使用 Google BigQuery 上的 Dimensions，在几秒钟内生成大规模科学图景网络分析。

Michele Pasin*、Richard J. Abdill**

**m.pasin@digital-science.com*，通讯作者，Digital Science，6 Briset St，伦敦，EC1M 5NR，英国

***richard.abdill@pennmedicine.upenn.edu*Digital Science，6 Briset St，伦敦，EC1M 5NR，英国。现地址：宾夕法尼亚大学，3400 Civic Center Boulevard，费城，宾夕法尼亚州 19104

摘要

大型且可通过编程访问的文献计量数据库不断发展，为出版物元数据的复杂分析带来了新的机遇。除了提供有关作者和机构的丰富信息外，Dimensions 提供的数据库还包含概念信息，并提供指向资助项目、资助方和专利等实体的链接。然而，评估学术研究模式时，数据并非唯一的挑战：这些大型数据集可能难以整合，尤其是对于不熟悉为容纳此类异质信息而设计的复杂架构的人而言；而最擅长数据挖掘的人，可能并不熟悉数据可视化。本文介绍了一个开源 Python 库，可简化访问 Google BigQuery 上 Dimensions 数据库子集并将其绘制成图表的流程，并展示了该库在免费提供的 Dimensions COVID-19 数据集上的应用。我们乐观地认为，这一工具将简化原本需要完成的多项复杂技术任务，从而扩大人们获取这类宝贵信息的渠道，使更多研究人员能够考察研究重点与合作模式随时间变化的情况。

引言

即使在差异最大的领域中，元研究也是评估和改进科学事业的关键要求。了解研究资助、开展和发表方式中的趋势与空白，有助于资助方、机构、企业和个体研究人员进行数据驱动的自我审视：研究的是什么？由谁开展？哪些合作最有成效？具体体现在哪些方面？对于具备提出问题所需技术能力的人而言，Dimensions、Crossref、Scopus 和 Web of Science 等综合数据库（Thelwall 2018）能提供无穷无尽的答案。尽管大多数此类数据库都提供用户友好的网络门户来访问信息，但通过编程方式访问数据才能实现强大的分析——也就是使用便于计算机处理的机制获取大量数据，再由第三方工具解析这些数据，完成现有网络应用程序未内置的任务。

其中一种机制是“Dimensions on Google BigQuery”数据集，它将 Dimensions 语料库（包括超过 1.27 亿份出版物、600 万个资助项目、69.5 万项临床试验等）（“The data in Dimensions”2022）整理为托管在 Google 强大的 BigQuery 平台上的关系数据库模型，能够快速执行查询，而这类查询在 MySQL 等较传统的数据库平台上可能规模过大或速度过慢。然而，

有关这些实体以及它们之间数十亿条链接的信息，必然分散在 10 张表中，每张表包含数十个字段（“Data SourceTables” 2022），其学习曲线远比 Dimensions 网络应用程序陡峭。

我们介绍了一个开源 Python 库，可简化生成与 COVID-19 相关的科学研究大规模网络可视化的流程。该库可在 GitHub[1]上获取，并依托公有领域的 BigQuery 上的 COVID-19 Dimensions 数据集[2]，提取并计算所关注科学实体之间的邻近关系；同时，它也可以轻松配置为使用完整的 Dimensions 数据。该库包含数据提取和处理组件，因此可供 VOSviewer[3]在线可视化工具使用。我们正在开发更多输出可视化，并将在未来几个月内陆续添加。

正如先前所论述的（Hook、Porter，2021），借助云计算基础设施（例如 Dimensions 在 Google BigQuery 上提供的基础设施），有可能从根本上改变科研分析的开展方式。这项工作的主要贡献是提供一种可重复使用的开源工具，切实展示如何利用此类技术，生成能够深入呈现科学研究图景中特定方面的网络表示，例如组织合作网络和主题共现网络。

数据集

该库默认使用免费提供的 COVID19 Dimensions 数据集（Hook 等，2020）。该数据集包含所有可在 Dimensions 上免费访问的 COVID-19 相关已发表文章、预印本、数据集、资助项目和临床试验。数据最初于 2021 年初以 CSV 格式发布，随后作为公有领域的 BigQuery 数据集发布，以帮助研究界及时掌握最新信息，并大幅减少从许多不同来源汇总这些信息原本所需的时间。该数据集每日更新；截至本文撰写时，其中包含超过 110 万份文档（注：我们使用该数据集是为了更方便地复现本文所述工作；不过，通过修改库的默认设置，也可以指定使用完整的 Dimensions.ai 数据集）。

表 1. COVID19 Dimensions 数据集所含数据摘要

实体	记录
出版物	1,031,972
临床试验	14,723
资助项目	16,703
专利	41,473
数据集	32,784
组织	36,670

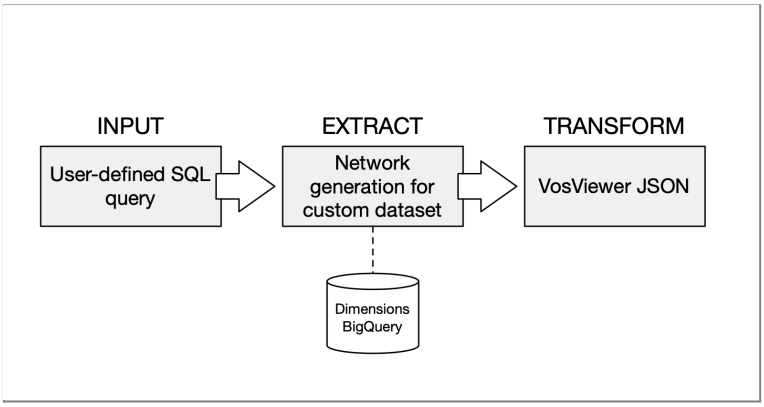
[1] https://github.com/digital-science/dimensions-network-gen[2] https://console.cloud.google.com/marketplace/product/digitalscience-public/covid-19-dataset-dimensions[3] https://www.vosviewer.com/

实现

该库使用 Python 进行数据处理，并使用 VOSviewer 展示网络数据。VOSviewer 是一款用于可视化网络数据的软件工具，可在多个平台上使用，包括基于浏览器的版本（van Eck，2010）。数据提取组件具有通用性，其输出的数据格式也可供其他可视化库读取；我们计划在未来几个月内将这些库纳入该库。

该库可分为三个主要组件：1）用户创建的 SQL 输入查询；2）BigQuery 数据提取与网络生成模块；3）数据转换与可视化组件（图 1）。

图 1：*dimensions-network-gen* Python 库的架构



1. 用户输入
可以针对整个 COVID-19数据库生成网络分析，也可以使用选定的数据子集。为此，用户可以输入任意 SQL 查询，以定义感兴趣的 COVID-19文档子集（例如一组期刊或一组国家）。用户可以将感兴趣的查询保存在文件夹中，建立查询库，然后对其中所有查询运行提取脚本。

例如，以下查询仅选择最近 30 天内添加到数据库的文档：

从 `covid-19-dimensions-ai.data.publications` 中选择 id，其中 EXTRACT(DATE FROM date_inserted) >= DATE_ADD(CURRENT_DATE(), INTERVAL -30 DAY)

请注意，输入查询只会返回一组记录ID——用户只需定义与其筛选条件相关的字段；而将有关这些出版物的信息进行联接和整理这一更复杂的任务则由该库处理。

2. 数据提取
数据提取步骤负责从 BigQuery 提取数据并计算网络表示。目前，我们包含了两种可能的网络计算方式：

1. 概念共现网络。此查询会生成双概念对，并统计这些概念共有多少篇出版物（注意：概念在

Dimensions 中是经过标准化并根据相关性评分加权的出版物级关键词）。2. 组织网络。此查询会生成两个组织组成的配对（取自作者的所属机构），并统计这些组织共有多少篇出版物。

数据提取和网络计算步骤均通过单个 SQL 查询完成。该查询包含用户输入的查询语句（`user-provided-subquery`）以及用于设置结果中最大节点数和边的最小权重的参数值（`@max_nodes`，`@min_edge_weight`）。该查询的核心在于双重 `CROSS JOIN UNNEST`。这种机制能够在几秒内遍历数量可能非常庞大的关系，并在同一数据结构中呈现共同署名组织的所有相关组合。

图 2：用于生成协作网络的 SQL 模板查询

```
WITH subset AS (
    {user-provided-subquery}
),
top_nodes AS (
    SELECT orgid, COUNT(p.id) AS pubs
    FROM `covid-19-dimensions-ai.data.publications` p
    CROSS JOIN UNNEST(p.research_orgs) orgid
    WHERE id IN (SELECT id FROM subset)
    GROUP BY 1
    ORDER BY 2 DESC
    LIMIT @max_nodes
),
links AS (
    SELECT
    CONCAT(g1.name, ' (', org1_id, ')' ) AS org1
    ,CONCAT(g2.name, ' (', org2_id, ')' ) AS org2
    ,COUNT(DISTINCT p.id) AS collabs
    FROM `covid-19-dimensions-ai.data.publications` p
    CROSS JOIN UNNEST(p.research_orgs) org1_id
    CROSS JOIN UNNEST(p.research_orgs) org2_id
    INNER JOIN `covid-19-dimensions-ai.data.grid` g1 ON org1_id=g1.id
    INNER JOIN `covid-19-dimensions-ai.data.grid` g2 ON org2_id=g2.id
    WHERE
    p.id IN (SELECT id FROM subset)
    AND org1_id > org2_id -- to prevent dupes
    AND org1_id IN (SELECT orgid FROM top_nodes)
    AND org2_id IN (SELECT orgid FROM top_nodes)
    GROUP BY 1,2
)
SELECT *
FROM links
WHERE collabs >= @min_edge_weight
```

3. 数据转换与可视化

在此步骤中，从 BigQuery 提取的数据会转换为 VOSviewer JSON 文件，并打包成可在浏览器中查看的 HTML 应用程序。Python 库还包含一个本地服务器组件，可用于在计算机上本地查看这些文件。下方展示了生成的 VOSviewer 网络示例图。

Fi图 3：VOSviewer 对组织网络的可视化呈现（针对过去 *30* 天的查询）

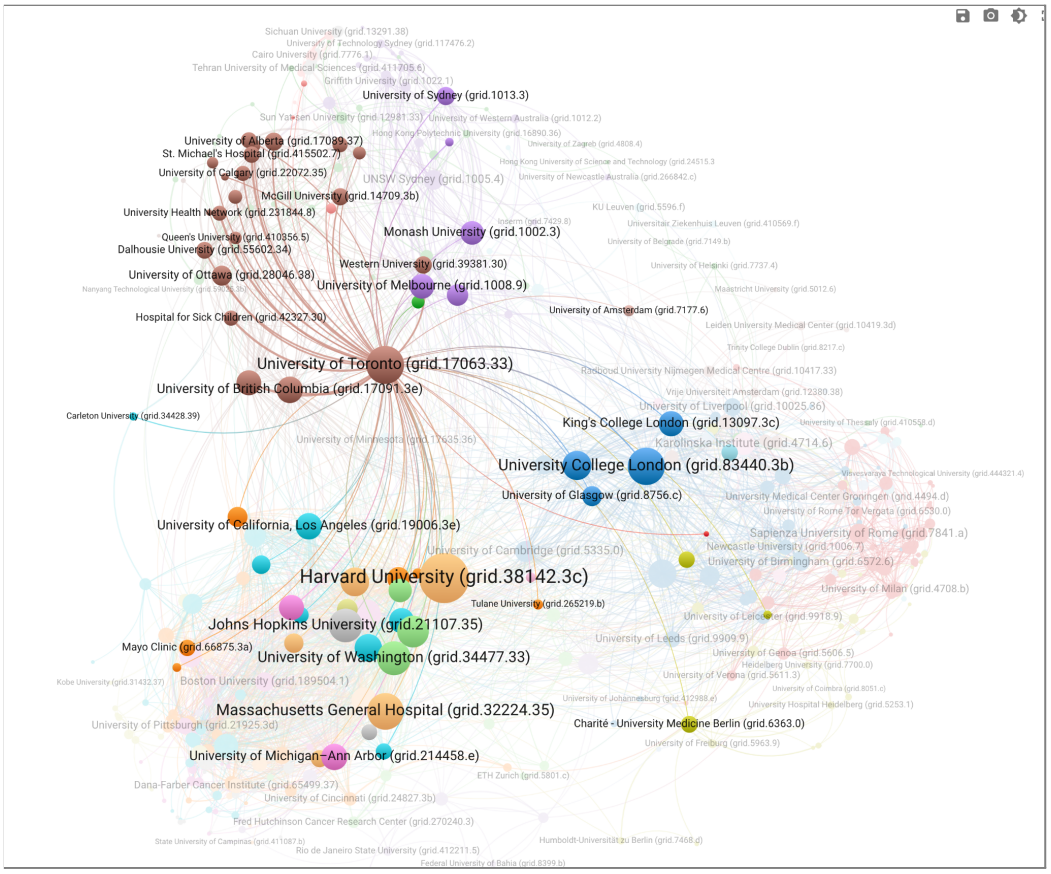
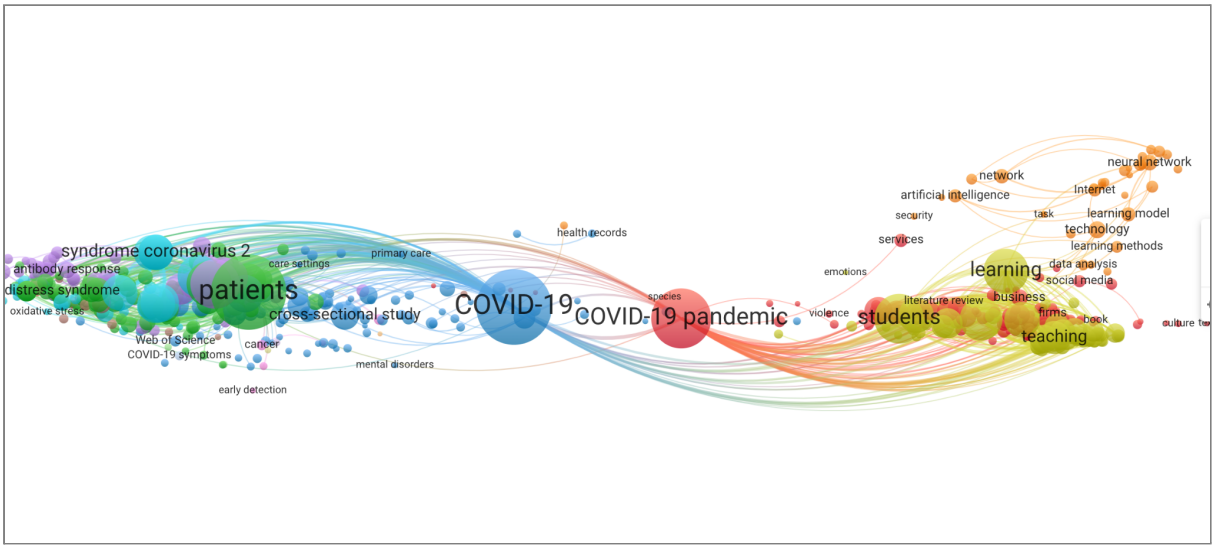


图 4：VOSviewer 对概念网络的可视化呈现（针对过去 *30* 天的查询） y



利益冲突

两位合著者分别是 Digital Science 的现任（MP）和前任（RJA）员工。Digital Science 是 Dimensions 的创建者和提供方。

参考文献

“数据源表”。Dimensions on Google BigQuery 文档。访问日期：2022年4月29日。https://docs.dimensions.ai/bigquery/data-sources.html Herzog, C., Hook, D., and Konkiel, S. (2020). Dimensions：打破科学计量学家与数据之间的壁垒。《定量科学研究》1, 387–395. Hook, D. W., Porter, S. J., (2021). 科学计量学扩展：作为大规模分析基础设施的 Google BigQuery 上的 Dimensions。《前沿研究计量分析》6. Hook, D. W., Porter, S. J., Draux, H., and Herzog, C. T. (2020). 实时文献计量学：Dimensions 作为分析 COVID-19 相关问题的资源。《前沿研究计量分析》5, 25. “Dimensions 中的数据”。Dimensions 网站。访问日期：2022年4月29日。快照见
h
t
t
ps://
web.archive.org/web/20220428173901/https://www.dimensions.ai/dimensions-data/
Thelwall, Mike (2018). Dimensions：Scopus 和 Web of Science 的竞争对手？《信息计量学杂志》。12(2): 430–435. van Eck NJ, Waltman L. 软件综述：VOSviewer，一款用于文献计量学映射的计算机程序。科学计量学。2010;84(2):523–538。