arXiv:2305.05403v2 [cs.AI] 6 Dec 2023

开放世界知识库中的完备性、召回率与否定：综述

SIMON RAZNIEWSKI,Bosch Center for AI，德国HIBA ARNAOUT,马克斯·普朗克信息学研究所，德国SHRESTHA GHOSH, 马克斯·普朗克信息学研究所，德国FABIAN SUCHANEK,巴黎电信学院，巴黎综合理工学院，法国

通用知识库（KB）是以知识为中心的人工智能的重要基石。其中许多知识库是通过网络来源以务实的方式构建的，因此远未达到完整状态。这给其内容的使用和整理都带来了挑战。尽管已有多项综述研究旨在解决不完整知识库的补全问题，但可以说，首要问题在于确定知识库是否存在不完整的情况、具体在哪些方面不完整，以及其不完整程度究竟如何。

在这篇综述中，我们讨论如何表达、抽取和推断有关知识库完备性、召回率和否定的知识。我们涵盖：(i) 部分闭世界语义下知识表示和查询的逻辑基础；(ii) 通过统计模式估算这类信息；(iii) 从知识库和文本中抽取有关召回率的信息；(iv) 识别有意义的否定陈述；以及 (v) 相对召回率的宽松概念。

本综述面向两类读者：(1) 希望跟踪知识库质量、聚焦抽取工作并构建质量感知型下游应用的实践者；(2) 希望了解开放世界假设之外知识库最新进展的数据管理、知识库和语义网研究人员。因此，本综述既介绍基础方法及其实际运作方式，也针对如何为具体问题选择不同方法提出面向实践的建议。

CCS 概念： • **通用与参照** → **调查与综述**； • **计算方法** →**知识表示与推理**； **人工智能**。

其他关键词和短语：知识库、数据完备性

ACM 参照格式：
Simon Razniewski、Hiba Arnaout、Shrestha Ghosh 和 Fabian Suchanek。2023年。开放世界知识库中的完备性、召回率与否定：一项综述。0，0，文章 0（2023年12月），42 页。

1 引言

动机。 维基数据 [124], DBpedia [13], NELL [21],或 YAGO [117] 等大规模知识库（KB）是语义网的基石。它们通常由网络资源务实地构建而成，重点在于表示肯定的知识，即真实的陈述。然而，它们通常只包含所有真实陈述中的一小部分，却没有说明这一子集究竟是什么。例如，某个知识库可能包含诺贝尔物理学奖的获奖者，但不一定包含所有获奖者。它甚至不会说明自己是否包含全部获奖者。例如，某个知识库可能缺少某位知名物理学家获得诺贝尔奖这一信息，但这并不意味着此人没有获奖——可能只是数据缺失。反过来，如果已知某位物理学家确实没有获得诺贝尔奖，那么在大多数

作者地址：Simon Razniewski，simon.razniewski@de.bosch.com，Bosch Center for AI，德国；Hiba Arnaout，harnaout@mpi-inf.mpg.de，马克斯·普朗克信息学研究所，德国；Shrestha Ghosh，ghoshs@mpi-inf.mpg.de，马克斯·普朗克信息学研究所，德国；Fabian Suchanek，巴黎电信学院、巴黎综合理工学院，法国，suchanek@telecom-paris.fr。

2023年。XXXX-XXXX/2023年/12-ART0 $15.00https://doi.org/

现有知识库中也没有办法表达这一点。知识库还可能完全不包含某位特定的科学家——却不表示该人物缺失。[1]

知识范围存在如此不确定性，给知识库的整理和应用带来了重大挑战：

(1) **知识库维护人员**需要了解知识库存在哪些不完整之处，以便优先安排补全工作。例如，在处理维基数据这类拥有超过 1 亿个条目的知识库时，系统地了解知识缺口和质量至关重要，这样才能知道该将有限的人力资源重点投入到哪里。

(2) **自动化知识库构建流程**也需要了解这些信息，以便知道如何调整接受阈值。例如，如果知识库中已有 218 位诺贝尔物理学奖得主，而且我们知道这就是预期数量，那么就应该拒绝所有后续候选者。相反，如果知识库中没有记录某位科学家的任何出版物，我们就应该接受自动化抽取的结果。这一点对于 NELL 等旨在自动补全自身的知识库项目尤其重要。

(3) **问答应用** 建立在知识库之上。它们需要知道数据在哪些地方不完整，以便提醒最终用户注意质量问题。例如，如果知识库中恰好缺少*GeoffreyMarcy*，那么查询“发现最多行星的天文学家”可能会返回错误答案。同样，用于问答的知识库应当能够意识到问题何时超出了其知识范围 [93]。对于“哈佛大学是否设有物理系？”这类布尔问题（回答“否”可能仅仅是因为缺少相关信息），以及“有多少所美国大学设有物理系？”这类分析型聚合问题（得到某个答案几乎无法说明其正确性），这一点尤其重要。

(4) **结构化实体摘要**需要了解的不仅是正向属性，还包括显著的负面信息。例如，瑞士最显著的属性之一，就是它并非欧盟 成员国（尽管它与欧盟国家经济联系紧密，且四周被欧盟国家环绕）。对于史蒂芬·霍金，他的荣誉摘要中应提及一件显著的事：他并未 获得所在领域的顶级诺贝尔物理学奖。

(5) **知识库和文本上的机器学习**，尤其是知识库链接预测和文本信息抽取任务，需要负训练样例来构建准确的模型。在处理这些任务时，获取高质量的负样例是一大难题，因此许多研究都聚焦于启发式方法，例如基于对正陈述进行混淆的方法。显式否定，或从完整性元数据推导出的否定，可能会对这些任务产生重大影响。

本综述介绍了近期文献中为解决这些问题而提出的方法。例如，研究人员已经开发出多种形式化方法，用于明确知识库中知识的覆盖范围，包括表达否定陈述的可能性，或标注完备性元数据。知识工程师可以记录他们已将所有诺贝尔物理学奖得主录入维基数据，也可以标记那些未曾获奖的杰出物理学家。如果无法进行这种人工标记，也可以采用自动推断知识库完备性的方法。例如，可以识别诸如“诺贝尔物理学奖共授予了*218*位人士”之类的语句，并将这一计数与知识库中的实体数量进行比较。还可以根据知识库中某一实体集合的增长情况或近乎停滞的状态、不同随机样本中诺贝尔奖得主集合之间的重叠程度，或不同知识库之间的重叠程度，推断知识库的完备性。此外，还可以观察到，大多数获奖者都拥有学位，因此可以将那些缺少母校信息的实体标记出来。

[1]我们不提供现实世界知识库中不完整实体的示例，因为在以往的出版物中，每当我们这样做时，热心的读者都会专门补全知识库中的这些实体，从而使我们的示例过时。

mater 可能是不完备的。因此，知识工程师手头有了日益丰富的方法库来应对完备性问题，而这正是本文综述所涵盖的内容。

本综述的重点。多篇全面的综述广泛讨论了知识库，尤其关注其构建、存储、查询和应用 [50, 55, 126]。关于知识库质量的专题研究探讨了如何提高召回率，以及如何通过评估准确性和来源来衡量质量 [17, 35, 85, 129]。近年来，还出现了用于描述召回率和否定知识的新型形式化方法 [10, 26, 71], ，以及用于估算召回率的统计方法和基于文本的方法 [15, 36, 52, 58, 61, 67, 97, 114] 和推导否定陈述的方法 [8, 12, 58, 108],，其中部分研究成果已被纳入一项系统性文献综述 [53]。本综述旨在对这一主题进行系统梳理、介绍和分析。本综述面向理论研究者和实践者，旨在向读者介绍完备性评估与否定领域的最新进展，并为他们提供一套方法，以便更好地表示和评估特定数据集的召回率。本综述基于在 VLDB'21、ISWC'21、KR'21 和 WWW'22 教程中介绍的内容。教程幻灯片可在以下网址获取：https://www.mpi-inf.mpg.de/knowledge-base-recall/tutorials。

大纲。本文综述的结构如下：第 2 节首先介绍知识库运行所依据的逻辑框架的基础、开放世界假设（OWA）、不完备性及其对查询回答的影响 [102], ，以及完备性和基数断言的形式语义 [26]。第 3 节讨论如何自动估算召回率。我们介绍监督式机器学习方法 [36], 、物种抽样等无监督技术 [67, 122], 、基于密度的估计量 [61], 、关于数量分布的统计不变量 [114],，以及关于人类对话中完备性的语言学理论 [97]。第 4 节重点介绍如何从知识库和文本中确定谓词的基数（即给定主体时对应的对象数量）。我们说明如何识别并利用知识库中的基数信息来评估召回率 [42], ，如何从自然语言文档中抽取这类信息 [71],，以及这与挖掘函数基数有何不同；函数基数对任意主体都保持不变（例如，所有人类都只有一个出生地） [43]。第 5 节重点介绍如何识别显著否定。我们说明为什么在开放世界环境中需要针对知识库实体的显式否定。我们介绍利用知识库自身推理结果识别否定的方法 [8, 11, 12, 107, 108]，以及从各种文本来源（尤其是查询日志）中抽取否定的方法 [8, 58], 。此外，我们还概述一些尚待解决的问题，例如精确率与显著性的权衡，以及本体建模与维护。第 6 节讨论相对召回率，即较为宽松的召回率概念，不以捕获现实世界的全部知识为目标。我们说明如何通过问答和实体摘要等外部使用场景衡量召回率 [52, 96], ，如何通过与开放信息抽取或外部参考资源进行比较来衡量 [39, 73], ，以及如何通过与知识库内部其他可比实体进行比较来衡量 [15, 70]。第 7 节总结近期议题、提高知识库召回率感知能力的建议，以及尚待解决的研究挑战。特别是，大语言模型（LLM）近期给以知识为中心的计算带来了巨大变化。尽管完备性和召回率研究尚未充分利用这些进展，我们仍指出大语言模型可能影响这一领域的若干途径。

2 基础

2.1 知识库

我们首先介绍基础概念，尤其是知识库、其语义，以及完备性、召回率和基数元数据的概念。知识库建立在三个两两不相交的无限集合 E（实体）、P（谓词）和 L（字面量）之上。实体指的是可唯一识别的对象，例如 玛丽·居里、诺贝尔物理学奖或巴黎这座城市。字面量是

玛丽·居里	
出生地	华沙
出生年份	1867
公民身份	波兰
指导	玛格丽特·佩雷 Óscar Moreno ..... </style》（另有 6 个姓名）.....
发现	钋
获得奖项	诺贝尔物理学奖 富兰克林奖章

表 1. 知识库运行示例。

字符串或数值，例如日期、重量和名称。谓词（也称为关系、属性或特征）将一个实体与另一个实体或字面量相连接。例如出生地或原子序数。元组 $s, p, o \in E \times P \times (E \cup L)$ 称为正向陈述（也称为三元组、断言或事实），其中 s 是主体，p 是谓词，o 是对象 [118]。陈述表示主体与对象具有给定谓词所表示的关系，例如 玛丽·居里, 获得奖项, 诺贝尔物理学奖。就本文而言，我们也会考虑否定陈述，它表示主体与对象不具有给定的关系，例如 ¬史蒂芬·霍金, 获得奖项, 诺贝尔物理学奖。实体可以归入诸如物理学家、城市或奖项等类。实体可以是某个类的实例，这可以用谓词“类型”构成的三元组表示，例如 玛丽·居里, 类型, 物理学家。类可以按子类层次结构组织。例如，物理学家是科学家，而科学家是人。这同样可以用三元组表示，并使用特殊谓词子类：物理学家, 子类, 科学家。某些知识库允许指定约束，例如谓词的定义域和值域约束、函数性约束，或类之间的不相交约束，通常使用一种称为描述逻辑的形式体系表示 [14]。在本文中，我们主要关注陈述和类别成员关系，而不太关注约束（也就是说，按照描述逻辑的术语，我们重点关注 A盒中的知识的断言部分，而较少关注 T盒中的术语知识）。

定义 1（知识库）。 知识库（或知识图谱）K 是由有限个正向陈述组成的集合 [60]。

> **示例：** 举例来说，考虑一个包含科学家传记数据的知识库。表1展示了该知识库的一个摘录，采用维基数据风格的布局，其中所有陈述的主体显示在顶部（玛丽·居里），谓词和对象则以表格形式显示在下方。该知识库包含各种各样的陈述，将实体与其他实体以及字面量联系起来；对于某些谓词，其中包含多个对象，但并非所有谓词都如此。

“知识图谱”（KG）一词也常被用作“知识库”的替代说法——自从 Google 在 2012 [110]推广这一术语以来。尽管有时会将知识图谱定义为知识库的特例，但总体而言，这两个术语一直被视为同义词。本综述也沿用这一做法，将二者互换使用。近年来，知识库（或知识图谱）得到了广泛应用 [50, 126]。从 YAGO [117]、DBpedia [13], 、Freebase [20], 或维基数据 [124], 等专门社群中的初步发展起步，如今，知识库已成为大多数大型科技公司的标准技术 [80]，并在其他领域得到广泛应用。尽管如此，理解其质量仍然

是一项长期挑战。通过本综述，我们详细探讨完备性、召回率和否定相关问题。

在本综述中，我们重点关注能够从文本中获取，并且能够以合理的结构化格式表示在知识库中的知识。无形知识，例如围绕运动、气味、节奏或触觉等概念的知识，不属于本文的研究范围。

2.2 不完备性

开放世界知识库本质上是不完整的，即它们只存储目标领域中所有真实陈述的一个子集 [100]。例如，我们在表 1 中的示例知识库就遗漏了玛丽·居里拥有的另外两种公民身份（俄罗斯、法国），甚至没有包含她获得诺贝尔化学奖这一事实。[2] 此外，它还缺少这样一条陈述：与一种常见的误解相反，她并非第一个发现放射性元素铀的人。

这种不完备性由多种原因造成：

(1) 来源不完整性：知识库通常根据其他来源构建，或者通过自动化方法 [126], ，或者由人工整理人员完成；他们会阅读网站、教科书、新闻文章等，或其他权威网站。这些来源本身也可能不完整。**(2) 访问局限性：**在某些情况下，只有部分相关文档可供自动化方法或人工整理人员轻易访问。例如，大量知识隐藏在深网（没有入链的网页） [19], 中，而且还有一些印刷文档没有数字形式。**(3) 抽取不完善：**即使陈述已有数字形式且可以访问，也可能被遗漏，因为抽取器并不完美。这一点对自动文本抽取方法 [126]尤其适用：在广受欢迎的 TACRED 基准测试 [125]中，表现最佳的模型召回率也仅达到 75%。**(4) 资源限制：**人工工作自然会受到可用工作时间的限制，自动抽取器在较小程度上也存在这一问题。这在知识长尾领域尤其明显，因为社交媒体内容不断涌现，速度很快。**(5) 无法穷尽地实例化所有真实否定：**对于否定断言，还存在另一项难题：真实否定陈述的集合近乎无限[3], ，因此，除了每个主体少数显著的否定陈述外，实例化其他否定陈述并不可行。

我们借助一个假设的理想知识库 K^i，对知识库（KB）的（不）完备性进行形式化定义。这个理想知识库包含现实世界中关于目标领域的所有真实陈述。理想知识库并不是一个容易界定的概念。在最简单直观的概念化方式中，K^i 仅包含现实中所有真实且能够通过给定谓词集合表达的陈述。对于“兄弟姐妹”或“出生地”等定义足够明确的关系而言，这种概念化方式是合理的。然而，对于其他关系（例如“爱好”），这种概念化方式可能定义不明确 [100]：尽管阿尔伯特·爱因斯坦的爱好之一是拉小提琴，但他可能还有数量不明确的其他“爱好”（例如散步或吃巧克力）。因此，对于这种关系，我们无法明确理想知识库应该包含哪些内容。实体的情况也是如此：拥有博士学位的人是否都算科学家？还是只有受雇于科学家岗位的人才算？如果一位自由职业科学家取得了重要发现，又该如何界定？这表明，只有对于定义明确的关系和实体集合，我们才能提出理想知识库这一概念：例如，在注册协会中开展的爱好、受雇于研究机构的科学家等。尽管这会限制可研究的范围，但许多关系和实体集合的定义足够明确，因此能够据此确定完备性：所有的国家……

[2]这使她成为有史以来仅有的两位两度获得诺贝尔奖的人之一。[3]若考虑无限域，则为无限；若考虑有限域（例如知识库的活动域），则为有限但难以处理。

联合国成员；某个国家境内所有海拔高于 1000 米的山脉；某个国家所有能够授予博士学位的大学；等等。在下文中，我们将重点讨论那些能够合理建立理想知识库 K^i 的领域。

我们称来自目标领域的陈述 st 在现实世界中为真，当且仅当 $st \in K^i$；来自目标领域的负陈述 $\neg st$ 在现实世界中为真（“确实为假”），当且仅当$st \notin K^i$。我们称另一个知识库 K是正确的，当且仅当 $K \subseteq K^i$。正确性通常是知识库构建的核心重点 [85]（也称为精确率、真实性或准确性）。尽管如此，知识库显然不可能在大规模情形下全部正确。不过，由于本综述聚焦于不完备性，我们将作出一个简化假设：所讨论的知识库是正确的。

选择条件 σ 是一个陈述 s，p，o，其中每个分量都可以是具体值，也可以是通配符（“$*$”）。例如，$*$，出生地，华沙 是一个选择条件。选择条件 σ 在知识库 K 上的结果（记作 $\sigma(K)$）是 K 中所有在非通配符位置与 σ 匹配的陈述构成的集合。例如，$*$，出生地，华沙 选出所有出生于华沙的人；玛丽·居里，获得奖项 $*$ 则选出玛丽·居里获得的所有奖项。

现在我们可以继续为本文综述定义主要概念。

定义 2（知识库完备性）。给定一个知识库 K 和一个选择条件 σ，如果 K 对于 σ 是完备的，则称其为完备。

$$\sigma(K) = \sigma(K^i).$$

定义 3 (知识库召回率**)**。给定一个知识库 K和一个选择条件 σ，K相对于 σ的召回率 定义为：

$$Recall = \frac{|\sigma(K)|}{|\sigma(K^i)|}.$$

由此可见，知识库召回率是一个实值概念，而知识库完备性是一个二元概念，当且仅当召回率等于1时才满足。

术语。现有文献对术语的使用并不一致。研究中经常将完备性和召回率这两个术语混用，也会使用覆盖率等其他术语。在本综述中，我们严格将完备性用于表示布尔概念，即在现有知识库上应用某一选择条件所得的结果，是否与在理想知识库上应用同一选择条件所得的结果相等；而将召回率用于表示这一概念的推广，即上述两个结果之间的定量比率。因此，从形式上讲，单独定义完备性这一概念是多余的，但我们发现布尔情形出现得足够频繁，因此有必要将其单独指出。

2.3 世界语义

在数据管理中，传统上有两种主要范式，用于解释肯定性知识库：封闭世界假设（CWA）认为，知识库中没有的陈述在现实世界中都是假的，即 $s, p, o \notin K \Rightarrow s, p, o \notin K^i$。与之相对的是开放世界假设（OWA），它认为这些陈述的真伪未知，也就是说，知识库中没有的陈述可能存在于 K^i，也可能不存在。在许多有限领域中，CWA 是合理的，例如所有员工、订单等信息都已知的企业数据库；要求学生完成正式注册流程才能成为学生的教育管理系统；或成员资格明确的职业体育主要联赛。然而，在许多其他场景中，收集某一领域的所有陈述和实体既不可能、不现实，也并非人们所愿，此时 OWA 更为合适。例如，要求 YAGO 这类知识库收录所有具有美国国籍的人既不现实，也不受欢迎。因此，大多数网络规模知识库都采用 OWA。

> **示例：** 再来看玛丽·居里。考虑查询“玛丽·居里是波兰公民吗？”。在 OWA 和 CWA 下，答案都是“是”，因为陈述 玛丽·居里, 公民身份, 波兰 在知识库中。现在考虑查询“玛丽·居里是澳大利亚公民吗？”在 CWA 下，鉴于知识库中没有这样的陈述，答案将是“否”。然而，在 OWA 下，该陈述在现实中仍可能为真，因此答案是“未知”。最后，考虑查询“非美国公民的诺贝尔和平奖得主”。在 CWA 下，该查询会返回玛丽·居里等许多人。在 OWA 下，该查询必须返回空集，因为对于任何一位获奖者，知识库都可能只是缺少其美国公民身份信息。

OWA 和 CWA 代表了两种极端立场，二者都存在严重缺陷。OWA 使得不可能对任何查询回答“否”，而 CWA 即使在知识库应当避免表明立场的情况下，也会给出“否”的答案。因此，需要采用被称为 部分封闭世界假设 (PCWA) [28, 74], 的中间设定，其中知识库的某些部分按照封闭世界语义处理，而其他部分则按照开放世界语义处理。PCWA 的一种实例化形式是 部分完备性假设 (PCA) [33, 38]。PCA 断言，如果某个主体对于给定谓词拥有至少一个 对象，那么现实世界中除了知识库中已有的对象之外，不再存在其他对象。例如，如果知识库中只知道 玛丽·居里获得过一个奖项，那么我们就假设她没有获得过其他奖项。然而，如果没有记录任何宠物，我们不会对她拥有多少宠物作出任何假设：她可能养过宠物，也可能没有。经验证据表明，PCA 在知识库中经常成立（见第 3.2 节——“弱信号”段落）。不过，PCA 是一种通用假设，不允许指定完备性/不完备性的具体区域。

2.4 完备性与基数元数据

接下来，我们介绍两种可用于指定知识库中完备区域的元数据断言 [74, 99]。

定义4 （完备性断言）。 给定一个可用知识库 K 和一个理想知识库 K^i，完备性断言 [74] 是针对已完整记录的元组的选择条件 σ。形式上，如果 $\sigma(K^i) = \sigma(K)$，则该断言得到满足。

> **示例：** 完备性断言 ⟨玛丽·居里，指导 ∗⟩ 规定，玛丽·居里 的所有指导关系都已记录在知识库中。根据表 1中的示例，这意味着除玛格丽特·佩雷、Óscar Moreno 以及列出的其他6人之外，居里没有指导过任何其他人。

完备性断言可以自然地扩展到选择-投影-连接查询 [74]。例如，可以使用连接来规定知识库知道所有来自英国的诺贝尔物理学奖得主的被指导者。投影同样可用，例如，可以断言所有具有 VIAF 标识符的主体都是完备的，但不记录标识符本身。连接和投影都存在一些细微问题，尤其是子查询应在可用知识库上求值，还是在概念上的理想资源 K^i 上求值。类似地，投影可以合并连接查询的多重性，而对结果集合或包的完备性断言具有不同的语义（知识库是否能够重构所有诺贝尔奖得主，还是还能够重构每位得主获奖的次数？）。有关更多详细信息，请参阅 [99]。

各种经过人工整理的资源都允许按照上述定义的理念指定完备性断言，如图1 （上）所示。左上方可以看到一条关于阿根廷诺贝尔奖获得者完备性的断言。这是一个典型的连接筛选器示例：谓词 获得奖项

[4]图片来源：https://wiki.openstreetmap.org/w/index.php?title=Abingdon&oldid=471369，https://www.imdb.com/title/tt0083987/fullcredits?ref_=tt_ov_st_sm，https://en.wikipedia.org/wiki/List_of_Argentine_Nobel_laureates 和https://en.wikipedia.org/wiki/Henrik_Wenzel

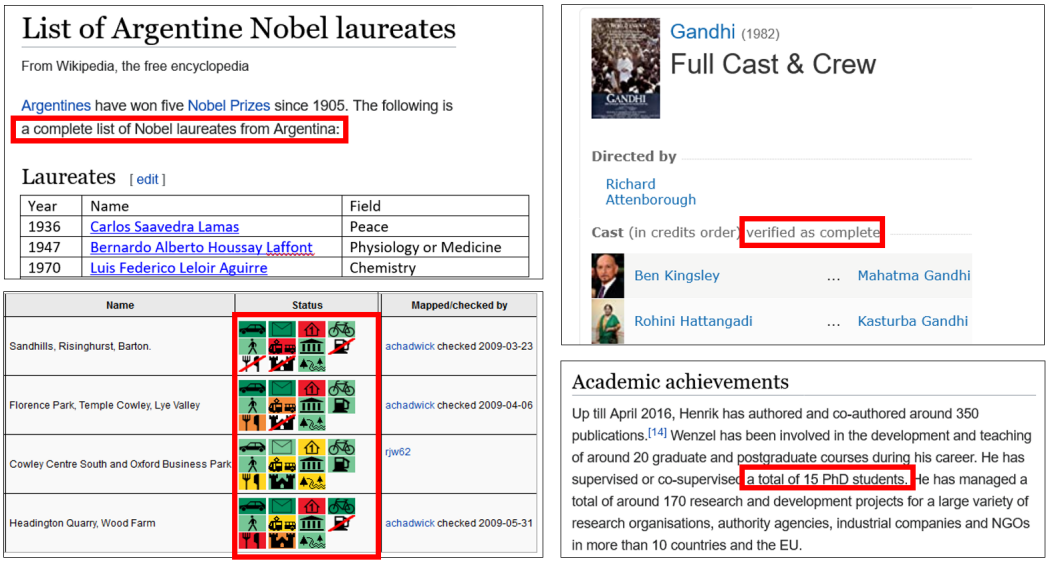

图 1. 网络上完备性断言和基数断言的示例。左上和右下：维基百科。右上：IMDb。左下：OpenStreetMap。[4]

需要与谓词公民身份连接。在右上方，我们看到一个关于电影演员的简单断言。这对应于类似 甘地, 出演于 $*$ 的选择。

知识库通常特别重视类（通过“类型”谓词来标识）。尽管两者都可以用相同的完备性断言来建模，但仍可进一步将完备性划分为实体完备性（如“我们是否拥有所有物理学家？”）和陈述完备性（如“我们是否拥有每位物理学家的所有奖项？”）。这种区分有其优势，因为应对这两类不完备性的方法各不相同，正如我们将在第 3 节中看到的那样。不过，这种区分往往并不明确，因为实体通常通过陈述来标识。例如，法国的城市由选择⟨*,*locatedIn, France*⟩来标识。因此，尽管后续章节讨论的研究在实践中可能侧重于这两类完备性中的某一类，但我们上文的定义并未在形式上加以区分。

基数断言受到描述逻辑中的数量限制的启发 [51]。它们具有以下形式：

定义 5 （基数断言）。 基数断言是形式为 $|\sigma(K^i)| = n$ 的断言，其中 σ 是一个选择条件，且 n 是一个自然数。它指定了 K^i中满足某种属性的元组数量。

> **示例：**继续来看选择条件 $\sigma =$ 玛丽·居里，获得奖项 $*$，基数断言 $|\sigma(K^i)| = 37$ 表示，实际上，玛丽·居里 总共获得了 37 个奖项。

可以通过根据定义 3, $|\sigma(K)|$ 计算商，利用基数断言推断给定选择条件的召回率。此外，在 $|\sigma(K)| = n$ 的特殊情况下，~~还~~可以利用它们推断完备性。

> **示例：** 假设我们知道，实际上 玛丽·居里 总共获得了 37 个奖项，即$|\sigma(K^i)| = 37$。在表 1 的示例中，对于玛丽·居里的奖项这一选择条件，$|\sigma(K)| = 2$。根据定义 3，这意味着知识库对她的奖项的召回率为 $\frac{2}{37} \approx 5.4\%$。

基数断言也见于网络资源，如图 1 （右下）所示，其中一篇维基百科冠词提及某位研究人员指导的博士生人数。左下方的片段没有明确的数字，但同样体现了记录详细程度的理念：在 OpenStreetMap 中，城市各区的信息记录到了什么程度。

完备性断言与基数断言密切相关：基数断言可以用于评估完备性，而完备性断言则可以确定基数。如果我们知道玛丽·居里指导过 8 名博士生，就可以推断图 1 中的知识库是完备的。如果我们知道她拥有 3 种不同的公民身份，就可以推断该知识库不完备。如果我们知道该知识库是完备的，就可以推断她指导的博士生的真实基数为 7。反过来，完备性还能帮助确立否定。如果我们知道玛丽·居里的指导学生名单是完备的，就可以推断她没有指导让·贝克勒尔、路易·德布罗意等人。

基于完整性元数据进行推理。给定复杂的知识库查询和有关完整性的元数据，一个自然的问题是推断查询答案的哪些部分（如果有）被认为是完备的（在涉及否定操作时，则是正确的）。这一问题在数据库研究 [28, 62, 64, 74, 98, 99, 109],以及知识库研究 [26, 27]中都受到了广泛关注。相关方法要么采用演绎路线（满足特定条件时，查询结果是完备的） [26, 64, 99], , 要么通过查询操作符归纳式地传播元数据 [62,74, 98], , 其计算复杂度与查询回答的复杂度相当或更高。

完备性断言也可以通过约束推断出来。尤其是，如果断言某个谓词具有函数性，那么对于每个主体，一个对象的存在就意味着信息完备。例如，如果存在玛丽·居里的一个出生地，那么玛丽·居里的出生地信息就是完备的。基数约束（例如每个人都有两位父母）自然地扩展了这一思路，而基数断言又进一步扩展了基数约束，使其针对特定主体。

3 预测性召回率评估

本节介绍能够自动确定知识库召回率的方法，即知识库所覆盖的现实世界目标领域的比例。我们将考察实体召回率（第3.1节）和陈述召回率（第3.2节）。

3.1 实体召回率

要估计知识库的“完整”程度，我们首先需要知道目标领域中缺失了多少实体。我们可以将这个问题形式化如下：

定义 6。缺失实体问题**输入：** 一个类 C 及其若干实例**任务：** 确定与 K^i 相比，C 缺失了多少个实例。

这个定义远没有人们希望的那么明确 [84, 100]: 拥有博士学位的人都是科学家吗？如果某个国家的一些大教堂 已被摧毁或重建，那么该国大教堂 的总数是多少？一个国家有多少岛屿 （我们是否也把小岛 和礁石计算在内）？一个国家的居民总数是多少（我们是否也计算已故的人，只计算名人吗）？因此，接下来我们将把讨论范围限定在清晰且定义明确的类上，例如截至2022年获得联合国承认的国家，以及某个国家海拔超过1000米的山脉等。现在让我们来看看解决缺失实体问题的各种方法。

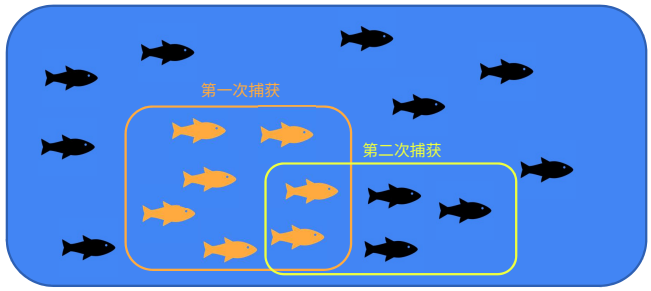

图 2. 标记重捕法：第二次捕获的鱼中有 40% 带有标记 ⇒ 我们在第一次捕获的鱼中标记了 40% ⇒ 鱼的总数约为 7/40% ≈ 17。

标记重捕法 是一种用于估算特定地形中动物总数的技术 [104]。例如，假设一位生物学家想知道 N 某个池塘里生活着多少鱼。为便于说明（图2），我们假设鱼的数量较少，且 $N = 18$。生物学家捕捉 $M = 7$ 条鱼，并给它们做上标记（例如，用颜色标记）。然后她将鱼放回水中（如果可能，应尽快放回，以免活鱼数量降至 $N - M = 11$）。第二天，她再次捕捉一些鱼，并查看其中有多少带有标记。假设只有被捕获的鱼中有40%带有标记。在假定重捕群体中带有标记的比例与整个群体中带有标记的比例相同的前提下，生物学家便可以通过 Petersen 估计量ˆ来估算鱼的总数，即 $N = M/r$（在我们的例子中为17）。此后，人们通过多种方法对这一估计量进行了改进 [104]。现在的问题是，如何将这一基本技术应用于知识库。人们或许会想从知识库中“捕捉并放回”实体，但这种技术只有在实体是从现实世界中捕捉的情况下才有效，而不能从已经被知识库“捕捉”过的样本中捕捉。

协作式知识库的估计器 [67] 将标记-再捕获方法扩展到维基数据等协作式知识库。其主要思想是，通过在维基数据上进行编辑，贡献者从现实世界中抽取一个实体。因此，用户贡献的任何陈述 s,p,o 都计作从现实世界抽取的 s 和 o。然后，我们可以将指定时间段内被编辑过且属于目标类的所有实体视为一个样本。如果 D 是知识库中目标类的当前实例数，n 是样本中观察到的实体数，而 f_1 是样本中恰好被观察到一次的实体数，则 Good-Turing 估计量将现实世界中目标类的实例总数估计为 $\hat{N} = D \times (1 - f_1/n)^{-1}$。该估计器背后的直觉如下：如果 $f_1 = n$，则样本中的每个实体都恰好被看到一次。这表明，只要我们继续抽样，就总会看到新的实体。这意味着现实世界总体的规模是无限的。反之，如果 $f_1 = 0$，则在抽样过程中我们已经至少两次看到所有实体，这表明ˆ该类不会大于我们当前拥有的数量（$N = D$）。其他可能的估计器包括刀切估计量、带有单例异常值缩减的估计器，以及基于丰度的覆盖率估计量 [67]。

在维基数据知识库上进行的实验表明，在测试的 8 个类中，有 5 个类的估计器收敛到了缺失实体数量的真实值。有趣的是，其中一些类是复合类，也就是说，它们由诸如“梵高的画作”这样的查询定义。

众包知识库的估计器 [1, 122] 目标设置：工作者通过向类或查询答案贡献实例来获得报酬。虽然每位工作者都必须贡献不同的实例，

不同的工作者可能会贡献相同的实例。工作者也可能随时来去，并且具有不同的工作风格。如果某位工作者（即“突击贡献者”）一次性添加了许多实体，经典召回率估计器可能会高估总体规模（主要是因为 f_1 的值较高，见上文）。解决这一问题的一种方法是，将每位贡献者可新增的实体数量限制在高于每位贡献者平均实体数两个标准差的范围内。

在 CrowdDB 平台上进行的实验[5] 表明，修改后的估计器在估算实体总数时更加保守。研究人员进行了两项具有已知真实值的实验：对于“联合国承认的国家”，新估计器更快地收敛到真实数量。对于“美国州”，这一效果不太明显，这很可能是因为这些州很容易枚举，因此可以一次性补全。

静态估计器适用于实体不会动态添加的目标知识库。在此类知识库中，不存在从现实世界进行的抽样过程。获取真实人口规模下界的一种方法是使用本福特定律 [18]。本福特定律指出，在许多现实世界的数值数据集合中，首位数字往往较小。更准确地说，数字“1”约出现在 30% 的案例中，数字“2”约出现在 18% 的案例中，而数字 d 约出现在 $100 \times log_{10}(1+d^{-1})$% 的案例中。这尤其适用于呈指数增长的数量（例如城市人口），因为反复乘以某个乘法因子的数量，其首位数字为“1”的次数会多于其他数字。要将此技术应用于知识库 [114], , 目标类必须具有数值谓词，例如城市的人口规模、河流的长度、产品的价格等。该谓词必须满足某些统计性质，本福特定律才能适用。然后，我们可以收集所有数值的首位数字，并检查其分布是否符合本福特定律。在 [114], 的研究中，这主要用于量化知识库中的实体样本在多大程度上能够代表现实世界分布。不过，这些数据也可用于计算至少需要向知识库添加多少实体，才能使分布符合本福特定律。假设现实世界分布遵循本福特定律，那么这个数量就构成了缺失实体数量的下界。

实验表明，本福特定律的参数化版本适用于许多截然不同的谓词，包括地点人口、地点的海拔和面积、水道的长度和流量、灾难的死亡人数和受伤人数，以及维基百科页面的出度。

3.2 陈述的召回率

在讨论完实体的召回率之后，现在让我们转向陈述的完备性。更确切地说：

定义 7. 缺失对象问题**输入：** 一个知识库 K、一个主体 s和一个谓词 p**任务：** 确定是否存在 o，使得 $s, p, o \in K^i$ 且 $s, p, o \notin K$。

例如，考虑玛丽·居里 和谓词获得奖项。我们想知道她获得的奖项是否多于知识库中记录的奖项。要界定我们认为缺失的对象 [100]并不总是那么容易：她高中时获得的奖项算不算？公开表彰是否等同于奖项？被拒绝的奖项仍算是她获得的奖项吗？等等。下文中，我们假设可以判定某种关系是否成立，例如通过众包工作者投票来确定。

关于缺失对象的一个初步线索来自**必需属性**。给定类的必需属性是指在当前时间点的现实世界中，该类的所有实例都必须具有的谓词。例如，出生日期 是人类的必需属性，而配偶关系 是

[5] http://www.crowddb.org/，自 2022 年 9 月 9 日起已无法访问。

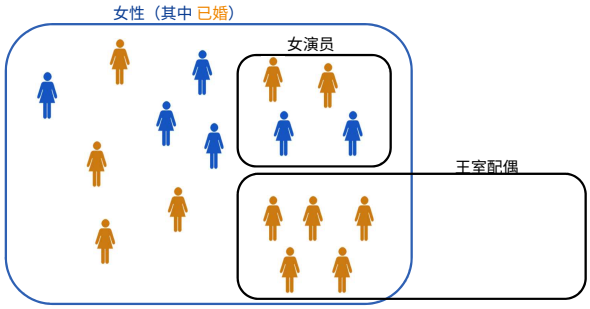

图 3. 为了确定已婚是否是类 女性的一项必需属性，我们检查其分布在知识库中向下遍历至子类时是否发生变化。

不是。如果我们知道必需属性，就可以利用它们解决部分缺失对象问题：如果某个类的一个实例没有某个必需属性对应的对象，那么该实例在该属性上必然是不完整的。事实证明，可以从不完整的知识库中自动确定某个类的必需属性 [61]。为此，我们考察知识库中具有该谓词的该类实例所占的比例。例如，我们可以考察已婚女性所占的比例（图3）。然后，我们检查当进入目标类的子类，或进入目标类与其他类的交集时，这一比例是否发生变化。在我们的例子中，我们可以进入女演员这一子类，或进入王室配偶这一交集类。如果比例发生变化，那么有一个定理告诉我们（在若干假设下），该谓词不可能是必需属性。在我们的例子中，进入王室配偶这一子类时，已婚女性的比例发生了变化。因此，已婚不可能是必需属性。如果我们假设其他谓词都是必需属性，就可以找出知识库中必然缺失对象的位置。

在 YAGO 和维基数据上的实验表明，该方法在所有谓词上的精确率可达 80%，召回率为 40%。该方法还可以确定，对于某个时间点之前出生的人，死亡日期 会成为一项“必需”属性。

弱信号可以结合起来，估计给定知识库 [36] 中某个主体与某个谓词是否完备。这些信号中最简单的是部分完备性假设（PCA），它指出，如果主体具有一些带有该谓词的对象，那么就没有其他缺失的对象。其他信号可以来自知识库的动态变化：无变化假设指出，如果知识库最近几版中的对象数量没有变化，那么对象数量就已“收敛”到真实数量，不再有对象缺失。流行度假设指出，如果主体很受欢迎（即拥有一个长度超过给定阈值的维基百科页面），那么就没有缺失的对象。更复杂的一种信号是类模式预言机。它假定，如果主体是某个类 c的实例，那么就没有缺失的对象。例如，如果主体是生存者类的实例，那么就不会缺少死亡日期。星型模式预言机对谓词也采用相同的做法：例如，如果一个人物有死亡地点，那么这个人物也应当有死亡日期。

这些信号可以如下组合：我们首先将简单信号作为陈述添加到知识库中。例如，如果已知埃尔维斯·普雷斯利 很受欢迎，我们就将*ElvisPresley,is, popular*添加到知识库中。如果埃尔维斯获得了多个奖项，对于较小的 k，我们就添加*ElvisPresley, moreThan*$_k$, 获得奖项。然后，我们为部分主体添加真实值。例如，如果我们知道埃尔维斯·普雷斯利 获得的奖项不超过知识库中已有的奖项，就将*ElvisPresley, complete, wonAward*添加到知识库中。最后，我们可以使用规则挖掘系统（例如AMIE [38]）来挖掘能够预测这些真实值的规则。此类系统可以发现，例如，受欢迎的人通常在其奖项方面是完备的，如下所示

方法	目标	假设
标记与再捕获 [104]	实体	从现实世界中采样。
协作式知识库估算器 [67]	给定类的实体	每个事实都是从现实世界中随机抽取的
众包式知识库估算器 [122]	给定类的实体	众包工作者独立贡献实体
静态估计器 [18]	给定类的实体	分布遵循本福特定律
义务属性估计器 [61]	现实世界中必不可少的关系	事实是从现实世界中独立同分布抽样得到的
弱信号估计器 [36]	给定主体和关系所缺失的对象	数量充足且质量可靠的弱信号
文本抽取 [97]	给定主体和关系的对象数量	文本中对对象或数字的明确提及

表 2. 预测性召回评估方法概览。

s, is,popular $\Rightarrow$ *s,complete, award*。接着可以使用这些规则预测其他主体的完备性，即预测玛丽·居里获得的奖项是否多于知识库中提及的奖项。

在 YAGO 和维基数据上使用众包真实值进行的实验表明，有些谓词很容易预测。这些是每个主体仅对应一个或少数几个对象的必需属性：出生地、具有国籍、性别等。在这些情况下，封闭世界假设和部分完备性假设效果很好。对于其他关系，流行度假设的精确率较高，但召回率较低，也就是说，它未能发现远多于预测数量的主体实际上是完整的。星形预言机和类别预言机的表现也不错。对于测试的全部 10 个谓词，组合方法都能达到 90%–100% 的 F1值，唯独配偶关系例外：仍然很难预测某人的配偶是否多于知识库中记录的人数。

文本信息也可用于发现知识库中不完整的部分。例如，假设我们遇到这样一句话：“玛丽·居里带她的女儿伊雷娜和伊芙去上学。”那么我们可以得出结论，居里至少有两个女儿。在日常对话中，我们甚至会认为她没有其他女儿，实际上也没有其他孩子（如果她还有其他孩子，我们会预期说话者会说明这一点，例如说“只带她的女儿去上学”）。这一结论是一种会话含义，即话语（或文本）所传达、但并未以字面形式明确表达的信息 [45]。接下来，我们考虑一个关于特定主体和特定谓词的给定句子，并假设有一个知识库在该主体和谓词方面是完整的（例如，维基数据中关于知名主体的数据）。如果使用一种简单的开放信息抽取方法，我们可以抽取句子中针对该谓词提及的对象，并将其与完整知识库中包含的对象进行比较 [97]。如果句子提及了所有对象，我们就认为它是完整的。这样，我们就可以针对给定谓词构建一个包含完整和不完整句子的训练集，并在该训练集上训练一个分类器。随后，该分类器可用于判断其他句子是否完整；如果是，我们便能确定该句子的主体在知识库中是否完整。

针对5个谓词的实验表明，该方法在段落上的效果优于在句子上的效果，因为在许多情况下，对象分散列举在多个句子中。F1值为45%–75%。

3.3 总结

可以估算知识库中实体和陈述的召回率。针对这两种情况，研究人员都提出了不同的方法（表 2），并取得了令人瞩目的结果。然而，针对实体的方法受限于真实值较少：只有极少数情况下，某个类的实例总数是已知的。这使得实验结果难以评判。

4 来自知识库和文本的基数

我们从计数实体和对象转向识别基数断言。如第2节所定义，基数断言指定理想知识库中满足 K^i

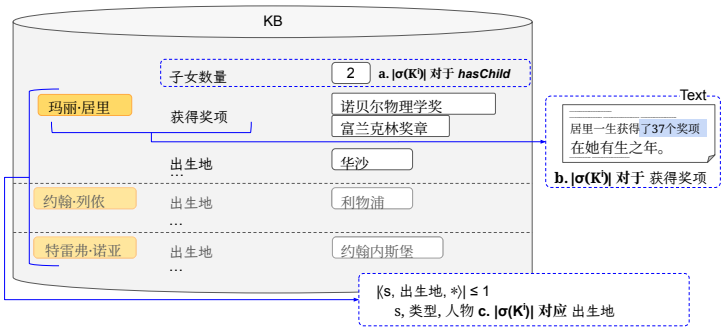

图4. 展示获取基数断言的三种方式：(a) 来自知识库谓词（子女数量），(b) 来自文本（奖项数量），以及 (c) 通过对知识库陈述进行推理（出生地数量）。

某个属性的记录数量。但是 K^i 是一个假设的理想知识库，因此通过选择条件计算基数断言并非易事。

知识库中的基数断言通过陈述显式表达，⟨玛丽·居里，子女数量，*2*⟩；文本中则通过“居里一生中获得了*37*个奖项”表达。~~正确识别此类~~陈述后，便可直接对知识库召回率进行数值评估。或者，如果我们知道某个谓词对于某个主体是完整的，就可以利用这一知识推导出基数断言。本节将探讨如何从知识库和文本中识别基数断言，如图4所示。我们将探讨识别显式存储基数断言的谓词的方法，例如，⟨s，子女数量，n⟩，以及将基数落实为对象的陈述，例如，⟨s，*hasChild*，o_i⟩，$i \in [0, n]$。从现在起，我们将显式存储基数的谓词称为计数谓词，p_c，并将存储基数所对应对象的谓词称为枚举谓词，p_e。

4.1 知识库中的基数信息

从知识库获取基数断言，最直观的方式是使用聚合查询。例如，要确定玛丽·居里 有多少个子女，我们可以统计满足选择条件 玛丽·居里，获得奖项 $*$ 的陈述数量。但是，正如上一节所述，知识库存在对象缺失问题。幸运的是，一些知识库会通过计数谓词显式表达基数断言。例如，在 DBpedia 中，我们可以找到进球数 （足球运动员进球数）和飓风总数 （气旋中的飓风数量）这两个谓词；而在维基数据和 YAGO 中，则可以找到剧集数量 （电影、电视或广播系列的剧集数）这一谓词。遗憾的是，这些基数断言的谓词通常不遵循任何命名规则：有些谓词以*numberOf...*为前缀，但另一些则是针对特定类的专用谓词，例如员工人数、人口或成员数量。此外，知识库通常不了解枚举谓词*hasChild* 与计数谓词子女数量之间的语义关系，而且这两个谓词并非对所有实体都同时存在。

通常，计数谓词（例如子女数量）将主体与相应枚举谓词（本例中为*hasChild*）的对象数量关联起来。不过，计数谓词也可能涉及某个枚举谓词的主体 数量。例如，枚举谓词 *worksAt* 将人物与其工作场所关联起来，而计数谓词*numberOfEmployees* 则将工作场所与在那里工作的人数关联起来。为处理此类情况，我们假设知识库中对于每个谓词 p，也都包含其逆谓词 p^{-1}，以及所有

三元组。在我们的例子中，我们假设知识库中也包含谓词$worksAt^{-1}$（我们称之为*hasEmployee*）。那么，*numberOfEmployees*就是枚举谓词*hasEmployee*的计数谓词。

这里我们的首要任务是识别这两类谓词。接下来，我们的第二项任务是确定这两组谓词之间的映射。

定义 8。基数检测问题**输入：**一个知识库 K**任务：**确定计数谓词 p_c和枚举谓词 p_e。同时确定哪些 p_e对应哪些 p_c。

基数界定。基数检测的第一步可以是界定谓词的基数。例如，如果我们知道知识库中有200 个国家，就知道一个人物的公民身份不会超过 200 个。然后可以利用这项知识帮助我们找到计数谓词。在某些情况下，基数界限来自知识库本身。例如，维基数据对恰好只有一个对象的谓词使用单值约束[6] 。[7]不过，这种情况很少见。

如果知识库也可能包含错误条目（这比第2章提出的朴素正确性假设更符合现实），一种自动界定基数的方法是，针对给定的类 [43]，挖掘谓词的显著最大基数。对于谓词，父母，如果我们发现类人物中有相当一部分实体拥有两个父母对象，那么就可以说该谓词的最大基数为 2。这一结论也适用于人物的子类，例如科学家、物理学家等，除非某个子类中有相当数量的成员具有更严格的界限。这与异常值检测密切相关，但重点在于挖掘通用约束；将这些约束用于异常值检测只是若干种可能应用之一。例如，考虑到上述约束，可以忽略类型为人物且有 6 个父母的实体（这类实体相对较少）。利用霍夫丁不等式约束 [48],，可以在给定置信水平和最小似然阈值的情况下，可靠地挖掘出某个类的显著最大基数。该方法采用自顶向下的方式，从类开始，逐步深入其子类，并剪枝掉非常具体、以致霍夫丁不等式不再成立的约束。这种方法很适合找出函数型谓词的最大基数，例如所有人类都只有一个出生年份，或所有足球比赛都有两支球队，因为这些基数在同一类的实体之间较为稳定：大多数人物有一两个父母，只有极少数人有两个以上的父母。下一种方法将介绍如何处理实体变异基数，例如某位作者所著的书籍数量或某家航空公司的目的地数量；这些基数可能因实体而异，即使实体属于同一类也是如此。

基数谓词检测旨在识别知识库中已有的基数断言及相应的枚举谓词。一种简单的方法是将所有对象为正整数的谓词识别为计数谓词，并将所有对象为一个或多个知识库实体的谓词识别为枚举谓词。这种简单方法因多种原因而行不通。首先，值为正整数是计数谓词的必要条件，但并非充分条件。例如，知识库中存储标识符（剧集编号、维基数据中的*VIAF* 标识符）、测量量（*riverLength*）或非实体计数（*floorCount*）的谓词，并不是计数谓词。其次，像出生地、*atomicNumber*等函数型谓词通常不会与计数同时出现。虽然这些谓词确实会枚举固定对象，但其计数并无任何意义

[6]维基数据中的单值约束定义：https://www.wikidata.org/wiki/Help:Property_约束_portal/Single_ 值。[7]一个元素只能有一个原子序数 https://www.wikidata.org/wiki/Property:P1086。

谓词：玛丽·居里的出生地数量为一这一事实并无信息价值。玛丽·居里主要是函数型、但也可能有多个值的许多准函数型谓词[8]。

谓词定义域、谓词值域和文本信息等信号可用于识别基数谓词 [42]。谓词名称能提供一些线索，帮助判断某个谓词是否可能是枚举谓词。例如，单词 奖项 的复数形式出现的频率几乎与其单数形式相同，甚至更高；但出生地（出生地）并非如此。枚举谓词的对象也是实体，对谓词主体和对象的类型信息进行编码也能提供线索：类型为 人物 （居里）的主体通常会与类型为 奖项 （诺贝尔物理学奖、富兰克林奖章）的多个实例共同出现。均值和百分位数等值域统计也是重要线索：在 DBpedia 中，谓词 获得奖项 每个主体对应的陈述数平均为 2.8，而谓词 博士生 的均值和第 10 百分位数分别为 28.3 和 5.5。

谓词对齐旨在将枚举谓词与相应的计数谓词进行对齐。这使我们能够将基数断言与为基数提供依据的陈述相匹配。对齐后的谓词可用于估算知识库召回率，因为计数谓词为枚举谓词定义了预期（或理想）计数。如果我们知道奖项与*numberOfAwards*相对齐，那么对于所有具有*numberOfAwards*的实体，我们都可以计算其奖项陈述的召回率。实际上，像上述情况那样的精确匹配仅存在于极少数实体和谓词中。更多情况下，枚举是不完整且相互重叠的。例如，某个研究所可能使用员工人数或*numberOfEmployees* 来表示相同的含义，而相应的枚举可能来自工作机构$^{-1}$ 和受雇于$^{-1}$。可以使用精确和近似共现度量，以及谓词标签的语言相似度等启发式方法，来确定适当对齐的谓词对 [42]。由此获得的知识库中的对齐数量远少于计数谓词和枚举谓词的数量。例如，不存在与枚举谓词获得奖项对齐的计数谓词。这些对齐关系也具有方向性。例如，如果(工作机构$^{-1}$, 学术人员)的得分高于(学术人员, 工作机构$^{-1}$)的得分，那么具有枚举谓词工作机构$^{-1}$ 的实体更有可能具有计数谓词学术人员，而反过来的可能性则较低。

4.2 从文本中提取基数信息

到目前为止，我们一直在研究用于基数检测的知识库，但知识库通常比较稀疏。即使我们已经确定博士生人数 是一个计数谓词，对于具有枚举谓词博士生 、但没有计数谓词博士生人数的实体，我们也无法预测其取值。在这种情况下，我们可以转向文本数据，即从文本来源中提取基数信息。例如，如果文本写道“玛丽·居里指导了 7 名学生”，我们希望为谓词 指导提取数字 7。更形式化地说：

定义 9。 计数量词抽取问题**输入：** 主体 s、关于 s 的文本和谓词 p**任务：** 从文本中确定与 s 构成谓词 p 的对象数量。

我们面临的第一个挑战是文本的多样性：同一个数量可以用数字（“3”）、单词（“three”）或短语（“三部书籍”）来表达。第二个挑战是组合性：数量可能需要相加，例如“……创作了三本书籍和*20*篇文章”，这表示谓词*numberOfPublications* 的值应为 *3+20=23*。最后，我们

[8]一生中拥有多重公民身份的科学家https://w.wiki/5UR3

必须确定目标谓词：句子“指导了7名学生”和“指导了7名学生”都映射到目标枚举谓词指导。我们可以采用多种方法来应对这些挑战。

开放信息抽取方法 [25, 69] 旨在从文本中抽取关系陈述，其中也可能包含基数信息。例如，从句子玛丽·居里有两个孩子中，这些方法可以抽取⟨玛丽·居里，有，两个孩子⟩。然而，这些信息并未与知识库谓词*hasChild*关联，而且要还原出数字 2 也并非易事。Roy 等人 [106] 进一步提出了一种针对数量的信息抽取系统。该系统会抽取形式为⟨值，单位，变化⟩的标准化数量陈述。这种方法效果良好，原则上也可应用于基数断言。永无止境学习范式（NELL） [21] 具有用于刻画谓词基数的种子谓词，例如*numberOfInjuredCasualties*，但由于这些谓词与特定基数值绑定，因此无法通过学习将其用于未来的抽取。

当前最先进的自然语言处理流水线大大简化了文本中的基数检测[9]。尽管开放信息抽取的结果中包含基数断言，但抽取正确的计数值并将其映射到知识库谓词仍是一项尚未解决的挑战。接下来，我们将讨论几种专门针对基数抽取的方法。

人工设计的模式 可用于从文本中抽取基数断言 [72]。其思路是捕捉文本中基数信息的组合性和多样性，而这超出了当前开放信息抽取方法的能力范围。 [72] 为维基数据中的*hasChild* 谓词提出了 30 个经人工整理的正则表达式模式。与一小组人工标注的金标准数据集以及规模更大的知识库银标准数据集相比，这些抽取结果的精确率超过 90%。研究从维基数据中为 86.2k 名人类抽取了基数断言，其中只有 2.65% 的人拥有完整的*hasChild* 陈述。

基数的自动抽取 可建模为序列标注任务，其中每个谓词都有自己的模型。 [71] 的研究探讨了一种基于特征的条件随机场模型和一种基于神经网络的条件随机场模型。输入句子中的基数信息会被替换为占位符，即基数词、序数词 和数值词 ，这样模型学习的便是概念，而非特定词元。随后，模型将占位符上下文窗口中的 n 元语法、词元等特征输入基于特征的条件随机场模型；对于双向长短期记忆模型，输入则包括词语、占位符以及词元的字符嵌入。模型输出标签：基数词对应计数 ，组合词元对应组合 ，其他词元对应O ，并给出置信度分数。一个句子可以有多个基数标签，或者多个句子可以包含给定 s 和 p的基数信息。因此，整合步骤会聚合组合线索周围的基数。同一基数断言可能在文本中多次出现，此时会采用启发式排序：基数词 > 数值词 > 序数词 > 冠词；若排序相同，则以置信度分数打破平局。这符合这样的直觉：如果文本中有一个句子是“居里发表了 *15* 篇文章。”，另一个句子是“居里于 *1890* 年发表了她的第一篇文章。”，那么预测居里发表文章数量时，基数应为 15。知识库中已有的基数谓词可有选择地用作真实值，以评估自动抽取的基数的精确率。例如，可以将知识库范围限制为热门实体，并排除高于第 90 百分位数的基数值，因为这些值很可能是异常值。选择时必须谨慎，因为知识库信息不完整会对整体训练和评估产生负面影响。

鉴于大语言模型已改变了众多任务，一个显而易见的问题是：它们是否能用于抽取基数断言？这一宽泛问题可分为两个方面：(i) 大语言模型能否帮助从给定文本中抽取基数断言？(ii) 基于预训练期间见过的文本，大语言模型能否直接输出基数断言？

[9] https://spacy.io/usage/linguistic-features

来源	Work	重点	优势	局限性
KBs	Giacometti 等人 [43]	提取.....的最大基数类与谓词对。	提供显著性保证。 可扩展至大型知识库。 通过高效剪枝缩小搜索空间。	已识别的大多数谓词具有最大基数——最小基数为 1。 用于发现非函数性断言的范围-关系有限。
	Ghosh 等人 [42]	提取计数与枚举谓词。	为识别提供重要特征基数谓词。 将计数谓词映射到匹配的列举谓词。	统计线索提供弱信号。 文本线索的信息量有限。
开放信息抽取	Mausam 等人 [69] Corro 等人 [25]	提取所有 *s, p,o* 陈述。	将其中包含的基数信息表示为计数。	纳入所需的规范化知识库中的陈述。
	Carlson 等人 [21]	提取新的陈述和规则。	永无止境学习范式（NELL）。 推断新的谓词。	无法从种子谓词。
	Roy 等人 [106]	提取数量陈述。	原则上可以提取基数断言—言。	仅用于数量状态—描述。
Text	Mirza 等人 [72]	计算基数断言。	手动模式在基数提取方面很有效。	扩展特定谓词模式。
	Mirza 等人 [71]		用于自动提取的序列建模基数信息。 跨组合整合以及多次提及。	对基数谓词的先验知识。 零计数的抽取仅限于临时预处理。

表 3。基数信息相关工作的比较。

对于问题 (i)，答案似乎是肯定的，因为多项研究表明，大语言模型可作为文本信息抽取中的强大组件 [5, 24, 57]。然而，这种用法仍需要围绕数据选择与准备、信息检索和输出整合开展全面工作，因此并未从根本上改变该任务。

在第二种范式中，可以实现根本性的转变。由于大语言模型在预训练期间接触了海量文本，因此可以尝试直接从这些模型中抽取关系，而无需在抽取阶段进一步输入文本。尽管早期工作得出了不一致的结果 [65], ，但当前的大语言模型（如 GPT-4）能够相当可靠地断言常识领域中的计数，例如不同动物的身体部位（脚、翅膀.....）的数量。对于命名实体的计数抽取，也有类似观察：只要这些计数经常在文本中被陈述，大语言模型就能返回相应结果。然而，一旦大语言模型需要自行进行聚合，它们就会退回到感兴趣关系中常见的数字 [113]。尽管大语言模型在基准测试中的能力正逐步增强，但基于 Transformer 的架构似乎存在原理性局限，使得在文本上正确解决计数任务变得困难 [46]。

4.3 总结

在本节中，我们定义了识别谓词基数的任务。我们介绍了如何从知识库和文本数据中提取基数。基数信息有许多应用，也面临着尚未解决的挑战。我们在表格中汇总了本节涉及的相关工作、其重点、优势和局限性 3。

应用。 CounQER 系统 [41] 展示了**对齐谓词** 在简单问答场景中的实用性：给定一个主体 *s* 和一个（枚举）计数谓词 *p* 后，系统会返回满足*s,p,**的对象，以及可用的排名前 5 的对齐（计数）枚举谓词所对应的陈述。例如，在 DBpedia 中，如果我们查找瑞典皇家科学院，枚举谓词工作地点$^{-1}$ 会返回在那里工作的八个实体，但我们还会了解到，对齐的计数谓词学术人员 和行政人员 没有填充。同样在 DBpedia 中，我们了解到，枚举谓词博士生 没有对应的计数谓词：对于主体玛丽·居里 和谓词博士生，系统返回了她的 7 名博士生，但没有对齐的基数断言。

Tanon 等人 [86] 利用基数信息改进规则挖掘。规则挖掘的任务是在知识库中的实体之间寻找有趣的关联，例如：如果 ⟨*?x,* 配偶关系, *?y*⟩ 和 ⟨*?y,livesIn,?z*⟩ 那么 ⟨*?x, livesIn,?z*⟩。随后可利用此类规则预测某人的居住地。基数信息可以通过降低违反（软）基数约束的预测分数，避免为每个人预测过多的居住地。作者将召回率感知分数与标准评分指标进行比较，发现随着知识库逐渐趋于完整，召回率感知分数与规则质量分数高度相关。另一项研究将基数信息用作神经链接预测中的先验 [75]。与 Tanon 等人 [86], 的工作类似，他们通过在预测数量违反给定关系类型的基数界限时惩罚模型，对高概率预测的数量进行正则化。

开放挑战。 原则上，大语言模型可以用于推断常识基数信息，例如某人有几位父母；也可以推断具体的基数断言，例如玛丽·居里指导过 7 位博士生。对较早期大语言模型进行数值常识探测的实验 [65]表明，微调能够提升模型性能，但模型仍无法超越人类。2023年版 LM-KBC 挑战赛 [111], 的重点是利用语言模型构建知识库，其中包含对两种关系的基数预测：人物的 子女数量 以及电视剧的集数 。即使是依赖 GPT-4 的最佳系统，在这两种关系上的 F1 分数也仅达到 69%。看来，在大语言模型的输出能够可靠地用于知识库之前，这方面还需要更多工作。无论是抽取基数断言，还是将计数与枚举谓词对齐，这两项任务都面临一个共同挑战：生成高质量的训练数据和真实数据。目前没有单一权威的真实数据来源：电影的演员与剧组信息可查 IMDb[10], ，地理位置则可查 GeoNames 数据集[11]。这些都是召回率很高的数据集的优秀范例，但转向科学家或纪念碑等其他主题时，情况就不那么理想了。众包是一种选择，但会带来若干挑战，其中最重要的是数据质量和可扩展性。另一种常用方法是采用启发式方法。正如我们在从文本中进行基数提取时所见， [71], 可以通过远程监督抽取真实陈述，但需要满足某些限制，例如依赖热门实体和知识库统计数据，并使用诸如 90th 百分位数谓词值之类的方法筛除可能的异常值。

5 识别显著否定

知识库通常存储肯定知识，几乎不存储否定陈述。这是出于原则性原因：可能的否定集合十分庞大，甚至可能是无限的（取决于是否假设常量集合有限或无限）。因此，添加完整的否定陈述集合几乎不可能成为目标。许多标准的人工智能应用，例如问答和对话系统，往往会受益于有关热门实体的陈述，也就是针对显著案例的显式否定。例如，烹饪聊天机器人应该知道某些民族食品不宜加热，例如鹰嘴豆泥 和西班牙冷汤；通用搜索引擎则必须能准确判断常见的事实性错误，例如著名人物并未在其领域赢得某些奖项，比如史蒂芬·霍金 并未获得诺贝尔物理学奖。完备性陈述能够支持推断否定，但这类陈述本身也很难获得。此外，尽管原则上否定陈述的数量远多于正陈述，但其中只有少数值得关注。因此，在本节中，我们探讨否定物化问题时，目标并非追求完备性，而是实现对显著否定的高召回率。

[10]IMDb: https://www.imdb.com/ [11]
GeoNameshttps://www.geonames.org/

显著性这一概念 在心理学 中有着悠久的历史 [120]，在社会语言学中也是如此 [91],; 近年来，也有人尝试在知识库中对其进行刻画 [59]。显著性的建模方法通常围绕频率、意外性或自身利益（针对行动主体）等概念展开，但尚未形成普遍共识。此外，这些模型都不易付诸实际应用。因此，我们接下来介绍的研究通常以人类（众包工作者）的判断作为衡量显著性的标准。

我们首先回顾开放世界知识库中现有的否定知识。然后，我们将展示如何从不完整的知识库内部以及通过文本抽取，自动收集显著否定陈述。

5.1 现有知识库中的否定

网络规模知识库遵循开放世界假设（开放世界假设）。这意味着，缺失的陈述并非为假，而只是未知。指定否定信息的唯一方式，是在知识库中显式实体化否定陈述，或断言能够隐含推出否定陈述的约束。本节聚焦于前一种情况。尽管大多数知识库构建项目不会主动收集否定陈述，但其中少数允许隐式或显式的否定信息：

- 否定谓词：少数知识库包含表达否定含义的谓词，即包含否定关键词。例如，DBpedia [13] 包含用于手机的谓词，如*carrierNeverAvailable*，以及用于飞机的 *neverExceedAltitude* 。医疗知识库 Knowlife [34] 包含*isNotCausedBy* 和*isNotHealedBy*等谓词。维基数据允许少数与类型无关的否定谓词，即*differentFrom* （827000 条陈述）、*doesNotHavePart* （535 条陈述）、*doesNotHaveQuality* （422 条陈述）、*doesNotHaveEffect* （36 条陈述）以及 *does-NotHaveCause* （13 条陈述）。关于否定谓词的一个更系统的例子可以在 Concept Net [115], 中找到，其中 6 个主要谓词都有对应的否定形式，即*NotIsA*、*NotCapableOf*、*NotDesires*、*NotHasA*、*NotHasProperty*和*NotMadeOf*。然而，否定知识所占比例不到 2%。此外，许多否定陈述并无信息价值，例如 *tree,NotCapableOf, walk*。
- 计数谓词：表达否定信息的一种微妙方式是将计数与枚举谓词相匹配（参见第4.1节）。例如，如果某个知识库断言⟨玛丽·居里，子女数量，*2*⟩，并附有两条 hasChild陈述，则表明对于这一主体-谓词对，对象列表是完整的。因此，没有其他实体是居里的子女。
- 具有否定极性的陈述：在 Quasimodo知识库 [105],中，每条陈述都会附加一个极性值，以表明它是肯定陈述还是否定陈述，例如，科学家，拥有，学位，其极性=为肯定。Quasimodo包含总计35.1万条否定陈述。
- 无值对象：维基数据 [124] 允许表达普遍否定陈述，即主体-谓词对的对象为空。例如，安格拉·默克尔，*hasChild*，无值[12]。维基数据中此类对象为无值的陈述总数为2.06万条
- 已弃用排名：像维基数据这样的知识库鼓励将某些陈述标记为不正确，而不是删除它们。这些通常是过时的陈述，或已知为错误的陈述[13]，共有 20.4k 条陈述。

尽管这些概念为表达否定陈述提供了一种方式，但我们仍然必须找到一种添加它们的方法——如果可能的话，最好是自动添加。关键挑战在于识别 显著的 否定。

567

[12]https://www.wikidata.org/wiki/Q[13]
https://www.wikidata.org/wiki/Help:弃用

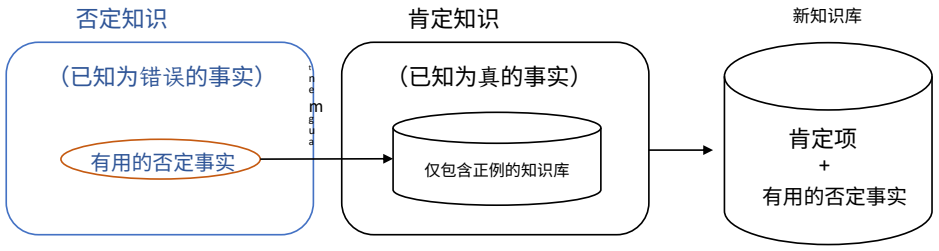

图 5. 主要挑战在于，从现有的仅包含肯定陈述的开放世界知识库中，识别出可添加的有用的否定陈述子集。

陈述，例如一位著名物理学家未获得诺贝尔奖，即 ¬⟨史蒂芬·霍金，获得奖项，诺贝尔物理学奖⟩。表 5展示了该研究问题的示例。错误事实或陈述的数量远远多于正面事实，例如，史蒂芬·霍金曾就读于 4 所教育机构，而他未曾就读的机构则有数千所。关键在于识别出可添加到现有开放世界知识库中的一部分有用的否定陈述。

如上所述，我们在此采用无模型方法，并将何为显著的判断留给人工标注员。我们仅假设频率和意外性可能是他们作出判断时的考量因素[91, 120] 。如果无法获得人工判断，则可以采用自动计算的目标指标，例如 [59]，中提出的指标，但这样做可能会为了优化便于计算的内容，而非真正显著的内容。

定义 10。显著否定问题**输入：**大型不完整知识库中的主体 *s*
任务：识别关于 *s*的准确且显著的否定陈述。

我们将方法分为三个主要类别：使用规范化程度较高的知识库的方法、使用松散结构知识库的方法，以及主要将文本语料作为否定陈述来源的方法。

5.2 结构良好的知识库中的显著否定

Wikidata、YAGO 和 DBpedia 等知名知识库由规范化程度较高的陈述（歧义极少）组成，即 A 盒，并配有人工编制的模式（又称 T 盒）。以下方法通过依赖结构良好的知识库中的正陈述，发现有关实体的有趣否定。

基于同伴的推理 [8] 是解决这一问题的最早方法之一。该方法提出，从<escape></*escape*>高度相关的实体<*escape*></escape>（即同伴）中推导出有关某个知识库实体的候选显著否定，然后根据相对统计频率对其进行排序。例如，著名<escape></*escape*>物理学家<*escape*></escape><Entity category="people" value="Stephen Hawking" disambig="英国理论物理学家"/>，从未获得过诺贝尔物理学奖，也未获得过奥斯卡奖。然而，高度相关的实体（其他获得诺贝尔奖的物理学家）表明，对于霍金而言，未获得诺贝尔奖比未获得奥斯卡奖更值得关注。因此，第一步必须是确定与输入实体相关的同伴。 [8] 的方法提供了三种相似度函数，用于为给定的输入实体收集同伴实体：（i）基于类的相似度 [15]：该方法利用知识库的类型系统，将至少共享一种类型的两个实体视为同伴；（ii）基于图的相似度：该方法依据两个实体共同拥有的谓词-对象对的数量进行判断；（iii）基于嵌入的相似度：该方法通过测量实体嵌入向量的余弦相似度，捕捉实体之间的潜在相似性 [128]。

> **示例：** 使用 (i)，史蒂芬·霍金 和马克斯·普朗克 以及阿尔伯特·爱因斯坦一样都是物理学家，因此，关于普朗克 和爱因斯坦 的信息可以帮助为霍金创建候选否定。使用 (ii)，史蒂芬·霍金 和鲍里斯·约翰逊 共有 9 个谓词-对象对，包括*(*性别，男性*)*、*(*公民身份，英国*)*和*(*母语，英语*)*。使用 (iii)，根据这一度量，与霍金 最接近的实体之一是他的女儿露西·霍金。

此处的否定还仅仅是候选项。由于知识库并不完备，每个候选项在现实世界中都可能为真，只是尚未被收录到知识库中。基于同伴的方法要求候选陈述满足部分完备性假设（主体对于该属性至少具有另一个对象 [37, 38]），才能将其纳入最终集合。具体而言，如果某些奖项被列为主体 史蒂芬·霍金 所获得的奖项，那么就假定该奖项列表是完备的，任何未列出的奖项都是因为其 不真实性而缺失的。

> **示例：** 我们知道霍金还获得过其他奖项，例如奥斯卡·克莱因奖章，也有包括露西·霍金在内的子女，但我们对他的爱好一无所知。因此，候选陈述 ¬史蒂芬·霍金, 有爱好, 阅读 被舍弃。

在对推断出的否定的正确性进行评估时，[8], 这一简单但强大的规则将结果的准确性提高了 27%。剩余的候选否定最终根据相对同类频率进行评分。例如，霍金 的 2 个同类中有 2 个赢得了诺贝尔物理学奖，但只有 2 个中的 1 个是爱德华·爱因斯坦的父母。

基于顺序的同伴推理 [10]是对前一种方法的扩展，利用知识库限定条件（如时间陈述）来获取更优的同伴实体，并提供带有解释的否定。在这种基于顺序的方法中，一个输入实体可以拥有多个同伴群组。例如，马克斯·普朗克 的 3 个独立同伴群组分别是诺贝尔物理学奖得主、科普利奖章得主和慕尼黑路德维希-马克西米利安大学校友。除了确定同伴外，排序还会考虑输入实体与同伴实体之间的接近程度。

> **示例：** 负面陈述 ¬⟨马克斯·普朗克，受教育于剑桥大学⟩ 其来源为“与此前*3* 位诺贝尔物理学奖获奖者中的 3 位不同”的陈述，优先级高于¬⟨马克斯·普朗克，公民身份，法国⟩ 其来源为“与此前*18* 位诺贝尔物理学奖获奖者中的 3 位不同”。

在这个例子中，时间上的新近性 会得到奖励。换句话说，在排序后的同伴列表中远远落后的相同同伴频率会得到较低的分数。此外，这项工作引入了条件否定这一概念，以区别于先前的简单否定。简单否定用 1 条否定陈述来表达：¬阿尔伯特·爱因斯坦，受教育于，哈佛大学；而条件否定则超越了单一陈述，用于表达仅 在某些条件下为真的否定信息。

> **示例：** ¬∃*o* ⟨阿尔伯特·爱因斯坦，受教育于，*o*⟩⟨*o*，位于，美国⟩（意思是爱因斯坦从未在美国上过大学。）

在一项评估 [10] 中结果否定质量的众包任务里，基于同伴的方法精确率达到81%，显著性达到44%；而面向顺序的推理则同时提升了精确率和显著性，分别达到91%和54%。

5.3 松散结构知识库中的显著否定

维基数据、YAGO 和 DBpedia 等百科全书式知识库经过了良好的规范化，而 ConceptNet [115] 和 Ascent [78, 79] 等常识知识库则使用未经规范化的短语来表达信息。例如，Ascent 中包含 产卵、产卵和 产卵等动作。因此，为结构良好的知识库设计的方法会导致许多错误的推理，例如，蝴蝶，

能够，产卵，但 ¬蝴蝶，能够，产卵以及 ¬蝴蝶，能够，产卵。此外，使用 PCA 规则（只有在同一谓词下至少存在另一个对象时，才将缺失的对象推断为否定）来提高推断出的否定的准确性并不足够，因为大多数常识知识库依赖于 ConceptNet 定义明确但数量有限且较为通用的谓词，例如 具有属性。与维基数据的 公民身份 谓词不同，后者预期只有少数对象，而 ConceptNet 的具有属性 可以有数百个被接受的对象短语。因此，假设某个实体至少存在一个一般属性，就认为对象列表已经完整，这并不是一个好主意。

NegatER [107, 108] 是一种近期提出的方法，用于识别常识知识库中的显著否定。给定一个主体 s（在本例中是一个日常概念）和输入知识库，使用知识库陈述的训练集和测试集对预训练语言模型（语言模型）进行微调。正陈述直接从知识库中查询得到，而负样本则在封闭世界假设下随机生成，也就是说，知识库是完整的，否定通过用其他任意随机概念或短语替换正陈述的部分内容来生成。随后训练语言模型，使其为每个谓词学习一个决策阈值，因此，微调后的语言模型可充当未见陈述的真假分类器。为了创建 s的信息性或主题性否定集合， s会被替换为知识库中邻近的实体短语。

> **示例：** 马，是一种，宠物被邻近主体骑马者替换，得到候选否定 ¬大象，是一种，宠物。

分类器（语言模型）随后判断此类候选项是否为假。构建候选项集合后，使用微调后的语言模型按负向似然从高到低对其排序。尽管 NegatER 会整理与主题相关的否定项列表，但它的一大局限是会生成许多类型不一致的陈述，这是由于缺少分类体系，例如， ¬马术骑手，属于，宠物。

Uncommonsense [12] 通过利用外部结构化分类体系（例如 WebIsALOD [47]）和潜在相似性（例如维基百科嵌入 [128]），计算可比概念（例如斑马、狮子），从而识别知识库中关于目标概念（例如大猩猩）的显著否定。与基于同伴的否定推断方法 [8], 类似，该方法使用部分封闭世界假设来推断候选否定（例如，没有尾巴、不具有领地意识）。一个关键区别在于检验候选否定准确性的技术。如前所述，在松散结构知识库中，部分完备性假设并不足够。因此，本研究引入了不同的审查步骤，以提高候选项的准确性。它会针对知识库本身进行语义相似度检查（使用 [101]），并进行外部来源检查（使用预训练语言模型）。语义相似度还用于将含义相同的否定短语归组，以提升它们在最终排序列表中的排名，即采用宽松兄弟频率。最终生成的高排名否定项还附有溯源信息，说明某些否定为何值得关注。例如，大猩猩没有尾巴，这一点不同于狮子和斑马等其他陆生哺乳动物。在一项众包任务中，研究者随机抽取了200个概念及其排名前二的否定项 [12], , 请工作者评估所给否定陈述的准确性和趣味性。UnCommonSense 的精确率为75%，显著性为50%；而 NegatER 的精确率相近，为74%，但显著性较低，仅为29%。

5.4 文本中的显著否定

大型文本语料库可以成为隐式和显式否定的良好外部来源。此外，由于现有知识库并不完备，基于文本的方法可以作为基于推理的方法的补充。

从查询日志中挖掘否定 [8] 是一种基于无监督模式的方法，可从查询日志中提取显著否定。其直觉在于，用户经常会询问某事为何不成立，例如“史蒂芬·霍金为什么没有获得诺贝尔物理学奖？”。随后可以利用这样的查询推断出霍金确实没有获得诺贝尔奖。该研究定义了 9 种否定式“为什么”问题，例如为什么 *<s>*？。

> **示例：** 给定模式 为什么史蒂芬·霍金没有……，搜索引擎（例如 Google）的自动补全 API 会生成获得诺贝尔物理学奖、接受爵士头衔……。

从文本中挖掘否定也是一种可能性，但迄今尚未取得显著成果，因此远远超出 [6]。原因在于，长篇文本（如报纸、博客和百科全书）很少提及显著否定，正如对 STICS [49] 语料库的分析所显示的那样。报纸和博客中具有否定含义的句子，大多涉及人们没有做过或没有说过的事情，例如“布拉德·皮特并未以罚款相威胁安吉丽娜·朱莉”，或“安格拉·默克尔从未为确保东德人产生归属感而付出过多努力”。另一方面，百科全书大多只包含正陈述。少数包含否定的句子通常包含双重否定、暂时性否定（例如“霍金最初在学业上并不成功”），或限定性否定（例如“如果没有奖学金的经济资助，他的家庭就无力承担学费”）。总体而言，这些来源均不包含带有否定关键词的简短知识性句子。

从编辑日志中挖掘否定信息 [58] 是一项利用维基百科等协作式百科全书的编辑历史作为隐式否定丰富来源的工作。编辑每天都会进行数千次修改，原因各异，包括修正拼写错误、改写句子、更新有争议话题的信息，以及纠正事实性错误。这项工作重点从最后一类修改中挖掘数据。具体而言，它考察维基百科中仅更改了一个实体或一个数字的句子编辑。

> **示例：** “铀的发现归功于玛丽·居里”被更新为“铀的发现归功于马丁·海因里希·克拉普罗特”，其中实体玛丽·居里 被马丁·海因里希·克拉普罗特所取代。前者随后被视为一种常见错误，即一条有趣的负向陈述。

为了判断某项更新是否应被标记为常见事实性错误或其他类别，需要应用多种启发式方法。这些方法包括：(i) 检查某个陈述被更新的频率（以排除不同编辑者持有不同观点的争议性话题）；(ii) 计算实体之间的编辑距离（以排除拼写纠正）；以及 (iii) 检查同义词（以排除对同一陈述的简单改写）。接下来，还需要检查此次编辑是否删除或引入了错误陈述。具体方法是统计网络上支持相关陈述的数量。

5.5 总结

本节定义了识别有关知识库主体的显著否定这一任务，并介绍了应对这一问题的不同方法。表4中总结了这些方法的重点、优势和局限性。

应用。 Wikinegata [9, 11] 是一款浏览工具，可用于浏览关于约 50 万个维基数据实体的 6 亿多条否定。它深入展示了基于同伴的方法 [8]，用户可以查看用于推断特定否定的不同同伴（见图 6）。在常识知识库中，Uncommonsense 系统 [12] 为日常概念提供了类似的体验。

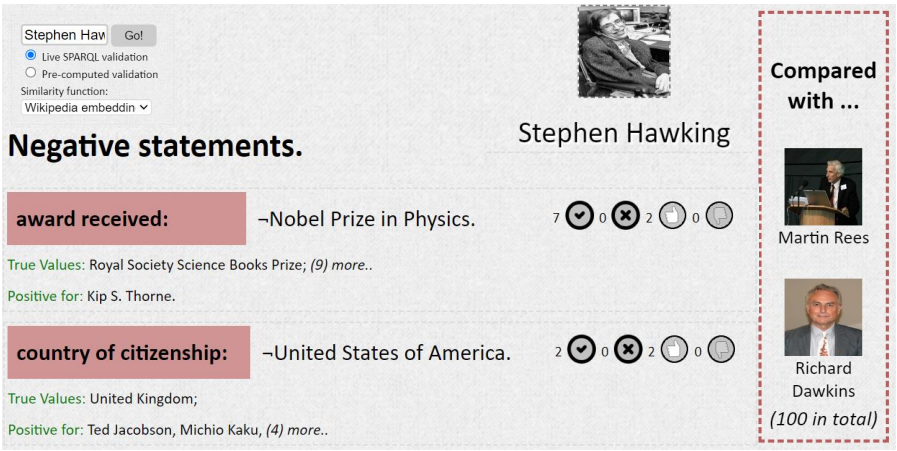

图 6. WikiNegata（https://d5demos.mpi-inf.mpg.de/negation）平台的用户界面。图中显示了系统自动计算出的有关史蒂芬·霍金的显著否定陈述，例如他没有获得诺贝尔物理学奖，而他的同事兼好友基普·索恩则获得了该奖。

尚待解决的挑战。对于基于文本的方法，主要问题似乎在于主体召回率。通常，只有当否定内容具有高度特殊性且涉及知名实体时，才会明确表达否定。对于基于知识库的方法，主要问题在于精确率与显著性之间的权衡。在封闭世界假设下，要获得近乎完美的精确率相当简单，因为大多数推断出的否定都是正确的，但却毫无意义，例如 史蒂芬·霍金的首都不是巴黎。根据 [12], 这一基线的精确率达到 93%，但显著性不足 7%。正如本节所示，在保持高精确率的同时提高显著性颇具挑战性，因为合理的候选项往往更难审查，尤其是在常识知识库中，例如 篮球难学吗？。此外，否定生成器还必须应对现实世界中的变化。这对于百科全书式知识库尤为重要，因为其中经常会添加新信息。例如，在 2016年之前，莱昂纳多·迪卡普里奥 从未获得过奥斯卡奖（这一否定如今已不再正确）。第二个挑战是实体同类度量和否定生成中的类层次结构问题。例如，存在噪声的分类体系会导致无关的同类实体。此外，大多数网络规模知识库都存在建模问题和不一致性，尤其是协作式知识库。例如，为了表达某个人是纯素主义者，编辑者应该使用 ⟨人物，生活方式，纯素主义⟩ 还是⟨人物，属于，纯素主义者⟩？虽然其中一种表述可以被明确断言，但另一种却可能被前面讨论的方法错误地否定。最后，本节介绍的方法旨在汇编一份关于常识实体和百科全书实体的简单否定陈述列表。另一方面，复杂否定仍需进一步研究，也更具挑战性。例如，在科学知识中，两个相互矛盾的事实可能在不同语境下都成立，例如 水不能扑灭所有类型的火灾，例如 汽油火灾，但这并不意味着 水不能灭火。此外，在社会文化知识中，同一陈述在不同文化因素影响下可能既正确又错误，例如 在婚礼上饮用葡萄酒 (在欧洲 与中东）。

用于否定生成的大语言模型。近期研究考察了大语言模型（LLM），例如 ChatGPT [82], , 生成显著否定陈述 [7, 22]的能力。 [22] 的研究结果表明，当比较针对同一知识内容的不同任务所得结果时，大语言模型的信念中存在矛盾。例如，大语言模型会生成句子“狮子生活在海洋中”，但在被问到“狮子生活在海洋中吗？”时却回答“否”。 [7] 带来的启示包括丁香五月

来源	Work	重点	优势	局限性
KB	Arnaout 等人 [8, 9, 11]	关于百科全书的有趣否定 实体的基于同伴的统 计推理。	主体召回率。 基于同伴频率的显著性措 mea- 施。	由知识库不完备性导致的精确率 & 建模问题。 超越简单否定（条件 否定）。
查询 logs	Arnaout 等人 [8]	通过百科全书 实体的基于模式的 查询抽取。	由于搜索引擎查询日志质量高，精确率较高 。	由于 API 访问受限，主体召回率较低。
Edit logs	Karagiannis et 等人。 [58]	常见事实性错误使用 挖掘出的文本变更日志。	由于采用包括网页 命中数计算。	侧重精确率而非显著性 挖掘出的 否定 需要 规范化 。
KB	Arnaout 等人 [12]	关于每一项的具有信息价值的否定- 使用微调语言模型处理日常概念。	由可比概念带来的显著性 可解释的 结果 通过 证明- 性生成。 可以处理未经规范化的知识库。	召回率取决于主体是否存在于 外部分类体系中。
知识库和 LM	Safavi 等人 [107, 108]	利用可比较的分类学兄弟姐妹，关于日常 概念提供信息丰富的否定 兄弟姐妹。	通过使用短语进行损坏来计算召回率 嵌入表示。	由于分类体系尚未得到 con- 考虑在内。

表4。知识库中显著否定相关不同工作的比较。

提示工程在此任务中的重要性。包含“否定陈述”“被否定陈述”和“否定语句”等表达的提示会返回非常不同类型的响应。此外，大语言模型难以准确理解所返回陈述的真正否定性，经常生成包含否定关键词但含义为肯定的陈述，例如：“咖啡桌不仅用于室内”。

6 相对召回率

到目前为止，召回率/完备性的衡量标准一直是现实世界：目标领域中的多少实体（或谓词或陈述）已被收录？ 虽然这通常是一个有意义的目标，但在某些情况下，这一概念并没有明确定义，或缺乏参考价值。接下来，我们将通过把绝对衡量标准放宽为相对标准，来探讨替代性表述：相对于其他实体/资源/使用场景，知识库的召回率如何？ 我们从三个方面探讨这个问题：（i）在同一知识库内，相对于其他实体的召回率；（ii）相对于其他资源（如知识库或文本）的召回率；以及（iii）相对于问答这一外部使用场景的召回率（见图7）。

6.1 相对于其他实体的实体召回率

我们将实体相对召回问题定义如下：

定义11。 实体相对召回问题**输入：** 一个实体 e和一个知识库 K**任务：** 确定 K 针对关于 e的陈述的召回率，并将其与 K中的相关实体进行比较。

例如，可以询问维基数据中关于阿尔伯特·爱因斯坦 的信息相对于相似实体而言有多完整。该问题的具体化包含两个主要组成部分：

（1）如何定义相关实体？（2）如何量化
和比较召回率？

实体相关性是数据挖掘领域中一个历史悠久的主题，现有多种基于文本、图结构和嵌入的相似性度量（见第 5 节及例如 [88, 127]）。类似地，召回率也可以通过多种方式量化，例如通过陈述、谓词、入链、出链等的数量。下面我们将介绍针对这两个问题的不同方法。

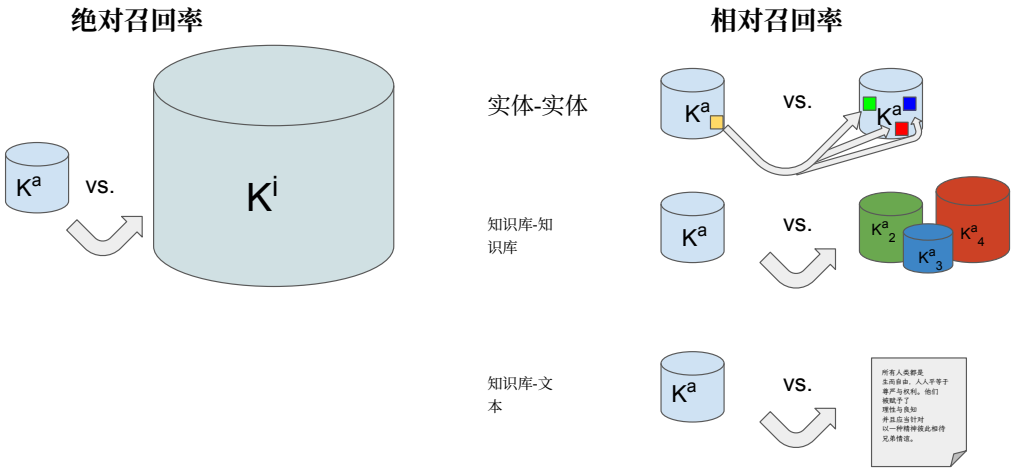

图7. 左侧为前文讨论的绝对召回率，右侧为相对召回率的各种概念，两者之间存在差异。

相对召回率指标。针对实体相对召回率问题，最早的方法之一是 Recoin（相对完整性指标）[2, 15]。它通过一种交通灯式召回率指标扩展维基数据的实体页面，用以表明与相关实体相比，该实体的信息有多全面。它对相关性的定义以类为依据：巴黎 会与其他首都城市比较，镭 会与其他化学元素比较，阿尔伯特·爱因斯坦则会与其他物理学家比较（将维基数据中的职业视为伪类别）。其召回率量化方法采用简单的频率聚合：在每个类中，计算频率最高的属性。例如，对于首都，频率最高的谓词是国家（99%）、坐标位置（97%）和人口（82%）。随后，将给定实体缺失的前五项属性的频率相加，再将总和与 5 个全局离散阈值进行比较，最终得出交通灯颜色。

相对召回率问题的另一种解决方案由维基媒体的 ORES 机器学习平台提供 [70]。该维基数据条目质量模块 专门为实体属于 5 个质量等级（A 至 E）之一的可能性分配概率。除召回率外，这些等级还涵盖完备性、参考文献、站点链接和媒体质量等多个质量维度。这些分数也具有相对性，因为条目质量优劣并无内在定义，只能通过与其他优质或劣质条目进行比较来理解。显然，ORES 采用了监督式机器学习，结合内容嵌入与潜在特征，但其具体工作方式仅有部分文档记录。我们展示了截至 2023年9月，Recoin 和 ORES 对玛丽·居里的示例输出，见图8。

属性排序。Recoin 和 ORES 都面临着一个共同问题：如何判断对于某个实体，预期应填写哪些谓词。这被称为属性排序问题。以玛丽·居里这样的实体为例，记录她的博士生导师是否比记录她曾效力的运动队更重要?

属性排序与陈述排序问题相关 [16, 63], ，尽管缺少对象值会使该问题更加困难。基于频率的排序（例如 Recoin 中采用的排序 [15], ）提供了合理的基线，但频率只能作为相关性的代理指标。例如，频率最高的属性通常涉及通用的传记陈述，例如出生

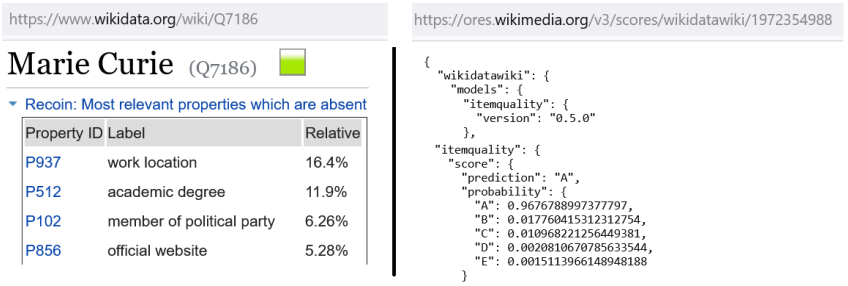

图 8. 玛丽·居里 在维基数据上的 Recoin（左）和 ORES（右）输出示意图。Recoin 输出显示缺失属性的类频率，而 ORES 质量等级 A-E 则通过监督回归模型计算得出。

地、出生日期，而科学发现、奖项或政治归属等有趣方面出现的频率较低。

因此，针对相对召回率评估的属性排序已通过多种方式得到改进：在[95], 基于文本的预测模型中，研究人员提出了一种模型，该模型基于维基百科中的实体描述进行训练，用于描述文本描述中伴随维基数据中某一特定结构化属性的可能性。在对成对属性相关性预测进行评估时，该方法与人工标注员的判断一致率最高可达 67%。Gleim 等人提出了 SchemaTree [44],，这是一种基于字典树的方法，用于捕获现有数据中的属性频率。该方法不仅考虑单个属性的频率，还考虑属性组合的频率，并针对罕见组合回退至基本情形。Issa 等人采用了关联规则挖掘 [54]。Luggen 等人 [66] 提出了一种基于多模态维基百科嵌入的方法，同时考虑多语言条目描述和图片。他们利用这些嵌入训练了一个多标签神经分类器，用于预测维基数据中当前存在的属性，并发现该方法的性能显著优于以往的方法。

分析。 所讨论的方法大多将实体视为属性集合——只区分属性是否存在，却不考虑存在多少个值。例如，尽管知识库中列出了爱因斯坦的一项奖项，但知识库可能还遗漏了许多其他奖项。忽略这一方面有其实际考量，源于部分完备性假设（见第 2 节）：某个属性存在一个值，可能意味着该属性的所有值都已存在。与此同时，若能进一步明确考虑多值属性，将是值得期待的改进。第二个局限是所有相对度量固有的：如果对比对象更差，那么即便对象本身表现不佳，相对度量也可能看起来很好，反之亦然。它与前几节所述的召回率概念并无确切联系；即使某个实体相较于相关实体覆盖得很全面，也无法据此对其相对于现实世界的召回率作出有力推断。

6.2 知识库召回率相对于其他资源

还有其他几类资源可用于知识库召回评估。较为直接的比较对象是其他知识库，而文本和引导 elicited 人类联想则提供了不同的关注重点：前者关注文本通常传达的信息的召回率，后者关注人类自发联想到相关概念的信息的召回率。

知识库到知识库。 比较知识库召回率的一种方法是比较它们的规模。几乎每个知识库项目都会从规模方面将自身与其他资源进行比较，通常通过统计实体、类和陈述的数量来衡量（例如，参见 Knowledge Vault 论文 [32] 中的表 1）。一种更

在 [103]中进行了细粒度分析。针对 25 个重要的类，即使是源自同一来源的知识库（如 DBpedia 和 YAGO），也发现了显著差异。部分差异可由不同的建模方法解释（例如，维基数据中的政治人物数量较少，是因为它通过属性来建模一个类）。然而，规模比较无法反映一个知识库的内容是否包含在另一个知识库中，或包含程度如何——资源的关注重点可能各不相同，而且较大的知识库相对于较小的知识库，召回率仍可能有限。

为避免此类问题，我们可以考察一方所包含的实体占另一方实体的比例。 [103] 中的工作也分析了这一点，通过简单的字符串距离函数匹配实体。合并知识库通常有助于提高召回率，即使对于规模已经更大的资源也是如此，因为它们的重点各不相同。

知识库与文本。文本是知识存储与共享的主要形式，因此自然会有人想知道，知识库对文本中的信息能召回多少。值得注意的是，知识库的诞生，部分源于文本在信息结构化方面的不足，因此研究知识库较高的结构化程度在多大程度上会导致召回率损失，非常有意义。此外，一套有原则的方法来比较知识库相对于文本的召回率，可以涵盖远多于知识库之间结构化比较的领域和子领域，因为文本的数量远比结构化资源丰富。目前已有多种方法可用于估算知识库相对于文本的召回率。

在 Aristo TupleKB 项目中， [73], 将自动构建的科学知识库的召回率与科学文本中包含的信息进行比较。为进行这项召回评估，Mishra 等人从基础科学教科书、维基百科和词典中汇集了一个包含 120 万个句子的语料库，并从中自动抽取关系陈述。随后，他们可以通过测量这些陈述中有多少比例包含在知识库中，来量化知识库的“科学召回率”。这项研究针对 5 个知识库（WebChild、NELL、ConceptNet、ReVerb-15M、TupleKB）开展。由于这些知识库中的谓词名称各不相同，因此分析仅限于 20 个通用谓词，并将它们人工映射到更广泛的表层名称集合。此外，在匹配主体和对象时，只考虑词头。在这些宽松条件下，发现这些知识库的科学召回率介于 0.1% 和 23.2% 之间。

OPIEC 项目中也有一项相关分析， [40], 该项目的语料库由在整个英文维基百科上运行开放信息抽取系统而抽取出的开放式陈述组成。在[39], 作者分析了 OPIEC 与 DBpedia 的关系，尤其是 OPIEC 陈述在多大程度上可以用 DBpedia 表达，以及实际上有多大程度已在其中表达。由于开放信息抽取能够合理近似文本中包含的关系陈述，因此这项评估可视为对 DBpedia 相对于维基百科的召回率评估。作者提出了几项重要见解：（i）29% 的开放式陈述可以在 DBpedia 中得到完整表达，29% 可以部分表达（例如，使用更具体或更一般的谓词），42% 的开放式陈述无法用知识库表达。[14] （ii）在知识库中加入合取甚至存在量化等更复杂的结构，会提升其召回潜力。（iii）在衡量实际召回率时，他们发现，18% 的开放式陈述完整地存在于 DBpedia 中，23% 部分存在，59% 完全不存在。

在时间维度上，知识库与文本召回率已使用远程监督假设进行研究：假设维基百科页面上的超链接实体与页面主体相关，那么它们在维基数据中的召回率会如何随时间变化？维基数据的修订历史使这种纵向研究成为可能，从而可以观察到召回率自2012年至2020年总体稳步提升[96]。

[14]这一方面也可以被视为召回率的另一个维度，称为模式召回率。 [76], 也对这一方面进行了研究，其中作者发现，在维基百科的几篇文章中发现的163个谓词样本中，流行的知识库模式包含其中的13至126个谓词。

知识库与人类联想。 除了其他知识库和文本之外，人类本身也是用于相对召回率评估的一种有趣资源。知识库对人类自发联想到某些概念的陈述具有多高的召回能力？这种召回率的界定方式在近期的常识知识库构建（KBC）项目中十分突出 [77, 78, 105]。前一个项目直接向人类众包工作者提问（“当你想到狮子时，脑海中会浮现什么？”），而后两个项目则依赖于 CSLB 属性规范语料库 [29], ，这是一个在心理学项目背景下收集的大型概念联想数据集。遵循这一方案的评估通常采用基于嵌入的启发式匹配技术，以判断测试知识库是否包含某条参考陈述，同时允许措辞存在细微差异。召回率通常处于 5%–13% 的范围内，这表明当前常识知识库项目在覆盖人类知识方面仍存在显著差距。

分析。 我们讨论了三种相对于其他资源估算知识库召回率的方法。它们共同面临的挑战是如何比较知识片段：参照资源中的陈述可能措辞不同或含义模糊，因此要找到语义等价的陈述并非易事。此外，语义等价本身必然是一个不精确的概念，需要在最大可接受差异程度上作出一定程度上任意的判断，同时还需要作出困难的技术决策，确定如何实际衡量语义相关性。目前常见的一种技术解决方案是将要比较的两个陈述表示在潜在嵌入空间中（例如， [77] 使用 S-BERT [101]），然后利用余弦距离等相似度指标来判断两个候选项是否被视为等价。请注意，即使在结构最完善的场景——知识库与知识库之间的召回率——中也会出现这个问题，因为即便主体和对象已消歧，谓词名称通常仍是文本字符串，例如 任职于、雇主、隶属于。

上述每一种比较都有不同的目的：只要存在相似领域的知识库，就应当尝试评估知识库相对于其他知识库的召回率，因为从结构化来源进行数据集成是一项相对简单的任务，并且具有显著提升召回率的潜力。评估知识库相对于文本的召回率可以发挥两种作用：对于提升知识库召回率而言，对于文本抽取知识库，相对于文本进行召回率评估可以帮助了解文本抽取过程中损失了多少信息。这可以指导我们改进流程。另一方面，这种评估还有助于理解知识库在下游使用场景中的潜力与局限性（例如，维基数据通常适合计算美国参议员的年龄、性别和职业统计数据，但不适合判断其治理质量）。当知识库只是支持某个下游使用场景的多个选项之一时，这一点尤其重要；例如，问答通常就是这种情况（另见第 6.3节）。知识库相对于人类联想的召回率是最内在的维度。

6.3 知识库召回率与问答需求的关系

评估知识库召回率的另一种相对方法是考虑某个使用场景，并量化知识库在多大程度上满足该场景的数据需求。自然语言问答 可以说是知识管理和自然语言处理领域的一项顶尖任务，因此尤其适合用来阐释这一观点。

静态分析。 有几篇论文探讨知识库在多大程度上能够回答常见问答数据集中的问题。最常见的情况是比较基于文本的问答系统与基于知识库的问答系统。在 [81] 和 [89]中可以找到两个特别有代表性的比较。第一项工作分析了最先进的问答系统在 4 个通用领域问题的热门基准测试（NaturalQuestions、WebQuestions、TriviaQA、CuratedTREC）上的表现。研究发现，基于知识库的系统能够正确回答这些查询中的 26% 至 43%。该研究还通过将上述数字与基于文本的问答系统的表现（召回率为 45% 至 62%）进行比较，比较了知识库召回率与文本召回率（见第 6.2节）。在其他

换言之，文本通常比知识库具有更高的召回率，不过还需要权衡结构化资源的其他优势。[89]的工作分析了 6 个问答基准数据集（LCQuad 1.0 和 2.0、WikiAnswers、Google Trends、QuALD、ComQA）中的复杂查询，发现最先进的系统仅使用知识库能够回答其中 10% 至 19% 的查询，而仅使用文本则能回答 18% 至 36%。此类分析还有其他变体，例如针对对话式问答的研究 [23]。

这些研究的共同结论是，知识库在问答系统中的应用具有令人鼓舞的召回率，但距离覆盖全部查询集仍有相当大的差距，因此通常需要构建结合知识库与文本的混合问答系统。与此同时，报告的分数通常混淆了知识库本身的召回率与系统从知识库中提取正确答案的能力。由于查询回答是一个基于启发式方法且并不完美的过程，因此，对于这些问答数据集而言，知识库的内在召回率可能更高。

预测问答用例的召回率要求。 Hopkinson 等人 [52] 为召回率分析提供了一种有趣的新思路：他们不再假设知识库（KB）是固定的，也不再仅仅衡量召回率，而是定义了一个期望的查询召回率（95% 的查询应当能够由知识库回答），并进一步探究需要向知识库中纳入哪些内容才能实现这一目标。这并非一个显而易见的问题，因为信息需求可能因实体类型的不同而存在很大差异，而且问题并非均匀地针对各类实体及其属性。这项工作起源于一家工业实验室（Amazon Alexa），在这一场景下，此类业务需求是合理的。问答服务提供商可以据此相应地规划自动化抽取工作，也可以安排付费的知识库维护人员来完善知识库中的特定领域。此外，商业服务提供商能够访问作为该技术基础的用户查询日志。从技术层面来看，该工作通过实体的类别成员关系来表示实体，从查询日志中抽取每个实体的属性使用频率，并利用回归模型或神经网络模型预测新实体的谓词使用模式。研究结果表明，该方法能够以较高的准确性预测所需属性。此外，还可以根据有关不同实体被查询频率的使用数据，将需求加权要求提升至整个知识库的层面。对于 Alexa 知识库中一组不具代表性的实体，作者发现，为实现 95% 的查询可回答率目标所需的谓词中，目前已有 58% 被纳入知识库。

纵向发展。 [96] 分析了知识库效用如何随时间变化。作者从三个搜索引擎日志（AOL、Google、Bing 查询）中选取问题，然后通过人工标注找出知识库（维基数据和 DBpedia）最早能够回答这些问题的时间。例如，问题“*Italian Job* 是在哪里拍摄的？”自 2015 年 10 月 15 日起便可由维基数据回答，当时实体*Italian Job* 新增了拍摄地点 这一属性。将可回答查询的数量按时间绘制成图后发现，2003 年至 2020 年间这一数量稳步增长，近几年仅略有放缓。

6.4 总结

以绝对值估计召回率通常是一项艰巨的任务。本节提供了一种务实的替代方案，展示了如何相对于其他知识库、其他资源和问答用例来衡量知识库实体和陈述的召回率。

我们在表 5 中总结了本节的主要见解。这些方法各有优缺点，主要源于相对召回率的衡量难题，以及参照中可能存在的系统性遗漏。

实体相对召回率 可以说是最容易分析的，量化也相对简单，因为在知识库中匹配谓词并不困难。定义参照实体时仍需斟酌。例如，将爱因斯坦与其他物理学家比较，可能会得出与将他和其他小提琴演奏者比较截然不同的结果。此外，实体相对召回率容易受到

相对召回的重点 call	方面	工作	优势	局限性
实体与其他实体的比较	实体相对召回率	[15, 70]	支持基于现实世界知识进行排序/优先级确定，即使没有关于现实的知识	如果许多属性是可选的/如果值只有部分内容存在
	确定 most 相关的 谓词 每个实体	[44, 66, 95]	生成可解释的建议 关于在哪里补全知识库	难以处理可选谓词
从知识库到其他资源	从知识库到知识库	[32, 103]	可以快速提供关于……的建议 何时集成数据	相似主题 知识库通常不会可用的是
	知识库转文本	[39, 40, 73, 96]	有助于识别文本中的问题。 抽取，或帮助做出选择——在知识库或基于文本的下游——应用之间	文本陈述何时被知识库覆盖的阈值并不明确——ous
	知识库与人类的关联——tions	[77, 78, 105]	所有评估的最高追求	实际影响尚不明确
知识库到问答使用场景	计数 #查询由知识库问答系统作答	[23, 81, 89]	能切实揭示知识库为某个应用场景提供支持的效果	难以区分知识库召回率与问答系统性能
	系统预测性问答召回率	[52]	可预测满足用例所需的内容 满足用例所需的知识库要求-ment	需要大量查询日志
	纵向发展 发展	[96]	表明知识库一直在稳步改进-已针对问答证明	依赖启发式匹配

表5. 相对召回率评估总结。

知识库中系统性缺漏的影响。例如，如果知识库中完全没有谓词以……闻名，那么也就无法为爱因斯坦推荐这一谓词。

相对于其他资源的召回率能从更外部的视角审视知识库，尤其是选择文本作为参照时，可开展许多有趣的分析（例如，可以分别以维基百科、简明维基百科、传记书籍等为参照，计算爱因斯坦知识库条目的召回率）。与此同时，以外部资源为参照来量化召回率更具挑战性，因为这需要处理模式匹配（知识库-知识库设置），或不精确且含义模糊的谓词和实体表面形式（知识库-文本设置）。

针对使用场景的相对召回率 提供了清晰的指标来衡量知识库在下游任务中的表现；如果构建知识库时有明确的主要使用场景，这一点就很有优势。然而，这种优势往往也会带来挑战，因为知识库构建通常是一项长期、跨职能的工作 [126],而使用场景也在不断变化。

7 讨论与结论

我们以讨论召回率的时间维度（第7.1节）、大语言模型的影响（第7.2节）、知识库召回评估的实施建议（第7.3节）、一系列值得牢记的经验（第7.4节），以及未来研究面临的挑战列表（第7.5节）。

7.1 完备性与召回率的时间维度

现实不断变化：曾经担任总统的人会卸任，人们会结婚或离婚，而曾经没有死亡日期的人可能会有了死亡日期。知识库也遵循这一过程，并且通常会得到更新。在本节中，我们将讨论这些变化对完备性和召回率估计任务的影响。

许多知识库都在稳步增长。例如，维基数据在2015年包含2000万条陈述，而在2023年则包含12亿条陈述，这表明其召回率大幅提高。另一方面，现实也在不断演变，这意味着过去完备的领域后来可能变得不完备。在本节中，我们将讨论时间标注的形式化方法、抽取与外推方法，以及有关召回率趋势的观测研究。

时间标注的形式化方法。 完备性和基数陈述可以通过添加有关其时间有效性的信息来扩展。Darari 等人 [27], 例如，引入了 带时间戳的完备性陈述这一概念。这些陈述为完备性陈述添加了一个“最新有效”日期（该日期通常是、但不一定是其创建日期）。示例包括 “诺贝尔奖获得者截至*2023*年是完备的”，或 “*XYZ* 的出版物截至*2018*年是完备的”。类似地，Arnaout 等人 [10] 通过引入时间前缀这一概念扩展了否定陈述推理，从而可以得出这样的结论：与她的7位前任不同，德国总理安格拉·默克尔并非男性。

在许多知识库中，时间标注也用于正向陈述。例如，在维基数据中，典型的陈述会限定阿尔伯特·爱因斯坦的德国公民身份于 1933 年终止，或者他于 1921 年获得诺贝尔物理学奖。[15] 因此，为完备性和召回率陈述添加时间限定似乎相当自然。

抽取和外推时间信息。 对于通过文本抽取得到的完备性和召回率信息，一种合理的基线做法是将抽取时间视为最新有效时间。还可以采用更细粒度的标注方式，例如，利用文档创建时间元数据或文本中的时间表达式，来估计最新的（确定的）有效时间[116]。

鉴于带时间标注的完备性或召回率元数据，一个独立的问题是如何解读其最新有效时间之后的情况。例如，如果关于阿尔伯特·爱因斯坦的出版物在2018年是完备的，那么合理的假设是，到2023年情况依然如此。然而，不应对关于阿尔伯特·爱因斯坦的书籍得出同样的结论。区分这两种情况意味着进入预测知识的时间稳定性的领域。这一问题取决于领域知识，而且对于结构化知识 [31, 94] 和文本 [4]而言，似乎都尚未得到充分探索。

召回率趋势。 维基数据、YAGO 和 DBpedia 等知识库正在积极开发中，人们可能会想知道它们的召回率如何变化。在第6.3节中，我们讨论了针对问答需求和维基百科内容分析 DBpedia 与维基数据召回率的研究 [96]。值得注意的是，相对于固定的信息需求（2000 年代的问答日志、维基百科页面的一个快照），召回率从 2003 年到 2020 年持续提升，近几年仅略有放缓。然而，现实和信息需求也会随着时间推移而变化，因此，知识库能否更快地捕捉现实，抑或现实的发展速度超过知识库的表述能力，仍是一个悬而未决的问题。

7.2 大语言模型的影响

近年来，Bert [30],(chat-)GPT [92],和 LLaMA [121]等大语言模型（LLM）对自然语言处理产生了重大影响。这种影响已延伸至知识密集型任务，尤其也涉及知识库 [83, 119]。尽管完备性和召回率研究尚未充分利用这些进展，但大语言模型很可能通过多种方式影响这一领域。

间接影响：召回率更高的知识库。大语言模型支持知识库构建流程中的多个步骤，从而能够构建规模更大的知识库，进而提高召回率

[15] https://www.wikidata.org/wiki/Q937

[3, 123]。具体而言，大语言模型既可用于直接进行知识预测，也可与检索到的文本结合使用，从而改进现有的文本关系抽取方法 [68]。

直接影响：文本与结构化模态之间的链接将更加容易。 将文本陈述与结构化陈述关联起来，是影响本文所讨论的若干方法的问题，其中尤以基于同伴的推理为甚（见第 5 节，其中需要将现有陈述与同伴相关的陈述以及文本证据相匹配），知识库到文本的相对召回评估也面临这一问题（第 6.2 节）。鉴于大语言模型在潜在表示和匹配不同表述形式及表示方式的断言方面能力突出，我们可以期待这些领域很快取得进展。

直接影响：会话准则。 大语言模型尤其擅长可泛化语言任务，这类任务不需要特定实例的知识。在召回评估的语境中，其中一项任务是估计文本抽取上下文是否意味着内容完备（见第 3.2 节）。例如，ChatGPT 能够针对以下示例给出令人信服的回答，而这些示例涉及十分微妙的语言细节：

> **用户：** 文本蕴含：请估计第一句话蕴含第二句话的可能性有多大。前提：约翰带女儿玛丽去了学校。假设：约翰只有一个孩子。**ChatGPT**：[..]可能性很高。**用户：** 文本蕴含：请估计第一句话蕴含第二句话的可能性有多大。前提：约翰也带女儿玛丽去了学校。假设：约翰只有一个孩子。**ChatGPT**：[..]不蕴含。

直接影响：生成元数据。 仿照直接提示语言模型生成陈述的方法，[3, 123], 原则上也可以提示大语言模型生成完备性或召回率元数据。下面是一个示例。

> **用户：** 根据维基数据，巴拉克·奥巴马有两个孩子，玛莉亚和萨莎。这真的是全部吗？**ChatGPT**：是的，根据公开信息，巴拉克·奥巴马有两个女儿 [..]。**用户：** 根据维基数据，阿尔伯特·爱因斯坦发表过 75 篇学术论文。这真的是全部吗？**ChatGPT**：不是，阿尔伯特·爱因斯坦发表的学术论文数量估计约为 300 篇。[..]

在上述示例中，ChatGPT 对两个问题都给出了正确答案，证实了关于巴拉克·奥巴马子女的完备性，并否定了关于爱因斯坦著作的完备性，同时还提供了召回率信息（75/300）。然而，以这种方式获得的元数据本身存在严重的可信度问题。众所周知，大语言模型会虚构信息，尤其是在长尾主题上，因此，如果没有恰当的答案来源，这类推导很可能难以用于下游任务。由于原则性原因，很难为文本中的大语言模型生成内容标注来源，这仍是一项开放的研究挑战 [130]。这一普遍问题可能存在一个例外，即否定常识知识（参见第5.3节）。

在这里，正确性无论如何都很大程度上取决于上下文，而对显著性的要求也会制衡对正确性的要求。在这种情况下，大语言模型可以合理地生成有趣的否定候选项 [7]。

7.3 知识库召回评估建议

迄今为止，我们力求均衡地覆盖各种方法，这或许会让注重实践的读者思考：面对具体的知识库召回评估问题，最佳的处理方式是什么。下文将更具体地说明，对于特定问题，我们认为合理的方法应用顺序应当如何。我们区分三种情形：（i）开放式的从零开始构建知识库，在这种情形下，召回率意识可以融入构建工作；（ii）由用例驱动的构建，在这种情形下，可以直接根据用例指标调整工作；以及（iii）知识库整理，在这种情形下，需要评估现有知识库。

情形 1：开放式的从零开始构建知识库。对于开放式的从零开始构建知识库，即为广泛用途而新建知识库，我们建议将数据获取过程与元数据获取结合起来。具体而言，如果知识库内容是从文本中抽取的，我们建议使用第3.2节介绍的文本召回率估计技术，为抽取结果标注完备性置信度。如果数据由人工作者创建，我们建议在数据编写工具中增加用于收集元数据的字段，例如复选框，以便作者注明某个主题的记录何时完成，类似于 Cool-WD [90]。

<escape>**设置 2：用例驱动的自下而上知识库构建**。对于由特定用例驱动的知识库构建，我们建议按照第 6.3 节所述，使用从实际用途推导出的指标来组织召回评估。首先，应建立知识库需要回答的查询概况，例如通过从用例中抽样来实现。如果样本的覆盖范围有限，则应采用插值方法来推导出更广泛的概况 [52]。随后，可以根据这一查询概况评估知识库填充工作，即针对每种具体的填充技术或领域，比较其成本与收益，并据此确定优先级。</escape>

设置 3：知识库整理。在需要评估现有知识库的情况下，我们建议首先检查知识库内部是否存在基数信息，如第 4.1 节所述。接下来，如果有维基百科等高质量文本可用，我们建议利用文本中的基数断言，最好采用基于简单模板的抽取方法，如第 4.2 节所述。相对召回率，尤其是通过比较知识库内部的实体来实现的相对召回率（参见第 6.1 节），也有助于发现信息缺口。我们在第 3 节讨论过的统计性质，只有在其他所有方法都已尝试之后才应使用，因为它们的可靠性最低。

7.4 核心启示

本综述的关键要点是：

(1) **知识库是不完备的**：尽管知识库构建、语义网和信息抽取领域历史悠久，但构建一个通用领域知识库本质上是一项模糊且不断演变的任务。因此，这类知识库始终是不完备的，我们必须能够应对这种不完备性，而不是寄希望于它会消失（第 2 节）。(2) **知识库几乎不包含否定信息（但应该包含）**：否定信息对下游任务非常有用，但遗憾的是，目前的知识库中这类信息少之又少。选择性地物化有意义的否定信息，可以显著提升知识库的效用（第2 节和第5节）。

(3) **预测技术适用于一系列出人意料的范式**：除了经典的监督式预测之外，数量分布、样本重叠和密度不变量等统计性质，即使没有典型的训练，也能实现召回率预测（第 3 节）。**(4) 计数信息是深入了解召回率的主要途径**：计数信息为召回评估提供了最直接的途径，既可以从现有知识库中获取，也可以从文本中获取（第4 节）。**(5) 显著否定可以通过启发式方法进行物化**：尽管否定知识近乎无限，但通过启发式方法物化相关部分，可以显著补充仅包含肯定知识的知识库（第5 节）。**(6) 相对召回率是绝对概念的一种切实可行的替代方案**：将知识库实体与其他知识库实体、外部资源或用例需求进行比较，可以从另一个有价值的角度审视知识库召回率（第6 节）。

7.5 挑战与机遇

在最后这一节中，我们概述了为全面理解知识库召回率仍有待解决的一些开放挑战，并指出了开展原创且可能产生重要影响的研究机遇。

(1) **开发高准确性的召回率估计器**。 本综述介绍的大多数估计器都只是概念验证，仅在有限领域中经过测试，或建立在非常具体的假设之上。如何构建切实可用且准确性高的估计器（可能通过结合多种互补的估计技术），仍是一项重大的开放挑战。**(2) 利用召回率估计来实现价值驱动的知识库补全**。 尽管二者显然密切相关，但迄今为止，召回率估计与知识库补全的研究基本上是独立发展的。量化知识的价值（如第 6.3 节所述），并制定优先策略，以提升召回率并最大化现有知识的价值 [52, 94],，是具有重要实际影响的研究良机。**(3) 估计预训练语言模型的召回率**。 从预训练语言模型中抽取知识，近期受到了广泛关注 [87], ，但这种方法在多值、可选和长尾谓词上能够获取多少知识，仍不清楚 [56, 112]。系统地测量语言模型的召回率，并将其与结构化知识库进行比较，是一项开放挑战。

近年来，知识库受到了广泛关注。尽管构建知识库时通常侧重于精确率，但要理解其召回率仍是一项重大挑战。在本综述中，我们系统梳理了知识库召回评估的主要途径，并概述了切实可行的方法和亟待解决的挑战。我们希望本综述能启发读者从新的角度审视知识库质量，并推动更多知识库项目系统地记录和反映其召回率。

参考文献

[1] Maribel Acosta、Elena Simperl、Fabian Flöck 和 Maria-Esther Vidal。2017。通过众包提升 SPARQL 查询的回答完备性。网络语义学期刊45（2017），41–62。[2] Albin Ahmeti、Simon Razniewski 和 Axel Polleres。2017。评估知识库中实体的完备性。载于语义网：*ESWC 2017* 卫星活动（*ESWC 2017*）。施普林格，7–11。[3] Dimitrios Alivanistos、Selene Báez Santamaría、Michael Cochez、Jan Christoph Kalo、Emile van Krieken 和 Thiviyan Thanapalasingam。2022。提示即探测：利用语言模型构建知识库。载于预训练语言模型知识库构建语义网挑战赛（*LM-KBC*）。[4] Axel Almquist 和 Adam Jatowt。2019。探索内容到期日期的判定：预测句子的有效期。载于信息检索进展：第*41*届欧洲信息检索研究会议（*ECIR 2019*）。施普林格国际出版公司，86–101。https://doi.org/10.1007/978-3-030-15712-8_6

[5] Christoph Alt、Marc Hübner 和 Leonhard Hennig。2019。将预训练变换器语言模型微调用于远程监督关系抽取。载于计算语言学协会年会（ACL）。1388–1398。[6] Hiba Arnaout。2023。用表达性否定陈述丰富开放世界知识图谱。博士论文。（即将发表）。[7] Hiba Arnaout 和 Simon Razniewski。2023。大语言模型能否生成显著否定陈述？载于ISWC 2023预训练语言模型知识库构建研讨会（KBC-LM）。[8] Hiba Arnaout、Simon Razniewski 和 Gerhard Weikum。2020。用有趣的否定陈述丰富知识库。载于自动知识库构建（AKBC ’20）。[9] Hiba Arnaout、Simon Razniewski、Gerhard Weikum 和 Jeff Z Pan。2021。开放世界维基数据中的否定知识。载于2021年万维网大会伴随论文集（WWW ’21）。计算机协会， 544 551 10 1145 3442442 3452339 – 。https://doi.org/ 。/ 。Hiba Arnaout、Simon Razniewski、Gerhard Weikum 和 Jeff Z. Pan。2021。被认为有用的否定陈述。Web Semant. 71, 2021 10 1016 2021 100661 C （） 。https://doi.org/ 。/j.websem. 。Hiba Arnaout、Simon Razniewski、Gerhard Weikum 和 Jeff Z Pan。2021。Wikinegata：包含有趣否定陈述的知识库。Proc. VLDB Endow. 14, 12 2021 2807 2810 10 14778 3476311 3476350 （） ， – 。https://doi.org/ 。/ 。Hiba Arnaout 、 Simon Razniewski 、 Gerhard Weikum 和 Jeff Z Pan。。UnCommonSense：关于日常概念的信息性否定知识。载于第31届ACM信息与知识管理国际会议论文集（CIKM ’22）。计算机协会，37–46。https://doi.org/10.1145/3511808. 3557484[13] Sören Auer、Christian Bizer、Georgi Kobilarov、Jens Lehmann、Richard Cyganiak 和 Zachary G. Ives。2007。DBpedia：开放数据网络的核心。载于语义网（ISWC）。Springer Berlin Heidelberg，722–735。[14] Franz Baader、Diego Calvanese、Deborah McGuinness、Peter Patel-Schneider、Daniele Nardi 等。2003。描述逻辑手册：理论、实现与应用。剑桥大学出版社。[15] Vevake Balaraman、Simon Razniewski 和 Werner Nutt。2018。Recoin：维基数据中的相对完备性。载于2018年万维网大会伴随论文集（WWW ’18）。国际万维网会议指导委员会， 1787 1792 10 1145 3184558 3191641 – 。https://doi.org/ 。/ 。Hannah Bast、Björn Buchhold 和 Elmar Haussmann。。类型关系三元组的相关性评分。载于第38届ACM SIGIR信息检索研究与发展国际会议论文集（SIGIR ’15）。计算机协会，243–252。https://doi.org/10.1145/2766462.2767734[17] Mohamed Ben Ellefi、Zohra Bellahsene、John G Breslin、Elena Demidova、Stefan Dietze、Julian Szymański 和 Konstantin Todorov。2018。RDF数据集剖析——关于特征、方法、词汇表和应用的综述。Semantic Web 9, 5（2018），677–705。[18] Frank Benford。1938。异常数字定律。美国哲学学会会刊（1938）。[19] Michael K Bergman。2001。白皮书：深网：发掘隐藏价值。电子出版杂志（2001）。[20] Kurt Bollacker、Robert Cook 和 Patrick Tufts。2007。Freebase：结构化通用人类知识的共享数据库。载于第22届全国人工智能大会论文集——第2卷（AAAI’07）。https: //dl.acm.org/doi/abs/ 10 5555 1619797 1619981 . / 。Andrew Carlson、Justin Betteridge、Bryan Kisiel、Burr Settles、Estevam Hruschka 和 Tom Mitchell。2010。迈向永不停歇的语言学习架构。载于第24届AAAI人工智能会议论文集（AAAI ’10）。AAAI 出版社，1306–1313。[22] Jiangjie Chen、Wei Shi、Ziquan Fu、Sijie Cheng、Lei Li 和 Yanghua Xiao。2023。说出你的本意！大语言模型对否定常识知识的表述过于正面。载于ACL。[23] Philipp Christmann、Rishiraj Saha Roy 和 Gerhard Weikum。2022。异构来源上的对话式问答。载于第45届ACM SIGIR信息检索研究与发展国际会议论文集（SIGIR ’ *22* 144 154 10 1145 3477495 3531815 ）。计算机协会，– 。https://doi.org/ 。/ 。Cuong Xuan Chu、Simon Razniewski 和 Gerhard Weikum。。Knowfi：从长篇虚构文本中抽取知识。载于自动知识库构建（AKBC）。1–19。[25] Luciano Del Corro 和 Rainer Gemulla。2013。ClausIE：基于分句的开放信息抽取。载于第22届国际万维网会议论文集（WWW ’13）。计算机协会，355–366。https://doi.org/ 10 1145 2488388 2488420 . / 。Fariz Darari、Werner Nutt、Giuseppe Pirro 和 Simon Razniewski。2013。关于RDF数据源的完备性陈述及其在查询回答中的应用。载于语义网—ISWC 2013：第12届语义网国际会议（ISWC *2013* 66 83 10 1007 978 3 642 41335 3_5 ）。施普林格，– 。https://doi.org/ 。/ - - - - Fariz Darari、Werner Nutt、Giuseppe Pirrò 和 Simon Razniewski。。RDF数据源的完备性管理。ACM Trans. Web 12, 3（2018）。https://doi.org/10.1145/3196248

[28] Marc Denecker、Álvaro Cortés-Calabuig、Maurice Bruynooghes 和 Ofer Arieli。2008年。迈向局部闭数据库理论的逻辑重构。ACM数据库系统汇刊。 35, 3 2008 10 1145 1806907 1806914 ( )。

[29] 2014 https://doi.org/ . / 。Barry J Devereux、Lorraine K Tyler、Jeroen Geertzen 和 Billi Randall。。言语、语言与大脑中心（CSLB）概念属性规范。行为研究方法 46（2014），1119–1127。https://doi.org/10.3758/s13428- 013-0420-4[30] Jacob Devlin、Ming-Wei Chang 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。载于2019年计算语言学协会北美分会会议论文集。 4171 4186 10 18653 1 19 1423 – 。https://doi.org/ 。/v /n -

[31] 2019 Ioannis Dikeoulias、Jannik Strötgen 和 Simon Razniewski。。墓志铭还是突发新闻？知识库属性的稳定性分析与预测。载于2019年万维网会议（WWW '19）伴随论文集。计算机协会，1155–1158。https://doi.org/10.1145/3308560.3314998[32] Xin Luna Dong、Evgeniy Gabrilovich、Geremy Heitz、Wilko Horn、Ni Lao、Kevin Murphy、Thomas Strohmann、Shaohua Sun 和 Wei Zhang。2014年。Knowledge Vault：一种面向网络规模的概率知识融合方法。载于第20届ACM SIGKDD国际知识发现与数据挖掘会议（KDD '14）论文集。计算机协会， 601 610 10 1145 2623330 2623623 – 。https://doi.org/ 。/ 。Xin Luna

[33] Dong、Evgeniy Gabrilovich、Geremy Heitz、Wiko Horn、Kevin Murphy、Shaohua Sun 和 Wei Zhang。2014年。从数据融合到知识融合。Proc. VLDB Endow. 7, 10 2014 10 14778 2732951 2732962 ( )。

[34] 2015 https://doi.org/ 。/ 。Patrick Ernst、Amy Siu 和 Gerhard Weikum。。Knowlife：一种构建大型生物医学科学知识图谱的多用途方法。BMC生物信息学 16（2015），1–13。[35] Michael Färber、Frederic Bartscherer、Carsten Menne 和 Achim Rettinger。2018年。DBpedia、Freebase、OpenCyc、维基数据和YAGO的关联数据质量。语义网 9, 1 2018 77 129 10 3233 170275 ( ), – 。https://doi.org/ 。/

[36] 2017 SW- Luis Galárraga、Simon Razniewski、Antoine Amarilli 和 Fabian M Suchanek。。预测知识库的完备性。载于第十届ACM国际网络搜索与数据挖掘会议（WSDM ' *17* 375 383 10 1145 3018661 3018739 ）论文集。计算机协会，– 。

[37] 2015 https://doi.org/ 。/ 。Luis Galárraga、Christina Teflioudi、Katja Hose 和 Fabian M. Suchanek。。使用AMIE在本体知识库中快速进行规则挖掘 + 24, 6 2015 707 730 10 1007 00778 015 0394 1 。

[38] 2013 VLDB期刊（ ），– 。https://doi.org/ 。/s - - - Luis Antonio Galárraga、Christina Teflioudi、Katja Hose 和 Fabian S uchanek。。AMIE：不完整证据下本体知识库中的关联规则挖掘。载于第22届万维网国际会议（WWW ' *13* 413 422 10 1145 2488388 2488425 ）论文集。计算机协

[39] 2020 会，– 。https://doi.org/ 。/ 。Kiril Gashteovski、Rainer Gemulla、Bhushan Kotnis、Sven Hertling 和 Christian Meilicke。。将开放信息抽取结果与知识库对齐：案例研究。载于第一届自然语言处理系统评估与比较研讨会论文集。计算语言学协会， 143 154 10 18653 1 2020 4 1 14 – 。https://doi.org/ 。/

[40] 2018 v / 。eval nlp- 。Kiril Gashteovski、Sebastian Wanner、Sven Hertling、Samuel Broscheit 和 Rainer Gemulla。。OPIEC：开放信息抽取语料库。载于自动知识库构建（AKBC）。[41] Shrestha Ghosh、Simon Razniewski 和 Gerhard Weikum。2020年。CounQER：用于发现和关联知识库中计数信息的系统。载于语义网：ESWC 2020卫星会议：ESWC 2020卫星会议（ESWC *2020* 84 90 10 1007 978 3 030 62327 2_15 ）。施普林格出版社，– 。https://doi.org/ 。/ - -

[42] 2020 - - Shrestha Ghosh 、 Simon Razniewski 和 Gerhard Weikum。。揭示知识库中集合信息的隐藏语义。网络语义学期刊 64 2020 10 1016 2020 100588 ( )。https://doi.org/ 。/

[43] 2019 j.websem. 。Arnaud Giacometti、Béatrice Markhoff 和 Arnaud Soulet。。挖掘知识库中的显著最大基数。载于语义网——ISWC 2019：第18届国际语义网会议。施普林格出版社，182–199。https://doi.org/10.1007/978-3-030-30793-6_11[44] Lars C Gleim、Rafael Schimassek、Dominik Hüser、Maximilian Peters、Christoph Krämer、Michael Cochez 和 Stefan Decker。2020年。SchemaTree：面向维基数据的最大似然属性推荐。载于欧洲语义网会议（ESWC 2020）。施普林格，179–195。[45] Herbert P Grice。1975年。逻辑与会话。载于言语行为。Brill。[46] Chadi Helwe、Chloé Clavel 和 Fabian Suchanek。2021年。使用基于变换器的模型进行推理：深度学习，但浅层推理。载于自动知识库构建国际会议（AKBC）。[47] Sven Hertling 和 Heiko Paulheim。2017年。WebIsALOD：将从网络中抽取的上位词关系作为关联开放数据提供。载于语义网——ISWC 2017：第16届国际语义网会议（ISWC）。施普林格出版社， 111 119 10 1007 978 3 319 68204 4_11 – 。

[48] 1994 https://doi.org/ 。/ - - - - Wassily Hoeffding。。有界随机变量之和的概率不等式。载于Wassily Hoeffding文集。施普林格， 409 426 10 1007 978 1 4612 0865 5_26 – 。https://doi.org/ 。/ - - - - Johannes

[49] 2014 Hoffart、Dragan Milchevski 和 Gerhard Weikum。。STICS：使用字符串、事物与猫进行搜索。载于信息检索研究与发展会议。[50] Aidan Hogan、Eva Blomqvist、Michael Cochez、Claudia d'Amato、Gerard De Melo、Claudio Gutierrez、Sabrina Kirrane、José Emilio Labra Gayo、Roberto Navigli、Sebastian Neumaier 等。2021年。知识图谱。ACM计算综述 54,

4（2021）。https://doi.org/10.1145/3447772
[51] Bernhard《概念语言中的固定数目约束》。载于第二届知识表示与推理原则国际会议（KR'91）论文集。Morgan Kaufmann Publishers Inc.，335–346.[52] Andrew Hopkinson、Amit Gurdasani、Dave Palfrey 和 Arpit Mittal。2018。《面向知识库的需求加权完备性预测》。载于2018年计算语言学协会北美分会会议：人类语言技术论文集，第3卷（工业论文）（NAACL '18）。计算语言学协会，
200 207 10 18653 1 18 3025–。https://doi.org/。/v /N - Subhi Issa、[53] Onaopepo Adekunle、Fayçal Hamdi、Samira Si-Said Cherfi、Michel Dumontier 和 Amrapali Zaveri。2021。《知识图谱完备性：系统性文献综述》。IEEE Access 9（2021），31322–31339。https: //doi.org/
10 1109 2021 3056622 . /ACCESS. 。Subhi Issa、Pierre-Henri Paris 和 Fayçal Hamdi。2017。[54]《评估 DBpedia 完备性的演变：案例研究》。载于《概念建模进展：ER 2017 工作坊 AHA、MoBiD、MREBA、OntoCom 和 QMMQ》。施普林格，238–247。[55] Shaoxiong Ji、Shirui Pan、Erik Cambria、Pekka Marttinen 和 S Yu Philip。2022。《知识图谱综述：表示、获取与应用》。神经网络与学习系统汇刊 33，2（2022），
494 514 10 1109 2021 3070843 –。https://doi.org/。/TNNLS.。Zhengbao Jiang、Jun [56] Araki、Haibo Ding 和 Graham Neubig。2021。《我们如何知道语言模型何时具备知识？论用于问答的语言模型校准》。计算语言学协会汇刊 9 2021 962 977 10 1162 _ _00407 ( )，–。
[57] https://doi.org/。/tacl a Martin Josifoski、Nicola De Cao、Maxime Peyrard、Fabio Petroni 和 Robert West。2022。《GenIE：生成式信息抽取》。载于计算语言学协会北美分会（NAACL）。4626–4643。[58] Georgios Karagiannis、Immanuel Trummer、Saehan Jo、Shubham Khandelwal、Xuezhi Wang 和 Cong Yu。2019。《从维基百科更新中挖掘"反知识库"及其在事实核查等领域的应用》。Proc. VLDB Endow.
13, 4 2019 561 573 10 14778 3372716 3372727 ( )，–。https://doi.org/。/。Nicholas Klein、[59] Filip Ilievski、Hayden Freedman 和 Pedro Szekely。2022。《识别维基数据中的意外事实》。维基数据工作坊（2022）。[60] Graham Klyne、Jeremy J Carroll 和 Brian McBride。2004。《资源描述框架（RDF）：概念与抽象语法》。W3C 建议标准。[61] Jonathan Lajus 和 Fabian M Suchanek。2018。《所有人都结婚了吗？确定知识库中的必需属性》。载于2018年万维网会议论文集（WWW '18）。国际万维网会议指导委员会，
1115 1124 10 1145 3178876 3186010 –。https://doi.org/。/。Willis Lang、[62] Rimma V Nehme、Eric Robinson 和 Jeffrey F Naughton。2014。《数据库系统中的部分结果》（SIGMOD '14）。计算机协会，美国纽约州纽约市，https://doi.org/10.1145/2588555.2612176[63] Philipp Langer、Patrick Schulze、Stefan George、Matthias Kohnen、Tobias Metzke、Ziawasch Abedjan 和 Gjergji Kasneci。2014。《为 DBpedia 事实分配全局相关性评分》。载于2014年 IEEE 第30届数据工程国际会议研讨会。IEEE，248–253。[64] Alon Y Levy。1996。《从不完整数据库中获取完整答案》。载于第22届超大型数据库国际会议论文集（VLDB '96）。402–412。[65] Bill Yuchen Lin、Seyeon Lee、Rahul Khanna 和 Xiang Ren。2020。《鸟有四条腿？！NumerSense：探测预训练语言模型中的数值常识知识》。载于2020 年自然语言处理实证方法会议（EMNLP）论文集。计算语言学协会，6862–6868。https://doi.org/ 10 18653 1 2020 557 . /v / .
[66] emnlp-main。Michael Luggen、Julien Audiffren、Djellel Difallah 和 Philippe Cudré-Mauroux。2021。《Wiki2Prop：一种从维基百科预测维基数据属性的多模态方法》。载于2021年万维网会议论文集（WWW '21）。计算机协会，
2357 2366 10 1145 3442381 3450082 –。
[67] https://doi.org/。/。Michael Luggen、Djellel Difallah、Cristina Sarasua、Gianluca Demartini 和 Philippe Cudré-Mauroux。2019。《协作知识图谱的非参数类完备性估计器——以维基数据为例》。载于《语义网——ISWC 2019：第18届国际语义网会议》。施普林格，453–469。[68] Alex Mallen、Akari Asai、Victor Zhong、Rajarshi Das、Daniel Khashabi 和 Hannaneh Hajishirzi。2023。《何时不应信任语言模型：探究参数化与非参数化记忆的有效性》。载于计算语言学协会第61届年会论文集（第1卷：长篇论文）（ACL）。计算语言学协会，
9802 9822 10 18653 1 2023 546 –。https://doi.org/。/v / .
[69] 2012 acl-long。Mausam、Michael Schmitz、Stephen Soderland、Robert Bart 和 Oren Etzioni。。《面向信息抽取的开放式语言学习》。载于2012年自然语言处理实证方法与计算自然语言学习联合会议论文集（EMNLP-CoNLL '12）。计算语言学协会，523–534。

[70] MediaWiki。2022。ORES。检索于2022年7月1日，网址：https://www.mediawiki.org/w/index.php?title=ORES [71] Paramita Mirza、Simon Razniewski、Fariz Darari 和 Gerhard Weikum。2018。《利用计数量词丰富知识库》。载于国际语义网会议（ISWC）。施普林格，179–197。[72] Paramita Mirza、Simon Razniewski 和 Werner Nutt。2016。《将维基数据中的父母信息扩充178%，或如何挖掘关系基数信息》。载于《语义网：ISWC 2016 海报与演示文稿集》。CEUR-WS.org。[73] Bhavana Dalvi Mishra、Niket Tandon 和 Peter Clark。2017。《领域定向的高精确率知识抽取》。计算语言学协会汇刊 5 2017 233 246 10 1162 _ _00058 ( )，–。[74] 1989 = + 14, 4 1989 480 502 https://doi.org/。/tacl a Amihai Motro。。《完整性、有效性与完备性》。ACM Trans. Database Syst. ( )，–。https: //doi.org/ 10 1145 76902 76904 . /。Emir Muñoz、Pasquale Minervini 和 Matthias Nickles。2019。《在神经链接预测器中嵌入[75]基数约束》。载于第34届 ACM/SIGAPP 应用计算研讨会论文集（SAC '19）。计算机协会， 2243 2250 10 1145 3297280 3297502 –。https://doi.org/。/。Ndapandula Nakashole、[76] 2012 Gerhard Weikum 和 Fabian Suchanek。《PATTY：具有语义类型的关系模式分类体系》。载于2012年自然语言处理实证方法与计算自然语言学习联合会议论文集（EMNLP-CoNLL '12）。计算语言学协会，1135–1145。[77] Tuan-Phong Nguyen 和 Simon Razniewski。2022。《从常识变换器构建物化知识库》。(2022 36 42 10 18653 1 2022 1 5)，–。[78] https://doi.org/。/v / .csrr-。Tuan-Phong Nguyen、Simon Razniewski、Julien Romero 和 Gerhard Weikum。2023。《从大规模网络内容中提炼常识知识》。IEEE Transactions on Knowledge and Data Engineering 35，8（2023），8431–8447。https://doi.org/ 10 1109 2022 3206505 . /TKDE.。Tuan-Phong Nguyen、Simon Razniewski 和 Gerhard Weikum。2021。[79]《常识知识抽取的高级语义》。载于2021年万维网会议论文集（WWW '21）。计算机协会， 2636 2647 10 1145 3442381 3449827 –。https://doi.org/。/。Natasha Noy、Yuqing Gao、Anshu Jain、Anant [80] Narayanan、Alan Patterson 和 Jamie Taylor。2019。《产业级知识图谱：经验与挑战：五家各具特色的科技公司展示实践之道》。Queue 17，2（2019），48–75。https://doi.org/10.1145/3329781.3332266[81] Barlas Oguz、Xilun Chen、Vladimir Karpukhin、Stan Peshterliev、Dmytro Okhonko、Michael Schlichtkrull、Sonal Gupta、Yashar Mehdad 和 Scott Yih。2022。《Unik-QA：面向开放领域问答的结构化与非结构化知识统一表示》。(2022 1535 1546 10 18653 1 2022 115)，–。https://doi.org/。/v / .[82] 2022 findings-naacl。OpenAI。。《推出 ChatGPT》。https://openai.com/blog/chatgpt。[83] Jeff Z. Pan、Simon Razniewski、Jan-Christoph Kalo、Sneha Singhania、Jiaoyan Chen、Stefan Dietze、Hajira Jabeen、Janna Omeliyanenko、Wen Z hang、Matteo Lissandrini、Russa Biswas、Gerard de Melo、Angela Bonifati、Edlira Vakaj、Mauro Dragoni 和 Damien Graux。2023。《大语言模型与知识图谱：机遇与挑战》。arXiv:2308.06374[84] Pierre-Henri Paris、Syrine El Aoud 和 Fabian M. Suchanek。2021。《名词短语中的模糊性之模糊》。载于自动知识库构建国际会议（AKBC *2021* 03344675 ）。[85] 2017 https://imt.hal.science/hal- Heiko Paulheim。《知识图谱精炼：方法与评估方法综述》。Semant. Web 8, 3 2017 489 508 10 3233 160218 ( )，–。https://doi.org/。/SW- Thomas Pellissier Tanon、Daria [86] Stepanova、Simon Razniewski、Paramita Mirza 和 Gerhard Weikum。2017。《从知识图谱中进行完备性感知规则学习》。载于《语义网——ISWC 2017》。[87] Fabio Petroni、Tim Rocktäschel、Sebastian Riedel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu 和 Alexander Miller。2019。《语言模型是知识库吗？》。计算语言学协会，中国香港，2463– 2473 10 18653 1 19 1250 。https://doi.org/。/v /D - Marco Ponza、Paolo Ferragina 和 Soumen [88] Chakrabarti。2017。《用于计算维基百科中实体相关性的两阶段框架》。载于2017年 ACM 信息与知识管理会议论文集（CIKM ' *17* 1867 1876 10 1145 3132847 3132890 ）。计算机协会，–。[89] 2021 https://doi.org/。/。Soumajit Pramanik、Jesujoba Alabi、Rishiraj Saha Roy 和 Gerhard Weikum。《UNIQORN：基于 R DF 知识图谱和自然语言文本的统一问答》。arXiv: 2108 08614 。Radityo Eko Prasojo、Fariz Darari、[90] 2016 Simon Razniewski 和 Werner Nutt。《使用 COOL-WD 管理和消费维基数据的完备性信息》。第七届关联数据消费国际研讨会 CEUR Workshop Proceedings 1666（2016）。[91] Péter Rácz。2013。《社会语言学中的显著性：定量方法》。第84卷。Walter de Gruyter。[92] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019。《语言模型是无监督的多任务学习者》。OpenAI 博客 1，8（2019），9。

[93] Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。《Know What You Don't Know：SQuAD 中的不可回答问题》。载于计算语言学协会第56届年会论文集（第2卷：短论文）。计算语言学协会，784 789 10 18653 1 18 2124 –。https://doi.org/。/v/P - [94] Simon Razniewski。2016。。《为衰减信息优化更新频率》。载于第25届 ACM 信息与知识管理国际会议（CIKM '16）论文集。计算机协会，1191 1200 10 1145 2983323 2983719 –。https://doi.org/。/。Simon Razniewski、[95] Vevake Balaraman 和 Werner Nutt。2017。。《博士生导师还是医学状况：迈向知识库属性的实体特定排名》。载于《高级数据挖掘与应用：第13届国际会议（ADMA 2017）》。施普林格，526–540。[96] Simon Razniewski 和 Priyanka Das。2020年。《结构化知识：我们取得进展了吗？对19年来知识库覆盖率的外部研究》。载于第29届 ACM 信息与知识管理国际会议（CIKM ' 20 3317 3320 10 1145 3340531 3417447 ）论文集。[97] 计算机协会，–。https://doi.org/。/。Simon Razniewski、Nitisha Jain、2019 Paramita Mirza 和 Gerhard Weikum。。《从句子和段落中抽取信息的覆盖率》。载于2019年自然语言处理实证方法会议和第九届自然语言处理国际联合会议（EMNLP-IJCNLP）论文集。计算语言学协会，中国香港，5771 5776 10 18653 1 19 1583 –。https://doi.org/。/v/D - [98] Simon Razniewski、Flip Korn、Werner Nutt 和 Divesh Srivastava。2015。。《确定部分完整数据库中查询答案的完备程度》。载于2015年 ACM SIGMOD 数据管理国际会议（SIGMOD '15）论文集。计算机协会，561–576。https://doi.org/10.1145/2723372。2750544[99] Simon Razniewski 和 Werner Nutt。2011年4月11日 2011年 749 760 10 14778 3402707。《不完整数据库上的查询完备性》。Proc. VLDB Endow.（），–。https://doi.org/。/。3402715 [100] Simon Razniewski、Fabian M. Suchanek 和 Werner Nutt。2016年。《但我们究竟知道些什么？》。载于第五届自动知识库构建研讨会（AKBC）论文集。40–44。[101] Nils Reimers 和 Iryna Gurevych。2019年。《Sentence-BERT：使用孪生BERT网络的句子嵌入》。计算语言学协会，中国香港，3982 3992 10 18653 1 19 1410 –。[102] https://doi.org/。/v/D - 1981 Raymond Reiter。。《论封闭世界数据库》。载于《人工智能读本》。Elsevier，–。[103] Daniel Ringler 和 Heiko Paulheim。2017年。《一个知识图谱统领一切？分析 DBpedia、YAGO、维基数据等之间的差异》。载于《人工智能进展：第40届德国人工智能年会论文集》（KI 2017）。施普林格，366–372。[104] D. S. Robson 和 H. A. Regier。1964年。《Petersen标记重捕实验中的样本量》。美国渔业学会会刊 3（1964），215–226。[105] Julien Romero、Simon Razniewski、Koninika Pal、Jeff Z. Pan、Archit Sakhadeo 和 Gerhard Weikum。2019年。《来自查询日志和问答论坛的常识属性》。载于第28届 ACM 信息与知识管理国际会议（CIKM '19）论文集。计算机协会，1411–1420。https://doi.org/ 10 1145 3357384 3357955。/。Subhro Roy、Tim Vieira 和 Dan Roth。2015年。《自然语言中的数量推理》。[106] 计算语言学协会汇刊 3 2015 1 13 10 1162 _ _00118 （），–。https://doi.org/。/ [107] tacl a Tara Safavi 和 Danai Koutra。2020。。《生成否定常识知识》。NeurIPS 知识表示与推理和机器学习研讨会（2020）。[108] Tara Safavi、Jing Zhu 和 Danai Koutra。2021年。《NegatER：在常识知识库中无监督发现否定知识》。载于2021年自然语言处理实证方法会议论文集。计算语言学协会，5633 5646 10 18653 1 2021 456 –。https://doi.org/。/v/。[109] emnlp-main。Ognjen Savković、Paramita Mirza、Alex Tomasi 和 2013 Werner Nutt。。《不完整查询的完备近似》。Proc. VLDB Endow. 6, 12 2013 1378 1381 10 14778 2536274 2536320 （），–。https://doi.org/。/。[110] Amit Singhal。2012。。《推出知识图谱：关注事物，而非字符串》。检索日期：2023年8月，网址：https: // blog.google/products/search/introducing-knowledge-graph-things-not/ [111] Sneha Singhania、Jan-Christoph Kalo、Simon Razniewski 和 Jeff Z. Pan。2023年。《从预训练语言模型构建知识库》。载于 ISWC 2023 挑战赛。[112] Sneha Singhania、Tuan-Phong Nguyen 和 Simon Razniewski。2022年。《LM-KBC：从预训练语言模型构建知识库》。载于 ISWC 2022 挑战赛。https://ceur-ws.org/Vol-3274/paper1.pdf [113] Sneha Singhania、Simon Razniewski 和 Gerhard Weikum。2023年。《从语言模型中抽取多值关系》。载于自然语言处理表示学习研讨会（RepL4NLP 2023）。[114] Arnaud Soulet、Arnaud Giacometti、Béatrice Markhoff 和 Fabian M Suchanek。2018年。《基于广义本福特定律的知识库代表性》。载于《语义网——ISWC 2018：第17届国际语义网会议》。施普林格，374–390。

[115] Robyn Speer、Joshua Chin 和 Catherine Havasi。2017年。《ConceptNet 5.5：通用知识的开放式多语言图谱》。载于第三十一届 AAAI 人工智能大会论文集（AAAI ’17）。AAAI 出版社，4444–4451。[116] Jannik Strötgen 和 Michael Gertz。2010年。《Heideltime：高质量的基于规则的时间表达式抽取与规范化》。载于第五届语义评估国际研讨会（SemEval ’10）论文集。计算语言学协会，321–324。[117] Fabian M Suchanek、Gjergji Kasneci 和 Gerhard Weikum。2007年。《Yago：语义知识的核心》。载于第十六届万维网国际会议（WWW ’07）论文集。计算机协会，697–706。https://doi.org/ 10 1145 1242572 1242667 。/。[118] Fabian M. Suchanek、Jonathan Lajus、Armand Boschin 和 Gerhard Weikum。2019年。《以实体为中心的知识库中的知识表示与规则挖掘》。Reasoning Web。《可解释人工智能：第15届国际暑期学校2019年教程讲义》，110–152。[119] Fabian M. Suchanek 和 Anh Tuan Luu。2023年。《知识库与语言模型：相辅相成》。载于 RuleML+RR 邀请论文。[120] Shelley E Taylor 和 Susan T Fiske。1978年。《显著性、注意与归因：注意力焦点现象》。载于《实验社会心理学进展》。第11卷。[121] Hugo Touvron、Thibaut Lavril、Gautier Izacard、Xavier Martinet、Marie-Anne Lachaux、Timothée Lacroix、Baptiste Rozière、Naman Goyal、Eric Hambro、Faisal Azhar 等。2023年。《Llama：开放且高效的基础语言模型》。arXiv 预印本 arXiv:2302.13971（2023）。[122] Beth Trushkowsky、Tim Kraska、Michael J Franklin 和 Purnamrita Sarkar。2013年。《众包枚举查询》。载于2013年 IEEE 第二十九届数据工程国际会议（ICDE）。IEEE，673–684。[123] Blerta Veseli、Sneha Singhania、Simon Razniewski 和 Gerhard Weikum。2023年。《语言模型知识库补全评估》。载于欧洲语义网会议（ESWC）。施普林格，227–243。[124] Denny Vrandečić 和 Markus Krötzsch。2014年。《维基数据：一个自由协作的知识库》。Commun. ACM 57，10（ 2014 78 85 10 1145 2629489 ），–。https://doi.org/。/ [125] Ruize Wang、Duyu Tang、Nan Duan、Zhongyu Wei、Xuan-Jing Huang、Jianshu Ji、Guihong Cao、Daxin Jiang 和 Ming Zhou。2021年。《K-Adapter：通过适配器将知识注入预训练模型》。载于计算语言学协会论文集（ACL-IJCNLP 2021）。1405–1418。[126] Gerhard Weikum、Luna Dong、Simon Razniewski 和 Fabian M. Suchanek。2021年。《机器知识：综合性知识库的创建与整理》。Foundations and Trends® in Databases 10，2-4，108–490。[127] I an H Witten 和 David N Milne。2008年。《利用维基百科链接获得的一种有效、低成本的语义相关性度量》。载于 AAAI 维基百科与人工智能研讨会论文集：不断演进的协同作用，2008年。AAAI 出版社，25–30。[128] Ikuya Yamada、Akari Asai、Jin Sakuma、Hiroyuki Shindo、Hideaki Takeda、Yoshiyasu Takefuji 和 Yuji Matsumoto。2020年。《Wikipedia2Vec：从维基百科学习并可视化词与实体嵌入的高效工具包》。载于2020年自然语言处理实证方法会议系统演示论文集（EMNLP）。计算语言学协会，23–30。https://doi.org/10.18653/v1/2020.emnlp-demos.4[129] Amrapali Zaveri、Anisa Rula、Andrea Maurino、Ricardo Pietrobon、Jens Lehmann 和 Sören Auer。2016年。《关联数据质量评估：综述》。Semantic Web 7，1（2016），63–93。[130] Haiyan Zhao、Hanjie Chen、Fan Yang、Ninghao Liu、Huiqi Deng、Hengyi Cai、Shuaiqiang Wang、Dawei Yin 和 Mengnan Du。2023年。《大语言模型的可解释性：综述》。arXiv 预印本 arXiv:2309.01029（2023）。