arXiv:2304.12417v1 [cs.DL] 24 Apr 2023

DONUT – 数据库的创建、发展与机遇

Barbara Giunti* J¯anis Lazovskis† Bastian Rieck‡

2023年4月26日

1 起源

DONUT [GLR22]是一个收录拓扑数据分析（TDA）实际应用及现实世界应用论文的数据库。其最初的种子是在2017年4月豪斯多夫数学研究所举办的HIM春季学校——应用与计算代数拓扑学校期间建立的群聊中播下的。



图1：2017年春季学校在波恩豪斯多夫数学研究所的参与者。HIM版权所有。

2019年，当时还是帕维亚大学博士生的Barbara Giunti 在群聊中询问，是否有人听说过拓扑在某个特定领域的应用。当时也还是博士生的 J¯anis Lazovskis（就读于芝加哥大学）在春季学校及之后的活动期间一直在收集相关论文，并分享了一份约含 10 篇论文的清单，展示拓扑数据分析的应用。在线电子表格的格式很快就显得过于受限，因此他们在 2020 年转而使用 Zotero [GLR20], ，这是一款专为管理书目数据库而设计的应用程序。当时，相关应用数量已增至约 30 个，J¯anis 和 Barbara 开始意识到需要周密规划：如果这个项目如我们所愿不断发展，收录数百条记录并供许多人使用，该怎么办？怎样才能让它便于检索且用途广泛？他们设计了一套标签 和类别 系统，与Zotero基础设施兼容，并据此对所有论文进行分类（有关该系统的更多信息，请参阅第 3.2 节）。与此同时，他们开始宣传这个数据库，并立即收到了社区的积极反馈。在支持者中，Mikael Vejdemo-Johansson 教授给予了他们极大帮助，提供了他个人数据库中的一百多篇论文。此外，次年他还承担了 Zotero 年度订阅费用，因为当时论文数量已超过免费存储空间的上限（此后费用一直由 Barbara 支付）。

2022年，Zotero 数据库中的条目已超过300条，所有条目都按照标签和风味系统进行分类，但它也开始显现出局限性：用户需要熟悉该应用程序，才能成功找到所需的参考文献。幸运的是，Barbara先是在萨尔茨堡举行的两年一度的TDA奥地利会议上，随后又在“罗马数据拓扑研讨会”期间结识了Bastian Rieck。Bastian Rieck很快爱上了这个项目，并通过贡献网页前端搜索引擎DONUT使其焕发新生。这个缩写

*Barbara Giunti 是格拉茨工业大学 Kerber 研究组的博士后研究人员。她的电子邮箱地址是 bgiunti@tugraz.at。

†J¯anis Lazovskis 是里加技术大学的一名 *docents* ，也是 RTU 里加商学院的教职人员。他的电子邮箱是 janis.lazovskis 1@rtu.lv。

‡BastianRieck 是亥姆霍兹慕尼黑中心的首席研究员，也是慕尼黑工业大学计算、信息与技术学院的教职人员。他的电子邮箱是 bastian.rieck@tum.de。



图2：2022年罗马研讨会“数据拓扑学”的参与者。版权归罗马大学所有。

代表“拓扑原创与非理论应用数据库”，同时指代具有非平凡拓扑的最基本形状之一——甜甜圈。

2 动机

最初的目标是开发一个工具，用于检索特定领域中拓扑学的所有应用，从而不仅能在具体应用层面提供概览，还能在更高层次展示所有应用领域。例如，该工具有助于推广这一研究领域，展现其丰富性和强大能力。这样的工具在准备论文和学位论文的引言时尤其方便；更重要的是，在撰写资助申请时，可以引用某一特定方法的相关应用。它还可以吸引更多研究人员加入这一领域；计算拓扑学是一个高度交叉且跨学科的领域，欢迎来自不同领域的新贡献。DONUT 也可用于创建或拓展项目，例如将拓扑数据分析应用于新的领域，或克服先前方法的局限性。

创建该数据库的另一项目标，是整理现有知识，这在信息过载的时代是一项迫切需求。因此，标签和类别不仅包括应用领域，还包括所使用的数学工具、数据的检索与预处理方式，以及结果在具体应用领域中的新颖程度。以这种结构化格式呈现信息，不仅有助于从业者实现其研究目标，也便于开展文献研究和跨领域研究。由于缺乏用于报告此类详细信息的结构化文献格式，DONUT 还承担着为该领域提供不断演进、动态分类体系这一重要职能。

3 编目工作原理

3.1 准入标准

要纳入 DONUT，条目必须使用拓扑数据分析技术分析数据。因此，我们排除了应用于数学或计算机科学其他领域的研究，以及采用不属于拓扑数据分析工具箱的数学（甚至是拓扑学）工具的研究。

我们编入索引的条目必须已经发表，或作为预印本发布在某个预印本服务器上（例如 arXiv 或 bioRXiv）。我们优先收录开放获取（OA）出版物和具有 DOI 的条目。对于后来正式发表的预印本，我们会以其正式发表的版本替代。如果后续发表的版本不是开放获取，仍会保留指向公开预印本的链接。只有已发表在会议论文集中的会议投稿才会被收录；仅包含摘要的会议投稿不予收录。

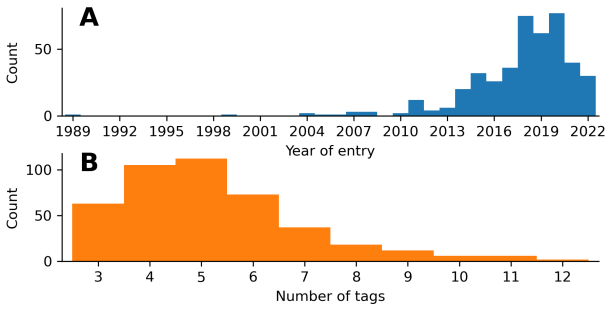

图3：每个条目的（A）出版年份和（B）标签数量直方图

3.2 标签与类别

标签分为三类（应用领域、所使用的数学工具和输入类型），风味标签分为两类（**创新**和**确认**）。每个已索引

条目必须为每一类至少设置一个标签；这是确保 DONUT 实用性的硬性要求。

应用领域。 这是最难添加的标签。我们希望利用社区的反馈，不断完善添加此标签的规则。理想情况下，此标签应涵盖不同层级，以确保最佳的搜索结果。例如，一篇关于癫痫的条目，其应用领域标签应为医学（作为一般领域）、神经病学（作为对该领域的细分），最后是最精确的标签癫痫。由于 DONUT 的工作方式（见第 4 节），搜索“epilepsy”而不是“tag:epilepsy”会返回所有提及该词的条目，即使该词只出现在参考文献中，因此搜索结果不够精确。

所使用的数学工具。 此类别最容易添加标签，因为作者通常会清楚说明技术描述、分析过程以及所使用的工具。我们的目标是为所有采用的工具添加标签，而不仅仅是来自拓扑数据分析的工具，从而忠实概括拓扑数据分析所应用的背景。

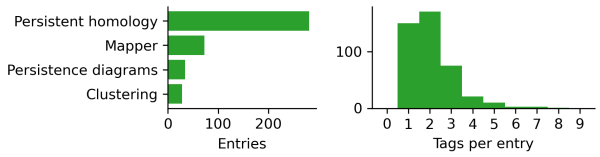

图 4：用于表示工具类型的最常用标签（左），以及每个条目包含此类标签数量的直方图（右）。

输入类型。 第三类标签在概念上非常简单，用于表示拓扑数据分析流程中所使用的数据类型，例如灰度图像、点云、时间序列等。然而，在实践中，我们发现该标签难以应用，因为作者通常不会明确说明输入数据的使用方式，或只是间接提及。分析应始终具有可复现性，而作者未能说明如何对原始数据进行预处理，使其适合用作拓扑数据分析输入，会严重妨碍结果复现。

类别。 类别并非必需，一方面是因为其分类更加微妙，另一方面是因为并非所有条目都能明确归入其中一种类型。**确认** 标签表示该条目的研究发现与现有方法的输出一致。这些条目承担着复现结果这一关键任务，并表明拓扑数据分析可以作为一种替代方法，而且可能是更优的方法。**创新** 标签涵盖了所有研究结果对于特定应用领域具有新颖性的条目。某项结果之所以新颖，可能是因为它来自更大数据集（其他方法无法处理的数据集），也可能是因为拓扑数据分析能够从相同数据中提取更多信息，或者因为拓扑数据分析能够分析其他方法无法处理的数据。由于该标签对于推广拓扑数据分析至关重要，我们决定对其采取严格标准：如果某项结果并非毫无疑问地具有新颖性，则不添加该类别（在目录收录的431个条目中，有58个标记为**创新**）。

4 技术细节

DONUT 基于一个由 Zotero 维护的文献条目数据库。使用 Zotero 的优势在于：(i) 它提供了一种搜索出版物并为其建立索引的简单方式；(ii) 所有文献条目都以 BibTEX 条目的形式存储。这意味着 DONUT 保持了灵活性，未来可以轻松切换到其他数据源，同时无需担心数据录入方面的问题。因此，DONUT 由三个相互独立的组件组成：

1. 用于在 Zotero 与条目数据库之间进行单向同步的导入器 。2. 用于处理查询并维护条目的全文搜索引擎。3. 用于与全文搜索引擎交互的 Web 前端。

导入器作为独立程序实现，通过 Pyzotero [H¨ug19] 使用 Zotero API。解析过程的结果是一系列 BibTEX 条目。随后，这些条目会被插入 Xapian，这是一款开源全文搜索引擎。Xapian 为文档建立书目信息索引，并通过定义良好的 API 使其可供访问。最后，一个基于 Flask 的 Web 界面

作为用于 Web 开发的 Python 前端，与数据库交互、呈现结果并处理所有查询。随后，我们将简要说明选择该搜索引擎和前端的原因。

4.1 全文搜索引擎

像 Xapian 这样的全文搜索引擎的优势在于，结构化文档数据的索引过程充满了隐藏的复杂性。例如，“high-dimensional”和“high dimensional”是同一个词吗？像将“simplical”误写为“simplicial”这样的简单拼写错误应如何处理？音译拼写（用“Pawel”代替 Pawe-l）或近似拼写（“Pavel”）又应如何处理？Zotero 的前端忽略了这些问题，只允许对给定字符串进行完全匹配的简单查询。相比之下，Xapian 能够识别语言，并可设置为针对查询提供替代拼写建议。由于 DONUT 的实用性取决于其对给定查询字符串返回结果的质量，我们选择尽可能多地为书目条目建立索引。因此，DONUT 查找文档的速度快于 Zotero （查询时间通常处于低毫秒级范围内），并且能够为查询提供更深入的结果。

4.2 前端

用户通常通过专门的查询界面与数据库交互；理想情况下，这些界面应像 Google 一样易于使用。我们使用基于 Python 的 Web 框架 Flask 提供了这样的界面（从某种意义上说，终端用户可能会认为 DONUT 是 Web 界面，但如上所述，DONUT 实际上由多个部分组成）。界面背后的设计选择首先着眼于速度和简洁，遵循极简主义的设计理念。搜索界面在大屏幕和小屏幕上都能良好运行，并且我们注意遵循官方的无障碍指南。

DONUT 不会通过 Cookie 或相关技术跟踪用户。常规的 Web 服务器日志会以匿名形式存储，因此无法识别用户。这些日志用于诊断

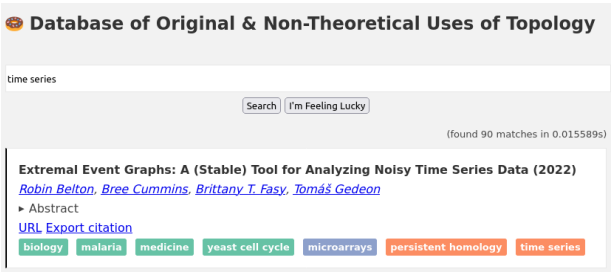

图 5：带有搜索词的用户界面（仅显示首条结果）。

问题，并提供有关访问数据库情况的汇总统计数据。日志会经过加密存储，并按滚动方式自动删除。因此，DONUT 完全符合《通用数据保护条例》的要求，并且远超当代网站的“最佳实践”。

4.3 代码

我们通过 GitHub 提供 DONUT 的代码，并采用 BSD 三条款许可证[1]。该许可证实质上允许任何人使用和修改代码，前提是保留原始版权声明。公开代码后，前端或后端中任何缺失的功能都可以轻松提出，并由社区或我们加以完善。我们也希望 DONUT 能启发类似的项目，因为这段代码并非专用于拓扑学应用，也很容易适用于其他科学领域。

4.4 示例查询

DONUT 的前端可以像使用常规搜索引擎一样使用。系统会在数据库的各个字段中搜索查询内容。因此，搜索 general 不仅会找到名为《The Euler Characteristic: A General Topological Descriptor for Complex Data》的条目，还会找到所有摘要中包含 general 的文档。其行为类似于搜索引擎的默认方式，因为搜索引擎并不关心

[1] https://github.com/Pseudomanifold/DONUT

搜索词在网站中的位置。为了满足研究人员的需求，DONUT 前端支持更精确的查询和各种运算符。例如，搜索 title:"general" 只会返回条目《The Euler Characteristic: A General Topological Descriptor for Complex Data》，因为（截至撰写本文时）没有其他条目的标题中包含 general。同样，也可以通过 tag: 和 author: 分别搜索文档标签和作者。

查询的自动规范化。 关于前述拼写差异问题，查询界面的一个突出特点是，它开箱即用地支持名称的音译拼写。因此，继续沿用原示例，搜索词 author:pawel 将显示同时包含“Pawe-l”和“Pawel”这两种拼写的结果。类似规则也适用于含变音符号及其他特殊字符的名称。此功能是 DONUT 独有的，Zotero 查询界面并不提供。

拼写建议。 此外，与 Zotero 不同，DONUT 能够根据相似性向用户建议其他搜索词。请注意，DONUT 绝不会自行更改搜索查询。不过，它会建议替代性的概念或拼写。例如，搜索同伦会显示同调这一潜在的相关查询。另一方面，搜索 homollogy 不会显示任何结果，从而促使 DONUT 建议：“您是不是想搜索‘同调’？”我们预计会随着时间的推移进一步改进此功能。

4.5 实验性功能

我们还将 DONUT 用作实验平台，探索各种让应用论文更易访问和查询的方式。例如，我们提供了一个“景观可视化”，按照自然景观隐喻展示所有已索引文档 [FMM10]。这为用户提供了一种与文档互动并有可能找到相似论文的方式。

另一项正在进行的改进涉及对开放获取出版物文本进行索引。这比整合书目信息（作者、摘要等）要复杂得多，因为这要求我们能够处理 PDF 文件。对于那些我们能够成功处理其文本的条目，我们会将文本纳入搜索引擎，这意味着仅出现在条目文本中的关键词或短语也将对读者开放。为确保符合版权要求，我们只会对开放获取出版物这样做。

5 未来机遇

我们首先希望 DONUT 成为一个真正有用的工具，服务于社区并由社区共同使用。因此，我们认为，最有价值的机遇在于扩展分类体系，也就是扩展标签系统。从“更一般”到“更具体”的转变，会形成自然的标签层级。例如，一条数据标签为 graphs 的条目，如果其上下文符合，可以被赋予形式为 graphs:directed 的更具体标签。随着时间推移，我们预计这一层级将变得更加精细，从而同时支持不具体和高度具体的查询。为了帮助用户与这一层级互动，我们计划实现标签的“树状可视化”。查看单个条目时，我们将显示该层级的摘录，使用户能够在树中导航。

我们希望 DONUT 能继续成为社区实用的工具。诚挚邀请每个人通过各种方式为 DONUT 做出贡献。我们欢迎有关纳入新内容、更新网页界面以及新增功能的更多建议。

致谢。 作者谨此感谢 Hausdorff 数学研究所组织了出色的数学活动，包括本数据库构想诞生时所举办的活动；同时感谢 Mikael Vejdemo-Johansson 教授和 Nina Otter 教授对数据库作出的贡献。B.R. 感谢与 Lukas Hahn、Maximilian Schmahl 和 Daniel Spitz 的讨论。B.G. 获奥地利科学基金（FWF P 33765-N）资助。

参考文献

[FMM10] Sara Irina Fabrikant、Daniel R. Montello 和 David M. Mark, *The natural landscape metaphor in information visualization: The role of commonsense geomorphology*, 《美国信息科学与技术学会杂志》**61** （2010）, 第 2 期, 253–270。[GLR20] Barbara Giunti、Janis Lazovskis 和 Bastian Rieck, *Zotero database of real-world applications of Topological Data Analysis* 2020, 。https://www.zotero.org/groups/tda-applications。[GLR22], *DONUT: Database of Original & Non-Theoretical Uses of Topology* 2022, 。https://donut.topology.rocks。[H¨ug19] Stephan H¨ugel, *Pyzotero* 2019 https://doi.org/10.5281/zenodo.7057503, Zenodo, 。。