通用 NFT 向量数据库：用于 NFT 相似性匹配的可扩展向量数据库

Samrat Sahoo Nitin Paul Agam Shah 佐治亚理工学院 佐治亚理工学院 佐治亚理工学院 美国佐治亚州亚特兰大 美国佐治亚州亚特兰大 美国佐治亚州亚特兰大 ssahoo61@gatech.edu npaul34@gatech.edu ashah482@gatech.edu

Andrew Hornback 佐治亚理工学院 美国佐治亚州亚特兰大 ahornback6@gatech.edu

Sudheer Chava 佐治亚理工学院 美国佐治亚州亚特兰大 sudheer.chava @scheller.gatech.edu

arXiv:2303.12998v1 [cs.DB] 23 Mar 2023

摘要

非同质化代币（非同质化代币）是一种数字资产，可作为特定数字物品（如艺术品、音乐或房地产）的所有权证明。由于非同质化代币具有不可替代性，重复代币不应具有相同的价值。然而，随着新的区块链不断涌现，且大量非同质化代币被创建，海量 NFT 数据随之产生，却没有相应的方法来追踪其相似性。这使人们能够仅通过更改一个像素或一个字节的数据，就创建出几乎完全相同的非同质化代币。尽管非同质化代币彼此相似，每个非同质化代币却都被分配了完全不同的代币 ID。为解决非同质化代币重复问题，我们开发了一套模块化、易于扩展、硬件无关且以云为中心的 NFT 处理系统，将非同质化代币表示为向量。我们建立了一个数据库，其中包含依据以太坊征求意见稿 721（ERC-721）代币标准生成的非同质化代币向量表示，以启动从不同区块链汇总 NFT 数据的流程。最后，我们开发了一款 NFT 可视化仪表板应用程序，配备用户友好的图形用户界面（图形用户界面），让非技术用户也能访问汇总后的 NFT 数据。通用 NFT 向量数据库是一个基于相似性汇总 NFT 数据的链下框架，为查询和分析 NFT 数据提供了一种有条理的方式，而这些数据此前无法通过链上解决方案获取。

PVLDB 引用格式：
Samrat Sahoo、Nitin Paul、Agam Shah、Andrew Hornback 和 Sudheer Chava。《通用 NFT 向量数据库：用于 NFT 相似性匹配的可扩展向量数据库》。PVLDB，17(1)：XXX-XXX，2023 年。
doi:XX.XX/XXX.XX

PVLDB 制品可用性：
源代码、数据和/或其他制品已在 https://github.com/gtfintechlab/Universal-NFT-Vector-Database 上公开。

本作品采用知识共享署名-非商业性使用-禁止演绎 4.0 国际许可协议授权。访问 https://creativecommons.org/licenses/by-nc-nd/4.0/ 查看此许可协议副本。对于超出本许可协议涵盖范围的任何用途，请发送电子邮件至 info@vldb.org 获取许可。版权归所有者/作者所有。出版权授权给 VLDB 基金会。VLDB 基金会论文集，第 17 卷，第 1 期，ISSN 2150-8097。doi:XX.XX/XXX.XX

1 引言

非同质化代币是区块链上的数字资产，与密码学密切相关，包含元数据和用于识别的独特哈希 [10]。非同质化代币旨在具备密码学唯一性，于 2012 年起源于比特币区块链，随后在以太坊区块链上广泛普及 [1]。非同质化代币通常以 ERC-721 标准表示 [4], , 非常适用于收藏品、数字密钥管理、演唱会门票等场景，因为它们能确立数字代币之间的唯一性。为了确保非同质化代币的真实性并提升非同质化代币生态系统的公平性，尤其是在图像类非同质化代币领域，研究人员和区块链开发者正在寻找检测并确保密码学唯一性的机制，使区块链系统和基础设施能够实现其预期功能 [2]。

目前，将原始文件数据存储在区块链上成本高昂。因此，区块链上的非同质化代币存储的是指向文件数据的引用（即 URL），而元数据则由链下存储解决方案托管。每个非同质化代币都会被分配一个唯一的代币 ID；虽然代币 ID 是唯一的，但链下数据却未必唯一。随着区块链基础设施的不断发展，仍缺乏健全的协议和标准来确保非同质化代币的真实性。这意味着某个实体可以复制原始非同质化代币的内容，只修改一个字节，就创建出一个全新的非同质化代币。因此，缺乏强制执行技术会阻碍依赖非同质化代币具备唯一性且无法被其他实体复制的数字创新，从而实际上使非同质化代币技术的初衷落空。

为应对非同质化代币面临的挑战，我们提出了一种模块化、易于扩展、与硬件无关且以云为中心的非同质化代币数据库。这样的基础设施有助于实现并促进代币唯一性的执行。它将数据标准化并存储到向量数据存储库中，用户可以通过网页界面轻松访问，以用于研究和实用目的。向量数据库通过判断非同质化代币特征向量之间的相似度，得出某个非同质化代币是否唯一。

首先，为了开发 NFT 向量数据库，我们使用 Graph Protocol 提取 ERC-721 合约。获取合约信息后，我们通过 Alchemy（一家区块链基础设施提供商）收集各个非同质化代币的数据。我们根据非同质化代币数据关联的媒体 URL 汇总图像，将其转换为向量，并将这些向量（每个都包含一个唯一 ID）存储在向量数据库提供商 Pinecone 中。

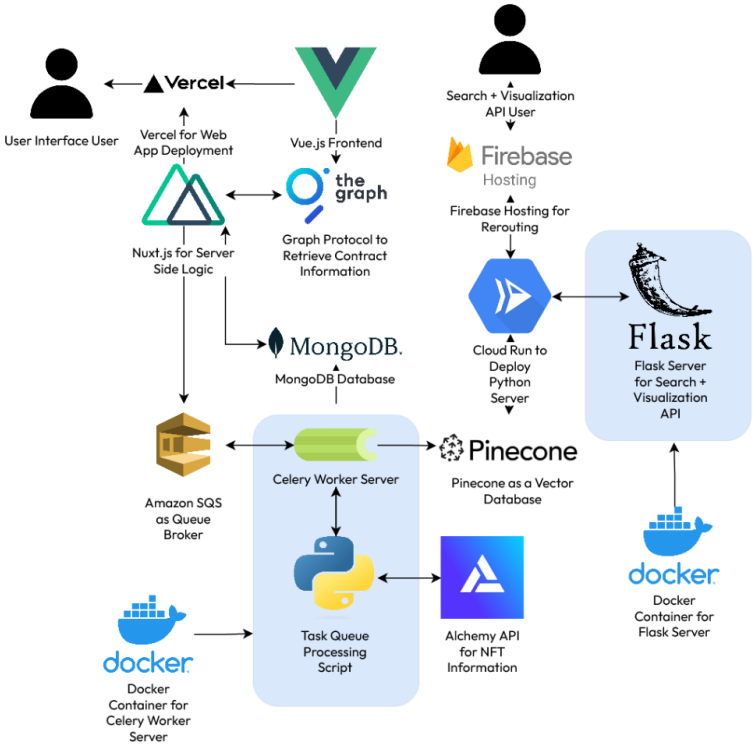


图1： 该图展示了我们设计的系统基础设施。左上方展示了客户端和服务器端逻辑，以及区块链数据收集工具。下方展示了任务队列处理系统，该系统负责处理 NFT 收藏集和非同质化代币。右上方展示了搜索与可视化 API 的基础设施。

接下来，我们将数据库连接到交互式图形用户界面，以获取有关非同质化代币合约数据的最新指标和关键信息。该 Web 应用程序使研究人员和消费者能够与向量数据库交互，并直接访问非同质化代币数据。此外，这套非同质化代币数据基础设施还可推广至各种需要独特数字资产的应用场景，例如身份认证和非同质化代币筛选系统 [5]。

通过这项研究，我们提出了一个确保非同质化代币独特性的框架，以便从这些代币中提取更多价值，供研究人员和消费者使用。此外，我们构建了一个搭载 NFT 向量数据库的 Web 应用程序界面，并提供易于使用的图形用户界面，以改善 NFT 数据的访问，并推动涉及非同质化代币的未来创新。本研究旨在展示该系统的可扩展性和模块化设计，使开发人员能够替换系统的部分组件，而不必针对每个应用领域都开发全新的基础设施。

2 数据收集与处理

2.1 区块链数据查询与收集

由于区块链固有的链表结构，查询以太坊区块链数据颇具挑战。传统上，若要查询区块链上的所有 ERC-721 合约和代币，需要启动一个以太坊 RPC 节点并遍历每个区块。我们的系统，则利用 Graph Protocol——一个构建在以太坊之上的去中心化查询层 [9]，，通过 GraphQL API 轻松快速地查询部分区块链数据（称为子图）。我们使用一个 EIP-721 子图 [7] 来监控所有 ERC-721 合约，并通过该子图收集合约地址。我们采用分页方式收集这些合约地址，每次处理 10 到 100 个合约，以确保子图基础设施和我们的基础设施都不会因过多的数据输入和输出操作而不堪重负。

我们测试了两种方法，以获取每个 ERC-721 合约中的单个 NFT 数据。第一种方法利用这些合约所使用的子图来获取非同质化代币及其各自的详细信息。第二种方法是将系统与 Alchemy 的非同质化代币基础设施集成；Alchemy 是一家 Web3 基础设施提供商。由于我们优先考虑尽可能快速地检索更多非同质化代币，因此进行了基准测试，以衡量哪种方法速度更快。我们发现，Alchemy 的非同质化代币基础设施显著缩短了处理时间，并且随着非同质化代币数量增加，其扩展性大致呈线性。这种线性扩展是因为 Alchemy 会将非同质化代币媒体信息缓存到数据库中，从而几乎可以即时检索。

2.2 数据处理

通过利用 Alchemy 的 NFT基础设施，我们获取单个非同质化代币的媒体 URL。然后，我们将此媒体 URL 输入数据处理系统，以获取该 URL 上的文件内容。我们的

系统不会对媒体内容进行任何预处理，以保留非同质化代币的固有属性。随后，系统通过 RegNetY- 080 嵌入 [6]，将图像直接转换为 2016x1 维向量。系统将数据分为两条数据库处理管线。第一条管线将生成的向量存储在名为 Pinecone 的向量数据库服务中。第二条管线则将非同质化代币的其他信息（例如链信息、NFT 元数据 URL 和 NFT 标准信息）存储在 MongoDB数据库中。这个双数据库系统使用户能够根据非同质化代币的视觉特征（通过向量数据库）获取其详细信息（来自 MongoDB数据库）。我们的设计不同于传统方法；传统方法仅依赖 NFT 集合合约 ID 和代币 ID 进行查询。

2.3 向量运算与降维

原始图像转换为向量后维度很高且不固定，这带来了两个我们试图解决的问题。首先，高维度会占用大量存储空间，即使是较小的图像，每个向量也需要数万个元素。通过利用图像嵌入，我们可以对图像进行降维，同时保留非同质化代币的固有属性。RegNetY-080 将图像表示为一个 2016x1 维向量，大幅减少了每张图像在数据库中占用的空间。

向量维度不固定还会带来第二个问题：向量比较。欧几里得距离等现有向量比较指标要求向量长度一致 [3]。然而，面对尺寸不同的图像，如果不进行标准化，就无法比较这些图像。将向量标准化后，我们便能够添加搜索相似非同质化代币的功能——我们选择余弦距离作为搜索指标，因为余弦距离通常更适用于高维嵌入 [8]。

对于可视化功能，我们需要一种有效的降维技术，将嵌入向量降至二维平面。为了确定哪种降维技术最有效，我们开展实验，对比多种方法，包括多维尺度分析、t 分布随机邻域嵌入、主成分分析、截断奇异值分解和等距映射。在本实验中，我们定义并计算了一项名为聚类比率的指标，以确定哪种低维降维技术最有效——聚类比率越高，表示降维技术越有效。我们将有效性定义为：在最大化非同质化代币收藏集簇中心之间距离的同时，最小化各个非同质化代币与其所属非同质化代币收藏集簇中心之间距离的算法。聚类比率、聚类距离和集合距离的定义如下：

- **聚类距离：** 这是两个非同质化代币簇中心之间的欧几里得距离。
- **集合距离：** 这是每个非同质化代币簇中的各个元素与该簇中心之间欧几里得距离的平均值。
- **聚类比率：** 这是两个聚类中心之间的距离与每个聚类中各元素到相应聚类中心的平均距离之比

元素中心。聚类比率更准确地定义为聚类距离与集合距离之比。

我们使用了一个包含来自两个不同非同质化代币收藏集的 50 个非同质化代币的非同质化代币数据集。每个非同质化代币收藏集中的非同质化代币都具有相似的美学属性。为了确定最佳算法，我们首先为数据集中的所有图像创建嵌入。随后，我们应用正在测试的降维技术，并使用 K 均值聚类算法对降维后的数据创建两个簇。根据结果，我们计算了聚类距离、集合距离和聚类比率。截断奇异值分解的聚类比率较高，因此被选为用于可视化的降维技术。

3 软件基础设施

3.1 ControlPanel、Dashboard 和后端服务器

我们的图形用户界面让非技术用户能够全面使用系统的各项功能。为了创建这一图形用户界面，我们采用了应用程序开发中常用的软件工程模型——模型-视图-控制器（MVC）模式，以确保用户能够有序、安全且便捷地访问数据：

- **模型：** 应用程序的模型部分主要涉及数据库架构，这些架构有助于更好地组织数据——我们使用 MongoDB 的 Mongoose 客户端编写并强制执行这些架构，以便将数据插入数据库。此外，使用 TypeScript 这类强类型语言，也让我们能够严格声明并强制执行类型。
- **视图：** 应用程序的视图部分让用户能够查看当前数据，并以抽象化的方式执行创建、读取、更新和删除（增删改查）操作。在图形用户界面中，用户通过查看分析数据、控制任务队列和更新分析数据等操作来执行增删改查操作。

- **控制器：** 系统的控制器是一个后端服务器，负责管理所有增删改查操作的逻辑。增删改查操作包括获取分析数据和更新任务队列等操作。此外，控制器增强了应用程序管理面板的安全性，用户通过 JavaScript 对象表示法（JSON）网络令牌进行身份验证，从而能够控制任务队列。控制器使用 TypeScript 编写，采用名为 Nuxt.js 的全栈 Web 框架，使客户端和服务器端代码能够置于单体代码库结构中，从而大大简化项目的组织与基础设施。

3.2 TaskQueue、工作进程与水平扩展性

使用图像嵌入将图像转换为向量在计算上非常耗费资源。因此，将此功能加载到 API 上会导致请求过载，需要垂直扩展性（增加计算资源），并使图像处理系统的健壮性降低。为此，我们实现了一个任务队列系统，可以将非同质化代币数据保存到单独的服务器上进行处理，从而实现水平扩展性。该系统的流程始于

管理后台，经过身份验证的用户可以在此将非同质化代币系列添加到任务队列中。管理员添加这些系列后，系统会在 MongoDB 任务队列集合中存储与已处理项目相关的补充信息，以及 ID 和状态。系统会将 ID 发送到消息队列。我们使用亚马逊云科技（AWS）的简单队列服务（SQS）作为消息队列。随后，这些项目会由运行 Celery 工作进程服务的独立服务器逐一处理，该服务包含两种不同的任务：

- **合约任务：** 第一个任务是合约任务，工作进程将通过 Alchemy 的 NFT基础设施获取某个合约下的所有单个非同质化代币，并按照第 2.2 节所述，为每个非同质化代币将任务项添加到任务队列中。

- **NFT任务：** 第二个任务是 NFT任务，工作进程将获取媒体 URL 中的数据，并按照第 2.2 节所述，通过 RegNetY-080 嵌入将其转换为 2016 维向量。

这种架构使系统不受机器物理硬件的限制，并着重于在不同设备间实现可复用性。

3.3 搜索与可视化API

我们提供公开的搜索与可视化API，让用户能够与向量数据库交互。这两个 API 在我们的 Web应用程序中也都配有交互式用户界面。

- **搜索API：** 搜索API 接受任意经过 Base64 编码的图像，并通过计算源图像向量（由 RegNetY-080 图像嵌入生成）与数据库中向量之间的余弦距离，找出与该图像最接近的非同质化代币。我们利用 Pinecone 的向量数据库搜索 API 来搜索合适的向量。
- **可视化API：** 可视化 API 的目标是通过 t 分布随机邻域嵌入对一组向量进行降维。可视化 API 接收一个向量列表，并将其降至二维。新生成的向量中的每个向量都对应笛卡尔坐标系上的一个点。

3.4 系统部署

结合使用云端和本地服务器进行部署，使我们能够借助专用的高性能硬件扩展规模，并每秒处理更多非同质化代币。部署系统时，我们的目标是降低成本，同时提高处理效率。我们需要部署五个系统：向量数据库、MongoDB数据库、任务队列、公共 API 和 Web应用程序。图1详细展示了部署基础设施。

- **向量数据库：** 借助 Pinecone 向量数据库，我们无需手动配置扩展、存储、分片及其他基础设施细节，从而大幅简化了架构。
- **任务队列：** 对于任务队列处理系统，我们部署了两个不同的系统。第一个是任务队列，用于存储其他任务 ID。为此，我们使用了亚马逊简单队列服务，作为 Celery 工作进程服务器的代理。第二个系统是 Celery 工作进程服务器，它被设置为系统服务，运行在 Docker 容器中，并部署在本地服务器上，全天候运行，随时等待处理任务队列中的项目。
- **公共 API：** 对于公共 API，我们将其打包为 Docker 容器，并作为无服务器函数部署到 Google Cloud Run。我们使用 Firebase 托管将 HTTP 请求转发到容器化应用程序中的函数。采用无服务器架构后，容器会根据 API 流量水平横向扩展或缩减，同时确保成本与实际使用量相匹配。

4 结论

本研究设计并开发了一种框架，通过以非同质化代币向量数据库为核心组件的系统来确保非同质化代币的唯一性，从而提升用户能够从非同质化代币中获取的内在价值。我们通过提取合约信息并查询各个非同质化代币数据，以获取媒体 URL，进而汇总非同质化代币图像，并将其作为向量相应地存入数据库，从而构建了非同质化代币向量数据库。我们还通过提供直观的 Web 应用程序界面，方便用户搜索数据库中的非同质化代币并获取相似度分数，从而改善了对非同质化代币数据及相关指标的访问。我们提供的数据访问方式和基于相似性的非同质化代币唯一性标准化框架解决方案，有助于让非同质化代币生态系统更加公平，使区块链开发者、消费者和研究人员能够在各种应用和领域中发掘其潜力，推动未来创新与研究。

参考文献

[1] Yoni Assia、Vitalik Buterin、M LiorhakiLior Hakim、Meni Rosenfeld、Peter Racz 和 Rotem Lev。2015年。*Colored Coins* 白皮书。技术报告。Colored Coins。[2] D. G. W. Birch。2022年。非同质化代币：新的欺诈目标。https://www.forbes.com/sites/davidbirch/2022/02/20/nfts-new-fraud-targets/?sh=1a9bbac34f37 [在线发布；发布于 20-February-2022]。[3] Ivan Dokmanic、Reza Parhizkar、Juri Ranieri 和 Martin Vetterli。2015年。欧几里得距离矩阵：理论、算法与应用简述。*CoRR* abs/1502.07541 (2015)。arXiv:1502.07541。[4] http://arxiv.org/abs/1502.07541 (2018)。William Entriken、Dieter Shirley、Jacob Evans 和 Nastassia Sachs。。https: //eips.ethereum.org/EIPS/eip-721 [在线；修改于24-January-2018]。[5] Rodolfo Mecozzi、Giuseppe Perrone、Dario Anelli、Nicola Saitto、Elia Paggi 和 David Mancini。2022年。区块链相关身份与访问管理挑战：（去）中心化数字身份监管。载于*2022 IEEE* 国际区块链会议（*Blockchain*）。443–448。https://doi.org/10.1109/ Blockchain55522.2022.00068[6] Ilija Radosavovic、Raj Prateek Kosaraju、Ross B. Girshick、何恺明和 Piotr Dollár。2020年。设计网络设计空间。*CoRR* abs/2003.13678（2020年）。arXiv:2003.13678。https://arxiv.org/abs/2003.13678。Ronan Sandford。。[7] EIP- Subgraph。https://thegraph.com/hosted-service/subgraph/wighawag/eip721-subgraph (2019)。[8] Erich Schubert。2021年。余弦相似度三角不等式。*CoRR* abs/2107.04071 (2021)。arXiv:2107.04071。[9] https://arxiv.org/abs/2107.04071 (2019)。Yaniv Tal、Brandon Ramirez 和 Jannis Pohlmann。。*The Graph*：区块链去中心化查询协议，第2版。技术报告。The Graph 基金会。[10] Qin Wang、Rujia Li、Qi Wang 和 Shiping Chen。2021年。非同质化代币（NFT）：概述、评估、机遇与挑战。*CoRR* abs/2105.07447（2021年）。arXiv:2105.07447 https://arxiv.org/abs/2105.07447