医共体共享平台下多源异构数据语义索引方法

冶伟伟 [1]
YE Weiwei

摘 要 医共体共享平台整合了多层级医疗机构的海量数据，现有的数据语义索引方法通过构建统一数据模型实现数据整合与检索，存在语义理解不准确、索引效率低等问题，难以满足医共体临床决策对数据精准获取的需求。对此，文章研究了一个医共体共享平台下多源异构数据语义索引方法。首先对多源异构数据特征分析，明确数据类型与格式。基于此利用数据语义相似度对多源异构数据聚类，将数据划分为不同簇。采用知识图谱和嵌入表示构建混合语义索引，为聚类后的数据生成语义索引。实验结果表明，该方法应用后，数据语义索引匹配度均达到了 96% 以上，同时，相较于传统方法，索引耗时较少，平均索引耗时为 107.8 ms，显著低于两种传统索引方法的 186.3 ms 和 173.7 ms，有效解决了多源异构数据语义冲突与检索效率低的问题，为医共体共享平台数据的高效利用提供了可靠技术支撑。

关键词 医共体共享平台；多源异构；数据；语义；索引

doi：10.3969/j.issn.1672-9528.2025.10.040

0 引言

医疗数据共享平台作为区域医疗资源整合的重要载体，其核心目标在于促进不同医疗机构之间的数据交互与业务协同，从而优化医疗服务质量与运营效率。在实际运行过程中，平台需要处理来自多个业务系统的异构数据，主要包括电子病历记录、数字化医学影像、实验室检测结果以及医疗保险信息等 [1]。这些数据在组织形式、存储格式和业务语义等方面表现出明显的差异性，给数据的标准化整合与高效利用带来了诸多困难。数据语义层面的不一致性问题不仅阻碍了系统间的互操作能力，更限制了基于数据的智能诊疗、临床决策支持等高级功能的开发与应用。

因此，构建一种高效的语义索引方法，以实现医疗共同体多源异构数据的统一表示与快速检索至关重要。现有研究表明，传统的数据语义索引技术在医疗领域的应用仍存在明显局限。其中，王思丽等人 [2] 提出的方法利用深度学习模型，通过向量化表示实现语义相似性计算实现索引，但医疗数据的专业性和复杂性使得通用模型难以直接适用，索引时效性存在不足，同时索引结果偏差较大。范国栋等人 [3] 提出的方法利用数据仓库技术，通过模式映射实现数据整合与索引，但该方法的灵活性较差，难以适应动态变化的医疗数据环境。因此，本文开展了医共体共享平台下多源异构数据语义索引方法研究，旨在为智慧医疗建设提供更高效的数据支撑方案。

1. 陇南市第一人民医院 甘肃陇南 746000

1 医共体共享平台下多源异构数据聚类

在医共体共享平台中，数据来源于各成员单位的数据库系统、文件存储系统以及外部合作机构的数据接口，其多源异构特性显著 [4]。首先建立统一的数据采集通道，利用 ETL 工具与 API 接口连接各数据源，并按照数据来源分类存储至分布式文件系统。根据表 1 所示的解析规则，对不同类型数据进行特征提取。

表 1 数据格式解析

序号	数据类型	解析方式
1	结构化数据	SQL Schema 解析提取字段模式
2	半结构化数据	JSONPath 解析标签 - 值对
3	非结构化数据	NLP 技术抽取实体

通过表 1 的解析，输出数据特征。通过对医共体共享平台下多源异构数据特征进行分析，以解析医共体数据的多源异构特性。在完成医共体共享平台下多源异构数据的特征分析后，已明确数据的类型、格式及语义关联特性。以此为基础，采用基于语义相似度的聚类方法对多源异构数据进行分组处理，为后续构建高效语义索引奠定基础。

首先，对数据进行向量化。针对结构化多源异构数据，将字段名和值经本体标注后转换为嵌入向量；针对非结构化多源异构数据，提取文本语义向量 [5]。为全面衡量数据间的相似性，提出结合结构与语义特征的混合相似度度量方法。用公式表示为：

$$S(A,B)=\alpha\cdot S_s(A,B)+(1-\alpha)\cdot S_t(A,B) \quad (1)$$

$$S_t(A,B)=\frac{\boldsymbol{v}_A\cdot\boldsymbol{v}_B}{\|\boldsymbol{v}_A\|\cdot\|\boldsymbol{v}_B\|} \quad (2)$$

式中：$S(A, B)$ 表示数据 A 和数据 B 之间的混合相似度；$S_s(A, B)$ 表示数据的结构相似度，根据数据在字段名、数据类型、取值范围等结构特征上的匹配程度得到；$S_t(A, B)$ 表示数据的语义相似度；$\boldsymbol{v}_A$、$\boldsymbol{v}_B$ 分别是数据 A 和数据 B 的向量表示；α 表示权重参数，用于调节结构相似度和语义相似度在混合相似度中的比重。通过该公式，能够更准确地衡量不同类型数据之间的相似程度。

利用 DBSCAN+ 语义约束对相似度度量后的数据进行聚类，根据数据点的密度将数据划分为不同的簇，而语义约束则确保聚类结果在语义上具有一致性 [6]。其公式为：

$$\left|N_\varepsilon(p)\right|\geq\min P \quad (3)$$

式中：$N_\varepsilon(p)$ 表示以点 p 为中心、半径为 ε 的邻域内的数据点集合；$|N_\varepsilon(p)|$ 表示该邻域内数据点的数量；min P 表示设定的最小样本数阈值。

加入语义约束，即只有当两个核心点不仅在空间距离上满足邻域条件，而且它们之间的混合相似度 $S(A, B)$ 也大于设定的阈值时，才将其划分到同一簇中 [7]。

通过上述流程的实施，将语义相近但结构不同的数据归入同一簇。输出带语义标签的数据簇集合，同一簇内数据可跨源但语义一致。

2 多源异构数据语义索引

基于聚类结果，提出一种融合知识图谱与向量嵌入的混合语义索引方法。以聚类生成的语义簇标签作为根节点，构建层次化子图谱。对于簇内实体 e_i 和 e_j，其关联强度计算为：

$$w_{ij}=\frac{f_c(e_i,e_j)}{\sqrt{f(e_i)\cdot f(e_j)}} \quad (4)$$

式中：f_c 表示医共体共享平台下多源异构数据共现频率；$f(e)$ 表示数据实体独立出现频次。将本体中的关系与现有图谱融合，补全隐含关联 [8]。

将医共体共享平台下多源异构数据语义的索引问题建模为语义决策系统，则数据语义索引的模糊参数指标集设定为 $G_k\in G$（$k=1, 2, \cdots, t$），在知识图谱中，得到医共体共享平台下多源异构数据语义索引的优先级控制向量集为：

$$\boldsymbol{R}_i=\frac{\boldsymbol{H}_i^-}{\boldsymbol{H}_i^+ + \boldsymbol{H}_i^-},\quad \boldsymbol{Q}_i=\frac{\boldsymbol{B}_i^-}{\boldsymbol{B}_i^+ + \boldsymbol{B}_i^-} \quad (5)$$

式中：$\boldsymbol{H}_i^-$、$\boldsymbol{H}_i^+$ 分别表示水平和垂直方向的语义索引控制向量；$\boldsymbol{B}_i^-$ 和 $\boldsymbol{B}_i^+$ 分别表示左、右对角线的语义索引控制向量。通过这 4 个控制向量，能够从多个维度综合控制数据语义索引的优先级 [9]。

在索引构建过程中，采用非度量多维标度方法进行自适应寻优。该方法能够在不考虑数据具体距离度量方式的情况下，根据数据之间的语义相似性，将数据映射到低维空间中，并优化数据点的位置，使得映射后的数据点之间的距离关系尽可能反映原始数据之间的语义相似性 [10]。用公式表示为：

$$\min S=\sqrt{\frac{\sum_{i<j}\left(d_{ij}-\hat{d}_{ij}\right)^2}{\sum_{i<j}d_{ij}^2}} \quad (6)$$

式中：d_{ij} 表示原始医共体共享平台下多源异构数据之间基于语义相似度计算得到的距离；$\hat{d}_{ij}$ 表示映射到低维空间后数据点之间的距离。

通过不断迭代优化，使应力函数值最小化，从而得到数据在低维空间中的最优布局。

在此基础上，计算出医共体共享平台下多源异构数据文本集 Y 的信息高效索引中心向量 $\boldsymbol{C}(Y)$，该中心向量是数据文本集语义特征的集中体现。结合数据文本集 X 的信息高效索引中心向量 $\boldsymbol{C}(X)$，得到数据的完整语义分布相似度函数：

$$S(X,Y)=\frac{\boldsymbol{C}(X)\cdot\boldsymbol{C}(Y)}{|\boldsymbol{C}(X)|\cdot|\boldsymbol{C}(Y)|} \quad (7)$$

式中：$S(X, Y)$ 的取值范围在 [−1,1] 之间，值越接近 1，表示两个数据文本集的语义分布越相似；值越接近 -1，表示语义分布越不相似；值为 0 时，表示两者语义分布相互独立。

通过该相似度函数，能够准确衡量不同数据之间的语义相似程度，为数据语义索引提供依据。当医共体共享平台新增数据时，利用上述聚类算法将其分配至已有簇或新簇。若新增数据和某个已有数据簇在语义层面上呈现出较高的相似性，那就把该数据归到这个簇中；若与所有簇的语义相似度均较低，则创建一个新簇。根据新增数据的具体信息，更新知识图谱中的实体关系和嵌入向量索引，确保索引始终保持最新状态。

根据上述索引结构，为聚类后的数据生成混合语义索引，能够支持基于关键词和语义相似度的混合查询，为医共体共享平台提供高效、准确的数据检索服务。

3 实验测试

3.1 实验准备

为检验本文方法在医共体共享平台多源异构数据语义处理场景下的实际应用性能，借助 Matlab 开展仿真实验分析。为保障实验结果的可靠性与严谨性，文献 [2]、文献 [3] 中提出的方法分别设置为对照组 A 与对照组 B，与本文所提的索引方法，即实验组，进行对比分析。

实验采用真实医共体环境下的多源医疗数据集，具体组成如表 2 所示。

表 2 多源医疗数据集组成

序号	数据类型	数据量
1	电子病历核心表	12.7 万条患者记录
2	检验结果表	54.3 万条记录
3	HL7 FHIR 格式的检查报告	8.6 万份
4	JSON 格式的医保结算数据	23.4 万条
5	临床文本	9.2 万份出院小结
6	DICOM 影像	6.8 万例

实验对硬件与软件环境均有明确要求。硬件方面，服务器硬盘容量必须超过 40 GB，以确保有足够的存储空间来支撑实验数据和运行过程中的临时文件存储；客户端需采用配备 768 GB DDR4 内存硬盘以满足大规模数据处理和复杂算法运行的内存需求。软件上，选用 Ubuntu 20.04 LTS 作为操作系统。

根据上述仿真环境和参数设定，进行医共体共享平台下多源异构数据语义索引测试分析。

3.2 索引结果分析

选择将医共体共享平台下多源异构数据语义索引匹配度作为实验的对比指标。针对表 2 中的每种数据类型，设计 200 个测试查询用例。对于每个查询用例，在 3 种方法中进行语义索引，返回与查询语义最相关的前 10 条结果。计算返回结果中正确匹配的数据条数占实际应匹配数据条数的比例，作为该查询用例的索引匹配度，对比结果如图 1 所示。

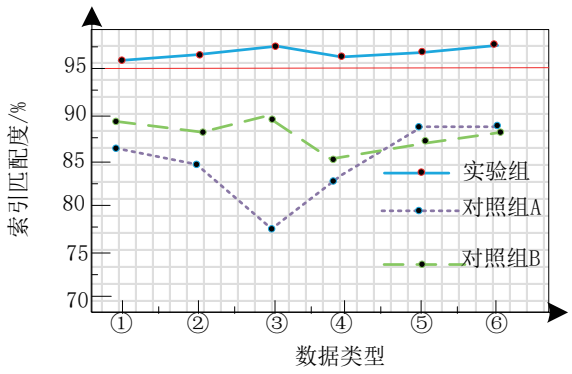

图 1 数据语义索引匹配度对比

通过图 1 的对比结果可以看出，在所有 6 种数据类型中，本文提出的语义索引方法的匹配度显著高于两种传统索引方法，均达到了 95% 以上，相比之下，两种传统索引方法在应用后，医共体共享平台下多源异构数据语义索引结果与目标数据语义的匹配度较不稳定，浮动范围较大。

由此可以证明相比传统的方法，本文设计的索引方法的应用效果良好，按照规范使用该方法进行医共体共享平台下多源异构数据语义索引，可以提高索引结果的匹配程度，提供更加优质的索引服务。

为了评估不同语义索引构建方法在处理大规模数据时的效率，通过持续增加索引数据量，分别应用 3 种方法完成语义索引构建，记录完成索引的耗时并进行对比分析。对比结果如表 3 所示。

表 3 数据语义索引效率对比数据

单位：ms

数据量 / 条	实验组	对照组 A	对照组 B
5 万	50.2	102.1	95.6
10 万	76.3	153.6	134.7
15 万	92.8	176.8	163.4
20 万	123.4	193.9	188.2
25 万	140.7	226.4	216.4
30 万	163.4	264.7	243.8

通过分析表 3 的对比数据可以得知，在医共体共享平台下多源异构数据量持续增加的情况下，本文设计方法的索引耗时较少，平均索引耗时为 107.8 ms，显著低于对照组 A 的 186.3 ms 和对照组 B 的 173.7 ms。

这一对比结果表明本文方法在处理大规模数据时，具有更好的索引查询效率稳定性，能够更高效地完成索引更新，保证索引的实时性，满足医共体共享平台在数据快速增长和实时查询需求下的性能要求。

4 结语

在医共体建设不断深化、医疗数据呈指数级增长的背景下，本文提出了医共体共享平台下的多源异构数据语义索引方法研究。通过系统分析医共体数据的多源异构特征，采用密度聚类算法实现数据语义分组，并构建层次化的混合索引结构，有效解决了传统方法在医疗数据整合与索引中的问题。实验结果表明，本方法在 6 类典型医疗数据中的索引匹配度均达到了 96% 以上，同时索引耗时较短，能够在保证索引准确性的前提下实现高效、实时索引查询的目标，为临床决策提供了可靠的数据支撑。

参考文献：

[1] 高峰，杨帆玉，顾进广．基于多级索引的前向语义数据流推理 [J]. 计算机应用与软件，2022, 39(1):24-29.

[2] 王思丽，李慧佳，孟庆洪，等．基于深度学习模型的领域知识库语义检索服务实现研究 [J]. 图书馆学研究，2024(10): 40-49.

[3] 范国栋，陈世展，肖建茂，等．面向 App 评论响应的语义检索和生成框架 [J]. 计算机学报，2022, 45(12):2528-2543.

基于同态加密的计算机隐私数据密码验证方法研究

郭 嘉[1] 黄玉帅[1] 吕兴胜[1] 徐迟阳[1] 韩 超[1]
GUO Jia HUANG Yushuai LÜ Xingsheng XU Chiyang HAN Chao

摘 要 随着信息技术的飞速发展，计算机隐私数据保护已成为全球关注的核心议题。在密码验证场景中，用户敏感数据面临潜在泄露与破解风险，如何构建安全可靠的验证机制以规避该问题，成为当前亟待解决的关键需求。文章聚焦基于同态加密的计算机隐私数据密码验证方法展开研究：首先阐述同态加密的基本原理与密码验证应用框架，明确密码存储策略与加密实现路径；其次详细梳理加密后验证的具体流程与操作步骤，进一步拓展多重身份认证机制设计及远程验证场景的应用模式；最后针对该方法实际应用中的性能瓶颈，提出针对性的性能优化与效率提升策略。研究旨在为计算机隐私数据的密码验证提供安全可行的技术方案，同时通过优化策略增强方法的实用性，为敏感数据保护领域的技术实践提供参考。

关键词 同态加密；隐私数据；密码验证

doi：10.3969/j.issn.1672-9528.2025.10.041

0 引言

随着信息技术的飞速发展，计算机系统的安全性和隐私保护问题成为全球范围内关注的焦点，尤其是在云计算、物联网、人工智能等技术的广泛应用下，隐私数据的保护显得尤为重要。随着互联网用户数量的激增和信息交换的日益频繁，个人数据和敏感信息在网络上传输和存储的风险极大增加。传统的密码验证方法，在确保数据安全的同时往往也伴随着隐私泄露的隐患，在数据存储和处理过程中，未加密的数据可能成为攻击者的攻击目标。

1 同态加密技术概述

同态加密技术是一种允许在加密数据上执行特定计算操作的加密方法，其核心优势在于能够在不解密数据的情况下对其进行处理，对于提升计算机隐私数据保护能力具有重要意义。传统加密方法如对称加密和非对称加密通常需要对加密数据进行解密后才可处理，这种操作不可避免地暴露了敏感信息，增加了数据泄露和篡改的风险。而同态加密技术通过其独特的数学结构，能够在密文状态下执行算术操作，如加法和乘法，使加密数据的隐私性在处理过程中持续保护。基于同态加密的计算机制具有可扩展性和灵活性，可以广泛应用于诸如云计算、分布式计算、隐私保护数据分析等领域，在计算机隐私数据密码验证中，能够实现无需解密密码即可完成身份验证、数据校验等任务，从而有效防止密码泄露、身份冒充及中间人攻击等安全威胁。具体而言，部分同态加密技术允许对加密数据进行加法操作，某些则支持加法与乘法操作的组合，使得其在不同场景下具有更广泛

1. 三未信安科技股份有限公司 山东济南 250000

[4] 赵征宇，罗景，涂新辉．基于多粒度语义融合的信息检索方法 [J]. 计算机应用，2024, 44(6):1775-1780.
[5] 旷灵，杨思超，刘谋苗．智能检索系统下语义检索的常见策略对比 [J]. 中国科技信息，2024(14):37-39.
[6] 杜鹏举．基于多模态数据语义检索的关键技术研究 [J]. 产业创新研究，2023(12):118-120.
[7] 李易玮，郭婉莹．基于智能化检索系统语义检索文本改写策略 [J]. 中国科技信息，2023(13):60-62.
[8] 滕少华，黄文彪，张巍，等．标签与样本双语义增强的跨模态检索 [J]. 江西师范大学学报：自然科学版，2023, 47(3):296-306.
[9] 周康宾，滕璐瑶，张巍，等．可判别性标签语义指导的域适应检索 [J]. 小型微型计算机系统，2024, 45(7):1639-1647.
[10] 王丽，蒋明，王伟，等．电力信息通信客服机器人特定语义数据检索优化 [J]. 电子设计工程，2024, 32(20):168-171.

【作者简介】
冶伟伟（1991—），男，甘肃陇南人，本科，工程师，研究方向：医院信息化。

（收稿日期：2025-05-14 修回日期：2025-09-29）