个知识图谱中的元素是否应该匹配，不考虑其他元素的匹配情况。结构级匹配不把各个元素作为孤立的资源，而利用知识图谱的结构，在元素匹配过程中考虑其他相关元素匹配情况的影响。

1、元素级匹配、字符串匹配：

前缀距离，后缀距离，编辑距离和n元语法距离，但这种方法忽略了语言符号的多义性，一词多义和一义多词。

词向量

将词表示为低维语义向量空间的一个点，语义相似度用点之间的距离来衡量。与字符串相比，有更强的捕获词背后的真正语义的能力。

可将其与其他编辑距离等结合，混合方法获得了很好的性能。

2、结构级匹配

不同元素的匹配之间也会有影响，例如属性的定义域和值域匹配度高，属性匹配度也高。基本思想是:相似的概念具有相似的概念结构。

基于图的技术

基于分类体系的技术

基于统计分析的技术

### 3.2.2 实体对齐

知识库对齐的目标是能够链接多个异构知识库，并从顶层创建一个大规模的统一知识库，从而帮助机器理解底层数据。

实体对齐可分为成对实体对齐和协同实体对齐两类不同的算法。成对实体对齐表示独立地判断两实体是否对应同一物理对象，通过匹配实体属性等特征判断他们的对齐程度。协同实体对齐，认为不同实体间的对齐是相互影响的，通过协调不同对象间的匹配情况，可以达到一个全局最优的对齐结果。

目前基于表示学习方法被用于支持对其通过知识库联合表示学习，将多个知识库表示在同一语义向量空间中，把知识库实体对齐的过程转化为两个知识库中的实体相似度计算问题。这是故乡量化之后，将两个知识和在同一向量空间中相近的实体视为相同实体成为一个对齐，这种对齐方法被称为基于支持向量联合学习的对齐方法。这种方法的优点是，不需要依赖人工设定的规则和特征，也不需要了解知识库的命名习惯，这种方法适应性强，很容易迁移到不同语言，不同领域的知识库对齐任务中。

### 3.2.3 冲突检测与消解

在框架匹配和实体对齐的基础上，知识融合还需要解决不同实例间的冲突。不同知识库对实体“姚明”的属性“身高”描述不同，对于这种冲突的检测和消解是知识融合的重要步骤，是多个知识图谱形成一个一致结果的最后步骤。如果检测冲突并进行消解是知识融合任务的主要研究问题冲突识别最简单的方法就是发现对于同样的属性和关系有不同的实例真的，但是对于某些属性，这种策略不一定有效。最简单的例子就是不同人在不同年龄段的身高以及体重等信息。

对于冲突的处理，目前常见的三那策略是冲突忽略、冲突避免、冲突消解。冲突忽略顾名思义就是在遇到冲突时，对其进行忽略处理。冲突避免不解决冲突，而是使用规则或者约束，对于数据来源进行过滤。 冲突消解关注于如何利用知识图谱本身的特征来消解冲突，这也是目前的主要研究方向。

冲突消解按照使用技术可以分为如下两类，基于投票的方法和基于质量估计的方法。

# 实体识别和扩展

实体（Entity）作为知识图谱的基本单元，也在文本中承载信息的重要单位。实体识别和分析是支持知识图谱构建和应用的重要技术。根据国际公开评测（Automatic Content Extraction，ACE）给出的定义，在文本中对实体的引用可以有三种形式：命名性指称、名词性指称以及名词性指称。狭义来说，命名实体指现实世界中具体或抽象的实体。命名实体的确切含义，只能根据具体应用来确定。

本章内容大多是任务驱动，因此，首先对需要做的任务进行概述，之后，介绍完成这些任务使用的不同方案。

## 4.1 实体识别和扩展

### 4.1.1 任务概述

    命名实体的任务主要是识别出文本中实体的命名性指称项，并标注其类别。一般来说，命名实体识别的任务就是识别出待处理文本中三大类、七小类命名实体。不同任务对不命名实体类别粒度的需求不同。细粒度实体是被的难点主要是类别多、类别具有层次、标注成本高。

### 4.1.1.1 实体识别的难点

    七类实体中的时间、日期、货币、百分比具有明显规律，较为容易识别，其他三类（人名、地名、机构名）比较灵活，大多需要配合上下文等其他信息配合完成识别。

命名实体的识别过程包括两部份：（1）识别实体边界；（2）确定实体类别（人名、地名、机构名等相关信息）。

    命名实体识别的主要难点在于以下几个方面：

   （1）命名实体形式多变。命名实体内部结构复杂。例如，人名中虽然对姓氏进行了限制，但是在名中用词多变，杜甫、杜少陵、杜工部、杜子美等都可以代指杜甫这个人。地名和机构名中也有类似情况。

   （2）命名实体语言环境复杂。同样的汉字序列在不同的上下文中可以代指不同的意境。例如，人名中名可以为彩霞，这个词在某些情况下也可以代指一种自然现象

### 4.1.1.2 相关评测

    相关评测会议推动了命名实体识别技术的发展。

### 4.1.2 基于规则的实体识别方法

    有关命名实体识别大致有两种方法，其中一种是基于规则的方法另一种是基于机器学习的识别方法。

    两种方法对比来说，对于基于规则的方法，其优势在于准确率高、接近人类的思考方式，表示直观，便于机器进行推理；其劣势在于成本较高，所制定的规则大多只用在本领域内，无法扩展到更大领域。对于基于机器学习的实体识别方法，其优势在于更加灵活、更加健壮，相对于基于规则的识别方式比较客观，不太需要人工干预，不限制可使用的领域；其缺陷是 在需要人工标注数据，数据稀疏问题严重。

    最具有代表性的方法：基于命名实体词典的方法。其他经典方法为有：采用字符串完全匹配或部分匹配的方式；基于正向最大匹配的方法；基于逆向最大匹配的方法；基于最短路径的方法

### 4.1.3 基于机器学习的实体识别方法——基于特征的方法

    基于机器学习的实体识别方法主要是利用已标注好的语料训练模型，通过已知的字段概率来推测候选字段的计算概率值，并由此判断其是否为命名实体。这种方法鲁棒性较好，构建代价小。

    基于特征方法主要是应用传统机器学习方法结合人工设计的大量特征进行实体识别。现在已经有多种机器学习模型被用于命名实体识别。但是，无论何种模型，都在试图充分发现和利用实体所在的上下文特征和实体的内部特征，包括词形，词性和角色级特征

    基于特征方法进行实体识别的基本步骤：

1.特征选取

例如，在英语中选择单词的词性，大小写，词缀信息等作为特征。

2.模型学习

根据具体的任务和需求选择。比如：支持向量机、最大熵模型，隐马尔科夫模型以及条件随机场等，多模型也可以联合使用。

3.样本预测

利用训练好的模型，对输入样本进行标注预测

4.后处理

最广泛的是基于字的模型，代表性的方法有CRF。

### 4.1.4 基于机器学习的实体识别方法——基于神经网络的方法

    基于神经网络的识别方法和基于特征的识别方法的不同点在于其实自动捕获特征进行识别的，不需要过多的人工干预。

    基于神经网络方法进行实体识别的基本步骤：

1.特征表示

主要是设计和搭建神经网络模型并利用其将文字符号特征训练为分布式特征信息；

2.模型训练

利用标注数据，度，优化网络参数，训练网络模型。

3.模型分类

利用训练好的模型对新样本进行分类，完成实体识别。

## 4.2 细粒度实体识别

### 4.2.1 任务概述

    细粒度实体识别大部分是针对特定任务的。这些细粒度实体类别可能会包含更多的知识，有助于相应任务性能的提升。例如,产品名、会议名、疾病名和赛事名。

细粒度实体类别分类:不同研究人员有不同的分类，其中自动内容抽取会议上将实体分为7大类以及45小类。其他人员分类有所不同。

细粒度实体类别特点:

    类型更多、类别具有层次结构

细粒度实体类别难点:

    类别制定、语料的标注、实体识别的方法

细粒度实体类别的制定:

    最直接的办法是人工制定，也可以使用人工构建的词典知识资源作为类别来源。

### 4.2.2 细粒度实体识别方法

    对于已经使用人工标注的语料，可以使用前面介绍过的方法。在没有标注的情况下，则可以使用聚类方法自动获得实体相关信息的集合，但是这样没有办法自动获得实体的类别标签。

## 4.3 实体扩展

### 4.3.1 任务概述

    实体扩展可以理解为给出一些种子实体，之后自动获得同类实体。其主要目标是从海量、冗余、异构、不规范的网络数据中大规模地抽取开放类别的命名实体，进而构建开放实类别命名实体列表。和实体识别相比，其具有目标实体类别开放、目标数据领域开放、以“抽取”代替“识别”等特点。

    实体扩展的一些应用如下所示：

学术界

知识图谱中的实体扩展。

提高问答系统的性能，尤其是List型问题。

提高垂直领域信息抽取的效果。

工业领域

知识图谱中同类实体的检索和推荐

提高查询分析的准确率

辅助文档分类

辅助用户行为分析

广告精准投放

### 4.3.2 实体扩展方法

    目前典型的实体扩展系统主要有以下三个模块组成：

1.种子处理模块

本模块主要负责选择高质量的种子，提高实体抽取的准确度。

2.实体抽取模块

本模块主要负责从语料中抽取属于目标类别的实体。通常来说，本模块会分为“候选抽取”和“打分排序”两个子模块。

3.结果过滤模块

对抽取出的实体集合进行过滤。

    其中，种子处理模块和结果过滤模块可以复用，不同的实体扩展方法的区别主要在于实体抽取模块。目前典型的实体扩展方法主要分为基于模板的实体抽取和基于统计的实体抽取两大类。

1.基于模板的实体抽取

基本思路：如果目标实体与种子同属于某一个语义类，则他们的上下文关系应该符合特定的模板。这里的模板可以是预先定义好的指示上下位关系的语义模板，也可以是通过分析种子实体所在的上下文得到的模板。

2.基于统计的实体抽取

基本思路：使用粗糙的方式获得候选实体，之后通过分析整个语料库的统计信息来得到候选的分布信息；最后计算候选实体和种子实体的分布相似度作为置信度，并对候选实体进行排序。

    同样的，种子处理和结果过滤两个方面也很重要，也有相关方案对其进行解决。

# 实体消歧

## 5.1 任务概述

    在实体识别的结果中，一方面，同一实体在文本中会有不同的指称（指称的多样性）；另一方面，相同的实体指称在不同的上下文中可以指不同的实体（指称的歧义性)。因此必须对实体识别的结果进行消歧才可以得到准确、无歧义的实体信息。实体消歧的目的是解决文本信息中广泛存在的名字歧异问题。

### 5.1.1 任务定义

M=N.E,D,O,K,δ.

### 5.1.2 任务分类

    按照目标实体的列表是否给定，实体消歧系统可以分为基于聚类的实体消歧系统和基于实体链接的实体消歧系统；按照实体消歧任务的领域不同，实体消歧任务可以分为结构化文本实体消歧系统和非结构化文本实体消歧系统。

### 5.1.3 相关评测

    两个主流的评测平台，一个是WePS主要是针对基于聚类命名实体消歧系统进行评测；第二个是TAC KBP的Entity Linking 评测，主要针对基于实体链接的命名实体消歧系统进行评测。

## 5.2 基于聚类的实体消歧方法

### 5.2.1 基于表层特征的实体指称项相似度计算

    传统方法中的大多数实体消歧系统只是利用实体指称项的表层特征值来计算相似度。这些方法主要是词袋子模型的自然延伸，难以取得良好的实体消歧性能。这类方式多是基于指称项上下文中表层特征关联来计算他们之间的相似度，没有考虑到上下文特征的内在关联影响了它们的聚类效果。目前已经有很多的系统使用这种方式进行实体消歧。

### 5.2.2 基于扩展特征的实体指称项相似度计算

    为了克服基于表层特征指称项相似度的缺陷，一些方法开始使用知识资源来提升实体消歧的性能，其中最直接的方式就是使用知识资源来扩展实体指称项的特征表示。这些抽取出来的属性信息一般具有两个作用：①作为实体指称项的扩展特征；②由于这些信息提供了更加准确的实体指称项信息，他们也可以用来重构聚类结果。其中一个重要的知识源就是Wikipedia。

### 5.2.3 基于社会化网络的实体指称项相似度计算

    基于社会化网络的实体指称项相似度通常使用基于图的算法，可以充分利用社会化网络的传递性，从而考虑到一些隐藏的关系，在一些情况下，可以取得更加准确的实体指称项相似度计算结果。但是其缺点就是它只用到了上下文的一些关系实体信息，无法完全利用到实体指称项的其他上下文信息。

## 5.3 基于实体链接的实体消歧方法

  基于实体链接的实体消歧方法，一般是将实体指称项链接到知识库中特定实体，也称为实体链接。其一般指的是将一个命名实体的文本指称项连接到知识库中相应实体的过程。

  一般来说，实体链接的输入有两个部分：①目标实体知识库。这一部分目前常用的知识库是Wikipedia，在其他的一些任务中也可以是一个特定领域的知识库，比如电影领域的IMDB等，这些知识库大部分包含一下的信息：实体表、实体的文本描述、实体的结构化信息、实体的辅助性信息等等。②待消歧实体指称项以及上下文信息。

  实体链接任务的主要目标是将文本中的实体指称项链接到目标实体知识库中的相应实体上。该任务假设目标知识库并不完备，一些情况下，目标知识库中可能没有待消歧实体。

  实体链接任务一般具有两个步骤：

1.链接候选过滤。

由于一个知识库中通常包含上百万的实体，在实体的实体链接任务中不会计算一个指称项和其他所有指称项链接的可能性，需要首先过滤掉大部分不可能指向的实体，仅保留少量链接实体候选。

2.实体链接。

给定一个实体，以及其链接候选，确定该实体指称项最终指向的目标实体。

  目前，大部分的研究重点在第二步。

### 5.3.1 链接候选过滤方法

  目前还缺少对于链接候选过滤方法的系统化研究和量化分析，大部分工作都是基于实体指称项词典：通过在词典中记录一个指称项所有可能之乡的目标实体来进行链接候选过滤。

### 5.3.2 实体链接方法

  实体链接的任务可以描述为，给定一个指称项mmm以及其链接候选E=e1,e2...。实体链接方式悬着和指称项具有最高一致性打分的实体作为其目标实体。

  实体链接的任务的关键就是如何计算实体指称项和目标实体之间的一致性打分 Score(e,m)Score(e,m)Score(e,m)。据此分为四种不同方法：（1）向量空间模型；（2）主题一致性模型；（3）协同实体链接；（4）基于神经网络的实体消歧方法；

### 5.3.2.1 向量空间模型

  在向量空间模型中，实体指称项与目标实体的效性打分主要基于实体指称项上下文与目标实体上下文中特征的共现信息来确定。在该模型中， 实体概念和实体指称项都被表示为上下文中Term组成的向量.基于Term向量表示，向量空间模型通过计算两个向量之间的相似度对实体概念和指你项之间的一致性进行打分。目前， 针对向量空间模型的研究集中在两个方面:（1）如何抽取有效的特征表示，传统的向量空间模型仅仅使用上下文中的词作为实体指称项的特征，通常难于准确表示实体指称项的信息；（2）如何更为有效地计算向量之间的相似度。

### 5.3.2.2 主题一致性模型

  在主题一致性模型中， 决定致性打分的是实体指称项的候选实体概念与指称项上下文中的其他实体概念的一致性程度。主题一致性模型通过利用主体一致性识别出一个实体指称项的目标实体，选择一个和指称项上下文中的实体具有最高一致性打分的实体作为指称项的目标实体。在计算致性打分时，通常需要虑如下两个因素:（1）上下文实体的重要程度；（2） 如何计算一致性。

### 5.3.2.3 协同实体链接

  上述向量空间模型和主题一致性模型都只能处理单个实体指称项的链接问题，而忽略了单篇文档内所有实体指称项的目标实体之间的关系。考虑到文档的主题一致性，一篇文本内的所有实体指称项的目标实体也应该是相互关联的。 因此，对单篇文档内所有实体指称项进行协同链接有助于提升实体链接的性能。

  Kullarni等提出了一种协同实体链接的方法， 它把单篇文档的协同实体链接看成是个优化任务。

### 5.3.2.4 基于神经网络的实体消歧方法

   无论是基于聚类的实体消歧方法还是基于实体链接的实体消歧方法，计算实体与实体、实体与文本，文本与文本之间的相似度都是核心问题。传统的计算方法主要是利用自然语言处理工具来抽取词性，依存句法等特征，尽管这些方法取得了不错的性能，但是特征可扩展性差，表示能力不足，容易造成误差传递。近年来很多工作利用深度学习的方法缓解上述问题，进一步提升实体消歧的性能。2013 年，He2等首先将神经网络应用到实体消歧上.他们将整个文档作为输人，并利用层叠降噪自动编码器通过预训练得到文档和实体的初始语义表示，然后通过有监督的消歧语料进行网络多数和文档与实体表示的微调，进而完成实体消歧。Sun3等在2015年提出利用检积神经网络生成指称项、实体和实体上下文语义表示的方法，在文本表示时间向量拼接了位置特征，井且在请义合成阶段使用了种神经张量网络， 进而完成实体消歧的任务。近些年，基于神经网络的方法不需要人工设计复杂的特征，易于捕获深层语义，取得了比较好的性能并具有良好的扩展性，占据了实体消歧的主导地位。

## 5.4面向结构化文本的实体消歧方法

  之前两大类实体消歧方法可以用在非结构化文本中，在非结构化文本中待消歧实体指称项的上下文可以为我们提供消歧所需要的关键信息。但是我们在构建图谱时大多是对大量结构化文本进行处理产生实体的。这类结构化文本所构建的实体，大多是含有较多信息的，使用上面的方法就可以很好的完成消歧任务，但是也有很多结构化数据只有实体名或者很少的结构化信息。对这类实体进行消歧时，相对比较困难。现有的方式主要是利用实体的类别信息/实体的流行程度和列表中的其他信息进行消歧。

# 关系抽取

前面的章节已经介绍了实体识别和消歧的任务及其方法。而实体之间的关系是知识图谱中不可或缺的部分，不同的关系将独立的实体连接在一起编织成知识图谱。如何从结构化或者非结构化文本中识别出实体之间的关系是知识图谱构建的核心任务之一。同时，关系抽取也是文本内容理解的重要支撑技术之一，能够将文本分析从语言层面提升到内容层面，对于回答系统、智能客服、聊天机器人、语义搜索等应用都十分重要。因此，这一任务得到了学术界和工业界的广泛关注，正成为越来越热门的研究课题。

## 6.1 任务概述

### 6.1.1 任务定义

  关系定义为两个或多个实体之间的某种联系，关系抽取就是自动识别实体之间具有的某种语义关系。根据参与实体的多少可以分为二元关系抽取和多无关系抽取，其中二元关系指两个实体间的关系，多元关系指三个及以上实体间的关系。二元关系抽取是其他关系推取研究的基础，因此本章主要介绍二元关系抽取，关注两个实体间的语义关系，relation表示实体间的语义关系。给定两个实体中国和北京，通过二元关系抽取得到它们之间的语义关系是首都，就能抽取出三元组（中国，首都，北京）。

### 6.1.2 任务分类

  根据处理数据源的不同，关系抽取可以分为以下三种：

1.面向结构化文本的关系抽取。结构化文本包括表格数据、XML文档以及数据库数据等，这类数据通常具有良好的布局结构，因此抽取比较容易，可针对特定网站编写特定模板进行抽取，抽取准确率也比较高。

2.面向非结构化文本的关系抽取。非结构化文本指的就是纯文本，例如，从句子姚明1981年出生于中国上海这句话中，我们希望识别出姚明和上海之间是出生地这一语义关系。但是，由于自然语言表达的多样性、灵活性，实体关系在文本中一般找不到明确的标识。相对来说，从非结构化文本中抽取关系的准确率较低，需要自然语言处理技术的支持。

3.面向半结构化文本的关系抽取。半结构化文本介于结构化和非结构化之间，数据的分布或布局具有一定的规律，但通常这种规律的类型是多样的，也是隐含的、或者说没有显式的标识，难以用人工的方法穷举各种类型的模板，需要对模板进行自动的学习。目前，针对模板相对连续的半结构文本，现有的技术也能达到较高的抽取准确率。

  同时，根据抽取文本的范围不园，关系抽取可以分为句子级关系抽取和语料（篇章）级关系抽取。

句子级关系抽取，也称为句子级关系分类，即从一个句子中判别两个实体间是何种语义关系。如上面的例子所示，我们需要从当前的这句话中识别出姚明和上海之间是出生地的关系。

语料（篇章）级关系抽取，该任务旨在判别两个实体之间是否具有其语义关系，而不必限定两个目标实体所出现的上下文。也就是只需要判别姚明和上海之间有出生地的关系，而并不必关注具体是哪一个句子表达了这两个实体有这样的关系。这一任务对于知识图谱构建十分重要，因为在知识图谱构建过程中，我们需要分析图谱中两个节点（实体）之间边上的语义标签（关系）而并不用特别关注它们所出现的具体文本。

  另外，根据所抽取领域的划分，关系抽取又可以分为限定域关系抽取和开放域关系抽取：

限定域关系抽抽取。是指在一个或者多个限定的领域内对实体间的语义关系进行抽取，通常情况下，由于是限定域，语义关系也是预设好的有限个类别。对于这一任务，可以采用基于监督学习的方法来处理，即针对每个关系类别标注充足的训练数据，然后设计关系抽取的模型，进行模型训练，最后利用训练好的模型抽取关系。但是在面对大规模的知识图谱构建时，人工标注的训练语料远远不够，所以很多工作利用弱监督学习解决训练语料的标注问题。

开放域关系抽取。和限定域关系抽取，开放域抽取并不限定关系的类别，依据模型对于自然语言句子理解的结果从开放式抽取关系三元组。

### 6.1.3 任务难点

  关系抽取是信息抽取中的一个关键环节和难点问题，相比较于实体识别和实体消歧任务而言,关系抽取任务更加复杂，其难点主要表现在下面几个方面：

1.一个关系可以具有多种不同的词汇表达表达方式。例如句子特朗普是美国总统和句子特朗普赢得了美国大选都可以说明特朗普和美国总统具有职位的关系。

2.同一个短话或者词可能表达不同的关系。特别是在不同的上下文中表示不同的关系。

3.同一对实体之间可能有不止一种关系。例如一个人的出生地和居住地可能是同一个城市。

关系抽取不仅涉及到两个或者两个以上的实体单元，还涉及实体周围的上下文，需要利用上下文中的一些结构化信息，这也使得问题复杂度呈指数级增长。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。关系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，但是实际情况下，许多针对这些工作的自然语言处理工具性能并不高，低性能工具引入的错误反而会降低关系抽取系统的性能。

### 6.1.4 相关评测

  针对这一任务，学术界组织了许多公开的技术评测，极大地推动了该领域的相关研究。

  除了消息理解会议(Message Under-standing Conference,MUC)、自动内容抽取(Automatic Content Extraction Evaluation,ACE)和文本分析会议(Text Analysis Conference,TAC)三大国际评测会议外，另一个比较有影响力的国际评测会议是语义评测会议（SemEval)。SemEval会议从1998举办，截止到2018年已经成功举办11届，影响力非常广泛，设有词语语义消歧、时间识别和关系抽取等多项任务。在使用最为广泛的SemEval 2010评测集中共定义了9类关系（工具-代理、原因-影响、内容-容器、产品-生产商、实体-出生、成分-整体、实体-目的地、成员-集体、消息-主题）和一个其他类。

## 6.2 限定域关系抽取

  限定域关系抽取指在一个或者多个限定的领域之内判定文本中所出现的实体指称之间是何种语义关系，且待判定的语义关系是预定义的。因此。已有研究常把这一任务看作是一个文本分类任务，即在输入一个句子以及标识句子中所出现的实体指称的条件下，系统将其分类到所属的语义类别上。早期针对这任务的研究多是采用模板的方式对文本中实体间的语义关系进行判别,随着统计机器学习的发展,越来越多的研究者采用有监督学习的方法,即针对每个关系类别标注充足的训练数据,然后设计关系抽取模型,其研究多关注于如何抽取有效的特征。根据从句子中所提取表征语义关系特征的方式的不同,我们可以把已有方法分为传统基于特征工程的关系分类方法和基于深度学习的关系分类方法。但是,人工标注语料耗时费力,成本高,因此很多情况下很难获得足够的训练数据,因此有很多研究利用弱监督学习的方法抽取关系。

### 6.2.1 基于模板的关系抽取方法

  基于模板的关系抽取方法通过人工编辑或者学习得到的模板对文本中的实体关系进行抽取和判别。

  当一个句子中所出现的实体指称的上下文文本满足上述模板时，就可以认为 两个实体指称在这个句子中具有收购(ACQUISITION)的关系。然而，人工方法不可能针对多类关系穷举所有的模板，那么就需要采用自动的方法学习抽取模板。问题是：①如何学习用于抽取关系的模板?②如何将学习到的模板进行聚类?

  其中关键步骤是抽取句子中的实体对之间表达关系的模板。模板可以是基于词汇的，也可以是基于句法或语义的。这一过程需要自然语言处理技术，包括句子边界探测、词性标注、名词词组块识别等。例如，我们可以首先使用句子边界探测工具 将给定的文本语料分割为句子，然后运行词性标注工具获得单词的词性。为了探测句子中的实体，可以使用名词短语块识别工具或命名实体识别工具提取合适的名词词组块和实体。然后在此基础上，分别抽取词汇级关系模板和句法级关系模板。下面是对一个句子进行处理和抽取的例子。

句子：The crime took place in View Royal on Vancouver Island.

词性标注：DT NN VBD NN IN NNP NNP IN NNP NNP

实体或词组块：The crime took place in [View Royal] on [Vancouver Island].

变量替换：View Royal=X, Vancouver Island=Y

字面形式：The crime took place in X on Y.

词性序列：DT NN VBD NN IN X IN Y.

词汇模式：Xon Y，X，Y

语法模式：XINY，X，Y

当模板板学习的另一个关键问题是不同的做板时能表示同一语义关系。因此，在抽取模板之后，需要对习得的模板进行聚类，将表示同一语义关系的模板聚在一起。矩阵的每一行表示实体对在模板空间上的分布，矩阵的每一列表示模板在实体对空间上的分布：分布的相似性用于识别表达同种语义关系的不同模式以及不同实体对之间存在的同种语义关系。Bolegala等人四在模板矩阵的基础上，提出了序列联合聚类算法（Sequential Co-clustering Algorithm),对所抽取的模板和实体对进行联合聚类。在模板矩阵中，每一行都是实体对在不同词汇-语法模板中的分布因此其行向量就是聚类的特征向量。每个聚类的类别看作一个类别的标签。另外，该算法能够识别特殊实体的不同字符串形式。例如Redmond software giant和Microsoft均指向同一个实体，这样会使得实体对(Microsoft,Powerset)和(Redmond Software Giant,Powerset)聚类在同一个类别中。

### 6.2.2 基于机器学习的关系抽取方法

  基于模板的关系抽取方法的最大问题在于其受限于模板的质量和覆盖度可扩展性不强。因此，随着机器学习方法的发展，很多工作将关系抽取看成是一个分类问题（通常看成是一个有监督学习或者弱监督学习的问题），开始尝试利用较为成熟的机器学习算法解决这一问题。主要可以分为有监督的关系抽取方法和弱监督的关系抽取方法，下面依次介绍。

### 6.2.2.1 基于机器学习的关系抽取方法

  有监督关系抽取的主要工作在于如何抽取出表征实体指称间语义关系的有效特征。通常情况下，特征抽取主要是使用自然语言处理工具包，从句子中抽取出如词汇、句法和请义等特征，作为关系分类的证据。为了缓解句法特征的稀疏性，有很多研究集中于利用核函数的方法进行关系抽取。近些年来随着神经网络的发展，有很多研究开始利用神经网络自动从文本中提取表征关系的特征，进而完成关系的抽取。下面对基于特征工程的方法、基于核函数的方法和基于神经网络的方法进行分别介绍。

1 基于机器学习的关系抽取方法基于特征工程的方法

  基于特征的方法的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。总结而言，该方法共三个步骤：①特征提取：提取词汇、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。②模型训练：利用提取的特征训练分类模型。③关系抽取：主要是利用训练好的模型对非结构化文本进行分类，进而完成关系抽取。常见的关系抽取特征举例如下：

  （1）词汇特征。这些特征包含实体本身的词语或者名词性词组块、两个实体（词组块）之间的词语和两个实体（词组块）两端的词语。

  （2）实体属性特征。这些特征包含实体或者名词性词组块的类型特征，包括人物（Person）、组织（Organization）、位置（Location）、设施（Facility）和地缘政治实体（Geo-Political Entity）等。

  （3）重叠特征。这些特征包括两个实体或者词组块之间词语的个数、它们之间部分包含其他实体或者词组块的个数、两个实体或词组块是否在同一个名词短语、动词短语或者介词短语之中。

  （4）依存句法特征。这些特征包含两个实体（名词性词组块）的依存句法分析树中的依存标签和依存路径。

  （5）句法树特征。这些特征包含连接两个实体（名词性词组块）的句法路径，不包含重复的节点，并且将路径使用头词（head words）标注。

2 基于核函数的方法

  不同于基于特征的方法，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提出，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)等分类算法来抽取实体间语义关系,在200篇来自新闻机构(如美联社、华尔街日报 等)的文本语料库中进行了两个简单任务的测试,取得了不错的结果。 Culotta等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（如主语依存于谓语、形容词依存于它所修饰的名词等）将包含实体对的成分句法树转换成依存句法树，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。Bunescu和Mooney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到达根节点的路径相同。最短依存树核函数主要缺点是虽然系统的准确率有一定的提高，但是召回率却较低。为解决最短依存树核函数召回率较低的问题，Buescu和 Mooney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将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。 Zhang等融合卷积树核函数(Convolution Tree Kernels,CTk)和线性核函数综合考虑了影响实体间语义关系的平面特征和结构化特征,利用卷积树核函数来计算包含实体对的句法树之间的相似度使用线性核函数计算实体属性（如实体类型等）间的相似度。Zhou等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。

3 基于神经网路的方法

  尽管上述基于特征工程和核函数的统计学习方法面对关系抽取任务能够达到一定的效果,但是在模型可扩展性上仍然存在很大的问题,限制了这些方法的应用和推广：①上述人工设计的特征（词汇特征、实体特征、依存句法树特征等）的提取均依赖于自然语言处理工具，同时特征抽取的过程也是一个串联 ( Pipeline）的过程，前一步自然语言处理的结果作为后一步的输入，例如：在依存句法分析前需要分词、词性识别甚至是实体识别。然而，目前已有的自然语言抽取工具的性能并不是百分之百准确。因此，这些自然语言处理工具容易造成错误累积和传递，使得抽取到的特征不精准。②面对中文、英语等富资源的语种来说,我们可以有丰富的自然语言处理工具用以抽取特征。然而，面对一些小语种，特别是那些资源贫瘠的语种，当没有可用的自然语言处理工具时, 就不能运用上述基于特征工程的关系抽取方法。因此：我们是否可以不依赖于自然语言描取工具，而是宜接从输入的，中自动学习有效的特征表示？随着深度学习方法的兴起，越来越多的研究者开始利用神经网络方法（例如：卷积神经网络、循环神经网络等）进行文本的学习、并逐步将其应用到关系抽取任务中来。该类方法主要包括如下几个步骤：①特征表示：主要是将纯文本的特征表示为分布式特征信息，例如：将单词表示为词向量。②神经网络的构建与高层特征学习：主要是设计搭建神经网络模型并利用其将上一步得到的基本特征自动表示为高层特征。③模型训练：利用标注数据，优化网络参数，训练网络模型。④模型分类：利用训练的模型对新样本进行分类，进而完成关系抽取。下面以Zeng等在2014年提出的县于娄积神经网络的关系抽取为例进行介绍。该模型主要包括3个部分：词表示(Word Representation)、特征抽取(Feature Extraction)和输出(Output)。整个模型的输入是一个句子以及给定的两个词（通常为名词、名词短语或实体）。输出是这两个词在该句子中所属的预定义的语义关系类别。

首先.输人的句子通过词向量表示，转化为向量的形式输人网络然后，特征抽取部分进一步提取词汇级别特征和句子级别特征，接下来将这两种特征拼接起来作为最终的特征进行关系分类。词汇级别特征的抽取是将句子中某些词向量挑选出来，并将挑选出的向量拼接起来作为这部分特征抽取的结果。句子级别特征是对整个句子学习出一个向量表示，由于语言表达多样，不同的句子个数理论上是无限的，这样就不能像学习词向量一样，以句子为处理对象，通过统计句子的须率、共现等信息学习出句子的向量表示。然而，一句话的语义通常是由组成这句话的词的含义和这些词的组合方式决定。因此，Zeng等使用卷积神经网络学习句子语义的组合规律和方式，在以词向量作为输入的基础上，将句子中包含的词向量组合起来，进而得到句子级别特征表示。相对于传统特征表示方法，其优点是不再依赖于传统的自然语言处理工具(POS，NER等），完全通过卷积网络直接从具有冗余信息的词向量中自动学习、挑选出有用特征信息，进面学习得到高质量的特征。

### 6.2.2.2 基于机器学习的关系抽取方法

  传统的有监督关系抽取方法需要依赖人工标注的数据,这限制了该方法的适用领域，使得它难以成为关系抽取的核心方法。实际上，带有标注的文本通常是稀缺的资源，距离监督( Distant Supervision)正是在这种背景下提出的方法。当前，在学术界和企业界的共同努力下，已经构建了许多开放可用的知识图谱，这些知识图谱以结构化三元组的形式存储实体和实体之间的关系，距离监督正是利用了这种结构化的数据，让知识图谱自动标注训练样本：由于标注过程不需要人工逐一标注，因此距离监督关系抽取也是弱监督关系抽取的一种。该方法启发式地对齐知识图谱和文本中的实体，然后根据这个对齐学习关系抽取器。该类方法主要基于如下的距离监督假设。

如果两个实体之间存在某种关系，则所有包含这两个实体的句子都表达了这种关系，这些句子的集合被称为一个“包”。这个假设允许研究者们从所有包含指定实体对的句子中抽取特征进行分类。 使用这种方法的共同点是在抽取文本特征时,都依赖于传统的自然语言处理工具,这些自然语言处理工具本身可能存在错误积累和传递的问题,因此会对最终的结果产生不利影响。由于基于弱监督学习的关系抽取 系统不需要人工标注数据，因此目前基于弱监督学习的关系抽取系统抽取的关系实例规模比较大，代表性的系统有:NELL和 Probase。

## 6.3 开放域关系抽取

  限定域关系抽取需要预先定义关系的类别，然而很多情况下预先定义一个全面的关系类型体系是很困难的。开放域关系抽取不需要预先定义关系,而是使用实体对上下文中的一些词语来描述实体 之间的关系。Relation-Words代表关系,就是上下文中描述关系的词或词序列。例如：“姚明出生于上海。”中,开放域关系抽取系统抽取的结果为（姚明，出生于，上海）。但是如果预定义关系抽取中定义了出生地的关系没有定义出生于的关系,则预定义关系抽取的结果应该是（姚明，出生地，上海）。

  华盛顿大学的人工智能研究组最早提出开放域信息抽取(Open Informaiton Extraction, Open IE)的概念，在这方面做了大量代表性的工作，并且开发了一系列原型系统： Text Runner、 Kylin、WOE、 Reverb等。 Banko等在2007年首先提出了开放域实体关系抽取并开发出一个完整的系统Text Runner，它能够直接从网页纯文本中抽取实体关系。Text Runner通过一些简单的启发式规则自动从宾州树库里面获取实体关系三元组的正负样本，根据它们的一些浅层句法特征训练一个分类器来判断两个实体间是否存在语义关系；然后将网络文本进行一定的处理后作为候选句子，提取其浅层句法特征，利用分类器判断所抽取的关系三元组是否可信，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。对于关系名称的抽取，Text Runner把动词作为关系名称，通过动词链接两个论元,从而挖掘论元之间的关系，其抽取过程类似于语义角色标注。

  下面以Text runner为例进行详细介绍:

  TextRunner 是第一个开放域实体关系抽取系统，它使用启发式规则在宾州树库中自动标注语料，不需要人工预先定义关系类别体系，主要分为三个模块：语料的自动生成和分类器训练、大规模关系三元组的抽取、关系三元组可信度计算。

语料的自动生成和分类器训练

语科的自动生成：主要是通过依存句法分析结合启发式规则自动生成语料，利用的启发式规则举例如下：

两个实体的依存路径长度不能大于指定值。

实体不能是代词。

关系指示词是两个实体之间依存路径上的动词或动词短语。

两个实体必须在同一个句子中。

分类器的训练：TextRunner利用朴素贝叶斯分类器进行训练，其使用的特征举例如下：

关系指示词的词性

关系指示词的长度

实体的类型

实体是否是专有名词

左实体左边词语的词性

右实体右边词语的词性

大规模关系三元组的抽取

利用上一步训练好的关系抽取器，在大规模的web文本上进行关系三元组的抽取，并将抽取的大量三元组存储起来。

关系三元组可信度计算

首先将存储起来的相似的三元组进行合并。然后根据网络数据的冗余性，计算合并后关系三元组在网络文本中出现的次数，进而计算相应关系三元组的可信度。

  在TextRunner之后，Wu等1在2007年提出了Kylin系统，该系统选取包含信息框(lnfobox)的Wikipedia页面，根据信息框中包含的条目属性及属性值回标产生训练数据，同时根据信息框中的属性名自动确定需要抽取的属性，不同的属性训练不同的CRF模型抽取属性值。Wu等在2010年提出了开放实体关系抽取系统WOE，该系统也是利用Wikipedia 页面信息框回标，通过一些规则挑选含有实体关系的高质量句子，然后使用依存句法分析树以及词性标注这样的浅层特征训练两个分类器，作为两个实体关系抽取器，以此来获得大量的实体关系三元组模板。最后对网络文本的句子做浅层句法分析处理后，同抽取器获得的模板进行比对，来判断实体关系三元组的可靠性。

   由于传统的信息抽取方法的局限性，开放域信息抽取得到许多学者的关注.

# 事件抽取

事件是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状态的改变。

几个要素是事件发生的时间，地点，参与事件的角色以及与之相关的动作或者状态的改变。

同一个类型的事件中不同粒度的时间、地点、角色代表了不同的事件实例。

事件抽取的定义

主要研究如何从描述事件信息的文本中抽取出的用户感兴趣的事件信息并以结构化的形式呈现出来。

相关概念：

事件指称

是指对一个客观发生的具体事件进行的自然语言形式的描述，通常是一个句子或句群。

事件触发词

指一个事件指称中最能代表事件发生的词，是决定事件类别的重要特征。如结婚，就职

事件元素

事件的参与者，组成事件的核心部分，它与事件触发词构成了事件的整个框架。

如成龙与林凤娇于1982年12月1日在洛杉矶举行婚礼。中的成龙，林凤娇，洛杉矶，1982年12月1日。

由实体，时间，属性值组成。

元素角色

事件元素与事件之间的语义关系。如成龙与林凤娇是夫妻角色。

事件类别

事件元素和触发词决定了事件的类别。

## 7.1限定域事件抽取

预先定义好目标事件的类型及每种类型的具体结构。 ACE定义了8大类事件，商务事件，冲突事件，交易事件等，事件又被细分为33类子事件。

### 7.1.1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与抽取的过程。

模式获取

模式匹配

有监督的事件模式匹配

语料的人工标注

模式的学习

模式的匹配

PALKA模式抽取系统

弱监督的事件模式匹配

只需要人工对语料进行一定的预分类或者制定少量种子模式。由机器根据预分类语料或者种子模式自动学习事件模式。

优缺点：

在特定领域中性能较好。

然而，依赖于文本的具体形式

获取模板的过程费时费力，具有很强的专业性

而且制定的模式很难覆盖所有的事件类型，当语料发生变化时，需要重新获取模式。可移植性不强，召回率低。

基于机器学习的事件抽取方法

建模成一个多分类问题，提取特征向量后再使用有监督的分类器进行事件抽取。

### 7.1.2有监督事件抽取方法

训练样本的表示

选择分类器并训练优化模型

利用训练好的模型从未标注数据中抽取事件实例。

基于特征工程的方法

两阶段分类问题，首先对句子中的每个词汇进行判断，判断其是否为事件触发词，如果是事件触发词，则进入第二个阶段事件元素分类，对句子中每个候选事件元素判断，判断是否是当前触发的事件中包含的事件元素并判断其在当前事件中扮演的角色。

事件触发词的特征

词汇特征

词汇，词汇小写形式，词干，词性标签，相邻词特征

句子级特征

依存路径，依存词汇，候选词在依存树的深度，依存词汇的词性标签，句子中的实体类型，最近距离范围内的实体类型等

外部知识

在wordnet中的同义词id

事件元素分类的特征：

触发词特征

词汇特征

句子级特征

### 7.1.3基于神经网络的方法

词向量学习

非监督学习的方式得到每个词的向量化表示

词汇级特征表示

利用词向量来捕获词汇级语义

句子级特征表示

利用动态多池化卷积网络来学习句子内部的组合语义特征

事件元素分类

利用softmax分类器为每个候选事件元素计算扮演不同角色的概率。

弱监督事件抽取方法

bootstrapping的事件抽取

基于distant supervision的事件抽取 自动生成事件标注样本

首先提出回标的假设规则，利用结构化事件知识去非结构化文本中进行回标，将回标文本当做标注样本，然后利用标注的样本训练模型，完成事件的抽取。

核心原色检测

事件触发词检测

事件触发词过滤核扩展

标注数据的自动生成

## 7.2开放域事件抽取

### 7.2.1基于内容特征的事件抽取方法

文本表示

事件聚类与新事件发现

关键在于如何寻找更好的文本表示方法，文本相似度衡量指标以及事件聚类模型。

### 7.2.2基于异常检测的事件抽取方法

通过检测文本的发布异常进行事件识别

开发域事件识别虽然可以自动发现新的事件，但其发现的事件往往缺乏语义信息，难以进行结构化。如果想获得准确的语义信息，则需要通过人工标注的方式为每个类别簇赋予特定的语义标签。

## 7.3事件关系抽取

事件并非单独存在，它们之间存在千丝万缕的联系。核心任务是以事件为基本语义单元，实现事件逻辑关系的深层检测和抽取。事件关系有：

### 7.3.1事件共指关系抽取

当两个事件指称指向同一个目标事件，具有共指关系。如联想正式完成摩托莫拉的收购。联想以29.1亿收购了摩托摩拉。

核心问题是计算事件指称之间的相似度。

### 7.3.2事件因果关系抽取

反映了事件间先后相继、由因到果的一种关系。有助于掌握事件演变的过程，从而为决策者提供重要的决策信息。

### 7.3.3子事件关系抽取

反映了事件之间的粒度和包含关系。比如地震事件包含伤亡、救援等子事件。

### 7.3.4事件时序关系抽取

事件在时间上的先后顺序。之前，之后，包含等。

# 知识存储和检索

知识图谱是一种有向图结构，描述现实世界中存在的实体、事件以及概念及其之间的关系。一般在图中使用节点表示实体、事件以及概念，图中的边表示相邻节点之间的关系。

在知识图谱是一种有向图结构，描述现实世界中存在的实体、事件以及概念及其之间的关系。一般在图中使用节点表示实体、事件以及概念，图中的边表示相邻节点之间的关系。

## 8.1 知识图谱的存储

  知识图谱中的结构是使用的RDF结构进行表示的，其基本构成单元是一个事实。每一个事实中是一个三元组（S，P，O）,其中S（Subject）代表主语，其取值可以是实体、事件、或者概念中的任意一个。P（Predicate）代表谓语，取值可以是关系或者属性; O（Object）代表宾语，取值可以是实体、事件、概念或者普通的值(例如数字、字符串等)。

  在实际系统中，按照存储方式的不同，知识图谱的存储可以分为基于表结构的存储和基于图结构的存储。

### 8.1.1 基于表结构的存储

  这种存储方式利用二维数据对知识图谱中的数据进行存储。

### 8.1.1.1 三元组表

   知识图谱中的事实是一个个的三元组，一种简单直接的存储方式是设计一张三元组表用于存储知识图谱中所有的事实。实际上目前已经有不少比较成熟的产品利用该形式存储知识图谐包括Jena、Oracle、Sosame、3store、SOR、 Rstar等。这种在储方式的优点是简单直接，易于理解: 然而缺点也非常明显，主要有以下两点:

整个知识图谱都存储在一张表中，导致单表的规模太大。对大表进行查询、插人、删除、修改等操作的开销很大，这将导致知识图谱的实用性大打折扣。

复杂查询在这种存储结构上的开销巨大由于数据表只包括三个字段，因此复杂的查询只能拆分成若干简单查询的复合操作，大大降低了查询的效率。

### 8.1.1.2 类型表

  为每种类型构建一张表， 同类型的实例存放在相同的表中。表的每一列表示该类实体的一个属性，每一行存储该类实体的一个实例。这种存储方式虽克服了三元组表的不足，但是带来了新的问题：①大量数据字段的冗余存储。假设知识图谱中既有“数学家”也有“物理学家”，那么同属于这两个类别的实例将会同时被存储在这两个表中，其中它们共有的属性会被重复存储。②大量的数据列为空值。通常知识图谱中并非每个实体在所有属性或关系上都有值，这种存储方式会导致表中存在大量的空值。一种有效的解决方法是，在构建数据表时，将知识图谱的类别体系考虑进来。具体来说，每个类型的数据表只记录属于该类型的特有属性，不同类别的公共属性保存在上一级类型对应的数据表中，下级表继承上级表的所有属性。

  基于类型表的存储方式根据类型对数据进行组织和管理，克服了三元组表面临的单表过大和结构简单的问题，但是也有明显的不足之处：

由于类型表的不同字段表示了不同的属性或者关系，因此在查询时必须指明属性或关系，无法做不确定属性或关系的查询。

由于数据表是和具体类型对应的，不同类型的数据表具有不同的结构，因此在查询之前必项知道日标对象的类型能确定查找的数据表。为了达到这一目的，系统需要维护一个“属性-类型”映射表，以便在进行属性查询时可以根据目标属性确定类型。

当查询涉及不同类型的实体时，需要进行多表的链接，这一操作开销巨大，限制了知识图谱对复杂香询的处理能力。

知识图谱通常包含丰富的实体类型，因此需要创建大量的数据表，并且这些数据表之间又具有复杂的关系，这为数据的管理增加了很大的难度。

  基于表结构的存储还可以有其他的设计方式，例如可以按照“属性”划分，为每个属性创建张数据表。不同的设计方式各有其优缺点，在实际应用中需要结合具体场景进行选择。

### 8.1.1.3 关系数据库

  目前常用的关系数据库都可以用来存储知识图谱，如MySQL，MS SQL Server等。

### 8.1.2 基于图结构的存储

  将实体看作节点，关系看作带有标签的边，那么知识图谱的数据很自然地能够满足图模型结构。因此，基于图结构的存储方式能够直接准确地反映知识图谱的内部结构，有利于对知识的查询。另外，以图的方式对知识进行存储，还可以借鉴图论的相关算法，有利于对知识的深度挖掘及推理。

### 8.1.2.1 基于图结构的存储模型

  基于图结构的存储模型用节点表示实体，用边表示实体之间的关系。值得注意的是，和基于表结构的存储不同，基于图结构的存储不按照类型来组织实体，而是从实体出发，不同实体对应的节点可以定义不同的属性。

  基于图结构的存储方法利用有向图对知识图谱的数据进行建模，因此无向关系需要转换为两条对称的有向关系。这种存储方式可以细致地刻画实体之间的关系。

### 8.1.2.2 常用图数据库介绍

  图数据库的理论基础是图论。通过节点、边和属性对数据进行表示和存储。具体来说，图数据库基于有向图，其中，节点、边、属性是图数据库的核心概念。

  节点：节点用于表示实体、事件等对象，可以类比于关系数据库中的记录或数据表中的行数据。例如人物地点、电影等都可以作为图中的节点。

  边:为是指图中连接节点的有向线条，用于表示不同节点之间的关系例如人物节点之间的夫妻关系、同事关系等都可以作为图中的边。

  属性:属性用于描述节点或者边的特性。例如人物(节点)的姓名，夫妻关系(边)的起止时间等都是属性。

  常见的图数据库存储系统有Neo4j、OrientDB、InfoGrid、Hyper-GraphDB、InfiniteGraph 等。

## 8.2 知识图谱的检索

  知识图谱上有各种各样的维护需求。查询，插入，删除，修改等都是常见操作。

### 8.2.1 常见形式化查询语言

  知识图谱的知识实际上是通过数据库系统进行存储的，大部分数据库系统通过形式化的查询语言为用户提供访问数据的接口。关系数据库和图数据库分别支持不同的查询语言，前者的标准查询语言是SQL,后者则是SPARQL。

### 8.2.1.1 SQL语言

  SQL是structured query language的缩写，中文翻译为结构化查询语言。它是一种介于关系代数与关系演算之间的语言，用于管理关系数据库。主要功能包括对数据的插人、修改、删除，查询四种操作。SQL是一个通用的、表达能力很强的数据库语言，目前已经成为关系数据库的标准语言。使用这种方式进行知识图谱的维护，可以直接使用SELECT，INSERT，UPDATE等语句进行维护。

### 8.2.1.2 SPARQL语言

  SPARQL是由W3C为RDF数据开发的一种查询语言和数据获取协议，是被图数据库广泛支持的查询语言。和SQL语言类似，SPARQL也是一种结构化的查询语言，用于对数据的获取与管理，主要包括数据的插人、删除和查询操作。由于RDF数据是以三元组的形式进行表示的，因此三元组在SPARQL语言中是一个非常重要的概念，几乎所有的SPARQL语句中都会包含三元组。

### 8.2.2 图检索技术

  由于知识图谱中的数据本身就是一种图结构，因此基于图结构的存储方式可以直观灵活地对知识进行表示和存储。然而利用图模型来建模知识图谱也有不可忽视的缺陷，标准的图查询算法复杂度较高，如何提高图责询效率成为知识图谱研究究的重要问题。

  一个图G可以用二元组(V,E)表示，记G=(V,E)。其中，V是顶点的集合。E是顶点之间边的集合。图查询的任务是在给定的图数据集中查找给定的查询图，其核心问题是判断查询图是否是图数据集的子图 ，因此也称为子得匹配问题。以下给出该向题的形式化定义:

子图区配问题是指在给定查询图Q和目标图集D= {G\_1,G\_2,G\_3…G\_n}的条件下，在D中找出所有与Q同构的子图。

子图同构的问题已经被证明是一个NP完全问题，目前尚不存在多项式时间复杂度内可解决的算法。知识图谱中的图结构具有半富的标签信息：图中的节点以及图中的边包含相关的标签信息。

  虽然子图同构判定问题的算法复杂度很高，但是在实际应用中匹配算法的运行时间通常都在可承受范围之内，主要有两方面的原因：一方面知识图谱中的图结构通常不会特别复杂，只有少数节点之间有边相连，因此并不会触发子图匹配算法的最坏情况；另一方面，利用知识图谱中丰富的标签信息可以有效降低算法的搜索空间。

  知识图谱是张大图，其中包含了大量的连通分支与子图。在进行图匹配时，需要将查询图与每个子图逐进行同构测试。 为了减少匹配的次数， 图数据库在进行子图匹配时会先按照定条件对数据进行筛选， 减小候选图的个数。

### 8.2.2.1 子图筛选

   图索引技术是实现子图筛选的有效方法。与信息检索中的索引机制类似，图索引是在数据预处理阶段进行的。其基本原理是首先根据图上的特征信息建立索引，根据查询图的特征能够快速地从图数据库中检索得到满足条件的候选子图，避免在全部图上进行匹配操作。基于路径的索引和基于子图的索引是两种常见的图索引方法。

### 8.2.2.2 子图同构判定

  子图同构是一个经典的NP难题，暂时无法确定是否存在多项式复杂度的解法。Ullmann算法是实践中常用的一种判定子图同构的算法，他可以枚举出所有同构子图，因此也叫枚举算法。

# 知识推理

## 9.1知识图谱中的推理任务

知识推理是人工智能应用迈向更高级认知智能的重要技术。包括知识补全和知识问答。

### 9.1.1知识补全

面向知识库或者知识图谱的事实补全

如图谱中给出了出生地但没有国籍，即可以通过推理的方法把实体或关系预测出来。称为链接预测。

它是利用已知知识预测未知的隐含知识，利于完善现有知识图谱。

包含两个评测任务：

1.三元组分类

判断是否正确 比如

首都（北京，中国）

首都（成都，中国）是错的，二分类问题，

过程是补全时可以选一条边连接任意两个实体，构成新的三元组，判断是否正确。

2.链接预测

预测三元组的头实体或尾实体

能预测出则能加关系。

### 9.1.2知识问答

简单推理

将问题转化为图谱上三元组的查询或者三元组序的查询。对于缺失的需要使用推理，例如唐朝开国皇帝是谁，若开国皇帝（唐朝，李渊）是缺失的。通过前文的链接预测开国皇帝（唐朝，X）。若问开国皇帝的父亲是谁，推理开国皇帝（唐朝，李渊），再对父亲（李渊，x）进行查询和推理。

复杂推理

表示成多个链接组成的非链式或有嵌套的复杂结构时，需要推理。

桃花开花后结果，问开花的目的是什么

答案是吸引蜜蜂采蜜，就是一个推理问题。

## 9.2知识推理分类

### 9.2.1归纳推理和演绎推理

归纳推理

从特殊到一般的过程。根据部分对象具有的性质，推出一类事物中所有对象都具有的这类性质的推理方式。

对资料进行观察、分析和归纳整理

得出规律性的结论，即猜想

检验猜想

例如<蓝鲸，可以喷射，水柱>

<抹香鲸，可以喷射，水柱>

观察整理得：

<xx鲸，可以喷射，水柱>

检验猜想：

<齿鲸，是一种，鲸鱼>

<齿鲸，可以喷射，水柱>

演绎推理

从一般到特殊

大前提，小前提，结论

大前提：

虎鲸背部有背鳍

小前提：

背部有背鳍的鲸鱼都属于海豚科

结论：

虎鲸属于海豚科

### 9.2.2确定性推理与不确定性推理

确定性逻辑推理

逻辑推理，具有完备的推理过程和充分的表达能力

不确定性推理

根据以往的经验和分析，结合专家先验知识构建概率模型，并利用统计计数、最大化后验概率等统计学习的手段对推理假设进行验证或推测。

概率图模型

有向图的贝叶斯网络以及无向图的马尔科夫网络

NP难题 主要的改进有：

基于和积变量消除的方法，通过对一个变量求和，并和其他因子相乘以消除变量，简化

基于概率图结构的置信传播或期望传播的方法，将原有的推理问题转为优化问题，优化 的方式设计好的能力函数或势函数求解概率最大以达到推理的目的

从所有实例触发，都其进行统计或采样以估计推理目标概率，如蒙特卡洛采样等

只对具有直接概率依赖的实例级元素，并没有对更高层次的语义框架进行抽象，但需要大量的重复的概率依赖关系，需要大量计算。

概率逻辑推理

弥补了概率图模型中缺乏可复用规则的特点。

结构学习又可以称为概率逻辑推理模型下的规则自动挖掘。迭代局部搜索代替全局搜素。

关联规则挖掘

路径排序算法是基于图模型上随机游走的启发式方法，通过枚举或抽样图上的两个节点间的路径，递归地计算两个点间的到达概率，对每个路径进行打分。

### 9.2.3符号推理和数值推理

传统的逻辑推理。特点是在知识图谱中的实体和符号上直接进行推理了操作。

## 9.3基于符号演算的推理

### 9.3.1归纳推理：学习推理规则

有三大方法

频繁子图挖掘

父亲（x,y）^母亲（y,z）->奶奶（x,z）

频繁子图规则挖掘是一个相反的过程，它是搜索知识图谱的规则实例，再将规则实例中的实体替换成变量，加约束，确定规则的实用性。

计算路径 剪枝

AMIE算法：

增加悬挂边

增加具体边

增加闭合边

归纳逻辑编程

更重视没有没有出现在图谱中的负三元组，认为 正例+负例+背景知识->假设。

结构学习方法

挑选整体概率最大的加入到推理模型中。 + 定义目标函数

创建逻辑规则结构

搜索逻辑规则的策略

### 9.3.2演绎推理：推理具体事实

确定性推理：λ推理

λ程序语言设计的模型。

λ三种操作：

α-置换

β-归约

η-变换

不确定性推理：马尔科夫逻辑网和概率软逻辑

马尔科夫逻辑是被看成一种通过为逻辑规则绑定权重的方式将一阶逻辑向概率逻辑进行扩展的方法，当权重调整到无限大时，为一阶逻辑。

软逻辑是一种基于一阶逻辑谓词和马尔科夫逻辑网络的统计学习关系票框架。

## 9.4基于数值计算的推理

将离散的符号表示成低维实数向量或矩阵以捕捉元素之间隐式关联的一种技术手段。

映射可以带来如下好处：

减少维度灾难

减少数据稀疏

通过数值计算填充了稀疏矩阵，在一定程度上解决了数据稀疏性问题。

使符号直接参与计算且计算速度非常快

有如下方法：

### 9.4.1基于张量分解

矩阵分解的基本思想是用多个低维的矩阵或张量的积代替原始的关系矩阵，从而用少量的参数代替稀疏而大量的原始数据。

### 9.4.2基于能量函数

目标不是恢复出原始的关系矩阵和张量，而是根据任务的不同，自定义能量函数使得三元组能量低，不成立的能量高，计算能量函数对事实是否成立进行推理。

# 知识问答与对话

知识图谱最直接和最重要的任务是满足用户的精确信息需求，提供个性化知识服务。

目前的问答只能事实型问题，不能很好地处理复杂问题。例如：为什么天是蓝色的。关注回答事实型问题的相关技术。

## 10.1自动问答概述

九十年代中期，系统的主要特点：利用浅层自然语言处理技术分析问题，并利用信息检索等技术从大规模文本或网页红抽取答案。 由于用户需求的多样性和自然语言的复杂性，没有得到广泛应用。

两大困难：

缺乏高质量的知识资源

高效的自然语言技术

近几年，这两点有了很大的突破。

问答系统分很多类型：

单文本问答系统：阅读理解式问答系统 如入学考试评测任务和斯坦福问答测试平台SQuAD

固定语料非结构化文本问答系统

从给定的文本语料中抽取答案

网络问答系统

从互联网中查找问题的答案，但是因为互联网动态变化，难以评测

知识库问答系统

知识问答和基于知识图谱的问答系统

从预先建立好的结构化知识库中查找问题的答案。本章主要是这两个方面：

## 10.2知识问答

### 10.2.1知识问答技术概述

结构化查询语言的有点事表达能力强，但是缺点是用户不但需要掌握结构化查询语言的语法，而且还要充分了解知识库中的资源表达形式。

使用自然语言进行人机交互是最简便、最直接也是最有效的一种模式。

按照技术分为两种类型：

语义解析类型

转化为结构化语句，就可以直接通过检索知识图谱得到精简答案。

搜索排序类型

首先通过搜索与相关实体有路径联系的实体作为候选答案，利用从问句和候选答案提取出来的特征进行比对，进而对候选答案进行排序得到最优答案，例如屠呦呦，得到宁波市，中国，诺贝尔奖等实体，进行匹配和排序得到答案。

### 10.2.2基于语义解析的方法

实质是语义解析。首先需要对问句中的词/短语与知识图谱中的资源项进行映射，然后对匹配到的资源项进行组合，最后对匹配和组合存在的歧义进行消解，得到结果。

语义解析分为：

基于训练数据的有监督方法

有哪些城市靠近大海？

哪些城市，靠近大海的语义和之间的组合关系

如何确定问句的字句（短语）和它们对应的形式化表示？

当已知各个子句的含义及其形式化表示，如何对它们进行语义组合？

1.语义组合模型

组合范畴语法的主要思想是把词的句法和语义信息组合在一起形成分析的基础词典，依据组合语法规则自底向上对自然语言句子进行解析。

2.语义辞典构造

逻辑表示语言

3.组合消歧模型

基于规则的无监督方法

语言分析工具模块

元素级匹配模块

三元组映射模块

融合和排序模块

### 10.2.3基于搜索排序的方法

主题词，链接到实体库上。 从候选答案中选择正确的答案。或者排序。

基于特征工程的方法

对问句和候选答案定义特征，并使用特征工程的方法抽取它们，最后基于特征分配的分类模型对问题和答案匹配度进行建模。

问句特征抽取

依存句法分析，抽取

问题词，如谁，哪

问句焦点词，如名字、时间

主题词，多个选一个出来

中心动词，和答案相关的一些信息

把依存语法树转为特征树

候选答案特征抽取

问句-候选答案匹配

基于表示学习的神经网络方法

## 10.3知识对话

知识对话，对话系统是更自然友好的知识服务模式，可以通过多轮人机交互满足用户的需求、完成具体的任务等。

对话系统有几大特征：

多角色切换

对话中通常有两个甚至多个角色，可以是提问者，也可以是问题回答者等，并且在对话中，各角色之间常常交替变化

连贯性

对话的前后内容是有关联的、有逻辑的

多模态

真正的对话中，包括语音、文字、图片等

### 10.3.1知识对话技术概述

语音识别

接收用户的输入信息，把输入转为计算机方便表示和处理的文字形式。

对话理解

对用户的输入信息进行分析处理，获得对话的意图

对话管理

根据对话的意图做出合适的响应，控制整个对话过程，使用户和对话系统顺利交互，解决用户的问题

任务管理

根据具体的任务管理对话过程所涉及的实例型知识数据和领域知识

对话生成

负责将对话管理系统的决策信息转换成文本结构的自然语言

语音合成

负责将文本结果的信息转换成语音数据发送给用户

分为两类：

任务导向型系统

用户在使用系统时有确定的目标，一般为确定任务，如：订机票、查路线等

通用对话系统

用户没有具体目标，可能在多个任务之间切换

### 10.3.2任务导向型对话模型

首先是系统引导对话，用户输入意图，通过用户输入和系统引导的方式交互地完善用户意图信息。

自然语言理解

目的是将文本数据表示的信息转换为可被机器处理的语义表示。

订机票有车次、类型、出发站、到达站、出发时间、到达时间、历时、票价

难度：

同样的意思有很多种不同的表达方式

自然语言表达常常存在不确定性

相同语言表达在不同语境下的语义可能完全不同

在自然语言中往往存在不规范、不流畅、重复、指代甚至是错误等情况 基于规则和模板，但是无法覆盖全。

槽抽取作为一个序列标注问题，通过对每个输入词的标注和分类找出各个槽对应的值。

对话管理

最重要的部分，也是体现其区别于问答系统的核心步骤。

用于控制对话的框架和结构，维护对话状态，通过与任务管理器的交互生成相应的动作。

方法有：

基于有限状态自动机的方法

最简答的对话管理方式。

把任务完成过程中系统向用户询问的各个问题表示为状态，而整个对话可以表示为状态之间的转移。

基于框架的方法

使用任务常常用框架表示，根据槽的情况控制对话的过程。由于可以一次获取框架中的多个槽值，因此，不需要重新询问用户已经提供过的信息。另外，填写各个槽值的时候也不需要按照固定的顺序进行，

基于概率模型的方法

以上两种均需要人工制定规则，耗费人力。基于概率统计的对话管理方法，使用数据驱动的方法自动学习对话模型。对话过程是个连续决策任务。

动作集合的目标是最大化完成任务回报等最小化损失。基于MDP求解的过程。

自然语言生成

得到具体的回复内容。

内容选择

管理模块决定

内容描述

生成模块

基于编码器和解码器的深度学习模型 + 知识检索

实体词 ，三元组

编码器

将问句和检索的知识实例编码成向量，以便后续深度生成模型利用。

解码器

根据答案和知识的编码向量生成自然答案。比如答案李连杰出生在北京

### 10.3.3通用对话模型

基于模板的任务

处理和生成的句式简单，也比较固定，能处理的领域有限，不能处理复杂的对话，处理的自然度差。

输出关键词，和符合语法的输出。比较生硬，不够自然。

端到端的方法

深度学习从原始对话中学习对话模型，从而对各种各样的消息进行回复。

如果把句子（消息-回复）看出词的序列，那么对话的建模就是一个学习一个词序列X=[x1,x2,…,x3]到词序列Y=[y1,y2,…y3]转化的任务。

基于深度学习的seq2aeq模型。

采用的是编码器-解码器的框架。基于编码器得到抽象编码表示，另一个RNN用于解码。

## 10.4评价方法

难以评价。人工评估和测试

用户满意度调查有几项：

文本识别模块的性能

任务易用度

交互速度

用户体验

系统范阔

期望行为

未来使用

评价指标基于以下两个准则:

最大化任务完成度

最小化成本开销