



万维网规模 搜索与推理的融合

Dieter Fensel • 奥地利因斯布鲁克大学

Frank van Harmelen • 阿姆斯特丹自由大学

最近我们获悉了一项要求在 100 毫秒内推理 100 亿个资源描述框架三元组合存储结构（主体，关系，对象的表述形式）的电信项目，其应用领域被定义在由新的内容敏感式描述以及个人移动电话服务产生的税务信息流。现有方法大致能以 100 毫秒内 1 亿个三元结构的速度处理资源描述框架模式的问题，但这个项目要求对大两个数量级的一系列三元结构进行复杂推理，所以需求显然会增加。实际上，级别要求的提升会随着时间的延长比任何在推理算法、智能编码领域取得的进步都要快，而改进的硬件可以对其进行弥补。

被迫推掉潜在客户使得我们思考这个问题存在的原因。问题通常由于不恰当地概念化而变得棘手——要求智能引入假设使问题变得可以解决。一方面没有限制它们，另一方面也确保了其有用性。所以问题就在于：为什么推理不能在网络规模中进行以及该如何来确定。

网络与推理之间的矛盾

大约于 1996 年，在向网页描述中添加语义要素的首个项目里，网络与推理开始相结合，二者的结合方式与在信息格式化中加入超文本链接标示语言的形式相类似。尽管在那之后，这个活跃而成型的研究领域出现了一些语义网成果，但对于推理是否真正向网络领域添加了某些实用要素的重大质疑仍旧存在。

研究者们开发了面向小型、封闭、可信、不变、静态领域的推理方法。他们通常规定出一小部分公理（包含有各种常识——一类特殊类型的公理）；一台不受干扰的仪器通常能推理出其中所含知识的完整而正确的结论。以自然

数为例：七条所谓的 Peano 公理（这是一种能够可计数地描述许多公理的概要）能够表示所有相关的知识属性。非常有趣的是，如同 Godel 那著名的不完备定理证明的一样，还没有开发出用于这个简单逻辑理论的完整而正确的推理方法。因此，在过去的 50 年中，我们目睹了为找到减少计算复杂逻辑行之有效的推理方法所作的重大努力（就是说逻辑不能对所有与自然数相关的知识进行编码）。比如说逻辑描述，它以一种能被发现和执行的可决定的进程来限制逻辑语言，以终止对一部分公理的推理演绎。另一个著名的例子是逻辑编程，它提取一阶逻辑的前端片段，并只对一个特殊模型（一种最小化模型）而不是全部进行推理。这两种方法都有自己的优点。逻辑描述推理机可以处理 10^5 数量级公理的推理（称为概念阐释），但该法在测算较大实例集合时表现较差。而逻辑编辑器能够处理相近大小的规则集合以及更大的实例集合（比如 10^6 数量级），然而这个方法仅能从这些理论当中得出简单的逻辑结论。

这两种流派都是研究中引人注目的领域，而且诸如何将二者结合起来的开放性论题得到了大量的关注。尽管如此，我们还是怀疑这是否是在网络级别上进行推理的现实途径，它意味着要对任意巨大数目的三元结构进行操作——多得不可计数，这里用了一种口语用法。然而 100 亿个三元结构是个什么概念呢？一个保守的估计认为，描述一个人要用大概一万个三元结构，这就有了一百兆个。错误匹配比有效的推理算法更能限制一阶逻辑的子集能解决什么。为了更加清楚，让我们回顾一下基本的逻辑假设概念：

- 小公理集合。描述自然数要求可计数的一定量的公理，而网络则绝不会需要得这么少。如果网络准备获取人类全部的知识，那么所需公理的数目将会非常巨大。
- 少量常识。假定谷歌计数了大概 300 亿个网页，而且保守估计每个网页上有 100 条常识，那么就可以说我们已然步入了充斥着几兆常识的生活空间。
- 结论规则的完整性。网络是开放的，没有规定任何界限。因此，完整性对于这个领域的推理过程来说是一个极其不可思议的要求。鉴于收集网络上所有相关信息既不可能，在很多时候也没有价值（通常你想要阅读的只是谷歌搜索的前 10 条信息，而并没有时间留给余下的两百万条），在已然如此不完整的常识

的对这些常识的获取和使用过程中得到改变。关于完整而正确推理的传统观念显然是基于将极端简化的世界观幼稚地应用于现实之中。一个著名的说法是，任何有关大型分布式系统状态的知识都是不完整或者过时的（这是错误的）。

讲得过分一点，当前的推理机从网络继承了成堆的句法结构（XML,RDF,以及 URI 等等），作为回报，网络接受了那些既不符合要求又不适合它规模的玩具仪器。基本上说，双方在一个极其肤浅的层次相结合，以至于无法运作出一些有用的东西。纯逻辑推理中最基础和根本的假设似乎没有和网络提供的现实情况相匹配。在科学一个不同领域里一个类似的不匹配现象也许为解决这一问题提供了途径。

关于完整而正确推理的传统观念显然是基于将极端简化的世界观幼稚地应用于现实之中。

上要求完整推理似乎毫无意义。

- 可信性，推理规则的正确性，以及一致性。传统逻辑将公理作为真实情况的反映，并试图推理出它所提供的隐含知识。这个过程的正确性确保了被公理记录下来的真实情况得以保存。在网络领域，信息从一开始就是不可靠的，这意味着即使是一台正确的推理机也无法确保其真实性；更糟的是，因为网络为“知识”提供的空间里充斥着各式各样的观点，所以任何不受干扰的仪器都会非常容易地推理出矛盾。
- 静态领域。网络是个活跃的实体：已知的常识将会在为了结论而进行

将有限理性进行推理是十分理智的

经典经济学理论所认为的完全理性的经纪人——是指那些将其决策建立在完整市场信息之上，并从中推断出最优选择的人。一方面，这会导致具有特定数学特性的值得关注的方程系统；但另一方面，它将会塑造出一个在认真考虑结婚之前，马不停蹄地进行 40 亿次约会的新郎模型。该理论由于其在全局优选方面的刚性，虽然略有数学价值，但它在预测和模拟现实情况方面的能力十分有限。收集信息并用上述理论进行推理实际上是一个被有限资源限制的过程。

Herbert Simon 于 1957 年提出了有限理性的概念来更好地模拟这个过程。按照这个观点，经纪人依据不完整的知识来作出决定，并且可能缺失一些能够得出所有潜在结论的资源（当然，如果前者极度不完整，那后者无论如何也不会有多大意义）。这种方法在启发式问题解决上开创了一个新的研究领域。

我们的目标是通过将网络级别的推理与搜索结合起来以实现一个与上面类似的范式转换。我们提出的，目前自称为“研究”的范式，没有那些赚人眼球的完整性及正确性的概念，取而代之的是在解决实际问题方面有一个更加具体的适用性概念。该观点同样基于当前的推理概念，当前的推理概念还不明确如何恰当地反映出获取并处理信息所需要的成果与资源，而实在是仅仅模拟了一些不合理的行为。

混淆推理与搜索

正如前面所述，一百兆及更多个三元结构理所当然地需要基于信息优先选择的不完整也不精确的推理，同时也需要统计学与逻辑的结合。基本原理是从任意数目的三元结构中抽取一个随机样本并对其进行推理。这种方法适用于任何规模。一个稍稍更加智能的方法将通过预处理来改进对于三元结构的选择，以找到对推理来说更加重要的一些资源。我们可以用如下方法描述这个研究算法：

do

抽取样本

对样本进行推理

if 您有更多时间

and/or 如果您对结果不信任
then 抽取一个更大的样本
repeat

接下来我们可以尝试基于如下条件的样本智能选择

- 已知三元结构的分布式资源
- 它们与问题之间的关系
- 诸如名誉或信任的来源属性，以及
- 以前问题的经验

这种算法并非仅仅在规模上优于经典算法。它还可以通过简单地更换属性的方式以适用于任意数量级（比如，对一个更合适的稍小一些的样本进行推理）。我们在这个领域的一些早期工作中，使用标准化的“谷歌距离”（这是一项用于测量在零至无穷大规模上任意两个单词之间距离的功能）。这是一项对于从巨大（以及全局不一致）知识库中抽取样本的有启发式的研究。

逻辑推理目前无法得知常识与公理源自何处。我们假定它们是真实的，并且对其进行保真性推理，但这既不合适，也不适用于网络领域。沟通分歧的唯一途径是将推理过程与通过补偿（排序或选择）和抽象（压缩信息）建立相关常识与公理的过程统一起来。那样的话，补偿与推理就成为了同一枚硬币的两面——一个目标是从网络上的数据中获取有用信息的过程。

相关工作的范围包括某些领域的的数据压缩，比如计算机图形，机器学习，以及数据存储，还有时效逼近推理。用这样的算法工作将必然提供

如下问题的答案（在其他的之中）：

- 什么是继承、一致性等等可能的概念？
- 这种推断令人满意的属性是什么？这可能包括：可重复性（如果你搜索两次，是否得到相同的答案？），单调性（如果你选取一个大一点的样本，是否得到了更好的结果？），以及时效可用性（计算时间与答案质量之间存在着什么样的互换关系？）。
- 何种问题语言及答案是必须的？
- 三元结构能否在以前推理任务的影响下自我组织，以选择相关的三元结构是将来的任务变得容易？

围绕这些要点的工作，必将会使逻辑与网络实实在在地融合。

不同于向语言中加入古怪的句法或者网络准则与推理肤浅结合出的不可改良的逻辑，我们尝试仔细思考最基本的标准。抛除那些不适合的东西，向剩余内容中加入新元素并将二者融合，以呈现出彻底的统一。Tim Berners-Lee 和他的同事或许在讨论推理的可选方法时也有同样的目标。

这种研究流派非常适合欧盟新的研究计划项目VII，它要求项目必须与“语义基础相关：有可行性，暂时性并有模拟形态且要有逼近推理能力。通过目标驱动研究来超越目前的形式主义。理论结果将与强大而日趋完善的执行参考相吻合。”

（<http://cordis.europa.eu/fp7/ict/>）。这个研究是需要的，而且它将会被投资。

致谢

我们对 Michael Kiefer, Atanas Kiryakov 和 Charles Petrie 非常有帮助的探讨表示感谢，并期待向他们详细阐述我们的初步设想。

参考资料

1. *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, D. Fensel et al., eds., MIT Press, 2003.
2. H. Simon, *Models of Man*, Wiley, 1957
3. R. Cilibrasi and P. Vitányi, “Automatic Meaning Discovery Using Google,” tech. report, 2004; www.arxiv.org/abs/cs.CL/0412098
4. S. Russell and E.H. Wefald, *Do the Right Thing: Studies in Limited Rationality*, MIT Press, 1991.
5. T. Berners-Lee at al., “A Framework for Web Science,” *Foundations and Trends in web science*, vol. 1, no.1, 2006, pp. 1-130.

Dieter Fensel 是奥地利因斯布鲁克大学的教授。他的研究方向包括数据与进程的语义通路层。他在因斯布鲁克大学取得人工智能博士学位。是 *Enabling Semantic Web Services* 的合作研究者：The Web Service Modeling Ontology, (Springer, 2006)。并且是 *Ontologies* 的作者：Silver Bullet for Knowledge Management and Electronic Commerce (Springer-Verlag, 2001, 2003)。联系方式 dieter.fensel@deri.org

Frank van Harmelen 是阿姆斯特丹自由大学的教授。他的研究防线包括网络制式表述与逼近推理技术。他在爱丁堡大学取得人工智能博士学位。他是语义网第一本教科书 *The Semantic Web Primer* (MIT Press) 的合著者。联系方式 Frank.van.Harmelen@cs.vu.nl; www.cs.vu.nl/~frankh.