知识图谱识别

**Jay Pujara**

**University of Maryland - College Park**

**Hui Miao**

**University of Maryland - College Park**

**Lise Getoor**

**University of Maryland - College Park**

**William W. Cohen Carnegie Mellon University,** [**wcohen@cs.cmu.edu**](mailto:wcohen@cs.cmu.edu)

知识图谱识别

Jay Pujara1, Hui Miao1, Lise Getoor1, and William Cohen2

1 Dept of Computer Science, University of Maryland, College Park, MD 20742

{jay,hui,getoor}@cs.umd.edu

2 Machine Learning Dept, Carnegie Mellon University, Pittsburgh, PA 15213

[wcohen@cs.cmu.edu](mailto:wcohen@cs.cmu.edu)

**概要。**大规模信息处理系统能够提取大量相互关联的事实，但不幸的是，将这些候选事实转化为有用的知识是一项艰巨的挑战。在本文中，我们将展示如何将提取出来的那些不确定的实体及其关系转化为知识图谱。提取形成一个提取图，我们参考去除噪声的工作，推断缺失的信息以及确定哪些候选事实应当包含在知识图谱中作为知识图谱识别。为了完成这项任务，我们必须共同推理候选事实及其相关的提取置信度，确定相关实体**并纳入本体约束**。我们提出的方法使用概率软逻辑（PSL），一种最近引入的概率模型框架，它可以轻松扩展到数百万个事实。我们演示了我们的方法在合成关联数据语料库上的强大功能。这些语料库源自MusicBrainz音乐社区和现实世界的一套来自NELL项目的提取，总共超过1M个提取和70K本体论关系。我们证明：与现有方法相比，我们的方法能够显着改善AUC和F1，并且运行时间更短。

**1 介绍**

网络是一个巨大的知识仓库，但是自动提取其中的大规模的知识已被证明是一项艰巨的挑战。最近的评估努力的重点是自动**知识库人群**(knowledge base population), 还有很多知名的广泛领域和开放信息提取系统，包括Never-Ending Language Learning（NELL）项目OpenIE [4]，以及谷歌的一些项目，它们使用各种技术从网络上以事实的形式来提取新知识。这些事实是相互关联的，因此，这种提取的知识被称为知识图谱。

生成知识图谱的一个关键挑战是从不同的来源以一致的方式引入噪声信息。信息提取系统可以在许多源文档（如网页）上运行，并使用一组策略从文档，句法，词汇和结构特征中生成候选事实。最终，这些提取系统产生候选事实，包括一组实体，这些实体的属性，以及实体之间的关系。我们称之为提取图。然而，提取过程中的错误在提取图引入了不一致，可能包含重复的实体并违反关键本体约束如包容，互斥，逆，域和范围限制。这样的噪音模糊了真实的知识图谱。真实的知识图谱是一个一致的集合，集合由实体，属性，以及其关系组成。

我们的工作是通过由信息提取系统生成的提取图来推断知识图谱。我们演示了信息提取系统遇到的错误需要对候选事实进行联合推理事实来构建一致的知识图谱。我们的方法进行实体决定，集体分类和链接预测，同时在知识图谱上执行全局约束，这个过程我们称之为知识图谱识别。

为了实现知识图谱的识别，我们使用概率软逻辑（PSL）[7]，最近推出的在连续值随机变量上概率推理框架。PSL提供了许多优点：使用带有一阶逻辑的声明性规则很容易定义模型语法，连续值变量提供了不确定性的方便表示，加权规则和权重学习捕捉模型的重要性

规则和高级功能（如基于集合的聚合和严格约束）被支持。另外，PSL中的推论是一个凸优化高度可扩展性使我们能够在几分钟内处理数百万事实。

我们为知识图谱识别开发了一个PSL模型，它既可以捕获事实之间的概率依赖关系又在实体和其关系之间执行全局约束。我们定义了解释的概率分布 - 或事实的真值分配 - 每一个都与可能的知识图谱相对应。通过使用提取图和本体进行推理，我们能够找到最可能的知识图谱。我们在两个大型数据集上建立了我们方法的优势：从MusicBrainz社区派生的合成数据集以及Music Ontology中定义的本体关系，以及大规模操作知识提取系统NELL的嘈杂提取。

我们在这项工作中的贡献是：1）制定知识图谱识别问题，支持在存在本体约束条件下推理多个不确定的提取源; 2）利用PSL有效地解决知识图形识别的凸优化问题; 以及3）通过在基准数据集上展示优于最先进方法的结果来展示知识图谱识别的能力，并在竞争系统不能得出结果的时间内生成大量知识图谱。

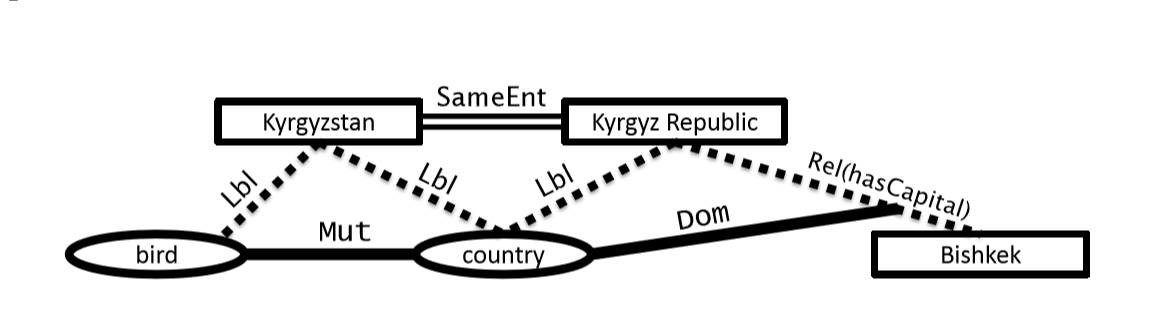
**2相关工作**

早期Cohen等人[8]考虑研究从一系列噪声事实中共同确定最佳潜在KB，但他们只考虑了KB错误的一小部分。江等人[9]通过使用本体关联候选提取并利用马尔可夫逻辑网络（MLN）[10]探索许多不同的建模选择，在更广的范围内执行知识库精简。。江等人。提供了对知识库中存在的本体论约束和候选事实的清晰解释，作为一阶逻辑中的规则，为我们在建模中采用的知识库提供了有吸引力的抽象。但是，选择MLN作为建模框架会带来一些限制。在MLN中，所有逻辑谓词都必须采用布尔真值，这使得融合置信度值很难。此外，布尔赋值随机变量的组合爆炸使MLN中的推理和学习难以解决优化问题。江等人。用一些近似值来克服这些障碍，并且与单独考虑每个事实的基线相比，证明联合推理的效用。通过使用PSL，我们可以避免这些代表性和可扩展性的限制，并且我们建立和改进了Jiang等人的模型。通过在我们的模型中包含多个提取器并推导出共同提及的实体。

**3动机：知识图识别**

在这项工作中，我们将来自信息提取系统的候选事实表示为知识图，其中实体是节点，类别是与每个节点相关联的标签，而关系是节点之间的有向边。信息提取系统可以提取这些候选事实，并且这些提取可以用来构建提取图。不幸的是，提取图通常是不正确的，存在诸如虚假和缺少节点和边的错误以及缺少或不准确的节点标签。我们的方法，知识图谱识别（KGI）结合了基于本体信息的规则介导的实体解析，集体分类和链接预测的任务。我们通过从现实世界的信息抽取系统 - 永无止境的语言学习者（NELL）[3]中挑战的例子来说明我们的方法的必要性。

实体提取是一个常见问题：许多初始看起来不一样的文本引用可能指向同一个真实世界的实体。例如，NELL的知识库包含 “kyrghyzstan”, “kyrgzstan”, “kyrgystan”, “kyrgyz republic”, “kyrgyzstan”, 和 “kyrgistan”的候选事实。它们都是涉及吉尔吉斯斯坦所有变体或拼写错误的实体。在提取的知识图中，这些不正确地对应于不同的节点。我们的方法使用实体决策来确定知识图谱中的相关对象实体，为每个解析节点生成一组一致的标签和关系。



图。1。 举例说明知识图形标识如何解析抽取图中的冲突信息。 实体用矩形表示，虚线表示不确定的信息，实线表示本体约束，双线表示用实体解析找到的同位对象实体。

知识图谱构造的另一个挑战是一致地推断标签。例如，NELL的提取将吉尔吉斯斯坦的标签分为“国家”和“鸟”。本体信息表明，一个实体不可能同时成为一个国家和一个鸟。使用知识图谱中相关实体的标签可以让我们确定实体的正确标签。我们的方法使用集体分类，用考虑到本体信息和相邻标签的方式来标记节点。

知识图谱中经常遇到的第三个问题是确定实体之间的关系。NELL也有很多关于吉尔吉斯斯坦的位置与其他实体有关的事实。 这些候选人关系包括吉尔吉斯斯坦位于哈萨克斯坦，吉尔吉斯斯坦位于俄罗斯，吉尔吉斯斯坦位于前苏联，吉尔吉斯斯坦位于亚洲，吉尔吉斯斯坦位于美国。 其中一些可能的关系是真实的，而另一些则显然是错误和矛盾的。 我们的方法使用链接预测以考虑到本体信息和推断结构的其余部分的方式预测边。

当我们考虑预测之间的相互作用并考虑到我们在提取中的可信度时，提取提取图变得更具挑战性。图1说明了这样一个复杂的例子。如前所述，NELL的本体包含了“鸟”和“国家”这两个互斥标签。。合理推理使我们能够解决这两个标签中哪一个更可能适用于Krygyzstan。例如，NELL高度肯定吉尔吉斯共和国拥有首都比什凯克。NELL本体规定，“hasCapital”关系的领域标有“国家”。 实体解决方案使我们可以推断“吉尔吉斯共和国”是指与“吉尔吉斯斯坦”相同的实体。现在决定吉尔吉斯斯坦现在是鸟还是国家涉及一个预测，其中包含来自共同指涉实体的相应“鸟类”和“国家”事实的置信度值，以及这些共同指称实体的本体论关系中的集体特征，例如“hasCapital “关系。我们将这个从嘈杂的抽取图中推断出知识图谱的过程称为知识图谱识别。与之前关于图形标识和知识库精简的工作不同，我们使用了一个非常不同的概率框架PSL，它允许我们在合并提取器置信度值并支持丰富的本体约束集合的情况下联合推断知识图谱。

4背景：概率软逻辑

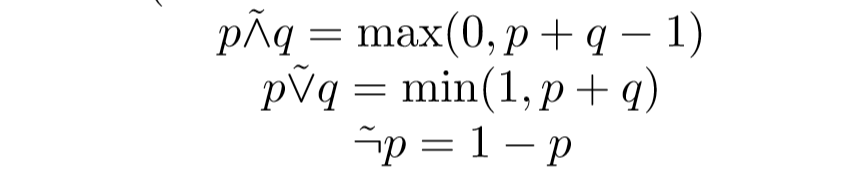
概率软逻辑（PSL）[7,14]是最近引入的框架，它允许用户指定连续值随机变量的丰富概率模型。 像其他统计关系学习语言（如马尔科夫逻辑网络（MLN））一样，它使用一阶逻辑来描述定义马尔可夫网络的特征。 与其他方法相比，PSL采用连续定值的随机变量而不是二元变量，并将最可能的解释（MPE）推断作为一个凸优化问题，比其组合对数（多项式对指数）要显着更有效。

PSL模型由一组加权的一阶逻辑规则组成，其中每个规则定义一组共享相同权重的马尔可夫网络的特征。 考虑公式



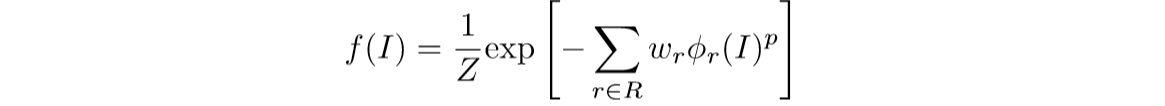
这是PSL规则的一个例子。 这里w是规则的权重，A，B和C是普遍量化的变量，P，Q和R是谓词。 规则的基础来自于将规则的原子中的常数量化变量替换为常数。 在这个例子中，将常数值a，b和c分配给上述规则中的各个变量将产生基本原子P（a，b），Q（b，c），R（a，b，c）。 每个基本原子在[0,1]的范围内取一个软真值。

PSL将数值距离与满足每个基本规则的数值距离相关联，以确定马尔科夫网络中相应特征的值。满足的距离是通过将基本规则视为规则中的基本原子的公式来定义的。特别是，PSL使用Lukasiewicz t-范数和共范数来放松逻辑连接词AND（∧），OR（∨）和NOT（¬），如下所示（松弛用连接词上的符号~表示）：



当p和q处于{0,1}时，这种松弛符合布尔逻辑，并且当p和q在数值范围[0,1]中时，提供软真值的一致解释。

一个由上面定义的模型组成的PSL程序Π与一组事实F一起产生一组基本规则R.如果我是一个解释（将软真值分配给地原子）和 r是一个规则的基础实例，那么r的满足距离φr（I）为1-Tr（I），其中Tr（I）是来自Lukasiewicz t-范数的软真值。 我们可以通过结合所有基本规则的加权满意度R和归一化来定义解释的概率分布，如下所示：



这里Z是标准化常数，wr是规则r的权重，并且{1,2}中的p允许规则的线性或二次组合。 因此，一个PSL程序（加权规则和事实集合）根据表示随机变量之间关系的逻辑公式定义概率分布。

PSL中的MPE推断使用已知基本原子的值以及由规则编码的原子之间的依赖性确定未知基本原子的最可能的软真值，其对应于潜在马尔可夫网络中的随机变量的推断。 PSL原子在区间[0,1]中采用软真值，与原子采用布尔值的MLN相反。 MLN中的MPE推理需要优化布尔真值的组合赋值。 相反，连续域的松弛大大改变了PSL中计算的易处理性：给定一组加权规则，发现最可能的解释等同于求解凸优化问题。 最近[15]的工作引入了适用于PSL模型的共识优化方法; 他们的结果表明共识优化与模型中的基本规则的数量呈线性关系。

**5利用PSL进行知识图谱识别**

知识图谱包含三类事实：关于实体的事实，关于实体标签的事实和关于关系的事实。 我们使用逻辑谓词Ent（E）来表示实体，并用逻辑谓词Lbl（E，L）来表示实体E具有标号L。关系用逻辑谓词Rel（E1，E2，R）表示，其中关系R成立，在实体E1和E2之间，例如，R（E1，E2）。

在知识图谱识别中，我们的目标是从一组嘈杂的提取中确定一组真实的原子信息。 我们的知识图谱识别方法包含三个部分：捕获不确定的提取，执行实体解析以及执行本体约束。 我们展示了如何创建包含这三个组件的PSL程序，然后将此PSL程序与可能的知识图谱上的分布相关联。