弥合联机分析处理与检索增强生成：语料库分区设计的多维方法

达里奥马约[1]，*，斯特凡诺 里齐[1]

[1]*DISI* - 博洛尼亚大学，复兴大道 *2* 号，博洛尼亚，*40136*，意大利

摘要

检索检索增强生成（RAG）系统日益广泛地应用于大规模文档集合，通常包含数百万份文档和数千万个文本块。在工业级检索平台中，可扩展性通常通过水平分片，以及近似最近邻搜索、混合索引和优化的元数据过滤相结合来实现。尽管这些机制从效率角度看效果显著，但它们依赖自底向上、由相似度驱动的组织方式，缺乏语料库分区的概念依据。本文认为，大规模检索增强生成系统的设计可以结合两种正交策略：语义聚类，用于优化嵌入空间中的局部性；以及多维分区，根据时间和组织背景等概念维度决定检索应在何处进行。尽管这些维度已隐含存在于当前系统中，但其使用方式随意且缺乏良好结构。我们提出维度事实模型（DFM），将其作为指导检索增强生成语料库多维分区设计的概念框架。DFM 为在面向检索的场景中分析事实、维度、层次结构和粒度提供了有原则的方法。该框架天然支持分层路由和可控的回退策略，确保即使元数据不完整，检索仍保持稳健，同时将搜索过程从“黑箱式”的相似度匹配转变为可治理、确定性的工作流程。本研究旨在提出一篇立场论文；其目标是弥合联机分析处理式多维建模与现代检索增强生成架构之间的差距，并推动对大规模场景下有原则、可解释且可治理的检索策略开展进一步研究。

关键词

检索增强生划分、多维模型、概念建模、元数据驱动的检索、检索治理
成

1. 引言

近年来，基于 Transformer 的架构取得了长足进步，使大语言模型（LLM）在自然语言理解与生成方面展现出卓越能力 [1, 2], ，支持广泛的应用场景，这些场景需要灵活访问规模庞大且异构的知识源。这些模型日益被视为基础模型，既能提供通用的语言能力，也带来了知识时效性、控制以及与外部数据源集成等方面的挑战 [3]。然而，大语言模型本质上是参数化且静态的，其知识仅限于训练期间编码的内容。为克服这一局限，检索增强生成（RAG）已成为主流架构范式，将生成模型与针对大型文档集合的外部检索机制相结合 [4]。

在 RAG 系统中，用户查询用于从语料库中检索相关文本片段，随后在生成过程中将其作为上下文输入提供给大语言模型。这种方法使系统能够基于最新且特定领域的信息生成回答，并已成为在企业级数据上部署基于大语言模型的应用的事实标准。随着 RAG 系统日益应用于包含数百万份文档以及数千万乃至数亿个文本块的集合，检索的可扩展性、鲁棒性和可治理性问题变得至关重要。

当前的大规模检索增强生成架构主要依赖于基于向量表示的相似度检索，通常还结合混合索引和近似最近邻搜索技术 [5, 6]。现代工业级 RAG 系统已将基于向量的相似度搜索与聚类和元数据过滤策略结合起来；然而，这些机制主要由自下而上的语义相似性驱动，并未为语料库划分提供明确的概念依据。

工业级 RAG 系统通常通过水平分片、混合索引策略以及基于元数据的筛选来实现可扩展性，并在可能的情况下借此限制检索空间。尽管这些方案在效率方面行之有效，但它们在很大程度上受实现方式驱动：语料库组织主要由嵌入空间的局部性决定，而对检索位置的逻辑约束则以临时且结构薄弱的方式处理。

与此同时，联机分析处理（OLAP）系统和数据仓库长期以来一直致力于解决大规模数据的组织与查询问题[7, 8]。多维建模为表示事实、维度、层次结构和粒度层级提供了严谨的框架，并支持对数据组织进行系统化推理。尤其是，维度事实模型（DFM）处于概念层面，可指导设计决策，同时不受物理存储和执行策略等实现细节的影响 [9]。

本文认为，大规模检索增强生成系统若具备类似的概念设计层，将会受益。我们提出将 DFM 重新诠释为一种框架，用于对检索增强生成所用的非结构化文档语料库进行多维分区。我们认为，有效的检索增强生成架构应明确区分两个正交的关注点：(i) 沿与领域相关的维度对语料库进行多维分区，以指导路由决策并确定检索位置；(ii) 在每个多维分区内进行语义聚类和子聚类，以支持基于向量表示的相似度检索。请注意，这里提及聚类和子聚类，仅是作为可在多维分区内采用的语义组织策略的代表性示例。所提出的框架既不强制规定任何特定的检索逻辑，也不

*通讯作者。dario.maio@unibo.it(D. Maio); stefano.rizzi@unibo.it(S. Rizzi) ID 0000-0002-0094-0022(D. Maio); 0000-0002-4617-217X (S. Rizzi)

会干预现有工业级检索系统所采用的内部过滤、索引或执行策略。相反，它通过使概念假设明确且可控，为检索提供了一种有原则的治理方式。

所提出的方法有意不依赖底层系统架构。多维分区的定义独立于物理分片，不过在实际部署中，多维分区与物理分片之间自然存在映射，并遵循标准的可扩展性与负载均衡实践。同样，是否存在数据仓库与该框架是否适用无关。当数据仓库或其他结构化存储库可用时，它们有助于设计维度，因为其多维模式可能已经存在。当缺少此类基础设施时，DFM 仍是一个纯粹的概念工具，用于推理语料库组织与检索策略。无论哪种情况，元数据提取与管理都是大规模检索系统固有的问题，并非由所提出的框架引入。

本文是一篇立场论文，探讨方法论问题。本文并未提出新的检索算法或实验评估，而是引入一种概念视角，以补充现有的检索增强生成架构，旨在推动对海量非结构化语料库检索的原则性、可解释且可治理设计策略的讨论。本文的贡献可概括如下：

- 我们指出，当前 RAG 架构存在一个概念性缺口：语料库分区和基于元数据的路由对于可扩展性和稳健性至关重要，但目前处理方式较为临时，受实现驱动，缺乏明确的概念设计层。
- 我们认为，高效的大规模检索增强生成系统需要明确结合两种相互正交的组织原则：通过语义聚类实现嵌入空间局部性，以及通过多维分区依据领域相关维度决定应在何处进行检索。
- 我们提出 DFM，作为指导大型非结构化语料库进行检索用途多维分区的概念框架。
- 我们表明，基于 DFM 的设计可自然地支持分层路由和检索过程中的受控回退策略，即使元数据不完整或部分缺失，也能实现确定性、可解释且可治理的检索工作流。
- 我们以复杂法律领域中的一个示例说明为例，涉及工业规模的司法判决或合同文件集合，展示多维建模如何沿着有意义的维度构建检索空间，同时不限制底层的基于相似度的检索机制。

本文其余部分的组织结构如下。第2 节回顾了有关检索增强生成架构、元数据驱动检索以及多维建模的相关工作。第3节介绍 DFM，并讨论其在面向检索的场景中的重新诠释。第4节给出了说明性示例这些示例取自复杂的应用领域，旨在说明所提出方法的必要性。最后，第5 节讨论了该框架的影响、局限性以及未来研究的一些开放方向。

2. 相关工作

2.1. 联机分析处理中概念建模与大语言模型

概念建模长期以来在数据仓库和在线分析处理系统的设计中发挥着核心作用。多维模型为在概念层面表示事实、维度、层次结构和粒度提供了严谨的框架，有助于系统地推理聚合路径、下钻和上卷操作，以及分析需求与物理实现之间的匹配 [7, 8]。DFM 在概念层面将这些概念形式化，并已被广泛用作数据仓库设计的参考模型 [9]。近期研究探讨了大语言模型在在线分析处理系统概念设计中的作用。Rizzi [10] 探讨了在多维立方体的供应驱动设计中使用 ChatGPT 来完善概念模式草案；Rizzi 等人 [11] 则通过实证研究考察了大语言模型辅助多维建模任务的潜力与局限。这些研究表明，大语言模型能够有效应用于联机分析处理设计的概念层面。

数据库自然语言接口的研究在数据管理领域有着悠久的传统，旨在将用户问题转换为 SQL 等形式化查询语言。近年来，神经网络和基于 Transformer 的语言模型取得了长足进展，使该领域重新焕发活力，推动了现代文本到 SQL 和模式感知语义解析方法的发展；这些方法在 Spider [12] 等复杂的跨领域基准测试中取得了出色的性能。这一研究方向主要旨在通过将用户意图映射为针对现有数据库模式的查询，提高结构化数据的可访问性。最近，大语言模型还被探索用作分析系统的灵活接口，支持在联机分析处理和商业智能平台 [13] 上进行对话式交互与解释。在这些方法中，大语言模型充当查询和探索多维数据的前端，同时假设底层的联机分析处理模式和数据组织方式已经给定。

尽管多维模型已经成熟并广泛应用于分析系统，但它尚未被明确应用于大规模非结构化文档集合的检索结构设计。在当前的检索增强生成架构中，概念建模基本缺失：检索单元、元数据和分区通常以临时方式定义，而语义聚类则被作为主要的组织原则。

2.2. 检索增强生成中的分区与元数据

RAG 领域的多项研究已认识到，对检索空间进行结构化对于提升生成质量十分重要。Wang 等人 [14] 提出了多分区检索增强生成架构（M-RAG），其中多个分区充当检索和生成的基本单元，并表明分区能够显著提升不同语言生成任务中的性能。然而，在这一方法中，分区被视为执行期间优化的操作单元

，其设计并没有明确的概念模式作为指导。

其他研究探讨了在检索增强生成流程中利用元数据过滤或限制检索空间。Di Oliveira 等人 [15] 提出了一种两阶段检索增强生成方法，先进行元数据过滤，再进行基于向量的检索，并证明结构化元数据能够有效降低噪声，改善对大语言模型输出的统计评估。

最后，已有早期尝试将联机分析处理风格的工作流与检索增强生成架构相结合。Ouafiq 和 Saadane [16] 提出了一种检索增强型联机分析处理架构，其中多维分析结构用于支持生成模型的推理过程。尽管这些方法证实了多维组织和元数据驱动的分区在检索系统中的重要性，但并未为非结构化文档语料库的分区设计提供通用的概念框架。

2.3. 结合结构化数据源的检索增强生成

前述研究主要关注检索空间的分区和元数据驱动的控制，而另一类研究则探讨将结构化数据源整合到检索增强生成流程中。若干研究将检索增强生成与关系数据库、知识库或图数据库相结合，使大语言模型能够在答案生成过程中同时检索非结构化文本和结构化事实 [4, 17, 18]。在这些系统中，结构化查询通常通过外部工具或专用查询引擎执行，而非结构化检索则依赖于对文本段落进行向量相似度搜索。

已有研究提出将语义检索与数据库查询交错结合的混合架构，以提升事实准确性、减少幻觉并支持多步推理 [17, 18]。然而，在这些方法中，非结构化语料库大多被视为扁平或结构较弱的集合，可能辅以简单的元数据筛选器。文本语料库的分区设计通常依赖启发式决策或基于相似度的聚类，并没有明确的概念模型来指导分区选择。

2.4. 本文的定位

与上述研究方向不同，本文聚焦于面向非结构化语料库检索的多维分区概念设计。我们并不引入新的检索算法或混合查询机制，而是主张，多维建模原则能够为大规模检索增强生成系统提供目前尚缺失的形式化设计层。据我们所知，本文首次明确提出将 DFM 作为概念框架，用于推理检索增强生成中的多维分区、分层路由和回退策略。

3. 将 DFM 应用于面向检索的分区

在在线分析处理系统中，DFM 从事实、维度、层次结构和粒度层级的角度，对分析数据进行概念表示，并独立于物理存储和查询执行策略 [8]。在此背景下，事实表示所关注的业务事件，例如发票或销售交易，并通过时间、客户、产品和地点等多个维度进行描述。

基于这一成熟的抽象，我们重新诠释了检索增强生成系统中的基本检索单元（例如文本块），将其视为事实的对应物；而维度则对应于刻画文档并指导检索决策的领域相关属性，例如时间、文档类型、司法管辖区或组织背景。维度内部的层次结构体现了有意义的抽象层级，并支持逐步细化检索空间。重要的是，所提出的模型并未规定这些维度应如何在物理层面建立索引或存储，而只说明如何在概念层面组织和组合这些维度。换言之，在本研究中，我们将这些概念重新诠释于检索增强生成的背景下，其首要目标并非聚合，而是以可控的方式从大型非结构化语料库中选取相关部分。

在这一框架下，多维分区与语义检索发挥着互补作用。沿一个或多个维度定义的多维分区用于路由查询，并确定应在何处执行检索。随后，在每个多维分区内应用语义聚类和子聚类，以支持基于向量表示的相似度检索。这种分离使基于嵌入的方法能够在局部运行，同时由明确的概念选择来控制全局检索行为。

该模型通过允许设置专门的回退分区来收纳维度值无法可靠确定的文档，从而明确处理元数据缺失或不完整的情况。从多维视角来看，这类分区对应于具有可选维度坐标的事实，并被视为概念设计中的一等要素，确保鲁棒性，而无需依赖隐式或临时行为。

总体而言，所提出的重新诠释将 DFM 定位为大规模检索增强生成系统的概念设计层，使人们能够以有原则的方式推理语料库组织、路由策略和检索范围，同时对底层架构和执行机制保持完全中立。通过明确分区假设，DFM 为管理路由、聚类策略和回退策略的演进提供了稳定的概念参照，以适应语料库和使用模式随时间发生的变化。

4. 示例说明

为说明所提出的方法，我们考虑一个大规模法律领域场景，其中涉及司法判决或合同文件集合。这类语料库具有高度的结构复杂性、鲜明的领域语义以及工业级规模；经常包含数百万份文档，而经检索增强生成处理后，文本块可达数亿个。在缺乏多维视角的情况下，此类场景中的检索通常基于扁平或结构较弱的语料库进行，并可能辅以临时的元数据筛选器。因此，来自不同司法管辖区、程序层级或时间背景的语义相似片段可能会被一并检索出来，尽管对于特定信息需求而言，它们在概念上并不兼容。

从多维视角来看，语料库可以

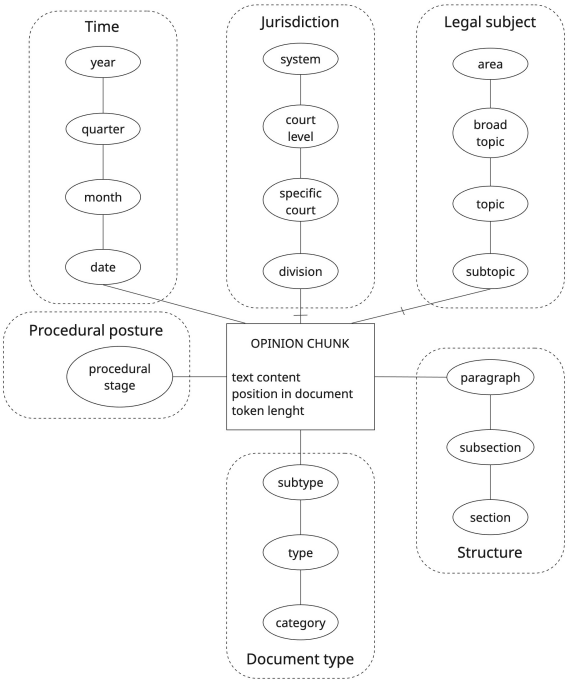


图 1： 法律领域中司法意见片段的 DFM 模式。方框表示基本检索单元及其定量指标；虚线区域表示维度（例如结构），圆圈则表示这些维度的层级，从最细的层级（附着于方框，例如段落）到最粗的层级（叶节点，例如节）。弧线上的虚线表示可选维度。

改为沿着具有法律意义的维度进行逻辑分区，例如时间、司法管辖区、程序状态和文档类型。这些维度定义了分区，用于指导查询路由并决定检索应在何处进行。例如，关于特定司法管辖区近期上诉裁决的查询，可以路由至相应的多维分区；而缺乏可靠时间或司法管辖区元数据的文档，则会被明确分配至专门的回退分区。在每个多维分区内，可以采用语义聚类和子聚类策略来组织向量表示，并支持不同粒度的基于相似度的检索。这样，基于嵌入的方法便可在概念上连贯的语料库子集中运行，同时由多维分区层保留对整个检索空间的全局控制。

图 1展示了美国法律体系中司法意见片段的一种 DFM 模式。基本检索单元对应一个 意见片段，即将法院意见分割成便于检索增强生成的文本片段后，从司法裁决中提取的一部分。意见片段源自司法意见、命令或备忘录等文档，这些文档通常以 PDF 或 HTML 等格式提供，并在数据摄取过程中处理为文本块。它们是法律话语的原子单元，可被检索并作为上下文证据提供给语言模型。每个意见片段都由多个正交维度描述，这些维度常用于法律研究与分析，包括：

- 司法管辖区，按（从最细到最粗的级别）分区、具体法院、法院级别（例如“最高级”和“上诉巡回区”）以及司法体系（“联邦”或“州”）来体现判决的制度背景；
- 法律主题，用于对意见所涉及的实体法律领域进行分类描述，层级包括领域（如“民事”“刑事”“宪法”）和广泛主题（如“合同”和“侵权行为”）；
- 文档类型，主要按类别（如“意见”和“命令”）及类型（如“多数意见”和“异议意见”）区分不同种类的法律文档和意见；
- 程序状态，描述作出裁决时的程序背景（如“审判”和“上诉”）；
- 结构，用于识别意见片段在意见中的修辞或功能作用（例如，作为某一节的“事实”“法律分析”或“裁判结论”）。

对于无法可靠确定维度值的意见片段，会明确将其关联到专门的回退分区；在 DFM 中，这些分区表示为可选维度（在我们的示例中为法律主题和管辖权）。

请注意，所提出的模式仅作为概念性参考，并未规定任何特定的物理存储、索引策略或系统架构。多维分区与物理设计选择（例如分片、索引和集群组织方式）之间的映射取决于具体部署，不在本文讨论范围之内。

在此框架中，多维信息用于指导查询路由，并在执行任何基于相似度的搜索之前确定检索范围。系统会分析查询，提取维度约束（显式或隐式），再据此选择语料库中相关的多维分区。例如，查询若要查找某项特定法律原则下的近期最高法院裁决，可能会对时间、司法管辖区和法律主题维度施加约束，从而将检索路由至包含近期最高法院意见且涉及该原则的分区。同样，旨在比较多个巡回上诉法院上诉裁决的查询，可能会有意跨越多个司法管辖区分区，以便进行结构化比较，而不是将语义相似但在制度上各不相同的裁决混为一谈。

维度及其层级进一步支持将检索聚焦于意见中的特定部分，例如法律分析部分；当事实背景或程序历史与信息需求无关时，可避免将其纳入检索范围。当维度约束较弱、含糊不清或不完整时，回退分区（通过可选维度表示）可确保检索保持稳健且受到明确管控，而非依赖隐式行为。

在每个选定的多维分区内，随后可应用语义检索技术（例如结合聚类或子聚类的向量相似度搜索），以识别最相关的意见片段。这样一来，维度路由决定检索应在何处 进行，而语义相似度则决定在该范围内哪些 内容最相关。

需要注意的是，所提出的多维分区不会对检索施加限制性的访问模式。在用户明确寻求广泛

覆盖范围的场景下，例如调查跨越多年、多个法院或司法管辖区的法律先例，查询可以合理地跨越多个或全部多维分区。在此类情况下，所提出的框架不会降低检索固有的计算成本，也无意如此。相反，它提供了一种明确且受控的方式来表达和管理这些检索范围，使广泛搜索成为一种经过审慎设计的选择，而不是扁平语料库组织所产生的隐性副作用。

类似的理由也适用于大规模生物医学和医疗保健相关文档集，例如科学文章、临床指南和技术报告。这些文档通常以异构格式提供，并经过处理后转化为检索增强生成（RAG）所用的文本块。这些语料库可能包含数百万份文档和数亿个文本块，其特点是领域语义鲜明，且元数据丰富但往往不完整。有意义的维度可能包括出版时间、疾病状况、研究或文档类型、目标人群和来源。沿这些维度进行分区，有助于将检索引导至语料库中概念上适当的子集（例如，近期指南与原始研究，或成人与儿科人群），同时在各分区内使用语义检索技术。与法律场景一样，对于无法可靠确定元数据的文档，可将其一致地分配至检索空间中的专用回退区域，从而确保鲁棒性，避免出现隐式或不受控的检索行为。

同样，尽管时间等某些维度看似普适，且已在实践中广泛使用，但 DFM 的贡献并不在于引入单个维度，而在于对这些维度进行系统化组织、组合与治理。通过在概念层面明确维度、层级和回退策略，所提出的方法能够支持灵活的路由策略，以适应不同的查询意图，同时不限制物理执行或性能优化机制。

5. 讨论与启示

本文并未提出新的检索技术或系统架构，而是倡导在设计大规模检索增强生成系统时，引入一种受成熟维度建模原则启发的概念视角。因此，本文的主要启示在于可治理性、可解释性和设计清晰度，而非性能优化。通过明确分区决策，并以多维模式为其提供依据，所提出的框架能够支持系统地推理检索应在何处进行、如何定义检索范围，以及如何处理回退策略。

从查询视角来看，RAG 系统需要支持各种不同的信息需求，范围涵盖以文档为中心的查询（例如查找和文档级摘要）、面向检索的查询（例如主题检索或多文档检索）、逐文档视图、语料库级聚合以及多步推理查询。这些不同类型的查询会以不同的粒度和范围与检索空间自然交互。因此，不同类别的查询与多维分区的交互方式可能各不相同：有些查询几乎不需要或完全不需要基于分区的限制，而另一些则会受益于对检索范围进行明确控制，以确保结果的完整性、正确性和可复现性。

这种方法的一个重要启示是，它能够灵活适应不同的查询意图。多维分区可以将检索限制在语料库中范围明确的子集中，但并不妨碍那些合理涵盖多个或全部分区的广泛查询，例如针对较长时间段或较大组织范围的探索性搜索。在这些情况下，检索的计算成本是任务本身固有的，所提出的框架并不会降低这一成本。相反，该框架提供了一种明确且可控的方式来表达此类检索范围，避免出现隐含或非预期的行为。尽管性能提升并非所提出框架的首要目标，但对于那些可在有限数量的多维分区内满足其信息需求的查询，缩小检索范围自然可能带来益处。

DFM 的拟议用法并不依赖于各个维度的新颖性，其中一些维度在实践中可能已被广泛采用（例如时间）。其贡献在于系统地组织维度、层次结构和粒度级别，并明确处理不完整或缺失的元数据。在现有系统中，这些方面往往是隐式处理的，而拟议方法则将它们提升为首要的设计决策。

需要承认一些局限性。我们的框架并未解决元数据提取问题；无论采用何种检索架构，在大规模文档处理中，这仍然是一项挑战。此外，框架没有规定如何将多维分区映射到物理分片、建立索引或执行，而是将这些选择留给现有的工业解决方案和优化策略。因此，应将该框架视为对当前检索技术的补充，而非替代。

总的来说，我们认为，本文提出的多维视角将有助于构建更具原则性、透明性和适应性的检索架构，并希望这能推动针对大型非结构化语料库检索的概念建模方法进一步发展。我们鼓励开发开源工具，以支持这一方向上的概念设计，尤其是协助定义和治理 RAG 系统中的多维分区；当数据仓库或其他结构化存储库已就绪时，此类工具理应与现有基础设施自然集成，而非取而代之。

致谢

我们感谢 PR 普利亚大区 FESR FSE+ 2021-2027 项目——欧洲区域发展基金、普利亚大区 Net Service 项目 QUICK SHIELD 的资助，项目编号：CUP:B85H24000920007。

参考文献

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, 《Attention Is All You Need》，载于：神经信息处理系统进展，2017 年，第 5998–6008 页。[2] T. Brown 等，《Language Models Are Few-Shot Learners》，《神经信息处理系统进展》第 33 卷（2020 年），第 1877–1901 页。

[3] R. Bommasani 等，《On the Opportunities and Risks of Foundation Models》，CoRR abs/2108.07258（2021 年）。[4] P. Lewis 等，《面向知识密集型自然语言处理任务的检索增强生成》，载于：神经信息处理系统进展，2020 年。[5] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang，《面向大语言模型的检索增强生成：综述》，《计算机科学前沿》第 18 卷（2024 年），第 1–20 页。[6] J. Chen 等，《检索增强生成中的大语言模型基准测试》，CoRR abs/2309.01431（2024 年）。[7] R. Kimball, M. Ross，《数据仓库工具包：维度建模权威指南》，Wiley，2013 年。[8] M. Golfarelli, S. Rizzi，《数据仓库设计：现代原理与方法》，麦格劳-希尔，2009 年。[9] M. Golfarelli, D. Maio, S. Rizzi，《维度事实模型：数据仓库的概念模型》，《国际合作信息系统期刊》第 7 卷（1998 年），第 215–247 页。[10] S. Rizzi，《在多维立方体供应驱动设计中使用 ChatGPT 优化概念模式草稿》，CoRR abs/2502.02238（2025 年）。[11] S. Rizzi, M. Francia, E. Gallinucci, M. Golfarelli，《使用大语言模型进行多维立方体概念设计：一项研究》，《数据与知识工程》（2025 年）。[12] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, D. Radev，《Spider：用于复杂跨领域语义解析和文本到 SQL 任务的大规模人工标注数据集》，载于：自然语言处理实证方法会议论文集，2018 年。[13] S. Bimonte, S. Rizzi，《文本转 MDX：借助大语言模型，根据用户问题生成 MDX 查询》，载于：国际概念建模会议论文集，法国普瓦捷，2025 年，第 165–181 页。[14] Z. Wang, S. Teo, J. Ouyang, Y. Xu, W. Shi，《M-RAG：通过多分区检索增强生成提升大语言模型性能》，载于：计算语言学协会年会论文集，2024 年，第 1966–1978 页。[15] V. Di Oliveira, P. C. Brom, L. Weigang，《用于元数据过滤和大语言模型统计评估的两阶段检索增强生成》，《IEEE 拉丁美洲汇刊》第 23 卷（2025 年）。[16] E. M. Ouafiq, R. Saadane，《检索增强型联机分析处理：面向智能系统与设备的生成式人工智能架构》，载于：AAAI 研讨会论文集，2025 年，第 304–312 页。[17] S. Yao 等，《ReAct：在语言模型中协同推理与行动》，载于：国际学习表征会议论文集，2023 年。[18] T. Schick 等，《Toolformer：语言模型能够自学使用工具》，载于：神经信息处理系统进展，2023 年。