中国科学: 信息科学 2020 年 第 50 卷 第 4 期: 551–575
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
评述

大规模知识图谱补全技术的研究进展

王硕[1,2*], 杜志娟[1], 孟小峰[1*]

1. 中国人民大学信息学院, 北京 100872
2. 河北大学机器学习与计算智能重点实验室, 保定 071002
* 通信作者. E-mail: wsterran@126.com, xfmeng@ruc.edu.cn

收稿日期: 2018–08–22 ; 接受日期: 2019–03–17; 网络出版日期: 2020–04–13

国家自然科学基金 (批准号: 61532010, 61532016, 91846204, 91646203, 61762082)、国家重点研发计划 (批准号: 2016YFB1000602, 2016YFB1000603)、中国人民大学科学研究基金 (批准号: 11XNL010) 和河南省科技开放合作 (批准号: 172106000077) 资助项目

摘要 随着谷歌知识图谱、DBpedia、微软 Concept Graph、YAGO 等众多知识图谱的不断出现, 根据 RDF 来构建的知识表达体系越来越为人们所熟知. 利用 RDF 三元组表达形式成为人们对现实世界中知识的基本描述方式, 由于其结构简单、逻辑清晰, 所以易于理解和实现, 但也因为如此, 当其面对现实中无比繁杂的知识和很多常识时, 往往也无法做到对知识的认识面面俱到, 知识图谱的构建过程注定会使其中包含的知识不具有完整性, 即知识库无法包含全部的已知知识. 此时知识库补全技术在应对此种情形时就显得尤为重要, 任何现有的知识图谱都需要通过补全来不断完善知识本身, 甚至可以推理出新的知识. 本文从知识图谱构建过程出发, 将知识图谱补全问题分为概念补全和实例补全两个层次: (1) 概念补全层次主要针对实体类型补全问题, 按照基于描述逻辑的逻辑推理机制、基于传统机器学习的类型推理机制和基于表示学习的类型推理机制等 3 个发展阶段展开描述; (2) 实例补全层次又可以分为 RDF 三元组补全和新实例发现两个方面, 本文主要针对 RDF 三元组补全问题沿着统计关系学习、基于随机游走的概率学习和知识表示学习等发展阶段来阐述实体补全或关系补全的方法. 通过对以上大规模知识图谱补全技术研究历程、发展现状和最新进展的回顾与探讨, 最后提出了未来该技术需要应对的挑战和相关方向的发展前景.

关键词 知识图谱, 知识库补全, 概念补全, 实例补全

1 引言

人类社会经过了漫长的发展和积累, 其中蕴藏了丰富的知识, 包括语言、文字、图像等, 其内涵极其丰富, 价值不可估量. 近年来, 随着互联网和人工智能技术的进步与结合, 包括社交网络系统、自动导航系统、自动问答系统、机器翻译系统、智能推荐系统等在内的智能系统取得了巨大进展, 而这些发展的背后都离不开人类知识和计算机技术的结合, 只有让机器也能够学习、使用和存储知识, 才能

引用格式: 王硕, 杜志娟, 孟小峰. 大规模知识图谱补全技术的研究进展. 中国科学: 信息科学, 2020, 50: 551–575, doi: 10.1360/N112018-00225
Wang S, Du Z J, Meng X F. Research progress of large-scale knowledge graph completion technology (in Chinese). Sci Sin Inform, 2020, 50: 551–575, doi: 10.1360/N112018-00225

表 1 知识图谱构建方式

Table 1 KB construction methods

Construction method	Schema (Y/N)	Typical KB
Artificial, experts	Y	OpenCyc, UMLS, WordNet
Artificial, volunteers	Y	Wikidata, Freebase
Automatic, semi-structured	Y	YAGO, DBPedia, Freebase
Automatic, un-structured	Y	Knowledge Vault, NELL, PATTY, DeepDive
Automatic, un-structured	N	ReVerb, OLLIE, PRISMATIC

更大地发挥机器的作用, 让其更好、更有效地为人类服务. 若想让知识在信息社会里成为真正的力量, 就需要更深、更广、更新和更加准确的知识库的构建和使用.

大数据是现在各行各业都在谈论的话题, 这种现象正好反映出数据的价值, 因为人类的知识大多数都是以数据的形式保存的, 比如前面所述的语言文字, 其中的信息量就非常丰富, 里面所包含的知识需要人们抽取出来才能够使用和发挥其应有的价值, 所以如何表示、存取和使用知识就成为了计算机科学家们需要解决的问题. 通常, 人们描述一个事物或事件时都是按照主谓宾的结构来叙述的, 也就是常见的 (S, P, O) 结构, 头尾是主体和客体, 一般都称为实体, 中间是谓词或者关系, 用来描述主客体间的作用关系, 比如 (姚明, isA, 篮球运动员), 一般都是三元组的形式. 而存储这些三元组的数据库就被称为知识库, 它是一个结构化的语义知识库, 能以符号化的形式来描述现实世界的概念及其相互关系, 其存储的内容主要就是现实世界的实体集合、关系集合和结合前两者的三元组集合. 这种结构往往可以用图结构来描述, 因此知识库也可以称为知识图谱, 图中的结点代表实体, 实体与实体间的连边就是关系.

构建知识图谱需要从外部数据源获取知识, 将其加工成三元组的形式存入数据库中, 该构建方式大体上经过了 4 个阶段. 早期的知识库需要人工来完成, 而且基本由小范围的专家组成 [1~3], 因此知识库的规模有限, 成本高昂; 后来, 随着众包技术的出现 [4], 人工工作由网络上的志愿者及少量专家的协作来完成, 成本下降但质量没有之前可靠, 但是知识库的规模得到了快速增长; 为了更好地保证知识库中数据的质量, 从半结构化的文档中自动化地抽取知识, 半结构化的自动提取方式产生了, 但其依然依赖大量的手工规则、学习规则和正则表达式, 这些规则会对最终产生的数据质量产生很大影响; 为了克服以上的不足, 利用人工智能方法和自然语言处理的最新技术, 自动地从无结构文档中抽取三元组成为可能 [5~8]. 表 1 [9] 中列举了针对技术发展各阶段的代表性知识库案例, 其具体实现细节可以参考有关文献.

现阶段知识图谱的构建流程 (如图 1 所示) 除了上述信息提取方式的变化之外, 也在不断进步和完善, 大致上可以分为 3 个步骤 [10~13]. (1) 语义信息抽取: 主要指从半结构或无结构化数据中通过自然语言处理等技术来完成语义识别, 完成实体抽取、关系抽取、共指消歧等功能; (2) 多元数据集成与验证: 主要指对获取的三元组进行验证, 必要时引入其他知识库来完成跨知识库间的实体匹配和关系类别识别, 从而更有效地达到知识融合的目的; (3) 知识图谱补全: 主要指通过已获取的知识来对实体间进行关系预测, 以达到对实体间关系的补全, 也可以是实体类型信息的补全, 该过程可以利用本知识库内部的知识, 也可以引入第三方知识库的知识来帮助完成.

2 问题的由来与描述

通过上面的基本描述, 我们可以看到无论是领域知识库, 还是自动构建的知识图谱, 在它们的

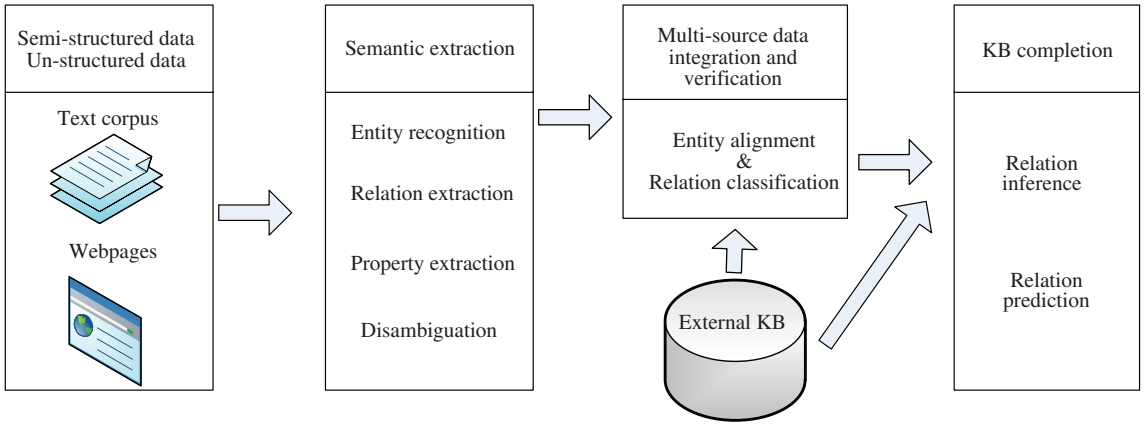


图 1 (网络版彩图) 知识图谱构建流程

Figure 1 (Color online) KB construction process

构建过程中会产生大量的实体和关系、实体和属性的三元组. 例如, 根据不完全统计, 医学领域的 UMLS [3] 就包含了 135 个实体, 49 种关系, 6800 多个三元组; NELL 中包含超过 519 万个实体, 306 种关系, 超过 5 亿个候选三元组; 而 Knowledge Vault 中的实体已超过 4500 万, 关系数量已达到 4469 种, 三元组数量达到了 2.7 亿个. 尽管从数量上已经达到了惊人的规模, 但是具体分析其中实体和关系的关联时会发现很多实体间本应存在的关系并没有建立起来, 比如对于人这种实体来说, 除了那些比较有名的人物, 依然有超过一半以上的人没有出生地关系或国籍的属性等, 甚至有些人物的关系存在明显错误, 这些现象都充分证明了知识图谱尽管体量越来越大, 但是其中的知识并不是完整和正确的.

在构建知识图谱的过程中, 大量知识信息来源于文档和网页信息, 在从文档提取知识的过程中往往会有偏差, 这些偏差来自于两方面: 一方面文档中会有很多噪声信息, 即无用信息, 它的产生可能来自于知识提取算法本身, 也可能和语言文字本身的有效性有关; 另一方面则是文档信息量毕竟有限, 不会把所有的知识都涵盖进去, 尤其是很多常识性知识不会在文本中明显地给出描述, 所有这些都导致知识图谱是不完整的. 因此知识图谱构建的第 3 个步骤 —— 知识图谱补全 —— 显得日益重要起来, 也逐步成为知识图谱构建的重要技术之一.

知识图谱补全问题可以分为两个层次的补全问题: 概念层次的知识补全和实例层次的知识补全. 这与知识图谱本身的知识组织结构密不可分, 前面提到的知识图谱构建的过程中只是提及了实体和关系的抽取, 然后就可以生成实体和关系组成的 RDF 三元组了, 但是仅仅获取三元组是不够的, 还需要考虑这些三元组所蕴含的知识模式 (有时称为数据模型), 因为三元组中的实体除了具有属性和关系之外, 还可以映射关联到知识概念层次的类型 (type), 而且一个实体的类型可以有很多. 举例如图 2 所示, 实体奥巴马的类型在不同关系中是有变化的, 比如在出生信息描述中其类型为人, 在创作回忆录的描述中其类型除了人以外还可以是作家, 在任职描述中除了人以外还可以是政治家, 从图 2 部分可以看出人、作家、政治家这些概念之间是有层次, 这里给出的正是这些概念的层次模型, 这也是知识图谱中知识模式的重要组成部分, 一个完整的知识图谱离不开概念层次的类型组织及其完整性.

上述过程中, 被获取的三元组就是实例的集合, 这些实例间也存在知识缺失的现象. 举例如图 3 所示, 比如有这样两个三元组 (Natasha Obama, child-of, Barack Obama) 和 (Michelle Obama, spouse-of, Barack Obama), 这两个三元组是从文本中发现的知识, 但是 (Natasha Obama, child-of, Michelle Obama) 这个知识没有在文本中出现过, 但是根据人们已有的知识可以推测出第 3 个三元组的关系应

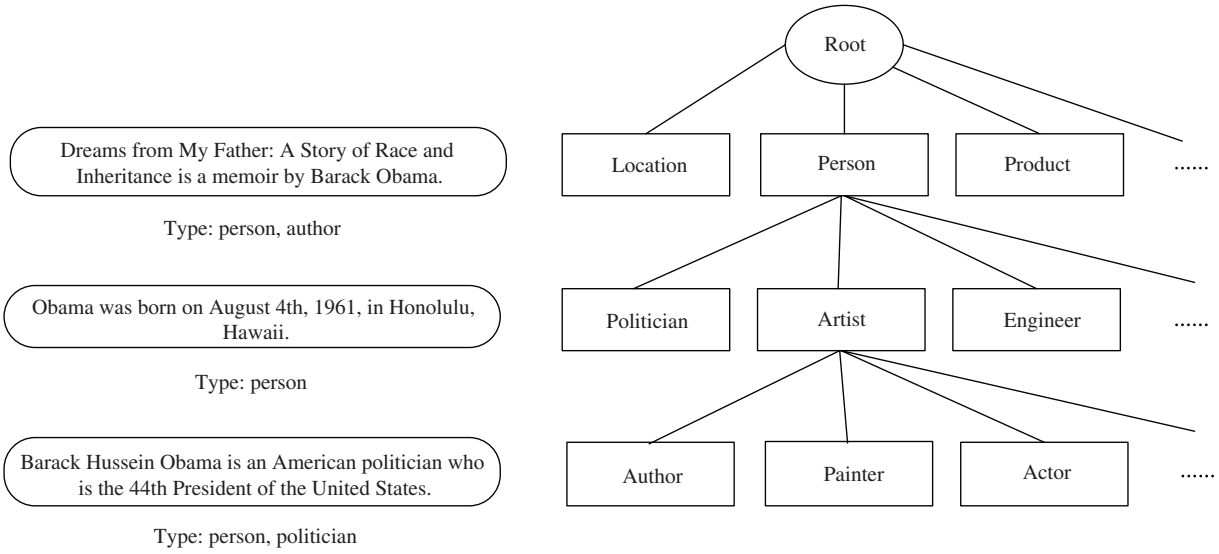


图 2 实体类型的概念层次模型

Figure 2 Conceptual hierarchy model of entity type

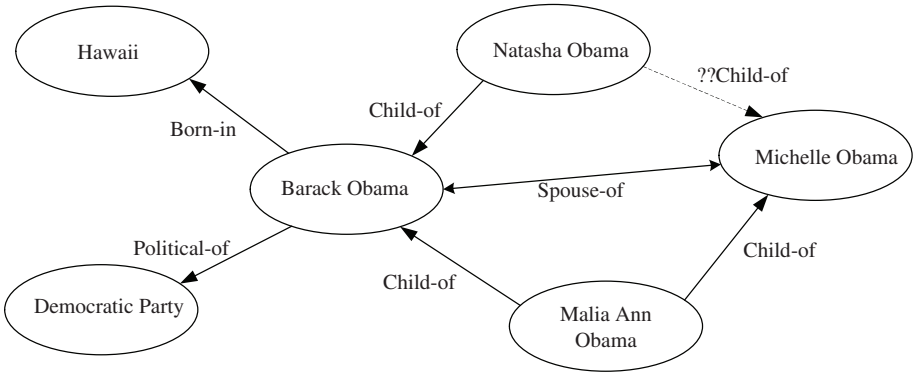


图 3 实例层次知识补全示例

Figure 3 Example for instance completion

该是存在的, 至少可能性会很大, 尽管它不存在于构建的知识库中, 这就是知识库的不完整问题, 有很多显而易见的知识对于人而言很轻松地就能推知, 但是对于存储在计算机中的知识库而言却没那么容易, 知识库的补全技术在此时就显得尤为重要.

在解决知识图谱补全问题上也要分为两个层次来解决. 针对概念层次而言, 主要是要解决实体的类型信息缺失问题, 正如前面的例子中所描述的那样, 一旦一个实体被判别为人这个类型, 那么在已构建好的知识模式中, 该实体除了人的类型外仍需要向下层概念搜索以发现更多的类型描述信息, 从而使该实体在概念层次的类型更加完整, 这一过程就是在概念层次对类型进行的补全, 将在第 3 节中进行描述. 而对于实例层次而言, 问题可以理解成如下过程: 针对一个实例的三元组 (SPO), 其中可能缺少了一个 S, 一个 P 或者一个 O, 即 (?, P, O)、(S, ? , O) 或者 (?, P, O), 这就如同知识库中不存在这个三元组, 此时需要预测缺失的实体或关系是什么. 根据所示例子可以看到, 很多缺失的知识

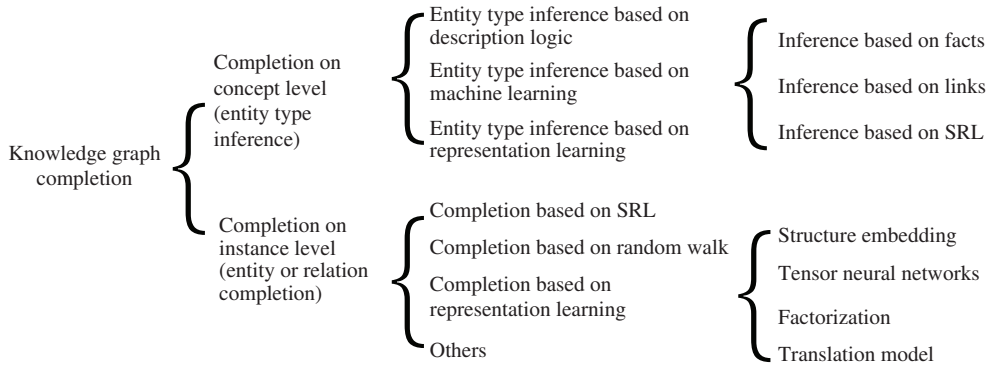


图 4 知识图谱补全方法分类

Figure 4 Classification of knowledge graph completion methods

是可以通过已经获得的知识来推知的, 有时这个过程也被称为链接预测, 此处的链接预测有别于社会计算中的社会网络的链接预测, 但是在某些情况下二者是有共通之处的. 此种情形就可以通过知识库自身已有的知识完成一定程度的补全, 这也是该问题在最初阶段的常见解决方案, 但随着研究的发展, 人们发现由单一知识库自己补全的效果有限, 于是产生了引入外部数据源的信息来补全和完善自身知识库的思想, 因而出现了跨知识库的补全和借助网络或其他文本数据进行知识库补全的方法, 这些多源补全算法依然是建立在之前的单源知识库工作基础之上的.

但有时知识不是缺失的, 而是新出现的, 即出现了新的三元组且该三元组不是原知识库所已知的知识, 此时需要将其作为新知识补全到知识库中, 但此种情形不是传统意义的补全, 而更像知识的补充, 因此不在本文中讨论. 这些将在第 3 节进行进一步总结. 本文述及的方法如图 4 所示, 后面章节也据此排列, 其中的统计关系学习在早期主要针对概念层次的学习, 但随后的发展使其也用于了实例层次补全问题的解决.

3 概念层次中的知识图谱补全

在知识库中存有大量的实体, 而这些实体除了具有属性和关系之外, 还可以映射关联到知识概念层次的类型 (type), 而且一个实体的类型一般会有很多, 这已从上一章节中的例子得到说明. 现有的知识库中大多数实体的类型信息都存在缺失或不完整, 因此需要对已有知识进行概念层次的补全, 即类型信息的补全, 才能使知识库中的知识更加完整和高效可用. 对于知识概念层次的补全问题研究已经由来已久, 最初的研究源于关于本体构建和模式发现的方法, 因此主要是基于描述逻辑的规则推理机制, 后来随着机器学习和自然语言处理相关技术的结合, 类型信息的判断被转化为了分类或者聚类等与机器学习类似的问题, 近年来, 随着表示学习的研究进展, 嵌入式方法和深度学习方法被引入进来, 使得实体类型的推理归结为解决一种分类问题.

3.1 基于描述逻辑的规则推理机制

在早期, 人们对知识的理解主要集中在如何针对现实世界来完成概念的抽象, 这就产生了本体论和模式的概念 [14], 即现实世界中的事物 (即实体) 都可以归结为一种本体, 而这种本体会具有一组模式来保证其独特性, 这组模式可以用规则来描述, 因此, 对于本体而言其可以由这组规则来描述. 比如

图 1 中所示, 奥巴马是个实体, 他的本体可以归为人, 而人的模式就是可以使用语言和工具、可以改造其他事物等等, 这些模式可以通过规则来描述, 因此基于描述逻辑的规则推理方法就出现了. 描述逻辑是一种常见的知识表示方式, 它建立在概念和关系之上. 比如可以将关于人的实体实例 (可以是文本) 收集起来, 从中提取出其中的模式并以规则的形式记录下来, 这样一来, 只要遇到一个新的实体实例, 只需将其代入到之前记录下的规则中进行比较即可做出判断, 如果符合规则, 就说明该实例可以归类为人的概念类型, 否则就判定为非此概念类型.

DL-Learner 系统 [15] 是此类方法的早期版本, 它直接利用描述逻辑中的 ILP (inductive logic programming) 方法来识别现实中出现的概念, 并可以基于此来构建本体, 其基本实现方式是从大量实例中推导出 T-box 和 A-box 等公理假设, 使其成为一种判别规则可以用来对概念类型进行判别.

另外一种 Formal Concept Analysis (FCA) 方法 [16], 也是对事实进行收集后可以产生出本体中缺失的公理假设, 同时还增加了交互式查询环节来增加公理的正确性. 以上两种方法的最大缺陷是面临着事实或者实体实例的不稳定性, 即存在大量的噪声数据会使据此产生的公理假设不正确或无效. 为了克服此种缺点, 引入了关联规则挖掘方法来改进之前的关系发现过程, 即通过加入规则约束的方式来降低噪声数据对本体构建和概念学习的影响 [17~21].

还有一些改进方式, 比如 SD-type [22] 是利用统计分布学习的方式结合传统的规则产生方法, 其本质是基于链接机制的, 通过将与实例相关的链接作为指示符, 来对每个链接做属性与待预测对象的统计分布, 并据此过滤规则, 进而可以对对象类型做出更为准确的预测. 再比如 OWLearner [23] 则能够通过表示学习构造本体 TBox 的公理, 即利用翻译模型向量化实体和关系后, 在这些向量上通过监督模型来预测本体的相关公理.

3.2 基于机器学习的类型推理机制

经过基于描述逻辑的规则推理的发展阶段后, 机器学习相关研究开始占据主流, 实体类型的推理 (或者说类型预测) 方法开始与机器学习的方法相结合, 即不是单纯地利用实例产生的规则等内部线索来进行判断, 同时也要利用外部的特征和线索来学习类型的预测. 这时一般假设对一个未知类型的实体 e1 而言, 如果能找到一个与其类似的且已知类型的实体 e2 的话, 那么就可以据此推知实体 e1 的类型应该与 e2 的类型一致或至少相似. 此类方法主要可以分为基于内容的类型推理、基于链接的类型推理和基于统计关系学习的类型推理几个方向.

3.2.1 基于内容的类型推理方法

基于内容的类型推理通常利用实体描述信息来对实体类型完成识别 [24~27], 这些描述信息主要来自于摘要、信息框、属性、文章片段等网站上的相关信息. 学习模型使用的特征也是出自这些从网站上直接获取的或者对其加工过的信息, 学习模型也是将自然语言处理技术和机器学习方法相结合来建立的.

比如训练过程可以先通过语法分析技术来产生概念句子的逻辑的 RDF 描述形式, 然后再产生基于图形模式的类型 —— 关系的图形表示, 在分类或聚类过程中就可以将此作为特征来完成类型推理. 机器学习的常见模型此时都可以设计用来完成此类任务.

3.2.2 基于链接的类型推理方法

基于链接的类型推理本质上就是将与实体相关的链接作为一种特征来帮助分类或聚类的过程 [28].

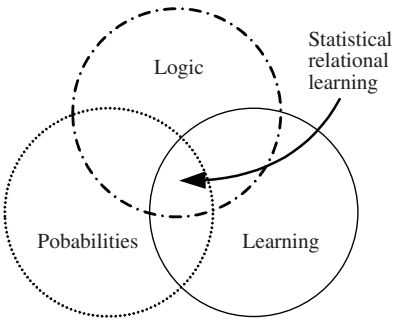


图 5 统计关系学习特点

Figure 5 Statistical relational learning characteristics

比如可以对这些链接做统计分布学习 [22], 从而将其转化为一种链接特征表示方式, 也可以利用 k 近邻方法来表示实体的链接特征 [24].

3.2.3 基于统计关系学习的类型推理方法

统计关系学习 (如图 5) 也称为概率归纳逻辑编程, 涉及在关系域中的机器学习和数据挖掘, 其中观察可能缺失, 部分观察或有噪声. 在这个过程中, 它解决了人工智能的核心问题之一 —— 概率推理与机器学习、一阶逻辑和关系表示的整合, 而且其还涉及推理, 参数估计和结构学习等所有相关方面. 在统计学习中, 统计关系事实上是通过统计数据, 正确采用统计方法去整理分析数据而得出的事物之间的统计规律性, 它通常以 "某些事物之间有关联" 的形式出现, 这正是知识库中对于实体间的链接预测所需要的.

统计关系学习是由似然关系模型和学习算法组成的, 前者一般是基于概率关系的似然估计模型, 需要将不同的概率似然估计模型和一阶逻辑表示结合到一起, 常用的概率模型有 Bayes 网、Markov 网、随机文法、Hidden Markov 模型等, 下面将 SRL 方法按以上模型分类介绍 [29,30].

(1) 基于 Bayes 网的统计关系学习方法.

通过对传统的 Bayes 网方法的扩展, 人们从不同角度、不同方式将其扩展为 SRL 方法. 首先是基于图结构的扩展方法, 研究者将数据的表示和数据关系的表示看作图形结构. 其中由 Koller 提出的似然关系模型 (probabilistic relational models, PRM) 是最有代表性的 [31], PRM 通过引入实体、实体属性和实体关系等来扩展 Bayes 网, 由图依赖结构 S 和与其关联的参数 θ 组成. S 由代表类的某个属性的结点和有向边构成, 结点与一个条件概率相关联, 而边则有两种: 一种是指向同一类中的其他相关属性, 一种是指向相关类中的其他属性. PRM 对类与关系的定义如下:

$$P\left(I|\sigma,S,\theta_s\right)=\prod_{X_i}\prod_{A\in A(X_i)}\prod_{x\in O^{\sigma}(X_i)}P\left(I_{x,A}|I_{\mathrm{pa}(x,A)}\right),\quad I\in I_S,$$

其中 σ 表示关系的框架, $O^{\sigma}\left(X_i\right)$ 表示在框架 σ 下 X_i 类的对象集合, $A\left(X_i\right)$ 表示 X_i 类的属性集合, $I_{x,A}$ 表示对象 x 的属性 A 对应的示例, $I_{\mathrm{pa}(x,A)}$ 表示对象 x 的属性 A 的父节点对应的示例. 根据该公式可以将 Bayes 网的学习和推理算法扩展到 PRM 方法中来.

另外一个有代表性的方法是 Ngo 和 Haddawy 在 1997 年提出的似然逻辑程序模型 PLP [32], 该模型结合了逻辑与 Bayes 网, 能将 Bayes 网直接提升为一阶逻辑. Kersting 等 [33] 在此基础上提出了贝叶斯逻辑程序模型 (Bayesian logic programs, BLP), 该模型建立了基原子和随机变量间的一一映射, 将

Bayes 网和正定子句逻辑结合了起来, 从而完成了对对象及其关系的描述. BLP 的基础源于知识模型构建 (knowledge based model construction) 的思想, 它能将 Bayes 网中的结点升级为 Bayes 原子, 是与领域 D 相关联的逻辑原子, 每个原子能表示一些变量的集合, 有直接因果影响的原子组成了 Bayes 正定子句, 每个正定子句有一个与之关联的条件概率分布. 所有的子句及其条件概率分布组成了 BLP, 它通过组合规则来解决对于不同的对象实例可能存在相同的子句头部的问题. BLP 推理采用两步策略, 首先从 BLP 中生成一个 Bayes 网, 再用已知的 Bayes 网推理算法进行推理; 关于 BLP 学习问题, 同样先学习逻辑子句以获得 BLP 的结构, 然后用借鉴 Bayes 网参数学习方法来求得 BLP 的概率分布.

(2) 基于 Markov 网的统计关系学习方法.

由 Markov 网扩展而来的 SRL 方法可以分为两类: 关系 Markov 网 (relational Markov networks, RMN) 和 Markov 逻辑网 (Markov logic network, MLN). 它们和基于 Bayes 网的方法相比可以更加灵活地表示现实中的关系特性, 事实上它们定义了一种 Markov 网模板, 即对于实例集合定义了一致的概率分布. 2002 年 Taskar 等 [34] 提出了关系 Markov 网的概念, RMN 结合数据库中的查询语句将 Markov 网中的团 (clique) 升级为关系团模板 (relational clique template), 它是一个三元组 (F, W, S), F 是不同类型的变量的集合, W 是 F 中对象属性间所满足的规则, S 是 F 中被选择的对象的属性集合. 模板会对每个团关联一个势函数, 每一个关系团模板的实例对应一个 Markov 网. RMN 学习过程实际是模板的参数估计过程, 其推理则是在实例化的 Markov 网上进行精确或近似推理.

2004 年 Richardson 等 [35] 提出了 Markov 逻辑网, 它融合了 Markov 网和一阶逻辑, 满足下面的两个条件: (1) 每个可能出现在 MLN 中的基原子, 网络中都有一个相应的二进位的结点与其对应, 当基原子取真时结点取值为 1, 否则取值为 0; (2) 每个在 MLN 中的基规则对应网络中的一个特征函数, 如果基规则取真值, 则这个特征函数值为 1, 否则为 0, 特征函数的权值 IV 与该规则相关. MLN 的学习由一阶逻辑子句的学习和权值的学习组成. 其推理方法分为两步: 首先转化为最小基的 Markov 网络, 然后再在该网络上进行推理. MLN 仍有很多值得研究的内容, 比如模型表示、推理算法效率和学习算法的有效性等.

(3) 基于随机文法的统计关系学习方法.

这是一种将随机文法提升到一阶逻辑的方法, 主要包括随机逻辑程序 (stochastic logic program, SLP) [36] 和统计建模程序设计 (programming in statistical modeling, PRISM) [37] 两类. 此种方法能对逻辑成分附加概率, 以处理关系和不确定性. 由于采用逻辑程序来描述模型结构, 因而其表示能力较强.

Muggleto 在 1996 年给出了随机逻辑程序框架. SLP 是随机文法的一个泛化, 一个 SLP 程序一般由带有标号的形如 "$p : C$" 的子句集合构成, 其中 p 是概率, C 是一个范围受限的一阶正定子句. 子句头部包括同一谓词符号的所有子句的概率之和不超过 1. 它通过对包含目标的逻辑子句序列进行归结操作来得到推理结果.

PRISM 是一个符号统计建模语言, 它对逻辑程序进行了概率扩展, 并且使用 EM 算法从例子中进行学习. 简言之, PRISM 程序在事实上附加概率分布的逻辑, 在事实集合上建立对应的联合概率分布, 通过采样过程求出所有基原子的联合概率分布. 该分部是在 1995 年由 Sato 提出的, 它是 PRISM 程序的理论 (语义) 基础, 被称为分布语义 (distributional semantics). PRISM 的逻辑程序使用证据树来进行推理.

(4) 基于 Hidden Markov 模型的统计关系学习方法.

Kersting 等 [38] 在 2002 年提出的逻辑隐 Markov 模型 (logical hidden Markov models, LOHMM)

表 2 统计关系学习主要方法比较

Table 2 Comparison of main methods of SRL

Comparative factors	Probabilistic relational models	Markov logic networks	Relational Markov networks	Bayes logic networks
Model class hierarchy	Unidirectional graph model	Logical clause	Bidirectional graph model	Bipartite monograph model
Parameter estimation	Maximum likelihood estimation filling CPT	Maximum likelihood estimation, learning weight	Bayes relational classifier, learning CPT	Maximum likelihood estimation, filling CPT
Structure learning	Score-based learning	ILP	Conditional relation learner	ILP
Inference graph	Bayes networks	Markov networks	Undirected model	Bayes networks
Inference method	Belief propagation	Quasi-likelihood estimation	Quasi-likelihood estimation	Bayes networks inference
Self-correlation	Self-cycling in class hierarchical model	Additional variables	Yes	Not involve
Multi-relational processing	Need integration	No need	No need	Need to combine rules

是主要的基于 (隐) Markov 模型的方法, 其在 Markov 模型的基础上进行了扩展, 允许状态具有不同类型, 可用来解决动态的序列化问题. 它用逻辑原子来取代 HMM 中的状态标识, 它是一个四元组 $M=(\Sigma,\mu,\Delta,r)$, 其中 Σ 为逻辑符号系统 (logical alphabet), μ 为 Σ 上的选择概率, 抽象标识 A 的选择概率是进行替换时 A 取某一具体实例的概率值; Δ 为抽象转移集合, 抽象转移是形如 $P:H\overset{o}{\leftarrow}b$ 的表达式, 其中 P 是概率, H, b 是逻辑原子, o 是抽象输出标识, r 是抽象转移的先验概率分布的集合; 设 B 为 Δ 中所有转移的转移体 b 所构成的集合, 满足

$$\forall b\in B:\sum_{(b\to H)\in\Delta}p\left(b\to H\right)=1. \tag{1}$$

LOHMM 拥有较强的表达能力, 且可以处理隐藏状态.

(5) 几种主要 SRL 方法的比较.

如表 2 中所示, 我们选取了以上 4 种统计关系学习中有代表性的 4 个方法, 分别从模型的类别、使用的参数估计和结构学习、推理机制及方法、自相关性和多关系处理等几个方面进行了比较. 从 2010 年之后, SRL 方法主要集中在 Markov 逻辑网的应用方面, 比如 Dietrich 等 [39] 将其应用于人脸识别问题中, Rettinger 等 [40] 将其用于社交网络中的信任学习, 应用于自然语言处理任务的有文献 [41~45], 应用于医学和生物领域的有文献 [46~48], 其他一些应用 [49, 50] 不再一一赘述了. 总之, 统计关系学习方法除了 Farnadi 等 [51, 52] 提出了在一阶逻辑中引入软性约束之外没有太大的新进展, 在知识图谱补全中的应用也多作为一种关联程度评价方法结合到预测过程中 [53~62].

3.3 基于表示学习的类型推理机制

近几年来, 随着表示学习的不断发展, 将嵌入式学习和深度学习两种方法引入到类型推理问题中成为可能 [26, 63~72]. 基于机器学习的类型推理方法大多假设数据中没有噪声, 而且其特征仍然需要人为选择和设计, 无法克服以上方法带来的缺陷. 引入深度学习方法可以避免特征工程, 而表示学习也

是对实体及关系等图形结构有较强的表达能力, 尤其在类型推理问题上不仅需要依据文本内容, 也需要链接结构等其他特征的支持, 此时嵌入式方法可以发挥其自身的优势将两方面结合起来, 使得问题的解决变得更加简单和高效. 因为实体类型的分类可以帮助命名实体识别 [66]、关系抽取 [66,67] 甚至知识库的构建等诸多方面, 所以其相关工作多是依据文本信息和已有的类型知识库来帮助判断实体的类型.

早期的方法一般是将实体与实体类型作为三元组的头尾实体, 谓词就是 type, 这样构成的 RDF 三元组可以利用嵌入式学习过程完成向量表示的学习, 从而完成类型预测任务. 但是仅仅简单地将实体的一个类型作为尾实体而言会损失很多信息, 比如实体所在文本的上下文环境信息, 外部知识库中对该类实体的描述等, 而且一个实体的类型是多样且有层次的, 所以对某个实体的相关文本也可以做嵌入式学习 [69,70], 比如将其本身及上下文环境都变为低维向量的表示形式, 然后将这些低维向量输入到深度学习模型中 [71,72], 从而使类型的推理简化为利用神经网络模型来执行的类别判断, 此时在网络中也还可以引入注意力机制, 该机制可以来自于自然语言处理技术中常用的条件约束, 也可以使用知识库中已知的类型层次结构信息来产生, 总之就是要引入外部信息 [72] 来改进嵌入式方法的效果.

3.4 上述方法的比较

对以上 3 大类实体类型推理机制的比较分别从特征选取、学习方法及特点、存在不足等几个方面来总结. 对于基于描述逻辑的类型推理方法而言, 以公理系统为基础, 主要通过实例学习方法来实现, 实体类型的预测基于逻辑推理技术, 取决于提供的实例数据, 在假设无噪声的前提下可以得到完美的公理系统, 并可以据此对新实例甚至未知实体做类型预测, 但是噪声数据对公理系统中的规则影响很大, 模型不够健壮, 类型预测精度较低; 对于基于机器学习的类型推理而言, 特征可以人为构造也可利用特征选取技术, 一般使用分类、聚类、统计学习等常见机器学习方法, 其所需要的实体类型特征主要来自基于内容的和基于链接的两类方法, 依靠自然语言处理和其他知识学习技术, 预测精度得到很大的提升, 但是其学习效果受到特征选取和学习模型自身特点等两方面的影响, 而且学习模型设计复杂; 对于基于表示学习的类型推理方法, 主要利用深度学习技术, 将实体类型预测归结为分类问题, 据此可以针对实例构建深度学习模型, 以便自动提取特征, 学习模型设计得到简化, 但是学习的可解释性较差, 仍需增加外部知识和一定的注意力机制设计来改善学习模型.

4 实例层次中的知识图谱补全

本文所涉及的实例层次的知识图谱补全方法主要针对之前提到的第 1 种情形, 即利用已知的知识来对知识库进行补全, 主要分为以下几个方面: (1) 基于随机游走的概率学习 (path ranking algorithm, PRA) 补全方法; (2) 基于表示学习的补全方法; (3) 其他方法. 其实还可以使用基于统计关系学习的补全方法来预测实例中三元组中缺失的实体或关系, 但是其主要方法已在前面介绍过了, 所以这里不再赘述.

4.1 基于随机游走的概率补全方法

知识图谱如第 2 节中所描述, 可以按照 “实体 – 关系 – 实体” 这样的结构来理解, 因此知识库实际上就是一个知识的图表示 (即知识图谱). 很自然地出现了基于图结构的知识查询和推理方法, 其最早是在科学文献等专业领域里研究的. 对于包含丰富元数据的科学文献领域来说, 使用标记有向

图可以很好地描述实体间的关系, 尤其对关于科学文献的 ad hoc 查询和命名实体识别 (named entity recognition, NER), 可以将它们转换为基于图的近邻查询, 而采用随机游走方法在标记有向图上来完成近邻查询成为了一个重要的方法分支. 其中具有代表性的算法是 Page 等 [73] 在 1998 年就提出来的 PageRank 算法和 Haveliwala 等 [74] 在 2003 年提出的关于主题的个性化 PageRank 算法. 在此基础上, 2005 年 Nie 等使用模拟退火原理在参数数量较少的前提下实现了每个边类型的局部搜索; Diligenti 等 [75] 使用后向传播方法来完成权值优化, 但需要非常多次迭代才能收敛; Minkov 和 Cohen [76] 在 2008 年提出了基于生成学习模型的随机游走策略以使路径上的实体更相关; 特别在 2010 年, Lao 等 [77] 提出了代表性的路径排序算法 PRA (path ranking algorithm), 其优化了边参数化随机游走模型, 并增加了约束以提高计算效率.

经典 PRA 方法具体描述如下. 定义一条关系路径 P 由一系列的关系 $R_1,\ldots,R_L$ 组成, 为了说明路径中每一步上的类型, P 可以写为 $T_0 \xrightarrow{R_1} T_1 \cdots \xrightarrow{R_L} T_L$, 此时 $T_i = \text{range}\,(R_i) = \text{dom}\,(R_{i+1})$, 其中 $\text{dom}\,(P) \equiv T_0$, $\text{range}\,(P) \equiv T_L$. 只有能够连接不同类型关系的结点才被称为概念, 比如 “the team certain player plays for” 和 “the league certain player's team is in” 这两句的语义可描述如下:

$$P_1 : \text{concept} \xrightarrow{\text{Athelete Playes For Team}} \text{concept},$$
$$P_2 : \text{concept} \xrightarrow{\text{Athelete Playes For Team}} \text{concept} \xrightarrow{\text{Team Plays In League}} \text{concept}.$$

对任意关系的路径 $P = R_1,\ldots,R_L$ 和源结点 $s \in \text{dom}\,(P)$, 由路径约束的随机游走来递归定义分布函数 $h_{s,P}$, 如式 (2) 和 (3) 所示:

$$\text{如果路径 } P \text{ 为空, 则 } h_{s,P}(e) = \begin{cases} 1, \text{ if } e = s, \\ 0, \text{ otherwise;} \end{cases} \tag{2}$$

$$\text{如果非空, 则令 } P' = R_1,\ldots,R_{L-1}, \text{ 有 } h_{s,P}(e) = \sum_{e' \in \text{range}(P')} h_{s,P'}(e') \cdot P\,(e|e';R_L)\,, \tag{3}$$

其中 $P\,(e\,|e';R_L\,) = \frac{R_L(e',e)}{|R_L(e',\cdot)|}$ 是结点 e' 在边类型为 R_L 的前提下经过一步到达结点 e 的概率, $R\,(e',e)$ 表示结点 e' 和 e 间是否有类型为 R 的边相连. 这样一来, 对于一个给定的路径集合 $P_1,\ldots,P_n$, 对结点 e 就可以用 (2) 和 (3) 中定义的 $h_{s,P_i}\,(e)$ 作为结点的路径特征来生成一个线性模型, 如下所示:

$$\theta_1 h_{s,P_1}\,(e) + \theta_2 h_{s,P_2}\,(e) + \cdots + \theta_n h_{s,P_n}\,(e)\,,$$

其中 θ_i 为路径的权重. 如果给定路径长度为 L (比如小于 4), 那么就可以计算出规定长度内的所有关系路径的集合 P_L, 从而得到能够通过式 (4) 来计算对结点 e 排序的 PRA 模型:

$$\text{score}\,(e;s) = \sum_{P \in P_L} h_{s,P}\,(e)\ \theta_P. \tag{4}$$

在 PRA 模型中, 其参数估计可以通过最大化下面的正则目标函数 (5) 来完成:

$$O\,(\theta) = \sum_i o_i\,(\theta) - \lambda_1|\theta|_1 - \lambda_2|\theta|_2/2, \tag{5}$$

其中 λ_1 控制 L1 正则化来改善结构选择, λ_2 控制 L2 正则化来防止过拟合, $o_i\,(\theta)$ 是每个实例的目标函数, 用于计算每个实例的重要程度.

从上述过程可以看到, PRA 方法在已知的图上通过随机游走过程计算出实体间对应于某种关系类型的概率从而完成预测任务, 这其实类似于在实体对间完成各种关系的概率计算, 只不过是在计算时增加约束来减少需要计算的关系数量. 但其实仍然需要大量的计算代价, 因此在针对大规模知识图谱的应用 PRA 方法时的效率不是十分理想. 所以在之后的研究中, Lao 等 [78] 在 2011 年通过引入 Horn 规则和 N-FOIL 算法来提高算法效率, 并在大规模知识库 NELL 进行了实验; Lao 等 [79] 在 2015 年提出反向随机游走 (backward random walks) 概念, 并给出了 Cor-PRA 模型, 能够处理更长的关系路径和一阶规则类; 同年, Gardner 等 [80] 提出了子图特征抽取 (SFE) 的方法, 其采用宽度优先搜索替代随机游走, 相当于只执行了 PRA 算法的第一步, 然后根据搜索的局部子图来计算, 进一步减少需要计算的关系数量, 这会十分明显地节省计算时间. Wang 等 [81] 在 2016 年提出了基于多任务学习框架的 Coupled-PRA 方法, 通过对路径的聚类和耦合, 能够多路径地学习实体间的关系预测.

此外, 还有对路径增加约束 [82]、对路径完成嵌入式学习 [83,84]、引入范式计算 [85] 等一系列改进措施. 总之, 在 PRA 模型的基础上, 很多人将其进行了扩展, 主要思路是通过引入嵌入式表示学习 [86] 来提高路径上特征的表示和提高算法的执行时间效率.

4.2 基于表示学习的补全方法

知识库的三元组表示形式决定了其图形化的结构, 因此也带来了其自身无法回避的问题: (1) 计算效率不高. 由于知识数据库中存储的是实体和关系信息, 所以在进行访问时需要设计基于图的算法, 加之不同知识库的数据结构不一致, 导致算法的移植性差, 计算复杂度高, 当数据数量过多时会无法及时获得结果. (2) 数据稀疏问题. 统计学中大数据的长尾问题在知识库中也不可避免, 在长尾部分上的实体和关系都是稀疏的, 这样会因实例不足导致计算准确率不高. 而表示学习利用机器学习的方法, 将实体和关系的语义信息表示为低维实值向量, 从而可以方便高效地实现向量间的数学计算, 比如利用余弦距离衡量两个对象的语义相似度, 同时从某种程度上增强了语义信息的表达, 从而有助于克服数据稀疏问题. 表示学习的结果是获得了研究对象的低维向量表示, 这是一种分布式的表示方式, 向量中的每一维并无确定含义, 但是组合在一起就能够表达出丰富的语义信息. 这些优势使得表示学习在近两年成为知识库补全的主流方法, 在自然语言处理领域中成为了研究的热点.

表示学习在求解低维向量表示时的方法模型很多, 本文主要介绍几个典型的方法, 有结构嵌入表示法、张量神经网络法、矩阵分解法、翻译法等.

4.2.1 结构嵌入表示法

这是一个典型的早期表示学习的方法, 由 Bordes 等 [87] 在 2011 年提出. 该方法将所有实体投影到同一个维向量空间中, 并为每个关系定义了两个矩阵 $M_{r,1}, M_{r,2} \in \mathbb{R}^{d\times d}$, 每个三元组的损失函数定义如下:

$$f_r(h,t) = |M_{r,1}h - M_{r,2}t|_{L_1}. \tag{6}$$

可以将知识库三元组作为学习样例, 通过优化模型参数使知识库三元组的损失函数 (6) 的值不断降低, 从而使实体向量和关系矩阵能够较好反映实体和关系的语义信息. 然后通过计算如下公式:

$$\arg\min_r |M_{r,1}h - M_{r,2}t|_{L_1} \tag{7}$$

就能得到两个实体间的距离最近的关系, 从而完成链接预测. 之后, Pasquale 等 [55] 引入能量方程的概念来改进学习过程, 提高参数学习的效率, 大大减少了式 (7) 的学习迭代次数.

4.2.2 张量神经网络法

张量神经网络 (neural tensor network) 由 Socher 等 [88] 在 2013 年提出, 其基本思想是用双线性张量取代传统神经网络中的线性变换层, 在不同的维度下将头、尾实体向量联系起来. 它为评价两个实体之间存在某个特定关系给出了如下定义:

$$g\left(e_1, R, e_2\right)=u_R^{\mathrm{T}} f\left(e_1^{\mathrm{T}} W_R^{[1:k]} e_2+V_R\left[\begin{array}{c} e_1 \\ e_2 \end{array}\right]+b_R\right), \tag{8}$$

其中 f 是一个 tanh() 函数, $W_R^{[1:k]}$ 是一个三阶张量, V_R 是关系 r 的投影向量. 张量神经网络在式 (8) 中引入张量后可以更精确地描述语义信息, 但是不足之处就是计算时间复杂度较高, 不适合大数据环境下的大规模知识库的应用.

4.2.3 矩阵分解法

矩阵分解可以帮助我们得到低维向量表示, 因此, 采用矩阵分解进行知识表示学习成为必然. 这方面的代表方法是 Nickel 等 [89] 提出的 RESACL 模型. 在该模型中, 知识库三元组构成一个大的张量 X, 如果三元组 (e, r, e') 存在则 $X_{ere'}$=1, 否则为 0. 张量分解旨在将每个三元组 (e, r, e') 对应的张量值 $X_{ere'}$ 分解为实体和关系的表示, 使得 $X_{ere'}$ 尽量地接近于 eM_re'. TuckER [90] 则是一种张量分解的全表达模式, 其他的分解模式可以视为其上的特殊情况.

4.2.4 翻译法

Mikolov 等于 2013 年提出的 word2vec 词表示学习方法及工具包使表示学习在自然语言处理领域受到广泛关注, 该方法发现词向量空间存在着平移不变现象, 如下例所示:

$$C(\text{king})-C(\text{queen}) \approx C(\text{man})-C(\text{woman}).$$

这说明词向量能够发现不同词语间隐含的语义关系. Bordes 等 [91] 受此启发提出第 1 种翻译方法 —— TransE. 其具体描述如下.

TransE 方法将实体间的关系看作是向量的平移, 即三元组 (h, r, t) 中关系 r 的向量作为头实体 h 的向量和尾实体 t 的向量之间的平移, 如图 6(a) 所示. 其公式化描述为: $h+r \approx t$. TransE 的损失函数定义为

$$f_r(h, t)=|h+r-t|_{L_1/L_2},$$

其优化函数定义为

$$L=\sum_{(h,r,t)\in S} \sum_{(h',r',t')\in S^-} \max\left(0, f_r(h, t)+\gamma-f_{r'}(h', t')\right),$$

其中 S 是合法三元组的集合, S^- 为错误三元组的集合, $\max(x, y)$ 返回 x 和 y 中较大的值, γ 为合法三元组得分与错误三元组得分之间的间隔距离.

TransE 方法简单高效, 非常适合大规模的知识学习, 它也成为表示学习的研究基础, 之后大量的工作都是以它为基础而来. Wang 等 [92] 提出了 TransH 方法, Lin 等 [93] 提出了 TransR 方法, Ji 等 [94] 提出了 TransD 方法, Xiao 等 [95] 提出了 TransA 方法, 等, 详细比较和描述见 4.2.5 小节.

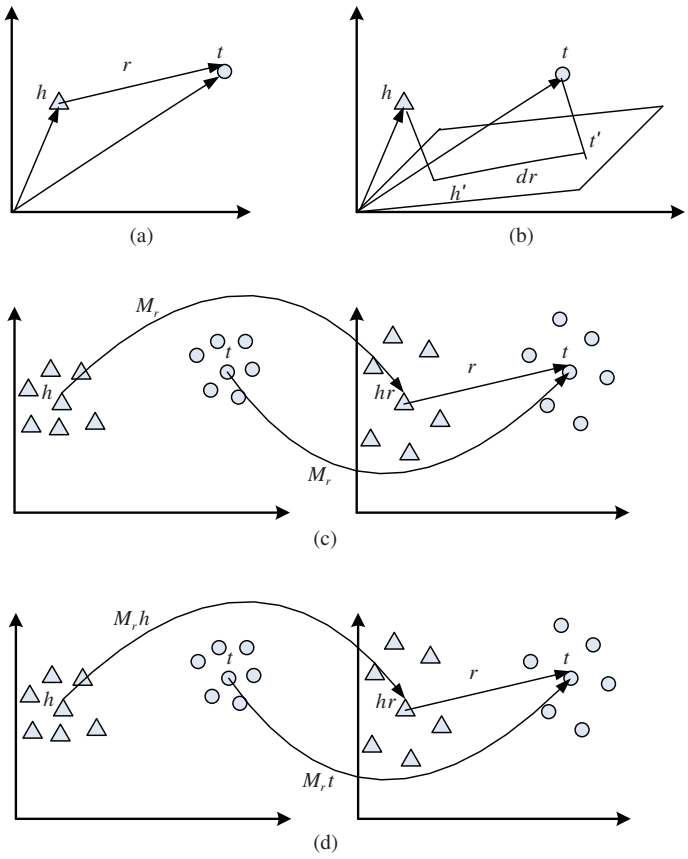


图 6 翻译法原理模型

Figure 6 Tranlation models. (a) TransE; (b) TransH; (c) TransR; (d) TransD

4.2.5 上述方法的比较

前两种模型相对于后面的方法提出时间较早, 因此其计算复杂度都比较高, 在规模较小的数据集上取得了一定效果, 但在大规模知识图谱上的计算代价比较大, 这也使得后面的翻译模型得到了快速发展, TransE 方法的计算代价较之之前的方法提高了不少, 在一对一的关系中表现很出色, 但是其自身也存在问题, 主要体现在针对多对一、一对多和多对多的关系描述上存在严重的缺陷 (如图 6(a)), 因此很多改进方法在其基础上产生出来, 改进策略也是针对其弱点来实现. TransH 引入了关系超平面来解决实体在不同关系中的不同表达问题 (如图 6(b)), 它可以通过将头尾实体投影到关系的超平面来解决同一个实体在不同关系中会得到不同的向量表示, 这对于 TransE 中实体表示固定的方式是一个很大的改进. 但 TransH 并没有完全打破实体和关系在同一个空间的假设, 因此 TransR (如图 6(c)) 的提出则将实体和关系建立在不同的语义空间下, 由于实体和关系处于两个不同的空间中, 因此需要一个映射矩阵将实体空间中的实体投影到关系空间中后再来完成类似于 TransE 中的计算过程, 该方法可以使具有相同关系的头尾实体在嵌入空间中比较邻近, 而没有相同关系的实体则会较远, 但这些实体尽管有相同的关系, 也还是会有差别的. 因此在 TransR 的基础上 Lin 等又提出了 CTransR 方法, 先通过聚类对头尾实体对进行分组, 然后通过每组实体对来学习上述的映射矩阵, 从而提高了实体与关系的映射准确性. 而 TransD (如图 6(d)) 则是通过分别学习头尾两个实体的不同映射矩阵来提高对应的关系投影的效果. 但是上述模型主要都是针对一对一的关系模型来学习, 而对于反射关系、一对

表 3 主要翻译模型比较

Table 3 Comparison of main translation models

Model name	Evaluation function	Optimization	Time complexity	Space complexity
Unstructured	$\left\|e^h - e^t\right\|_2^2$	SGD	$O\left(N_t\right)$	$O\left(N_e m\right)$
NTN	$r^{l\mathrm{T}}\tanh(e^{h\mathrm{T}} M_r e^t + W_{r,1} e^h + W_{r,2} e^t + b_r)$	L-BFGS	$O(((m^2+m)s+2mk+k)N_t)$	$O(N_e m + N_r(n^2 s + 2ns + 2s))$
SE	$\left\|W_{r,1} e^h - W_{r,2} e^t\right\|_2$	SGD	$O(2m^2 N_t)$	$O\left(N_e m + N_r\right)$
SME	$(W_{1,1} e^h + W_{1,2} r^l + b_1)^{\mathrm{T}} \cdot (W_{2,1} e^t + W_{2,2} r^l + b_2)$	SGD	$O\left(4mkN_t\right)$	$O(N_e m + \mathrm{N}_r n + 4mk + 4k)$
RESCAL	$\left\langle e^h \middle\| R^l \middle\| e^t \right\rangle$	SGD	$O(mnN_t)$	$O(N_e m + N_r n^2)$
TransE	$\left\|e^h + r^l - e^t\right\|_2$	SGD	$O\left(N_t\right)$	$O\left(N_e m + N_r n\right)$
TransH	$\left\|(e^h - \left\langle w^l \middle\| e^h w^l \right\rangle) + r^l - (e^t - \left\langle w^l \middle\| e^t w^l \right\rangle)\right\|_2^2$	SGD	$O\left(2mN_t\right)$	$O\left(N_e m + 2N_r n\right)$
TransR	$\left\|\left\langle e^h \middle\| M_l \right\rangle + r^l - \left\langle e^t \middle\| M_l \right\rangle\right\|_2$	SGD	$O\left(2mnN_t\right)$	$O\left(N_e m + N_r\left(m+1\right)n\right)$
CTransR	$\left\|\left\langle e^h \middle\| M_l \right\rangle + r^l - \left\langle e^t \middle\| M_l \right\rangle\right\|_2$	SGD	$O\left(2mnN_t\right)$	$O\left(N_e m + N_r\left(m+d\right)n\right)$
TransD	$\left\|(I + r_p h_p^{\mathrm{T}}) e^h + r^l - (I + r_p t_p^{\mathrm{T}}) e^t\right\|_2$	AdaGrad	$O\left(2nN_t\right)$	$O\left(2N_e m + 2N_r n\right)$
TransA	$-(\lvert h + r - t\rvert)^{\mathrm{T}} M_r (\lvert h + r - t\rvert)$	SGD	$O(m^2)$	$O(N_e m + N_r m^2)$
TranSparse	$\left\|W_r^h(\theta_r^h) e^h + r^l - W_r^h(\theta_r^h) e^t\right\|_2$	SGD	$O(2(1-\hat{\theta})mnN_t)$	$O(N_e m + N_r(1-\hat{\theta})\left(m+1\right)n)$
LFM	$\left\langle y \middle\| R^l \middle\| y' \right\rangle + \left\langle e^h \middle\| R^l \middle\| z \right\rangle + \left\langle z \middle\| R^l \middle\| e^t \right\rangle + \left\langle e^h \middle\| R^l \middle\| e^t \right\rangle$	SGD	$O(N_e m + N_r n^2)$	$O(N_e m + N_r n + 10n^2)$
CirE	$\left\|e^h + r^l - e^t\right\|_2$	SGD	$O\left(m \log m N_t\right)$	$O\left(N_e m + 2N_r n\right)$

多、多对一和多对多等关系模型无法有效处理, 尤其是一对多和多对多的关系在这些模型上的预测精度并不十分理想.

还有一些方法 [96~109] 主要集中在改进嵌入学习后更好地表达词向量间的语义距离, 一般是采取引入额外信息的方式来增加翻译模型的翻译效果, 使词语间的平移不变性更好地得到体现. Fan 等 [110] 在嵌入学习中将实体和提及一起进行训练来补充语义信息. Choi 等 [111] 引入了 Web 中的统计信息和知识库的模式信息. Huang 等 [112] 则引入了 PRA 方法中的路径可能性信息. Guo 等 [113] 则是在训练中加入了基数规则以增加出度和入度的约束信息. Yang 等 [57] 通过定义逻辑规则来增加逻辑约束信息. Zhao 等 [114] 通过自然语言中的词语对概念来利用其中的语义信息. Oh 等 [115] 利用实体及其周围邻居的拓扑结构信息来增强其学习能力. Niu 等 [116] 则在采样上定义了过滤原则来提高正负样本的采样效果.

除了上述翻译模型之外, 采用能量方程和神经网络方法的比较典型的学习模型还有语义匹配能量模型 SME [23]、非结构化模型 UM、潜在因素模型 LFM [117]、循环矩阵模型 CirE [118]、胶囊网络模型 [119] 等. SENN [120] 则将头实体、尾实体和关系三者分开, 在共享向量的同时做分别的预测学习即同时训练不同的 3 个神经网络. 表 3 中对以上基本模型从评价函数进行了必要的比较, 其中 N_e, N_r 分别是实体和关系的数量, N_t 是知识图谱中三元组的数量, e^h, e^t 分别表示头实体和尾实体向量, m, n 分别表示实体和关系在嵌入空间中的维数, d 为关系的平均聚类数量, s 为张量的片数, k 为神经网络中隐层中的结点数量.

4.3 其他补全方法

4.3.1 跨知识库补全方法

前文也提到随着知识库的规模变大, 不同知识库应运而生, 而没有任何一个知识库是完备的, 单一知识库的补全更无法做到完备, 因此跨知识库的补全过程不可避免, 这也是知识融合的必然趋势和要求.

跨知识库的补全需要在不同的知识库中找到对方没有的知识. 比较有针对性的方法是 He 等 [121] 在矩阵分解 (matrix factorization) 的基础上提出对不同知识库进行知识库补全, 该方法将三元组中的关系描述为谓词, 而谓词对主体和客体都有类型要求, 比如谓词 BirthIn 就会要求主体是人, 客体是地理位置或名称, 如果不满足类型要求就说明该谓词无法正确联系主客体, 此处等同于增加了谓词的类别约束. 尽管知识库在构建时无法做到从定义到实现的完全一致, 但是它们都会遵从三元组和知识模型的基本框架, 因此可以比较容易地得到谓词的类别描述和谓词的特征, 从而对不同知识库中的谓词相似性进行判断, 该方法使用式 (9) 来衡量谓词的相似性:

$$\min_{U} \sum_{i=1}^{m} \sum_{f \in S(i)} \operatorname{Sim}(i,f) \left\| U_i - U_f \right\|_F^2, \tag{9}$$

其中 $S(i)$ 表示与谓词 ri 相似的谓词集合, $\operatorname{Sim}(i,f)$ 是用来描述两个谓词 ri 和 rf 之间的相似度函数, 定义如下:

$$\operatorname{Sim}(i,j) = \frac{\sum_{j \in I(i) \cap I(f)} r_{ij} \cdot r_{fj}}{\sqrt{\sum_{j \in I(i) \cap I(f)} {r_{ij}}^2 \cdot \sum_{j \in I(i) \cap I(f)} {r_{fj}}^2}},$$

而其目标函数定义如下:

$$\begin{aligned} \min_{U,V}(X,U,V) =& \frac{1}{2} \sum_{r=1}^{m} \sum_{e=1}^{n} I_{er} \left(X_{er} - g\left(U_r^{\mathrm{T}} V_e \right) \right) \\ &+ \frac{\varphi}{2} \sum_{i=1}^{m} \sum_{f \in S(i)} \operatorname{Sim}(i,f) \left\| U_i - U_f \right\|_F^2 + \frac{\lambda_1}{2} \left\| U \right\|_F^2 + \frac{\lambda_2}{2} \left\| V \right\|_F^2, \end{aligned}$$

其中的 φ 是相似性度量的权重. 该方法在衡量相似谓词后可以发现相似谓词在不容知识库中的不同三元组, 从而在类别约束下发现丢失的知识, 即不存在的三元组. 该方法在知识库 DBpedia3.9 和 YAGO -2s 上进行了测试并取得了较好的效果.

4.3.2 基于信息检索技术的知识库补全方法

信息检索 (information retrieval, IR) 技术由来已久, 近年来发展快速的智能问答系统就需要 IR 技术的支持, 它可以从 Web 上获取与问题相关的知识 [122], 比如文本、图片等. 而这些信息中就会包含很多隐含的知识, 比如当你问一个人的出生地时, 返回的结果可能不仅仅有他的出生地信息, 甚至还会有他家人的信息、他的学校、他的社区等. 据此, West 等 [123] 在基于 Freebase 的问答系统上开发了一套基于问答搜索的知识库补全系统, 其基本思想是针对某类或某个关系进行实体发现, 从而建立起知识即三元组表示. 该系统主要由 4 部分组成. (1) 线下训练: 该部分主要是通过对查询的分析来构建关系的模板, 从而发现相似的关系并对这些模板的质量进行评估; (2) 查询模板选择: 使用 Heatmap 技术发现查询模板与相关知识间的关联性, 从而获得高效的查询模板; (3) 查询结果处理: 此时根据给定的关系生成查询模板, 然后对返回的结果即片段进行排序和分析, 从中找出有关的实体; (4) 结果处

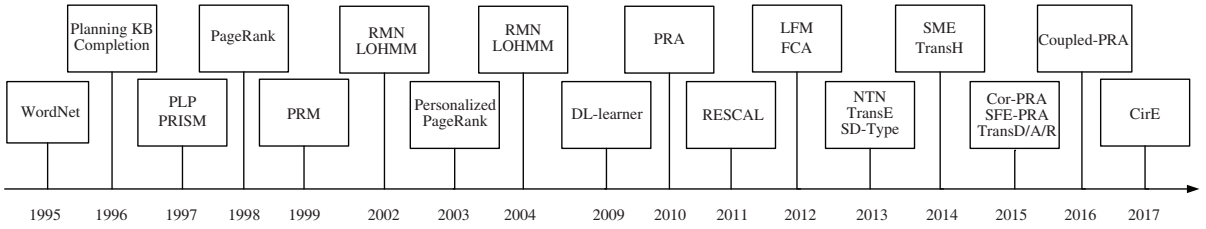


图 7 知识图谱补全技术发展时间轴

Figure 7 Time axis of the development of KB completion

理与纠正: 这是系统的最后一步, 需要对之前发现的实体集合进行集成和筛选, 利用关系类别等规则来完成判断和选择. 该方法实际上是将大量已知的信息检索、实体识别和关系发现等诸多方法综合运用, 在谷歌的 Freebase 框架下来完成的补全过程, 如果加上在其他知识库中的查询结果, 实际上也可以完成跨知识库的补全过程, 只不过其计算任务会更加复杂和艰巨.

4.3.3 知识库中的常识知识补全

现有的大多数知识库中的知识都是来自文本中提取的实体和关系组成的三元组, 但是人类的很多知识属于常识性知识, 并不会出现在文本之中. 针对此类问题, Angeli 等 [124] 提出来一个查询系统, 其能够通过将查询的事实与知识库上已有的事实通过相似性比较来发现潜在的事实以补全原有知识库, 不过该过程需要人工参与来完成查询的确认; 为了自动地发掘常识知识, Li 等 [125] 利用概念网络 (ConceptNet) 来寻找与某个限定关系相关的词汇, 据此生成该关系与其有关的词汇构成的三元组 (LeftTerm, Relation, RightTerm), 然后通过嵌入学习和深度学习中的 LSTM 模型来构建一个评价函数, 该函数可以针对未出现过的上述三元组队进行打分来确定其是否真实, 此种方法可以针对一个常识性的三元组 (概念网络中未出现的) 进行评价, 当然这需要从概念网络中生成大量训练数据, 这本身会对该方法造成一定的影响. 总之, 关于常识知识的产生和补全仍未达到人们的预期, 相关工作仍需大量人力参与.

5 总结

通过以上描述, 我们可以看到知识库补全技术的发展是与人们对知识的认知和知识表示方法的进步紧密结合在一起的, 其发展历程总结如图 7 所示. 随着大数据时代的到来, 海量知识的表示、存储和有关计算必然面临巨大的挑战, 对于知识图谱补全技术而言, 今后面临的挑战和主要发展的方向应该体现在如下几个方面:

(1) 实体和关系的稀疏性更加突出. 在统计学中, 数据的长尾现象本身就十分普遍, 为此统计学中会有很多方法来解决此种问题. 而在大规模知识图谱构建中, 也会出现此种长尾现象, 其主要体现在有很多实体和关系会反复出现, 频率很高, 比如美国总统或是体育明星, 关于此类实体的新闻或文章都是十分丰富的, 因此其相关的关系实例也会很多; 但是对于另外一些实体和关系却实例很少, 比如普通的民众, 虽然出现频率不高, 但是数量众多, 导致与其相关的关系实例也是十分稀疏, 而且在数据量不断增加的情况下, 这种情况会更加明显. 由前述知识图谱补全问题可以看到, 长尾实体的关系和属性实例很少, 会导致其关系和属性缺失问题严重, 会对知识图谱补全过程造成严重影响. 因此在大规模知识图谱补全问题上, 解决长尾实体及其关系的稀疏性会变得越来越迫切, 这可能需要借助信息检索、关系发现、实体链接等多种技术的结合来综合解决.

(2) 实体关系的一对多、多对一和多对多问题变得更加严重. 通过前面方法的介绍可以看到, 早期的知识图谱构建和补全技术都是以一对一的关系为基础来设计的, 随着知识的不断增加和丰富, 一对多、多对一和多对多的关系会越来越多, 这在针对领域知识图谱的构建中会十分突出. 比如在生命科学领域, 某一种基因会和几百甚至上千种蛋白质相关, 某一反应路径会重复出现在成千上万组反应序列中, 某些属性会在大多数的基因组中出现, 此时的一对多、多对一和多对多关系相较于开放领域中的问题而言, 不是一对十几或几十数量级那么简单, 而是成百上千的数量级, 传统的解决方案无法有效甚至根本无法解决此种数量级别的关系学习问题, 这需要针对之前已存在解决方案的基础上增加新的控制变量和约束, 甚至需要提出完全不一样的解决思路或表示模型.

(3) 三元组的动态增加和变化导致的知识图谱的动态变化加剧 [126]. 在大规模知识图谱中, 新的知识会源源不断地产生出来以更新旧的知识, 尤其是我们之前认为正确的知识或不确定的知识可能会在后来证明是错误的或者需要修正改变, 那么在之前认识的基础上完成的知识补全过程也会需要修正改变, 由此而引发的知识图谱补全的更新过程也会变得十分复杂, 如何令知识图谱补全技术适应知识图谱持续加快的动态变化会变得越来越重要, 目前这方面的技术还未引起足够的重视.

(4) 知识图谱中的关系预测路径长度会不断增长. 在 PRA 等方法中我们会看到从一个实体到另一个实体的路径长度往往超过 1, 而现有的主要方法中该路径长度一般不会超过 4, 也就是说关系预测能推理的长度是有限的, 但是在大规模知识图谱上, 实体间的关系路径序列长度会变得越来越长, 比如某类疾病和某种基因间的关系序列长度往往会超过 5, 例如在微生物数据集上, 微生物的某个物种与基因组间的关系路径长度往往会大于 4, 甚至有可能会达到 6, 所以知识图谱补全过程需要更高效的模型来描述更复杂的关系预测模型.

总之, 大规模知识图谱补全技术会伴随着知识图谱构建技术的发展而不断前进, 虽然在知识不断积累变化的过程中会遇到新的问题和挑战, 但其最终会结合人类认知领域和知识工程的技术不断完善和发展.

参考文献

1 Lenat D B. CYC: a large-scale investment in knowledge infrastructure. Commun ACM, 1995, 38: 33–38
2 Miller G A. WordNet: a lexical database for English. Commun ACM, 1995, 38: 39–41
3 Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res, 2004, 32: 267–270
4 Bollacker K D, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of International Conference on Management of Data, 2008. 1247–1250
5 Fan J, Ferrucci D A, Gondek D, et al. PRISMATIC: inducing knowledge from a large scale lexicalized relation resource. In: Proceedings of North American Chapter of the Association for Computational Linguistics, 2010. 122–127
6 Fader A, Soderland S, Etzioni O. Identifying relations for open information extraction. In: Proceedings of Empirical Methods in Natural Language Processing, 2011. 1535–1545
7 Nakashole N, Weikum G, Suchanek F M. PATTY: a taxonomy of relational patterns with semantic types. In: Proceedings of Empirical Methods in Natural Language Processing, 2012. 1135–1145
8 Niu F, Zhang C, Ré C, et al. Elementary: large-scale knowledge-base construction via machine learning and statistical inference. Int J Semantic Web Inf Syst, 2012, 8: 42–73
9 Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs. Proc IEEE, 2016, 104: 11–33
10 Schmitz M, Soderland S, Bart R, et al. Open language learning for information extraction. In: Proceedings of Empirical Methods in Natural Language Processing, 2012. 523–534

11 Hoffart J, Suchanek F M, Berberich K, et al. YAGO2: a spatially and temporally enhanced knowledge base from wikipedia. Artif Intell, 2013, 194: 28–61

12 Dong X L, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In: Proceedings of Knowledge Discovery and Data Mining, 2014. 601–610

13 Nemoto Y, Akasaka F, Chiba R, et al. Establishment of a function embodiment knowledge base for supporting service design. Sci China Inf Sci, 2012, 55: 1008–1018

14 Völker J, Niepert M. Statistical schema induction. In: The Semantic Web: Research and Applications. Berlin: Springer, 2011. 124–138

15 Lehmann J. DL-Learner: learning concepts in description logics. J Mach Learn Res, 2009, 10: 2639–2642

16 Gangemi A, Nuzzolese A G, Presutti V, et al. Automatic typing of DBpedia entities. In: Proceedings of International Semantic Web Conference, 2012. 65–81

17 Chien S. Static and completion analysis for planning knowledge base development and verification. In: Proceedings of the 3rd International Conference on Artificial Intelligence Planning Systems, Edinburgh, 1996. 53–61

18 Chien S A. Static and completion analysis for knowledge acquisition, validation and maintenance of planning knowledge bases. Int J Human-Comput Studies, 1998, 48: 499–519

19 Baader F, Ganter B, Sertkaya B, et al. Completing description logic knowledge bases using formal concept analysis. In: Proceedings of International Joint Conference on Artificial Intelligence, 2007. 230–235

20 Sertkaya B. Explaining user errors in knowledge base completion. In: Proceedings of the 21st International Workshop on Description Logics (DL2008), Dresden, 2008

21 Baader F, Sertkaya B. Usability issues in description logic knowledge base completion. In: Proceedings of the 7th International Conference of Formal Concept Analysis (ICFCA 2009), Darmstadt, 2009. 1–21

22 Paulheim H, Bizer C. Type inference on noisy RDF data. In: Proceedings of International Semantic Web Conference. Berlin: Springer, 2013. 510–525

23 Zhang L, Zhang X, Zhao L, et al. An embedding-based system to constructing OWL ontologies. In: Proceedings of the 16th International Semantic Web Conference, California, 2018

24 Nuzzolese A G, Gangemi A, Presutti V, et al. Type inference through the analysis of wikipedia links. In: Proceedings of the LDOW 2012, 2012

25 Wu T, Ling S, Qi G, et al. Mining type information from Chinese online encyclopedias. In: Proceedings of the 4th Joint International Conference, 2014. 213–229

26 Kelloumenouer K, Kedad Z. Discovering types in RDF datasets. In: Proceedings of International Semantic Web Conference, 2015. 77–81

27 Ma C, Yan D, Wang Y P, et al. Advanced graph model for tainted variable tracking. Sci China Ser F-Inf Sci, 2013, 56: 112105

28 Wang P, Xu B W, Wu Y R, et al. Link prediction in social networks: the state-of-the-art. Sci China Ser F-Inf Sci, 2015, 58: 011101

29 Khosravi H, Bina B. A survey on statistical relational learning. In: Advances in Artificial Intelligence. Berlin: Springer, 2010. 256–268

30 Getoor L, Mihalkova L. Learning statistical models from relational data. In: Proceedings of International Conference on Management of Data, 2011. 1195–1198

31 Koller D. Probabilistic relational models. In: Lecture Notes in Computer Science. Berlin: Springer, 1999

32 Ngo L, Haddawy P. Answering queries from context-sensitive probabilistic knowledge bases. Theor Comput Sci, 1997, 171: 147–177

33 Kersting K, de Raedt L. Adaptive Bayesian logic programs. In: Proceedings of International Conference on Inductive Logic Programming, 2001. 104–117

34 Taskar B, Abbeel P, Koller D. Discriminative probabilistic models for relational data. In: Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence, 2002. 485–492

35 Richardson M, Domingos P. Markov logic networks. Mach Learn, 2006, 62: 107–136

36 Muggleton S. Stochastic logic programs. In: Proceedings of the 5th International Workshop on Inductive Logic Programming. Amsterdam: IOS Press, 1996. 254–264

37 Sato T, Kameya Y. PRISM: a symbolic-statistical modeling language. In: Proceedings of 15th International Joint

Conference on Artifical Intelligence, 1997. 1330–1339
38 Kersting K, Raiko T, Kramer S, et al. Towards discovering structural signatures of protein folds based on logical hidden Markov models. Biocomputing, 2003, 2003: 192–203
39 Dietrich D, Schulz E. Relational learning for collective classification of entities in images. In: Proceedings of Association for the Advancement of Artificial Intelligence Workshop on Statistical Relational Ai, 2010. 79–110
40 Rettinger A, Nickles M, Tresp V. Statistical relational learning of trust. Mach Learn, 2011, 82: 191–209
41 Rios M, Specia L, Gelbukh A, et al. Statistical relational learning to recognise textual entailment. In: Proceedings of International Conference on Computational Linguistics and Intelligent Text Processing. New York: Springer, 2014. 330–339
42 Wang W Y, Cohen W W. Joint information extraction and reasoning: a scalable statistical relational learning approach. In: Proceedings of Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing, 2015. 355–364
43 Yang S, Korayem M, Aljadda K, et al. Application of statistical relational learning to hybrid recommendation systems. 2016. ArXiv: 1607.01050
44 Natarajan S, Soni A, Wazalwar A, et al. Deep distant supervision: learning statistical relational models for weak supervision in natural language extraction. In: Solving Large Scale Learning Tasks. Challenges and Algorithms. Berlin: Springer, 2016
45 Yang S, Korayem M, AlJadda K, et al. Combining content-based and collaborative filtering for job recommendation system: a cost-sensitive statistical relational learning approach. Knowledge-Based Syst, 2017, 136: 37–45
46 Stefano T. Statistical Relational Learning for Proteomics: Function, Interactions and Evolution. Dissertation for Ph.D. Degree. Trento: University of Trento, 2013
47 Montoya L A, Pluth M D. Hydrogen sulfide deactivates common nitrobenzofurazan-based fluorescent thiol labeling reagents. Anal Chem, 2014, 86: 6032–6039
48 Cilia E, Teso S, Ammendola S, et al. Predicting virus mutations through statistical relational learning. BMC BioInf, 2014, 15: 309
49 Renkens J, Shterionov D, Broeck G V D, et al. ProbLog2: from probabilistic programming to statistical relational learning. In: Proceedings of the NIPS Probabilistic Programming Workshop, 2012
50 Farnadi G. Statistical relational learning towards modelling social media users. In: Proceedings of International Conference on Artificial Intelligence, Buenos Aires, 2015. 4365–4366
51 Farnadi G, Bach S H, Blondeel M, et al. Statistical relational learning with soft quantifiers. In: Proceedings of International Conference on Inductive Logic Programming. Berlin: Springer, 2015. 60–75
52 Farnadi G, Bach S H, Moens M F, et al. Soft quantification in statistical relational learning. Mach Learn, 2017, 106: 1971–1991
53 Popescul R, Ungar L H. Statistical relational learning for link prediction. In: Proceedings of the Workshop on Learning Statistical Models from Relational Data at IJCAI-2003, 2003
54 Wei Z, Zhao J, Liu K, et al. Large-scale knowledge base completion: inferring via grounding network sampling over selected instances. In: Proceedings of Conference on Information and Knowledge Management, 2015. 1331–1340
55 Minervini P, d'Amato C, Fanizzi N. Efficient energy-based embedding models for link prediction in knowledge graphs. J Intell Inf Syst, 2016, 47: 91–109
56 Galárraga L, Razniewski S, Amarilli A, et al. Predicting completeness in knowledge bases. In: Proceedings of the 10th ACM International Conference on Web Search and Data Mining, Cambridge, 2017. 375–383
57 Yang F, Yang Z, Cohen W W. Differentiable learning of logical rules for knowledge base completion. CoRR abs/1702.08367 (2017)
58 Aggarwal C C, Xie Y, Yu P S. On dynamic link inference in heterogeneous networks. In: Proceedings of SIAM International Conference on Data Mining, 2012. 415–426
59 Kliegr T, Zamazal O. Towards linked hypernyms dataset 2.0: complementing DBpedia with hypernym discovery and statistical type inference. In: Proceedings of Language Resources and Evaluation Conference, 2014
60 Davis J, Ong I M, Struyf J, et al. Change of representation for statistical relational learning. In: Proceedings of International Joint Conference on Artificial Intelligence, 2007. 2719–2726
61 Rossi R A, Mcdowell L K, Aha D W, et al. Transforming graph representations for statistical relational learning.

2012. ArXiv: 1204.0033
62 Schulte O, Khosravi H, Kirkpatrick A E, et al. Modelling relational statistics with Bayes nets. Mach Learn, 2014, 94: 105–125
63 Neelakantan A, Chang M. Inferring missing entity type instances for knowledge base completion: new dataset and methods. In: Proceedings of North American Chapter of the Association for Computational Linguistics, 2015. 515–525
64 Miao Q, Fang R, Song S, et al. Automatic identifying entity type in linked data. In: Proceedings of Pacific Asia Conference on Language, Information and Computation (PACLIC 30), 2016. 383–390
65 Xu B, Zhang Y, Liang J, et al. Cross-lingual type inference. In: Proceedings of International Conference on Database Systems for Advanced Applications. Berlin: Springer, 2016. 447–462
66 Kirschnick J, Hemsen H, Markl V. JEDI: joint entity and relation detection using type inference. In: Proceedings of Acl-2016 System Demonstrations, 2016. 61–66
67 Zhu H Y, Zeng Y, Wang D S, et al. Relation inference and type identification based on brain knowledge graph. In: Proceedings of International Conference on Brain and Health Informatics. Berlin: Springer, 2016. 221–230
68 Kuhn P, Mischkewitz S, Ring N, et al. Type inference on wikipedia list pages. In: Lecture Notes in Informatics. Berlin: Springer, 2016. 2010–2111
69 Zhou H, Zouaq A, Inkpen D. DBpedia entity type detection using entity embeddings and N-Gram models. In: Proceedings of International Conference on Knowledge Engineering and the Semantic Web, 2017. 309–322
70 Abhishek, Anand A, Awekar A. Fine-grained entity type classification by jointly learning representations and label embeddings. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. 797–807
71 Shimaoka S, Stenetorp P, Inui K, et al. Neural architectures for fine-grained entity type classification. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. 1271–1280
72 Murty S, Verga P, Vilnis L, et al. Finer grained entity typing with TypeNet. In: Proceedings of Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, 2017
73 Page L, Brin S, Motwani R, et al. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford: Stanford University, 1998
74 Haveliwala T, Kamvar A, Jeh G. An analytical comparison of approaches to personalizing PageRank. Stanford, 2003
75 Diligenti M, Gori M, Maggini M. Learning web page scores by error back-propagation. In: Proceedings of International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc. 2005. 684–689
76 Minkov E, Cohen W W. Learning graph walk based similarity measures for parsed text. In: Proceedings of Empirical Methods in Natural Language Processing, 2008. 907–916
77 Lao N, Cohen W W. Relational retrieval using a combination of path-constrained random walks. In: Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery, 2010. 81: 53–67
78 Lao N, Mitchell T M, Cohen W W. Random walk inference and learning in a large scale knowledge base. In: Proceedings of Empirical Methods in Natural Language Processing, 2011. 529–539
79 Lao N, Minkov E, Cohen W W. Learning relational features with backward random walks. In: Proceedings of Meeting of the Association for Computational Linguistics, 2015. 666–675
80 Gardner M, Mitchell T M. Efficient and expressive knowledge base completion using subgraph feature extraction. In: Proceedings of Empirical Methods in Natural Language Processing, 2015. 1488–1498
81 Wang Q, Liu J, Luo Y, et al. Knowledge base completion via coupled path ranking. In: Proceedings of Meeting of the Association for Computational Linguistics, 2016. 1308–1318
82 Lao N, Subramanya A, Pereira F, et al. Reading the web with learned syntactic-semantic inference rules. In: Proceedings of Empirical Methods in Natural Language Processing, 2012. 1017–1026
83 Gardner M, Talukdar P P, Krishnamurthy J, et al. Incorporating vector space similarity in random walk inference over knowledge bases. In: Proceedings of Empirical Methods in Natural Language Processing, 2014. 397–406
84 Lin X, Liang Y, Guan R. Compositional learning of relation paths embedding for knowledge base completion. CoRR abs/1611.07232 (2016)
85 Gardner M, Talukdar P P, Kisiel B, et al. Improving learning and inference in a large knowledge-base using latent syntactic cues. In: Proceedings of Empirical Methods in Natural Language Processing, 2013

86 Shi B, Weninger T. Discriminative predicate path mining for fact checking in knowledge graphs. Knowledge-Based Syst, 2016, 104: 123–133
87 Bordes A, Weston J, Collobert R, et al. Learning structured embeddings of knowledge bases. In: Proceedings of National Conference on Artificial Intelligence, 2011. 301–306
88 Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion. In: Proceedings of Neural Information Processing Systems, 2013. 926–934
89 Nickel M, Tresp V, Kriegel H. A three-way model for collective learning on multi-relational data. In: Proceedings of International Conference on Machine Learning, 2011. 809–816
90 Balazevic I, Allen C, Hospedales T M. TuckER: tensor factorization for knowledge graph completion. 2019. ArXiv: 1901.09590
91 Bordes A, Usunier N, Garciaduran A, et al. Translating embeddings for modeling multi-relational data. In: Proceedings of Neural Information Processing Systems, 2013. 2787–2795
92 Wang Z, Zhang J, Feng J, et al. Knowledge graph embedding by translating on hyperplanes. In: Proceedings of National Conference on Artificial Intelligence, 2014. 1112–1119
93 Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of National Conference on Artificial Intelligence, 2015. 2181–2187
94 Ji G, He S, Xu L, et al. Knowledge graph embedding via dynamic mapping matrix. In: Proceedings of Meeting of the Association for Computational Linguistics, 2015. 687–696
95 Xiao H, Huang M, Hao Y, et al. TransA: an adaptive approach for knowledge graph embedding. 2015. ArXiv: 1509.05490
96 Minervini P, d'Amato C, Fanizzi N, et al. Efficient learning of entity and predicate embeddings for link prediction in knowledge graphs. In: Proceedings of the 11th International Workshop on Uncertainty Reasoning for the Semantic Web (URSW 2015) Co-Located With the 14th International Semantic Web Conference (ISWC 2015), Bethlehem, 2015. 26–37
97 Garcia-Duran A, Bordes A, Usunier N, et al. Combining two and three-way embedding models for link prediction in knowledge bases. J Artifical Intell Res, 2016, 55: 715–742
98 Neelakantan A, Roth B, Mccallum A. Compositional vector space models for knowledge base completion. In: Proceedings of Meeting of the Association for Computational Linguistics, 2015. 156–166
99 Yang B, Yih W, He X, et al. Embedding entities and relations for learning and inference in knowledge bases. In: Proceedings of International Conference on Learning Representations, 2015
100 Wang Q, Wang B, Guo L. Knowledge base completion using embeddings and rules. In: Proceedings of International Joint Conference on Artificial Intelligence, 2015. 1859–1865
101 Aiguier M, Atif J, Bloch I, et al. Some algebraic results in description logics: free model and inclusions, finite basis theorem, and completion of knowledge bases. 2015. ArXiv: 1502.07634v2
102 Feng J, Huang M, Wang M, et al. Knowledge graph embedding by flexible translation. In: Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning, Cape Town, 2016. 557–560
103 Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data. Mach Learn, 2014, 94: 233–259
104 Xiao H, Huang M, Zhu X. From one point to a manifold: knowledge graph embedding for precise link prediction. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, 2016. 1315–1321
105 Li M M, Jia Y, Wang Y, et al. Hierarchy-based link prediction in knowledge graphs. In: Proceedings of International World Wide Web Conference, 2016. 77–78
106 Shijia E, Jia S, Xiang Y, et al. Knowledge graph embedding for link prediction and triplet classification. In: Proceedings of the 1st China Conference, Beijing, 2016. 228–232
107 Nguyen D Q, Sirts K, Qu L, et al. Neighborhood mixture model for knowledge base completion. In: Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, Berlin, 2016. 40–50
108 Tay Y, Luu T A, Hui C S. Non-parametric estimation of multiple embeddings for link prediction on dynamic knowledge graphs. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, 2017. 1243–1249
109 Hayashi K, Shimbo M. On the equivalence of holographic and complex embeddings for link prediction. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 2: 554–559

110 Fan M, Zhou Q, Abel A, et al. Probabilistic belief embedding for knowledge base completion. 2015. ArXiv:1505.02433
111 Choi S J, Song H J, Yoon H G, et al. A re-ranking model for accurate knowledge base completion with knowledge base schema and web statistic. In: Proceedings of IEEE Congress on Evolutionary Computation, CEC 2016, Vancouver, 2016. 4958–4964
112 Huang W, Li G, Jin Z. Improved knowledge base completion by the path-augmented TransR model. In: Proceedings of International Conference on Knowledge Science, Engineering and Management. Berlin: Springer, 2017. 149–159
113 Guo S, Ding B, Wang Q, et al. Knowledge base completion via rule-enhanced relational learning. In: Proceedings of the 1st China Conference (CCKS 2016), Beijing, 2016. 219–227
114 Zhao Y, Gao S, Gallinari P, et al. Knowledge base completion by learning pairwise-interaction differentiated embeddings. Data Min Knowl Disc, 2015, 29: 1486–1504
115 Oh B, Seo S, Lee K. Knowledge graph completion by context-aware convolutional learning with multi-hop neighborhoods. In: Proceedings of Conference on Information and Knowledge Management, 2018. 257–266
116 Niu J, Sun Z, Zhang W. Enhancing knowledge graph completion with positive unlabeled learning. In: Proceedings of International Conference on Pattern Recognition, 2018. 296–301
117 Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data. In: Proceedings of Neural Information Processing Systems, 2012. 3167–3175
118 Du Z J, Hao Z H, Meng X F, et al. CirE: circular embeddings of knowledge graphs. In: Proceedings of International Conference on Database Systems for Advanced Applications, 2017. 148–162
119 Nguyen D Q, Vu T, Nguyen T D, et al. A capsule network-based embedding model for knowledge graph completion and search personalization. 2018. ArXiv: 1808.04122
120 Guan S, Jin X, Wang Y, et al. Shared embedding based neural networks for knowledge graph completion. In: Proceedings of Conference on Information and Knowledge Management, 2018. 247–256
121 He W, Feng Y, Zou L, et al. Knowledge base completion using matrix factorization. In: Proceedings of Asia-Pacific Web Conference, 2015. 256–267
122 Bing L D, Zhang Z M, Lam W, et al. Towards a language-independent solution: knowledge base completion by searching the web and deriving language pattern. Knowl Based Syst, 2016, 115: 80–86
123 West R B, Gabrilovich E, Murphy K, et al. Knowledge base completion via search-based question answering. In: Proceedings of International World Wide Web Conference, 2014. 515–526
124 Angeli G, Manning C D. Philosophers are mortal: inferring the truth of unseen facts. In: Proceedings of Conference on Computational Natural Language Learning, 2013. 133–142
125 Li X, Taheri A, Tu L, et al. Commonsense knowledge base completion. In: Proceedings of Meeting of the Association for Computational Linguistics, 2016. 1445–1455
126 Luan S M, Dai G Z, Li W. A programmable approach to revising knowledge bases. Sci China Ser F-Inf Sci, 2005, 48: 681–692

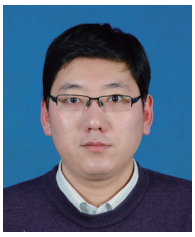
Research progress of large-scale knowledge graph completion technology

Shuo WANG[1,2*], Zhijuan DU[1] & Xiaofeng MENG[1*]

1. *Information School, Renmin University of China, Beijing* 100872*, China*;
2. *Key Laboratory of Machine Learning and Computational Intelligence, Hebei University, Baoding* 071002*, China*
* Corresponding author. E-mail: wsterran@126.com, xfmeng@ruc.edu.cn

Abstract With the continued growth of various knowledge graphs, such as Google Knowledge Map, DBpedia, Microsoft Concept Graph, and YAGO, the knowledge representation system, constructed based on RDF, has become more well-known. The RDF triple format has become the basic description of knowledge in the real world. Due to its simple structure and clear logic, it is easy to understand and implement. Nevertheless, when faced with extremely complicated knowledge and common sense, complete knowledge can become difficult to describe. The construction process of knowledge graphs is bound to lead to incomplete knowledge contained in the graphs. At this point, the knowledge-based completion technology is particularly important for managing such situations. Any existing knowledge graph must be improved continuously through completion technology and newly inferred knowledge. Beginning with the construction of a knowledge graph, this paper divides the problem of knowledge graph completion into two levels: concept completion and instance completion. (1) The concept completion level primarily focuses on the completion of entity types. It is described in terms of three development stages: a logical reasoning mechanism, based on description logic, a type inference mechanism, based on traditional machine learning, and a type inference mechanism, based on representation learning. (2) The instance completion level can be further divided into an RDF triple completion and new instance discovery. This paper focuses on RDF triples completion learning, which includes entity completion or relationship completion and is described in three development stages, such as statistical relational learning, probability learning based on random walks, and knowledge representation learning. Through the review and discussion of the research process, the development status, and the latest progress in the above-mentioned large-scale knowledge graph completion, we present the challenges that the technology will face and the development prospects of future work.

Keywords knowledge graph, knowledge base completion, concept completion, instance completion



Shuo WANG was born in 1981. He is a Ph.D. candidate at the Renmin University of China and a teacher in the Key Laboratory of Machine Learning and Computational Intelligence at Hebei University. His research interests include natural language processing, machine learning, data fusion, and knowledge fusion.

Xiaofeng MENG was born in 1964. He is a professor and Ph.D. supervisor at the Renmin University of China, as well as a fellow of the China Computer Federation. His main research interests include cloud data management, web data management, flash-based databases, and privacy protection.

Zhijuan DU was born in 1986. In 2018, she obtained her Ph.D. degree in computer software and theory from Renmin University of China, Beijing. Her research interests include web data management and cloud data management.