Atitit 图3.知识获取的基本步骤

知识获取

图3.知识获取的基本步骤

知识的获取是个系统工程，流程复杂，内涵丰富，涉及到知识表示、自然语言处理、数据库、数据挖掘、众包等一系列技术。知识获取的基本步骤如图3所示：

第一步是模式（Schema）设计。这一步是传统本体设计所要解决的问题。基本目标是把认知领域的基本框架赋予机器。在所谓认知基本框架中需要指定领域的基本概念，以及概念之间subclassof关系（比如足球领域需要建立“足球运动员”是“运动员”的子类）；需要明确领域的基本属性；明确属性的适用概念；明确属性值的类别或者范围。比如“效力球队”这个属性一般是定义在足球运动员这个概念上，其合理取值是一个球队。

此外，领域还有大量的约束或规则，比如对于属性是否可以取得多值的约束（比如“奖项”作为属性是可以取得多值的），再比如球队的“隶属球员”属性与球员的“效力球队”是一对互逆属性。这些元数据对于消除知识库不一致、提升知识库质量具有重要意义。

第二步是明确数据来源。在这一步要明确建立领域知识图谱的数据来源。可能来自互联网上的领域百科爬取，可能来自通用百科图谱的导出，可能来自内部业务数据的转换，可能来自外部业务系统的导入。应该尽量选择结构化程度相对较高、质量较好的数据源，以尽可能降低知识获取代价。

第三步是词汇挖掘。人们从事某个行业的知识的学习，都是从该行业的基本词汇开始的。在传统图书情报学领域，领域知识的积累往往是从叙词表的构建开始的。叙词表里涵盖的大都是领域的主题词，及这些词汇之间的基本语义关联。在这一步我们是要识别领域的高质量词汇、同义词、缩写词，以及领域的常见情感词。比如在政治领域，我们需要知道特朗普又被称为川普，其英文简称为Trump。

第四步是领域实体发现（或挖掘）。需要指出的是领域词汇只是识别出领域中的重要短语和词汇。但是这些短语未必是一个领域实体。从领域文本识别某个领域常见实体是理解领域文本和数据的关键一步。在实体识别后，还需对实体进行实体归类。能否把实体归到相应的类别（或者说将某个实体与领域类别或概念进行关联），是实体概念化的基本目标，是理解实体的关键步骤。比如将特朗普归类到政治人物、美国总统等类别，对于理解特朗普的含义具有重要意义。实体挖掘的另一个重要任务是实体链接，也就是将文本里的实体提及（Mention）链接到知识库中的相应实体。实体链接是拓展实体理解，丰富实体语义表示的关键步骤。

第五步是关系发现。关系发现，或者知识库中的关系实例填充，是整个领域知识图谱构建的重要步骤。关系发现根据不同的问题模型又可以分为关系分类、关系抽取和开放关系抽取等不同变种。关系分类旨在将给定的实体对分类到某个已知关系；关系抽取旨在从文本中抽取某个实体对的具体关系；开放关系抽取（OpenIE）从文本中抽取出实体对之间的关系描述。也可以综合使用这几种模型与方法，比如根据开放关系抽取得到的关系描述将实体对分类到知识库中的已知关系。

第六步是知识融合。因为知识抽取来源多样，不同的来源得到的知识不尽相同，这就对知识融合提出了需求。知识融合需要完成实体对齐、属性融合、值规范化。实体对齐是识别不同来源的同一实体。属性融合是识别同一属性的不同描述。不同来源的数据值通常有不同的格式、不同的单位或者不同的描述形式。比如日期有数十种表达方式，这些需要规范化到统一格式。

最后一步是质量控制。知识图谱的质量是构建的核心问题。知识图谱的质量可能存在几个基本问题：缺漏、错误、陈旧。

先谈知识库的缺漏问题。某种意义上，知识完备对于知识资源建设而言似乎是个伪命题，我们总能枚举出知识库中缺漏的知识。知识缺漏对于自动化方法构建的知识库而言尤为严重。但是即便如此，构建一个尽可能全的知识库仍是任何一个知识工程的首要目标。既然自动化构建无法做到完整，补全也就成为了提升知识库质量的重要手段。补全可以是基于预定义规则（比如一个人出生地是中国，我们可以推断其国籍也可能是中国），也可以从外部互联网文本数据进行补充（比如很多百科图谱没有鲁迅身高的信息，需要从互联网文本寻找答案进行补充）。

其次是纠错。自动化知识获取不可避免地会引入错误，这就需要纠错。根据规则进行纠错是基本手段，比如A的妻子是B，但B的老公是C，那么根据妻子和老公是互逆属性，我们知道这对事实可能有错。知识图谱的结构也可以提供一定的信息帮助推断错误关联。比如在由概念和实例构成的Taxonomy中，理想情况下应该是个有向无环图，如果其中存在环，那么有可能存在错误关联。

最后一个质量控制的重要问题是知识更新。更新是一个具有重大研究价值，却未得到充分研究的问题。很多领域都有一定的知识积累。但问题的关键在于这些知识无法实时更新。比如电商的商品知识图谱，往往内容陈旧，无法满足用户的实时消费需求（比如“战狼同款饰品”这类与热点电影相关的消费需求很难在现有知识库中涵盖）。

经历了上述步骤之后得到一个初步的知识图谱。在实际应用中会得到不少反馈，这些反馈作为输入进一步指导上述流程的完善，从而形成闭环。此外，除了上述自动化构建的闭环流程，还应充分考虑人工的干预。人工补充很多时候是行之有效的方法。比如一旦发现部分知识缺漏或陈旧，可以通过特定的知识编辑工具实现知识的添加、编辑和修改。也可以利用众包手段将很多知识获取任务分发下去。如何利用众包手段进行大规模知识获取，是个十分有意思的问题，涉及到知识贡献的激励机制，我前几年有个题为《未来人机区分》的报告，专门讨论如何利用知识问答形式的验证码来做知识获取，可以搜索此文获取更多信息。

---------------------

作者：AI科技大本营

来源：CSDN

原文：https://blog.csdn.net/dQCFKyQDXYm3F8rB0/article/details/83451280

版权声明：本文为博主原创文章，转载请附上博文链接！