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

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

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。

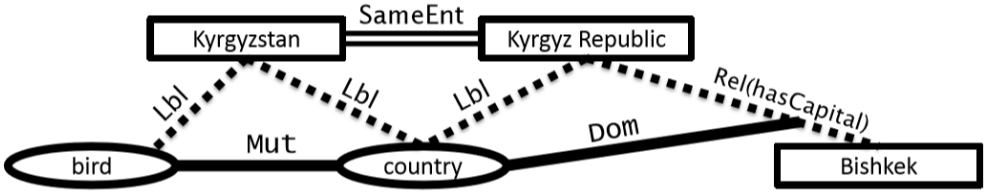
2 相关工作

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

$$\mathrm{P}(A,B)\tilde{\wedge}\mathrm{Q}(B,C) \overset{w}{\Rightarrow} \mathrm{R}(A,B,C)$$

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）：

$$p\tilde{\wedge}q = \max(0, p+q-1)$$
$$p\tilde{\vee}q = \min(1, p+q)$$
$$\tilde{\neg}p = 1-p$$

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：

$$f(I) = \frac{1}{Z}\exp\left[-\sum_{r\in R} w_r \phi_r(I)^p\right]$$

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。

5.1 表示不确定的提取

我们通过引入候选谓词，使用类似于[9]的公式，将信息提取系统中的嘈杂提取与上述逻辑谓词联系起来。对于每个候选实体，我们引入一个相应的谓词 CandEnt（E）。信息提取系统生成的标签或关系对应于我们系统中的谓词 CandLbl（E，L）或 CandRel（E1，E2，R）。给这些谓词分配一个等于来自提取器的置信度值的软真值来捕获这些提取中的不确定性。例如，提取系统可能会产生一个关系，hasCapital（吉尔吉斯斯坦，比什凯克），其信度为.9，我们将表示为 CandRel（吉尔吉斯斯坦，比什凯克，hasCapital），并赋予它一个.9 的真值。

信息提取系统通常使用许多不同的提取技术来生成候选者。例如，NELL 从词法，结构和形态模式等方面产生单独的提取。我们通过使用 CandRelT 和 CandLblT 形式的每种技术 T 的单独谓词来表示关于用于提取候选者的技术的元数据。这些谓词与我们试图使用加权规则推断的属性和关系的真实值有关。

$$\textsc{CandRel}_T(E_1, E_2, R) \quad \overset{w_{CR}-T}{\Rightarrow} \quad \textsc{Rel}(E_1, E_2, R)$$
$$\textsc{CandLbl}_T(E, L) \quad \overset{w_{CL}-T}{\Rightarrow} \quad \textsc{Lbl}(E, L)$$

我们通过使用提取系统的输出将上述规则基础化而产生候选集合，作为集合 C.

5.2 实体决策

在我们的 PSL 程序中，我们还利用与本体相对应的规则，其基础标记为 O. 我们的本体规则基于[9]中提出的逻辑公式。每种类型的本体关系都被表示为一个谓词，而这些谓词表示了关于标签和关系之间关联的本体知识。例如，本体论谓词 Dom（hasCapital，country）和 Rng（hasCapital，city）指定 hasCapital 关系是从具有标签国家的实体到具有标签城市的实体的映射。谓词 Mut（国家，城市）指定标签国家和城市是相互排斥的，因此实体不能同时拥有国家和城市的标签。我们同样使用谓词来包含标签（Sub）和关系（RSub），以及与之相关的函数（Inv）。为了使用这种本体论知识，我们引入了将每个本体论谓词与代表知识图的谓词相关联的规则。我们在使用加权规则的实验中指定了七种类型的本体约束：

$$\begin{aligned}
&\textsc{Dom}(R, L) && \tilde{\wedge} \ \textsc{Rel}(E_1, E_2, R) && \overset{w_O}{\Rightarrow} \ \textsc{Lbl}(E_1, L)\\
&\textsc{Rng}(R, L) && \tilde{\wedge} \ \textsc{Rel}(E_1, E_2, R) && \overset{w_O}{\Rightarrow} \ \textsc{Lbl}(E_2, L)\\
&\textsc{Inv}(R, S) && \tilde{\wedge} \ \textsc{Rel}(E_1, E_2, R) && \overset{w_O}{\Rightarrow} \ \textsc{Rel}(E_2, E_1, S)\\
&\textsc{Sub}(L, P) && \tilde{\wedge} \ \textsc{Lbl}(E, L) && \overset{w_O}{\Rightarrow} \ \textsc{Lbl}(E, P)\\
&\textsc{RSub}(R, S) && \tilde{\wedge} \ \textsc{Rel}(E_1, E_2, R) && \overset{w_O}{\Rightarrow} \ \textsc{Rel}(E_1, E_2, S)\\
&\textsc{Mut}(L_1, L_2) && \tilde{\wedge} \ \textsc{Lbl}(E, L_1) && \overset{w_O}{\Rightarrow} \ \tilde{\neg}\textsc{Lbl}(E, L_2)\\
&\textsc{RMut}(R, S) && \tilde{\wedge} \ \textsc{Rel}(E_1, E_2, R) && \overset{w_O}{\Rightarrow} \ \tilde{\neg}\textsc{Rel}(E_1, E_2, S)
\end{aligned}$$

5.4 不确定知识图上的概率分布

将本节介绍的逻辑规则与原子（如来自信息提取系统的候选人（如 CandRel（吉尔吉斯斯坦，比什凯克，hasCapital））的候选人，来自实体解决系统（如 SameEnt（吉尔吉斯斯坦，吉尔吉斯共和国））的共同参考信息和本体信息（如 Dom（hasCapital，country））结合起来，我们可以定义一个 PSL 程序 Π。这个程序的输入实例化了一组基本规则 R，它由来自不确定候选者 C，共同指涉实体 S 和本体论关系 O 的基础联合组成。由 PSL 生成的解释分布 I 对应于知识图上的概率分布 G：

$$P_{\Pi}(G) = f(I) = \frac{1}{Z}\exp\left[\sum_{r \in R} w_r \phi_r(I)^p\right]$$

推理的结果为我们提供了最可能的解释，或者对构成知识图谱的实体，标签和关系的软实质分配。通过在解释中选择软真值的阈值，我们可以选择一组高精度的事实来构建知识图谱。

6 实验评估

6.1 数据集和实验设置

我们在两个不同的数据集上评估我们的方法：一个来源于 LinkedBrainz 项目[16]的综合知识库，它使用 MusicOntology 的本体信息[17]映射来自 MusicBrainz 社区的数据，以及来自 Never-Ending 的 Web 抽取数据语言学习（NELL）项目[3]。我们的目标是评估知识图谱识别的效用，将其表述为 PSL 模型，从噪音数据推断知识图谱。另外，我们对比了两种不同的评估设置。首先，正如前面的工作[9]所使用的，推理仅限于从测试或查询集生成的知识图

谱的一个子集。在第二种评估设置中，推论产生完整的知识图，它不受测试集限制，但对原子采用软真值阈值。我们提供文档，代码和数据集以在 GitHub3 上复制我们的结果。

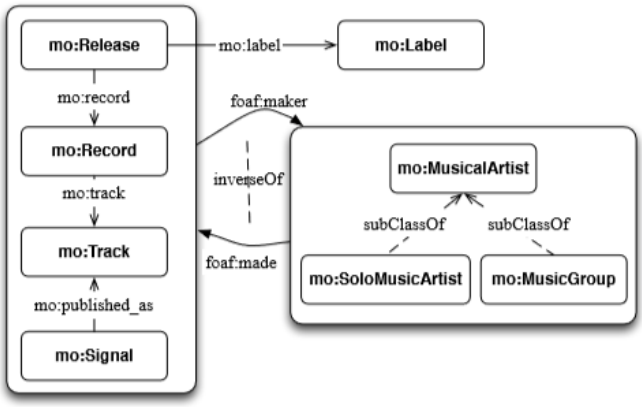


图 2。 Music Ontology 的子集使用 LinkedBrainz 映射到我们的综合数据集中的 MusicBrainz 数据

MusicBrainz MusicBrainz 是一个社区驱动的开放源代码的音乐元数据结构化数据库，其中包括有关艺术家，专辑和曲目的信息。Music Ontology 建立在众多知名本体之上，如 FRBR [18]和 FOAF [19]，并广泛使用，例如 BBC Music Linked Data 站点[20]。但是，可从 MusicBrainz 获得的关系数据以专有模式表示，并不直接映射到音乐本体。为弥合这一差距，LinkedBrainz 项目使用 D2RQ 发布了免费提供的 MusicBrainz 数据和音乐本体之间的 RDF 映射[21]。图 2 显示了我们在数据中使用的标签和关系的总结。我们使用本体关系直接映射到 PSL 谓词，使用来自 Music Ontology 使用的 FRBR 和 FOAF 类的本体信息。具体而言，我们将 rdfs：domain 转换为 Dom，将 rdfs：range 转换为 Rng，将 rdfs：subClassOf 转换为 Sub，将 rdfs：subPropertyOf 转换为 RSub，将 owl：inverseOf 转换为 Inv，并将 owl：disjointWith 转换为 Mut。

我们的综合知识图使用 MusicBrainz 项目 4 的 LinkedBrainz 映射中的数据样本，并添加噪声以生成一个真实的数据集。为了生成 LinkedBrainz 数据的一个子集，我们使用 MusicBrainz 数据集中一组曲目的雪球采样生成一组录音，发行版，艺术家和标签。接下来，我们通过随机删除已知事实并添加不一致的事实以及为这些事实生成随机置信度值，将噪声引入此图中。这种噪音可以被解释为由 MusicBrainz 用户错误地输入艺术家姓名引起的错误，意外切换输入字段或在生成知识库时遗漏信息。

我们通过扭曲真实输入数据的百分比来对这些错误进行建模。对于标签，我们省略了已知标签并为输入数据中 25％的事实引入了伪造标签。在处理关系时，我们关注的是 foaf：maker 和 foaf：艺术家和创意作品之间的关系。我们在 25％的时间内随机删除这两个关系中的一个。最后，我们有 25％的时间取消了作品与艺术家之间的关系，并在作品与生成的艺术家之间插入了新的关系，为这两位艺术家添加了 SameEnt。输入中发现的事实的置信度值是从正态（.7,2）分布产生的，而不一致的事实具有由正态（.3,2）分布产生的较低置信度值。这些分布的高度变化确保了重要的重叠。对于 SameEnt，相似度值由 Normal（.9，.1）分布生成。在所有情况下，分布都被限制在[0,1]范围内。

我们总结了表 1 中的重要数据统计数据。在我们的实验中，我们将知识图的嘈杂关系和标签用 PSL 谓词 CandLbl 和 CandRel 表示为候选事实。在评估过程中，我们使用 PSL 程序进行知识图谱识别，以推断最可能的知识图谱。

在这种情况下，我们对所有规则使用静态权重的二次组合，其中 wCL = wCR = 1，wEL

= wER = 25，wO = 100。我们通过比较用于生成数据的真实知识图来评估我们的结果，包括与我们介绍的虚假数据相对应的假标签。

NELL NELL 的目标是反复生成知识库。在每次迭代中，NELL 使用从前一次迭代中学到的事实和一组网页来生成一组新的候选事实。NELL 有选择地促进那些对提取器有高度置信度的候选者，并且服从现有知识库的本体约束来构建高精度的知识库。我们在 NELL 的第 165 次迭代中展示实验结果，使用候选事实，推广 NELL 在迭代过程中使用的事实和本体关系。我们在表 1 中总结了这个数据集的重要统计数据。由于网络的多样性，来自 NELL 的数据更大，包含更多类型的关系和类别，并且比我们的综合数据具有更多的本体论关系。

NELL 使用不同的提取来源，在我们的实验中，我们使用不同的谓词 CandLblT 和 CandRelT 作为来源 CBL，CMC，CPL，Morph 和 SEAL，而剩余的来源不贡献显着数量的事实，用 CandLbl 和 CandRel 谓词。除了候选事实之外，NELL 还使用启发式公式在系统的每次迭代中将候选人“推介”为知识库，但这些提升往往是嘈杂的，因此系统会为每个提升指定一个置信度值。我们用前面的迭代来代表这些被推荐的候选者，作为具有相应候选谓词的额外源。

除了来自 NELL 的数据外，我们还使用 YAGO 数据库中的数据[22]作为实体解决方案的一部分。我们的模型使用 SameEnt 谓词来捕获两个实体的相似性。为了纠正数据中发现的众多变体拼写，我们使用 NELL 实体到维基百科文章的映射技术。然后，我们使用相似性作为 SameEnt 谓词的软真值来定义文章 URL 中的相似度函数。

YAGO 数据库包含与维基百科文章，这些实体的变体拼写和缩写以及相关的 WordNet 类别相对应的实体。我们的实体解析方法将 NELL 中的实体名称与 YAGO 实体进行匹配。我们在 NELL 实体上执行选择性词干，在候选标签上使用阻塞，并使用不区分大小写的字符串匹配来找到相应的 YAGO 实体。一旦我们找到一组匹配的 YAGO 实体，我们就可以生成一组映射到相应 NELL 实体的 Wikipedia URL。我们可以通过计算与实体关联的维基百科 URL 上的集合相似性度量来判断两个实体的相似度。对于我们的相似性分数，我们使用 Jaccard 指数，设置交集的大小与集合的大小之比。

在我们使用 NELL 的实验中，我们考虑了两种情况。第一种类似于[9]中的实验设置，其中使用训练数据学习规则权重，并且在测试集的有限邻域上进行预测。。先前工作中使用的邻域试图通过产生测试集的接地来提高可扩展性，并且仅包括在该接地中不平凡的原子。实际上，这会产生一个通过省略可能与测试集中的原子相矛盾的原子而被扭曲的邻域。例如，如果 Sub（国家，地点）和 Mut（国家，城市）等本体论关系存在，测试集原子 Lbl（吉尔吉斯斯坦，国家）将不会引入 Lbl（吉尔吉斯斯坦，城市）或 Lbl（国家，地区）即使相互矛盾的数据出现在输入的候选人中，也是如此。通过消除推理矛盾信息的能力，我们认为这种评估方式减少了问题的真实难度。我们验证了我们在这个设置上的方法，但也提供了一个更现实的设置结果。在第二种情况下，我们执行独立于测试集的推理，懒惰地为输入数据支持的原子生成真值，使用软真值阈值.01。第二个设置允许我们推断一个类似于 MusicBrainz 设置的完整知识图谱。

6.2 MusicBrainz 的知识图鉴定结果

我们对 MusicBrainz 数据的实验尝试恢复完整的知识图，尽管增加了噪声，这会导致事实的不确定性，消除真实信息并添加虚假标签和关系。我们评估了许多变体恢复这种知识图谱的能力。我们使用许多度量标准衡量绩效 :精确度 - 回忆曲线下的面积（AUC），精确度，召回率以及 0.5 的软真度阈值下的 F1 分数，以及最大的 F1 分数数据集。由于置信度值的高度变化以及大量的真的事实，在所有变体中，最大 F1 值出现在软真值阈值 0 处，此时召回被最大化。这些结果总结在表 2 中。

表 1.NELL 和 MusicBrainz 的数据集统计摘要，包括（a）输入数据中候选事实的数量，

不同关系和标签的存在，以及（b）这些关系和标签之间定义的本体关系的数量。

(a)

	NELL	MusicBrainz
Cand. Label	1.2M	320K
Cand. Rel	100K	490K
Promotions	440K	0
Unique Labels	235	19
Unique Rels	221	8

(b)

	NELL	MusicBrainz
Dom	418	8
Rng	418	8
Inv	418	2
Mut	17.4K	8
RMut	48.5K	0
Sub	288	21
RSub	461	2

表 2. MusicOntology 数据的知识图谱识别方法比较表明，知识图谱识别有效地将图谱识别和推理的优势与本体信息相结合，并产生出色的结果。

Method	AUC	Prec	Recall	F1	Max F1
Baseline	0.672	0.946	0.477	0.634	0.788
PSL-EROnly	0.797	0.953	0.558	0.703	0.831
PSL-OntOnly	0.753	0.964	0.605	0.743	0.832
PSL-KGI-Complete	**0.901**	**0.970**	**0.714**	**0.823**	**0.919**

我们考虑的第一个变体只使用输入数据，将软真值设置为生成的置信度值，作为数据中潜在噪声的指示。基线结果只使用我们在 5.1 小节介绍的候选规则。我们通过添加 5.2 节中介绍的实体解析规则（我们将其作为 PSL-EROnly 报告）或利用加权规则来捕获本小节 5.3 中介绍的本体约束来改进这些数据。最后，我们将第 5 节中介绍的所有知识图谱标识元素进行组合，并将这些结果报告为 PSL-KGI-Complete。基线结果显示了输入数据中的噪声幅度；知识图谱中不到一半的事实可以被正确推断。在图谱识别中进行共同参照实体的联合推理可以改善结果。使用本体论约束，因为以前在改善这个领域提取方面的工作，也提高了结果。比较这两项改进，增加实体决策具有更高的 AUC，而本体论约束显示 F1 分数更大的改进。然而，当这两种方法相结合时，就像知识图谱的识别一样，结果显著提高。与更具竞争力的基准方法相比，知识图谱识别增加了 AUC，精密度，召回率和 F1 基本上优于其他变体，将 AUC 和 F1 提高 10％以上。。总的来说，我们能够推断 71.4％的真实关系，同时保持 0.97 的精度。此外，.901 的高 AUC 值表明知识图谱识别平衡了广泛参数值的精度和召回率。

6.3 NELL 的知识图鉴定结果

与以前工作的比较虽然综合噪声数据的结果证实了我们的假设，但我们对大型嘈杂的真实世界数据集的结果特别感兴趣。我们将我们的方法与 NELL 第 165 次迭代的数据进行比较，使用之前报告的手动标记评估集[9]的结果。这些结果的总结如表 3 所示。我们比较的第一种方法是与 MusicBrainz 结果中使用的基线类似的基线，其中候选者被赋予等于提取器置信度的软真值（在合适的情况下跨越提取器进行平均）。结果为 0.45 的软真值阈值，其使 F1 最大化。

我们还比较了 NELL 项目使用的默认策略，以选择候选事实以包含在知识库中。他们的方法使用本体来检查每个提议的候选人与已经在知识库中的以前推动的事实的一致性。不违背先前知识的候选人使用基于提出事实的提取器的置信度分数的启发式规则进行排序，并且根据评分和排名阈值选择最高候选者进行提升。请注意，NELL 方法包含所有输入事实的判断，而不仅仅是测试集里的那些判断。

我们比较的第三种方法是来自文献[9]的表现最好的 MLN 模型，它表达了本体论约束，并通过与我们模型中的逻辑规则类似的逻辑规则提出候选和推动事实。MLN 使用额外的谓词，这些谓词含有从使用手动标记的数据训练的逻辑回归分类器中提取的置信度值。MLN 使

用硬本体约束条件，独立地考虑规则的规则权重并且使用逻辑回归，通过提取器置信度来对权重进行缩放，并且使用具有受限制的原子集的 MC-Sat 来执行近似推断，以 0.5 的边际概率输出。这将最大化 F1 分数。如前所述，MLN 方法仅生成通过对查询集的值进行调节而生成的 2 跳邻域的预测。

我们的方法 PSL-KGI 使用 PSL 和二次加权规则进行本体约束，实体解析，候选和推广事实以及并入先验知识。我们还将 MLN 方法生成的谓词合并为一个更平等的比较。我们学习所有规则的权重，包括先验，使用投票感知器学习方法。权重学习方法通过对训练数据进行推理和调节来生成一组目标值，然后在没有训练数据的情况下选择最大化与这些目标的一致性的权重。由于我们将提取器置信度值表示为软真值，因此我们不会缩放这些规则的权重。使用学习权重，我们对由 MLN 方法使用的查询集定义的同一邻域执行推理。我们反映了这些结果，使用.55 的软阈值最大化 F1，作为 PSL-KGI。如表 3 所示，知识图谱识别在 F1 和 AUC 方面都产生了适度的改善。

表 3.与之前关于 NELL 数据集的工作相比，使用 PSL 进行知识图识别显示出了实质性的改进。

Method	AUC	Prec	Recall	F1
Baseline	0.873	0.781	0.881	0.828
NELL	0.765	0.801	0.580	0.673
MLN	0.899	**0.837**	0.837	0.836
PSL-KGI	**0.904**	0.777	**0.944**	**0.853**

表 4.比较 PSL 图谱识别的变体显示了本体信息的重要性，但是当知识图谱识别的所有组成部分被组合时，取得了最佳效果。

Method	AUC	Prec	Recall	F1
PSL-NoSrcs	0.900	0.770	**0.955**	0.852
PSL-NoER	0.899	0.778	0.944	**0.853**
PSL-NoOnto	0.887	**0.813**	0.839	0.826
PSL-KGI	**0.904**	0.777	0.944	**0.853**

表 5.生成完整的知识图会降低测试集的性能，这表明在生成有限的推理集时会掩盖问题的真实复杂性。

Method	AUC	Prec	Recall	F1
NELL	**0.765**	**0.801**	0.580	0.673
PSL-KGI-Complete	0.718	0.709	**0.929**	**0.804**
PSL-KGI	0.904	0.777	0.944	0.853

分析知识图谱识别的变体 为了更好地理解我们模型各个组成部分的贡献，我们研究了忽略知识图形识别模型的一个方面的变体。PSLNoSrcs 删除不同候选资源的谓词 CandLblT 和 CandRelT，用单一的 CandLbl 或 CandRel 替代它们，其平均置信度值跨源。PSL-NoER 删除第 5.2 小节中的规则，用于推理共同指称实体。 PSL-NoOnto 删除了使用本体关系约束知识图谱的 5.3 小节的规则。虽然源信息和实体解析都提供了好处，但本体信息显然是知识图形识别成功的主要原因。我们与以前的工作进行比较的一个缺点是将模型限制在一小组推理目标上。这个集合的构建模糊了现实世界中的一些挑战，例如冲突证据。为了评估我们的方法在推理目标不限制潜在矛盾推论的情况下的性能，我们还使用相同的学习权重运行知识图

谱识别，但没有预定义的目标集，允许惰性推理生成完整的知识图谱。由此产生的推论总共产生了 4.9M 个事实，这些事实包含在测试集中。我们将测试结果报告为 PSL-KGI-Complete。允许模型在完整知识图上而不仅仅是测试集上进行优化，降低了由特定测试集测量的性能，这表明由冲突证据引入的噪声确实对结果有显着影响。与 NELL 评分方法相比，KGI 具有较低的 AUC 和精确度，但较高的召回率和 F1。对于这种黯淡的表现，一种可能的解释可能是使用为不同设置学习的权重。例如，在权重学习过程中，Mut 规则的权重显着下降。但是，根据 MusicBrainz 数据显示的结果，在恢复完整的知识图谱时，知识图谱识别可以非常强大。

可扩展性 使用 PSL 进行知识图谱识别的一个优势是能够将复杂联合推理框架化为凸优化。在 PSL 中实现的知识图形标识可以处理像 NELL 这样的真实世界的数据集中的问题，其中包含数百万个候选事实。当给出明确的 70K 事实的查询集合（PSL-KGI）时，推断仅需要 10 秒钟。我们比较的 MLN 方法需要几分钟到一个小时才能运行相同的设置。当推断一个没有已知查询目标的完整知识图时，如在 LinkedBrainz 和完整的 NELL 实验中，MLN 的推断是不可行的。相反，NELL 数据集上的知识图形识别可以在 130 分钟内生成包含 4.9M 事实的完整知识图。在这些现实环境下生成完整知识图谱的能力是我们实施知识图谱识别的一个重要特征。

7 结论

我们已经描述了如何制定知识图谱识别问题：通过确定共同参照实体，预测关系链接，共同分类实体标签和执行本体论的组合过程，从信息提取系统的噪声输出中联合推断知识图谱。使用 PSL，我们说明了我们在两个知识图推理问题上的方法的好处：来自 MusicBrainz 的合成数据和来自 NELL 的嘈杂的实际网络提取。在两个数据集上，知识图的识别通过将本体论推理的优势与图形识别相结合，产生了优越的结果。此外，我们的方法通过有效的凸优化解决，从而允许以分钟为单位解决先前不可行的问题。将来，我们希望将知识图谱识别应用于规模更大，本体关系更丰富，更多变的问题

致谢 我们要感谢 Shangpu Jiang 和 Daniel Lowd 分享他们的数据并提供热心的帮助。这项工作得到了 NSF CAREER 拨款 0746930 和 NSF 拨款 IIS1218488 和 CCF0937094 的部分支持。 本资料中所表达的任何观点，结论和结论或建议均为作者的观点，并不一定反映国家科学基金会的观点。

参考

1. Ji, H., Grishman, R., Dang, H.: Overview of the Knowledge Base Population Track. In: Text Analysis Conference. (2011)

2. Artiles, J., Mayfield, J., eds.: Workshop on Knowledge Base Population. In Artiles, J., Mayfield, J., eds.: Text Analysis Conference. (2012)

3. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E.R., Mitchell, T.M.: Toward an Architecture for Never-Ending Language Learning. In: AAAI. (2010)

4. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open Information Extraction from the Web. Communications of the ACM 51(12) (2008)

5. Pasca, M., Lin, D., Bigham, J., Lifchits, A., Jain, A.: Organizing and Searching the World Wide Web of Facts-Step One: the One-million Fact Extraction Challenge. In: AAAI. (2006)

6. Singhal, A.: Introducing the Knowledge Graph: Things, Not Strings (2012) Official Blog (of Google), see: http://goo.gl/zivFV.

7. Broecheler, M., Mihalkova, L., Getoor, L.: Probabilistic Similarity Logic. In: UAI. (2010)

8. Cohen, W., McAllester, D., Kautz, H.: Hardening Soft Information Sources. In: KDD. (2000)

9. Jiang, S., Lowd, D., Dou, D.: Learning to Refine an Automatically Extracted Knowledge

Base Using Markov Logic. In: ICDM. (2012)

10. Richardson, M., Domingos, P.: Markov Logic Networks. Machine Learning 62(1-2) (2006)

11. Namata, G.M., Kok, S., Getoor, L.: Collective Graph Identification. In: KDD. (2011)

12. Memory, A., Kimmig, A., Bach, S.H., Raschid, L., Getoor, L.: Graph Summarization in Annotated Data Using Probabilistic Soft Logic. In: Workshop on Uncertainty Reasoning for the Semantic Web (URSW). (2012)

13. Yao, L., Riedel, S., McCallum, A.: Collective Cross-Document Relation Extraction Without Labelled Data. In: EMNLP. (2010)

14. Kimmig, A., Bach, S.H., Broecheler, M., Huang, B., Getoor, L.: A Short Introduction to Probabilistic Soft Logic. In: NIPS Workshop on Probabilistic Programming. (2012)

15. Bach, S.H., Broecheler, M., Getoor, L., O'Leary, D.P.: Scaling MPE Inference for Constrained Continuous Markov Random Fields with Consensus Optimization. In: NIPS. (2012)

16. Dixon, S., Jacobson, K.: LinkedBrainz - A project to provide MusicBrainz NGS as Linked Data see http://linkedbrainz.c4dmpresents.org/.

17. Raimond, Y., Abdallah, S., Sandler, M.: The Music Ontology. In: International Conference on Music Information Retrieval. (2007)

18. Davis, I., Newman, R., DArcus, B.: Expression of Core FRBR Concepts in RDF (2005) see http://vocab.org/frbr/core.html.

19. Brickley, D., Miller, L.: FOAF Vocabulary Specification 0.98 (2010) see http://xmlns.com/foaf/spec/20100809.html.

20. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media Meets Semantic Web–How The BBC uses DBpedia and Linked Data to Make Connections. In: ESWC. (2009)

21. Bizer, C., Seaborne, A.: D2RQ–Treating Non-RDF Databases as Virtual RDF Graphs. In: ISWC. (2004)

22. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A Core of Semantic Knowledge. In: WWW. (2007)