graphrag的原理、源码及应用介绍

作者：acedar(C大)

目录

为什么会有graphrag?

rag解决的问题：使用检索增强生成（RAG）技术从外部知识源检索相关信息，使大型语言模型（LLMs）能够回答涉及私有或之前未见过的文档集合的问题

传统rag：

解决外部知识的详情介绍，某些细节的检索。

比如刘德华的成名作电影是？(彩云曲)

无发解决QFS（query focused summarization）：

文章中的主要主题是什么？ -> 跨多个段落，甚至多个文章

graphrag的基本概念

Document(文档)- 系统中的输入文档。这些文档要么代表 CSV 中的单独行，要么代表单独的 .txt 文件。

TextUnit(文本块)- 要分析的文本块。这些块的大小、重叠以及它们是否遵守任何数据边界可以在下面配置。一个常见的用例是设置CHUNK_BY_COLUMNS为id，以便文档和 TextUnits 之间存在一对多关系，而不是多对多关系。

Entity(实体)- 从 TextUnit 中提取的实体。这些实体代表人物、地点、事件或您提供的其他实体模型。

Relationship(关系)- 两个实体之间的关系。这些关系由协变量生成。

Covariate(协变量)- 提取的声明信息，其中包含可能受时间限制的实体的陈述。

Claim(声明)- 代表具有评估状态和时间限制的积极事实陈述，以协变量(Covariates)的称呼在各处使用

Community Report(社区报告)- 一旦生成实体，我们就对它们执行分层社区检测，并为该层次结构中的每个社区生成报告。

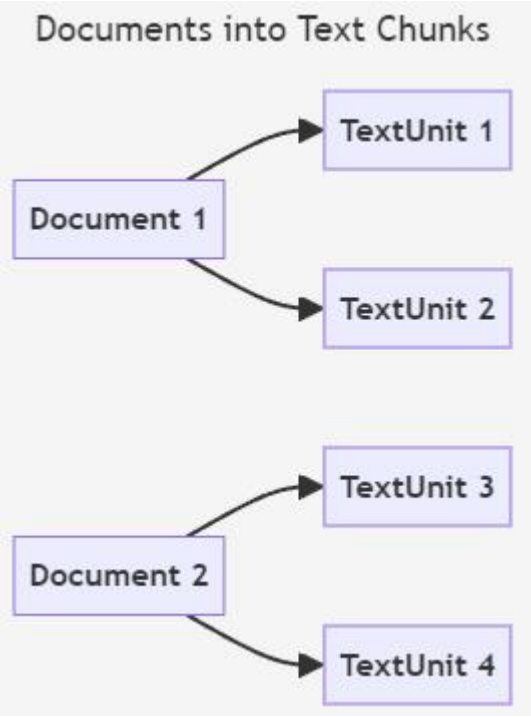
Node（节点） - 包含已嵌入和聚集的实体和文档的呈现图形视图的布局信息。

文档转换为TextUnits

说明：

概念 -> 追溯原文

文档和文本单元之间存在严格的一对多关系



切换技巧：

- 切换大小chunk size: 1200 token
- 较大的块会导致输出保真度较低，参考文本意义较小；使用较大的块可以大大缩短处理时间。

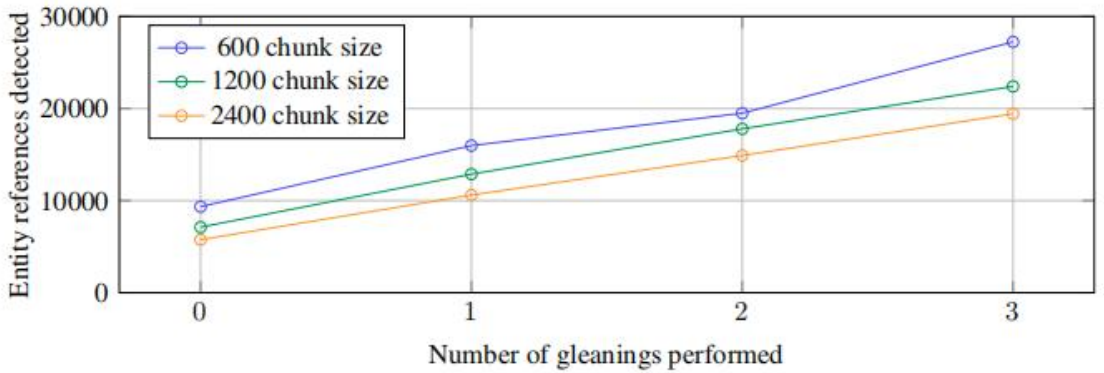


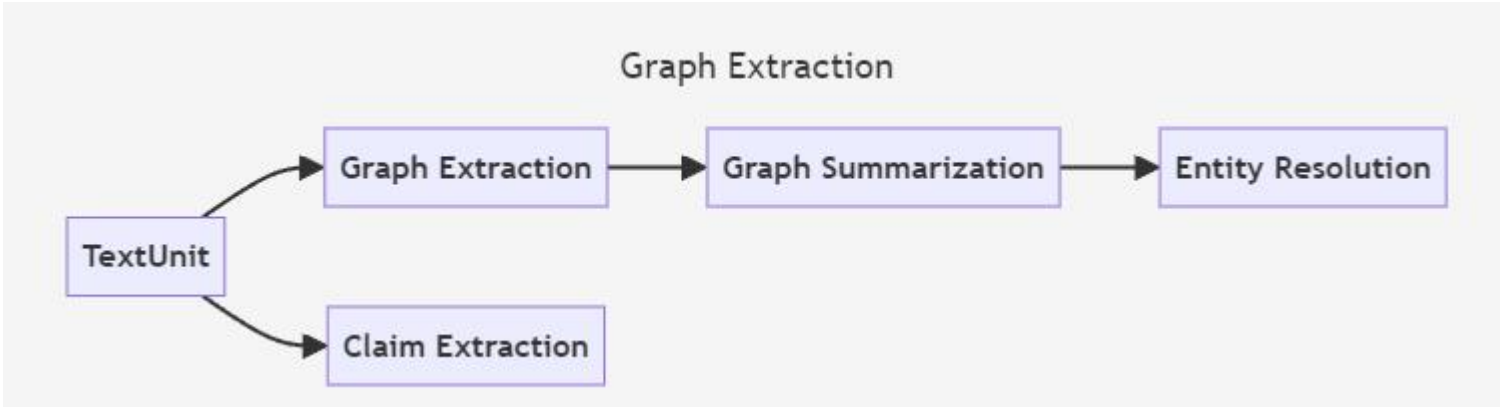
Figure 2: How the entity references detected in the HotPotQA dataset (Yang et al., 2018) varies with chunk size and gleanings for our generic entity extraction prompt with `gpt-4-turbo`.

图提取

功能：分析每个文本单元并提取图形基元：实体、关系和声明

实体和关系提取:

使用 LLM 从原始文本中提取实体和关系,包含具有名称、类型和描述的实体列表，以及具有源、目标和描述的关系列表。



实体和关系摘要:

通过 LLM为每个实体和关系提供简短的摘要描述

Claim Extraction & Emission：

声明代表具有评估状态和时间限制的积极事实陈述，以协变量(Covariates)的称呼在各处使用

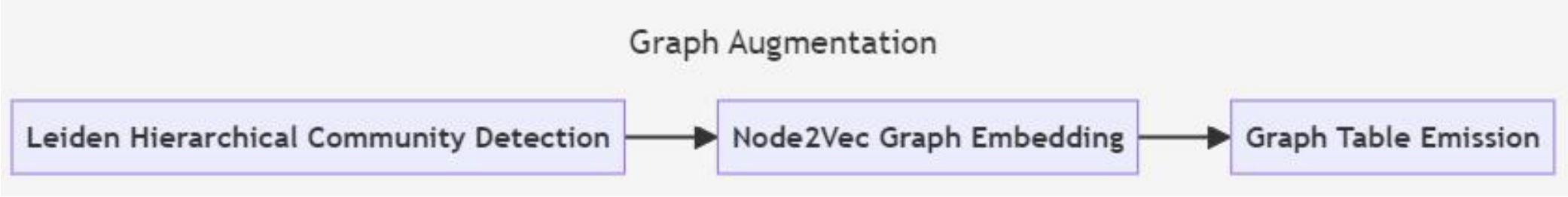
图增强(构建社区)

社区检测：

使用分层莱顿算法生成实体社区的层次结构，此方法将对我们的图应用递归社区聚类，直到达到社区规模阈值。这将使我们能够了解图的社区结构，并提供一种在不同粒度级别上导航和总结图的方法。

图嵌入：

使用 Node2Vec 算法生成图的向量表示。这将使我们能够理解图的隐式结构，并提供额外的向量空间，以便在查询阶段搜索相关概念。



Graph Tables Emission： the final Entities and Relationships tables are emitted after their text fields are text-embedded. -> 理解构建的社区和实体及关系存在关联关系

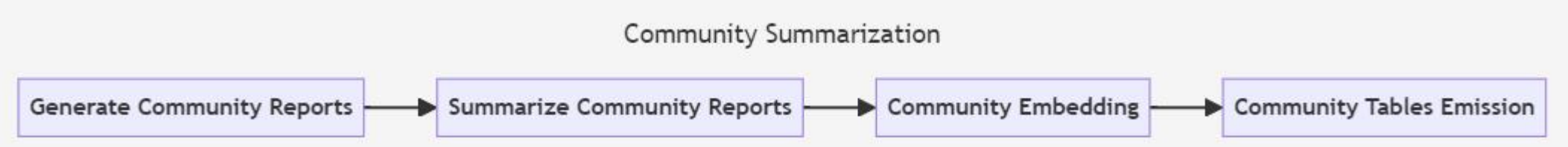
社区总结

功能：基于社区数据并为每个社区生成报告，这让我们可以从多个粒度点对图表有一个高层次的了解。例如，如果社区 A 是顶级社区，我们将获得有关整个图表的报告。如果社区是较低级别的，我们将获得有关本地集群的报告。

总结社区报告：每个社区报告都会通过 LLM 进行总结，以供速记使用。

生成社区报告: 使用 LLM 生成每个社区的摘要,引用社区子结构中的关键实体、关系和声明

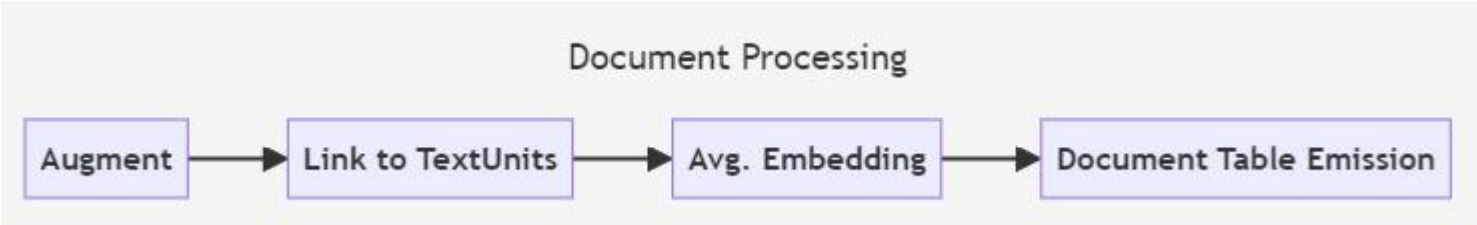
社区嵌入：通过生成社区报告、社区报告摘要和社区报告标题的文本嵌入来生成我们社区的向量表示。



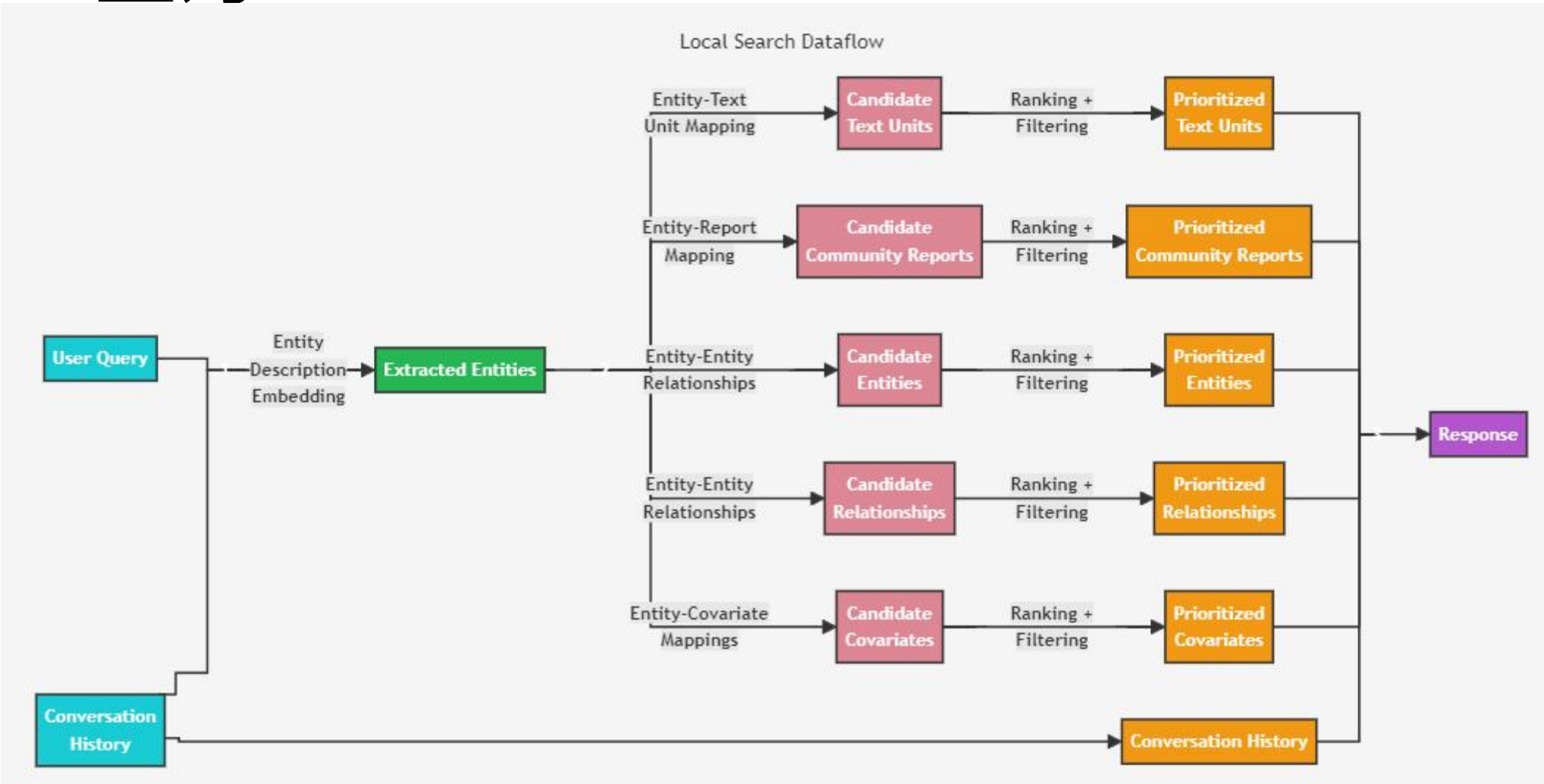
文档处理

链接到 TextUnits：将每个文档链接到第一阶段创建的文本单元，了解哪些文档与哪些文本单元相关

文档嵌入：文档切片的平均嵌入来生成文档的向量表示，能够理解文档之间的隐式关系

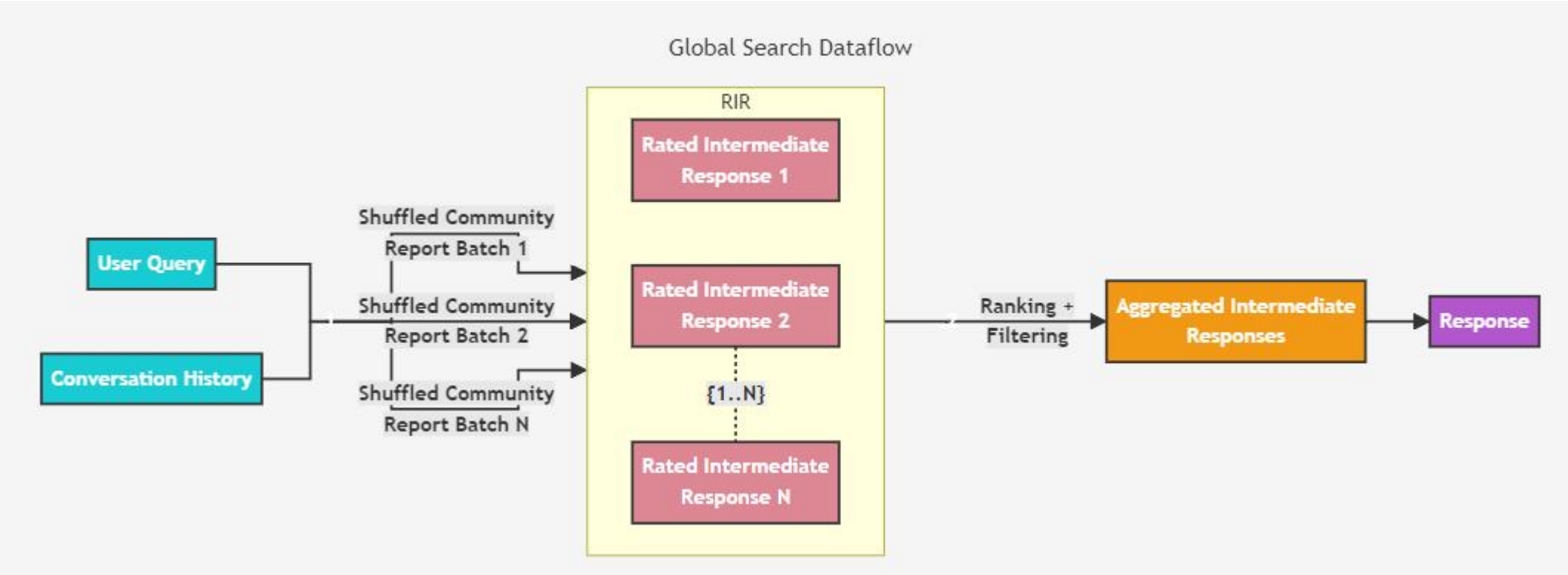


local查询



图中的概念在构建索引中都已创建

global查询



图中的概念在构建索引中都已创建

后续交流

进群交流

文章分享

视频课程