Atitit 人工智能与知识图谱的关系

、知识图谱的重要性

知识图谱是实现机器认知智能的基础。机器认知智能的两个核心能力：“理解”和“解释”，均与知识图谱有着密切关系。首先需要给机器“理解与解释”提出一种解释。我认为机器理解数据的本质是建立起从数据到知识库中的知识要素（包括实体、概念和关系）映射的一个过程。

比如如果我说到“2013年的金球奖得主C罗”这句话，我们之所以说自己理解了这句话，是因为我们把“C罗”这个词汇关联到我们脑子中的实体“C罗”，把“金球奖”这个词汇映射到我们脑中的实体“金球奖”，然后把“得主”一词映射到边“获得奖项”这个关系。我们可以仔细体会一下我们的文本理解过程，其本质是建立从数据（包括文本、图片、语音、视频等）到知识库中的实体、概念、属性映射的过程。

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

看人类是如何“解释”的。比如我问“C罗为什么那么牛？”，可以通过知识库中的“C罗获得奖项金球奖”以及“金球奖地位影响力最大的足球奖项之一”这两条关系来解释这一问题。

这一过程的本质就是将知识库中的知识与问题或者数据加以关联的过程。有了知识图谱，机器完全可以重现我们的这种理解与解释过程。有过一定计算机研究基础的，是不难完成上述过程的数学建模的。知识图谱对于机器认知智能的重要性也体现在下面几个具体方面。

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

识图谱对机器认知智能的必要性还可以从若干具体问题来进行阐述。首先，我们来看机器认知的核心能力之一：自然语言理解。

我的观点是机器理解自然语言需要类似知识图谱这样的背景知识。自然语言是异常复杂的：自然语言有歧义性、多样性，语义理解有模糊性且依赖上下文。机器理解自然语言困难的根本原因在于，人类语言理解是建立在人类的认知能力基础之上的，人类的认知体验所形成的背景知识是支撑人类语言理解的根本支柱。

我们人类彼此之间的语言理解就好比是根据冰山上浮出水面的一角来揣测冰山下的部分。我们之所以能够很自然地理解彼此的语言，是因为彼此共享类似的生活体验、类似的教育背景，从而有着类似的背景知识。冰山下庞大的背景知识使得我们可以彼此理解水面上有限的几个字符。我们可以做个简单的思想实验，假如现在有个外星人坐在这里听我讲报告，他能听懂么？我想还是很困难的，因为他没有在地球上生活的经历，没有与我相类似的教育背景，没有与我类似的背景知识库。

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

个很多人都有体会的例子，我们去参加国际会议时，经常遇到一个尴尬的局面，就是西方学者说的笑话，我们东方人很难产生共鸣。因为我们和他们的背景知识库不同，我们早餐吃烧饼、油条，西方吃咖啡、面包，不同的背景知识决定了我们对幽默有着不同的理解。所以语言理解需要背景知识，没有强大的背景知识支撑，是不可能理解语言的。要让机器理解我们人类的语言，机器必需共享与我们类似的背景知识

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

现机器自然语言理解所需要的背景知识是有着苛刻的条件的：规模足够大、语义关系足够丰富、结构足够友好、质量足够精良。

以这四个条件去看知识表示就会发现，只有知识图谱是满足所有这些条件的：知识图谱规模巨大，动辄包含数十亿实体；关系多样，比如在线百科图谱DBpedia包含数千种常见语义关系；结构友好，通常表达为RDF三元组，这是一种对于机器而言能够有效处理的结构；质量也很精良，因为知识图谱可以充分利用大数据的多源特性进行交叉验证，也可利用众包保证知识库质量。所以知识图谱成为了让机器理解自然语言所需的背景知识的不二选择。

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

知识图谱使能可解释人工智能

知识图谱对于认知智能的另一个重要意义在于：知识图谱让可解释人工智能成为可能。

“解释”这件事情一定是跟符号化知识图谱密切相关的。因为解释的对象是人，人只能理解符号，没办法理解数值，所以一定要利用符号知识开展可解释人工智能的研究。可解释性是不能回避符号知识的。

我们先来看几个解释的具体例子。比如，我若问鲨鱼为什么可怕？你可能解释说：因为鲨鱼是食肉动物，这实质上是用概念在解释。若问鸟为什么能飞翔？你可能会解释因为它有翅膀。这是用属性在解释。若问鹿晗关晓彤前些日子为什么会刷屏？你可能会解释说因为关晓彤是鹿晗的女朋友。这是用关系在解释。我们人类倾向于利用概念、属性、关系这些认知的基本元素去解释现象，解释事实。而对于机器而言，概念、属性和关系都表达在知识图谱里面。因此，解释离不开知识图谱。

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

知识引导将成为解决问题的主要方式

 知识图谱的另一个重要作用体现在知识引导将成为解决问题的主要方式。前面已经多次提及用户对使用统计模型来解决问题的效果越来越不满意了，统计模型的效果已经接近“天花板”，要想突破这个“天花板”，需要知识引导。

举个例子，实体指代这样的文本处理难题，没有知识单纯依赖数据是难以取得理想效果的。比如“张三把李四打了，他进医院了”和“张三把李四打了，他进监狱了”，人类很容易确定这两个不同的“他”的分别指代。因为人类有知识，有关于打人这个场景的基本知识，知道打人的往往要进监狱，而被打的往往会进医院。但是当前机器缺乏这些知识，所以无法准确识别代词的准确指代。很多任务是纯粹的基于数据驱动的模型所解决不了的，知识在很多任务里不可或缺。比较务实的做法是将这两类方法深度融合。

▌2.5 知识将显著增加机器学习能力

知识对于认知智能又一个很重要的意义就是将显著增强机器学习的能力。

当前的机器学习是一种典型的“机械式”学习方式，与人类的学习方式相比显得比较笨拙。我们的孩童只需要父母告知一两次：这是猫，那是狗，就能有效识别或者区分猫狗。而机器却需要数以万计的样本才能习得猫狗的特征。

我们中国人学习英语，虽然也要若干年才能小有所成，但相机器对于语言的学习而言要高效的多。机器学习模型落地应用中的一个常见问题是与专家知识或判断不符合，这使我们很快陷入进退两难的境地：是相信学习模型还是果断弃之？机器学习与人类学习的根本差异可以归结为人是有知识的且能够有效利用知识的物种。

我相信，未来机器学习能力的显著增强也要走上充分利用知识的道路。符号知识对于机器学习模型的重要作用会受到越来越多的关注。这一趋势还可以从机器智能解决问题的两个基本模式方面加以论述。

机器智能的实现路径之一是习得数据中的统计模式，以解决一系列实际任务。另一种是专家系统，专家将知识赋予机器构建专家系统，让机器利用专家知识解决实际问题。如今，这两种方法有合流的趋势，无论是专家知识还是通过学习模型习得的知识，都将显式地表达并且沉淀到知识库中。再利用知识增强的机器学习模型解决实际问题。这种知识增强下的学习模型，可以显著降低机器学习模型对于大样本的依赖，提高学习的经济性；提高机器学习模型对先验知识的利用率；提升机器学习模型的决策结果与先验知识的一致性。

因此，知识将成为比数据更为重要的资产。前几年大数据时代到来的时候，大家都说“得数据者得天下”。去年，微软的沈向洋博士曾经说过“懂语言者得天下”。而我曾经论述过，机器要懂语言，背景知识不可或缺。因此，在这个意义下，将是“得知识者得天下”。如果说数据是石油，那么知识就好比是石油的萃取物。如果我们只满足卖数据盈利，那就好比是直接输出石油在盈利。但是石油的真正价值蕴含于其深加工的萃取物中。石油萃取的过程与知识加工的过程也极为相像。都有着复杂流程，都是大规模系统工程。知识工程的鼻祖，费根鲍姆曾经说过的一句话“knowledge is the power inAI”。

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

知识管理

知识图谱的管理主要图谱的存储、检索等问题。通常这些问题的解决需要数据库系统的支撑，因而系统的选型也是知识图谱管理的一个重要问题。这里主要讨论能用于知识图谱管理的数据库系统选型以及知识图谱查询语言。知识图谱存储是个较为专业化的问题，此处不再深入讨论。

知识图谱管理系统的选型。知识图谱本质上在表达关联，天然地可以用图加以建模，因而很多人想到用图数据库对领域知识图谱加以存储。图数据库的确是知识图谱存储选型的重要选择，但是不是唯一选择。传统关系数据库，近几年充分发展的其他类型的NoSQL数据库在很多场景下也是合理选择。那么数据库的选择考虑的要素是什么呢？有两类重要的选型要素：图谱的规模以及操作复杂度。

从图谱的规模角度来看，百万、千万的节点和关系规模（以及以下规模）的图谱对于图数据库的需求并不强烈，图数据库的必要性在中等或者小规模知识图谱上体现并不充分。但是如果图谱规模在数亿节点规模以上，图数据库就十分必要了。

从操作复杂性来看，图谱上的操作越是复杂，图数据库的必要性越是明显。图谱上的全局计算（比如平均最短路径的计算），图谱上的复杂遍历，图谱上的复杂子图查询等等都涉及图上的多步遍历。图上的多步遍历操作如果是在关系数据库上实现需要多个联结（Join）操作。多个联结操作的优化一直以来是关系数据库的难题。图数据库系统实现时针对多步遍历做了大量优化，能够实现高效图遍历操作。

除了上述因素之外，还应该充分考虑系统的易用性、普及性与成熟度。总体而言图数据库还是发展中的技术，对于复杂图数据管理系统的优化也是只有少部分专业人员才能从事的工作。在数据库选型时需要充分考虑这些因素。我们实验室在实现CN-DBpedia（2000万实体、2.2亿关系）在线服务系统时先后采用了RelationalDB、Graph DB、MongoDB，最后出于综合考虑选用的是MongoDB，已经稳定运行了三年，累计提供10亿多次API服务。

知识图谱查询语言。通常对于表达为RDF形式的知识图谱，可以使用SPARQL查询语言。SPARQL语言针对RDF数据定义了大量的算子，对于推理操作有着很好支撑，因而能够适应领域中的复杂查询与复杂推理。从应用角度来看，也可以将知识图谱仅仅表达为无类型的三元组。对于这种轻量级的表示，关系数据库与传统NoSQL数据库也是较好选择。那么此时，SQL语句就是比较好的选择。SQL十分成熟，语法简单，用户众多且有着几十年的成功应用基础。

很多领域图谱上的查询是相对简单的，以单步或者两到三步遍历居多。此时，SQL完全能够胜任。但是不排除有一些特定场景，特别是公共安全、风控管理等领域，通常需要进行复杂关联分析，需要较长路径的遍历，需要开展复杂子图挖掘，此时SQL的表达能力就显得相对较弱了。

四、知识图谱的发展现状及应用

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

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！