# 知识图谱的基本知识

## 基本概念

1. 知识图谱是什么

知识图谱本质上是一种语义网络，用图的形式描述客观事物，这里的图指的是数据结构中的图，也就是由节点和边组成的，这也是知识图谱（Knowledge Graph）的真实含义。知识图谱中的节点表示概念和实体，概念是抽象出来的事物，实体是具体的事物；边表示事物的关系和属性，事物的内部特征用属性来表示，外部联系用关系来表示。很多时候，人们简化了对知识图谱的描述，将实体和概念统称为实体，将关系和属性统称为关系，这样就可以说知识图谱就是描述实体以及实体之间的关系。实体可以是人，地方，组织机构，概念等等，关系的种类更多，可以是人与人之间的关系，人与组织之间的关系，概念与某个物体之间的关系等等。

1. 知识图谱是怎么组织数据的

知识图谱是由实体和实体的关系组成，通过图的形式表现出来，那么实体和实体关系的这些数据在知识图谱中怎么组织呢，这就涉及到三元组的概念，在知识图谱中，节点-边-节点可以看作一条记录，第一个节点看作主语，边看作谓语，第二个节点看作宾语，主谓宾构成一条记录。

1. 知识图谱的数据结构

从知识图谱数据组织的架构来看，可以把知识图谱的数据分为两个层次，一个是数据模型层，数据模型是按照本体论的思想，勾画出来的数据组织模式，数据模型可以展示数据的组织方式，数据之间的相互关系。数据模型除了确定对象之间的分类，关系，还要明确对象的属性，针对不同的知识图谱，需要收集的数据的内容也不相同，内容范围由对象的属性确定。数据模型的分类，关系反映了数据之间的关系特征，数据模型的属性反映了数据的内在特征；另一个就是具体数据层，具体数据是一条条的知识，它是依据数据模型组织起来的。在构建知识图谱的时候，是先确定数据模型再收集具体数据，还是先收集具体数据再确定数据模型，这就形成了两种构建知识图谱的方式。

# 知识图谱的构建方式

1. 自顶向下的方式

自顶向下的构建方式，是指先确定知识图谱的数据模型，再根据模型去填充具体数据，最终形成知识图谱。数据模型的设计，是知识图谱的顶层设计，根据知识图谱的特点确定数据模型，就相当于确定了知识图谱收集数据的范围，以及数据的组织方式。这种构建方式，一般适用于行业知识图谱的构建。

1. 自底向上的方式

自下向上的构建方式，是指先按照三元组的方式收集具体数据，然后根据数据内容来提炼数据模型。 采用这种方式构建知识图谱，是因为在开始构建知识图谱的时候，还不清楚收集数据的范围，也不清楚数据怎么使用，就是先把所有的数据收集起来，形成一个庞大的数据集，然后再根据数据内容，总结数据的特点，将数据进行整理、分析、归纳、总结，形成一个框架，也就是数据模型。

# 本体的构建

1. 基于本体论的工程构建方法
2. 骨架法，IDEF5法，七步法，五步循环法，METH-ONTOLOGY法，TOVE法，KACTUS法，SENSUS法。具体差距如图。
3. METH-ONTOLOCY法、KACTUS法、SENSUS法和七步法,主要用于构建领域知识本体,它们的不同之处在于:METHONTOLOCY法是以化学领域的本体构建方法为基础,经过改进发展而来的,构建方法更为常用；KACTUS法主要是对已有本体的提炼、扩展,难以用于构建新的本体; ;SENSUS法遵循自上而下的层级结构,可操作性较强；七步法是基于本体构建工具Protégé的本体构建方法,较为实用,应用广泛。
4. 七步法

主要用于领域本体的构建，七个步骤分别是

1. 确定本体的专业领域和范畴
2. 考查复用现有本体的可能性
3. 列出本体中的重要术语
4. 定义类和类的等级体系
5. 定义类的属性
6. 定义类的分面
7. 创建实例
8. METH-ONTOLOGY法

这个本体开发方法更接近软件工程开发方法。它将本体开发进程和本体生命周期两个方面区别开来，并使用不同的技术予以支持。

它的流程包括：

1. 管理阶段：这一阶段的系统规划包括任务的进展情况、需要的资源、如何保证质量等问题。
2. 开发阶段：分为规范说明、概念化、形式化、执行以及维护五个步骤。
3. 维护阶段：包括知识获取、系统集成、评价、文档说明、配置管理五个步骤。
4. 以上各常用方法成熟度排名

七步法、Methontology方法、IDEF-5法、TOVE法、骨架法

1. 综合上述本体构建的工程思想，归纳并总结出构建领域本体的步骤如下
2. 确定领域本体的专业领域

领域知识往往十分庞大的，本体不可能包括所有的概念，因此，在建立本体前必须先确定本体将覆盖的专业领域、范围和应用目标，本体应该在哪些方面发挥作用以及它的系统维护者与应用对象。不同的应用领域，领域概念肯定是不同的，即使是同一个领域，由于应用的不同，本体表示的概念的侧重点肯定也会有所不同。因此，建立本体之前一定要明确本体建立的领域和应用目标。本体是一个复杂的知识体系，确定每个阶段的范围和目标有助于对本体模型的范围作一个限定，有利于复杂系统的实现。

1. 考虑复用现有的本体

本体的主要作用就是解决知识的共享和重用问题。所以在设计和建立自己的领域本体之前，应该考虑重用已经存在的本体。

1. 列出本体涉及领域中的重要术语

领域本体是描述概念以及概念与概念之间的关系，首先要列举出该领域中的所有概念以及对该概念的详细解释。在特定领域，这些概念就是与领域相关的专业术语。把领域中一些重要术语列举出来，有利于知识工程师更好地理解本体建立的目标，明确方向。除此之外，针对每个概念，要列出它所有可能的属性，每个属性都有对应的属性值。

1. 定义分类概念和概念分类

概念分类层次将领域概念进行分类组织，用于描述领域概念间的类属关系，并将本体中的概念模块化。建立一个分类概念的层次结构有3种可行的方法: 自顶向下法、自底向上法和综合法。

一般领域概念分类层次对应着一棵树，树中的节点体现了领域概念间的层次结构关系。树有四类元素组成：根节点，枝节点，树枝，叶节点。

建立领域概念的分类关系后，将分类概念的属性值添加到分类概念中，这样就把领域概念通过树形结构形象地描述出来，并且通过树结构清晰地体现了领域概念间的类属关系。每一个子树都对应着领域中独立的、模块化的知识模型。

领域分类概念应该包括:概念名称，语义描述，该概念可能的同义词、缩略语。定义分类概念，就是对这些信息进行描述。同时，要对所建立的概念分类层次进行检验，保证没有重复的概念，防止冗余定义。

1. 定义概念之间的关系

概念的分类层次结构体现了分类概念之间的一种继承关系(kind-of)，但是在领域本体中，概念和概念之间通过关系来交互，除了继承关系，在我们构建的领域本体中还可以根据需要，定义其他的关系。

1. 将叙词表转换成本体
2. 基于叙词表的领域本体构建

手工创建：手工转换主要通过手工操作构建本体模型,将叙词表中的专业词汇转换为本体概念, 再将词汇关系及注释转换为本体语义关系,最后添加领域知识等完善本体。手工转换方式费时费力、效率低、过程繁琐,所以在叙词表转换研究中应用较少,与之相关的研究也没有太多进展。

具体方法：

1. 联合国粮农组织在农业本体研究项目中,提出一个由叙词表向本体转化的方法模型:首先需要细化转换规则,丰富语义关系;其次增加推理规则;最后采用程序辅助填词,进行概念抽取,按照转换规则构建本体。
2. 一种基于概念的转换方法，

首先需要对叙词表中术语的明确性进行分析,确定之后在进行语法转换,主要是原始语法格式转化为RDF(S),最后进行语义转换。该方式是叙词表转化本体最基本的方法,可将任意形式的叙词表转换为本体,不需要提前构建本体模型,具有较强的实用性和通用性。

1. 基于OWL的网络叙词表转换方法

首先在准备阶段需要对叙词表进行分析;其次确定类和属性,建立类间关系层次结构;接着建立语法和语义关系参照表;最后使用OWL进行标准化表示和描述。使用此方法构建的本体具有更好的推理能力,更易更新和维护。此方法是在上述方法基础上的进一步优化改进。具有较强的实用性。

# 本体构建工具

1. **可视化手工构建工具**

可视化手工构建工具包括Protégé、Apollo、Web Onto、Web ODE和Onto Edit。

各构件工具优缺点对比见https://kns.cnki.net/KXReader/Detail?TIMESTAMP=637314524750557500&DBCODE=CJFQ&TABLEName=CJFDLAST2016&FileName=QBZZ201606028&RESULT=1&SIGN=CJutYrU3%2fLnprXTHPMb7%2boZcP0c%3d

1. **半自动化构建工具**

基于Java语言的Jena也仅属于半自动本体构建工具。Jena提供实现本体形式化的方法,可通过程序调用方法来实现本体的自动构建。

总之,目前常用的本体构建工具仍存在诸多问题,还未实现高效快捷的本体自动化构建。虽然多数工具提供了有好的图形界面和错误检查机制, 避免了多数人为错误的发生,但这些工具仍需要手工输入和编辑大量的数据信息,费时费力,并且出错率较高,难以实现大规模的本体构建。

# 构建流程

### 数据处理

1. 结构化数据处理

结构化数据主要是指关系型数据库中存储的数据。把关系型数据库中的数据转换为RDF数据（linked data），普遍采用的技术是D2R技术。D2R主要包括D2R Server，D2RQ Engine和D2RRQ Mapping语言。d2rq最主要的功能是将mysql映射到mapping上，通过修改，最终获取RDF形式的数据，即重在数据格式或者内容的转化。

D2RQ Engine的主要功能是使用一个可定制的 D2RQ Mapping 文件将关系型数据库中的数据换成 RDF 格式。D2RQ Mapping 语言的主要功能是定义将关系型数据转换成 RDF 格式的 Mapping 规则。通过mapping语言可以把关系型数据库中的数据映射到我们定义好的本体中。

1. 半结构化数据处理

半结构化数据，主要是指那些具有一定的数据结构，但需要进一步提取整理的数据。

自动的从标注好的训练样例集合中学习数据抽取规则,用于从其他相同标记或相同网页模板抽取目标数据。另一种情况下，可以通过包装器归纳这种基于有监督学习的方法，自动的从标注好的训练样例集合中学习数据抽取规则,用于从其他相同标记或相同网页模板抽取目标数据。

1. 非结构化数据处理

概念的获取和概念间关系的获取。

概念的获取，常见的有三种方法，基于语言学的方法，基于统计学的方法，以及将上述两种方法结合起来。

1. 基于语言学的方法。主要根据语法规则、词法结构来获得术语字符串。由于不同的语言的语法规则, 词法结构是不一致的, 因此, 基于语言学的概念获取模式针对不同的语言要做不同的处理。
2. 基于统计学的方法。主要根据领域概念具有不同于一般词汇的统计特性来鉴别领域概念。它在领域的文本集上进行词频的统计, 然后运用互信息, 似然对数比等方法来获得搭配。由于搭配, 术语, 技术术语和专有名词往往有很多交叉的地方，所以从某个领域来讲, 搭配的获取已经隐式地包含了术语的获取。
3. 结合方法。一般先对文本进行语法分析, 获得候选的术语集, 然后通过统计的方法对候选集合进行过滤;也可以先对文本进行统计处理, 再用语法过滤, 从而更加精确地获取某个领域的术语。

### 概念间关系抽取的方法，

1. 基于模式抽取的方法需要判断文本中词的序列是否匹配某个模式, 如果匹配, 则可以识别出相应的关系。
2. 概念聚类的方法是利用概念之间的语义距离, 对概念进行层次聚类, 聚类的结果就是概念间的分类关系。
3. 关联规则挖掘的方法常用于获取概念间的非分类关系, 其基本思想是, 如果两个概念经常出现在同一文档 (或段落, 或句子) 中, 则这两个概念之间必定存在关系。

# 本体的构建原则

1. 一致性

原则1：避免类层次中的循环。类间的“超类——子类”关系不能存在循环。

原则2：描述具体事物的类的子类中没有描述抽象事物的类。本体分类结构中的类从上到下是逐渐具体化的。

原则3：语义距离最小：有相同直接超类的类之间的语义距离最小。

原则4：子类不相交原则：类按照确定依据分出的子类的交集为空。

原则5：类名唯一原则：本体类中的命名唯一，即不存在同名的类。

1. 完备性

原则1：多分类标准原则。一个类可以按多种标准分类。

原则2：个体有类型原则。任何个体都是一定类的实例，都具有类型。

原则3：类与本体定义完备。本体中出现的任何类和个体都有定义。

1. 减少冗余

原则1“超类—子类”关系无冗余原则：层次关系的传递原则。即如果本体中已有知识“C1是C2子类”和“C2是C3的子类”，则知识“C1是C3子类”是不必要的。

原则2类知识位置最“高”原则：在本体系统中位置最可能“高”的类中描述个体共有的知识。在本体模型中，子类继承父类所有的属性，因此属性应当被尽可能定义在通用类中。

原则3类的子类个数不为一原则：如果一个类只有一个直接子类，那么这个类是不完善的，此时应将子类与超类合并。

# 课程中的原则

1. 业务原则

一切从业务逻辑出发，也就是说通过观察知识图谱的设计也很容易推测其背后业务的逻辑，而且设计时要想好未来业务的变化。

例子：张三 。在前面这张图中，我们把申请人张三作为申请的属性保存，一来我们无法看出这幅图表示的实际业务。二来这样的构建方式意味着借款人本身的重要性低于贷款申请这个实体，这与实际业务的矛盾的。

1. 分析原则

知识图谱中任何一个实体都是为关系分析而服务的，如果一个实体对分析网络结构没有帮助，则可以设置成属性或者不要放在知识图谱中。

例子：比如在金融知识图谱中，电话网络是很重要的一部分，很多欺诈行为可能隐藏在电话网络之中。而在这种业务场景下，如果把电话作为属性，则很难从电话网络中获得一些有效信息。

1. 冗余原则

避免把超级节点放入知识图谱中，这会导致系统的性能急剧下降。同时避免存储任何重复的信息。

1. 效率原则

知识图谱尽量轻便，只存储关键信息，剩下的可以存储在传统数据库中。