一种面向数据场景化使用的跨数据空间搜索系统

姜海鸥[1]，胡子逸[2]，刘冉[2]，罗超然[1]，杨婧如[1]
1. 数据空间技术与系统全国重点实验室，北京 100195；
2. 北京大学 软件与微电子学院，北京 102600

摘要

随着国家数据要素政策的推进，如何在分散于各行业、各领域形成的数据空间之间，实现数据资源高效发现和便捷使用，成为数据要素有效流通的关键前提。研究提出一种面向数据场景化使用关系的数据空间搜索系统，针对各个数据空间数据不出域、不汇聚的安全要求，提出了基于数字对象架构的分布式搜索框架；基于数据场景化使用形成的数据语用关系，提出了基于查询相似度和数据语用关系的数字对象排序算法——DO-Rank算法，为上层应用提供了跨数据空间数据资源搜索发现的能力。实验表明跨空间搜索系统在不汇聚数据的情况下精确率、召回率、nDCG等指标接近数据集中汇聚的检索方法。

关键词

数据空间；数据流通；数字对象架构；搜索系统；搜索排序

中图分类号：TP311　　文献标志码：A　　doi：10.11959/j.issn.2096-0271.2026018

A cross-data-space search system for the scenario-based data usage

Jiang Haiou[1], Hu Ziyi[2], Liu Ran[2], Luo Chaoran[1], Yang Jingru[1]
1. National Key Laboratory of Data Space Technology and System, Beijing 100195, China；
2. School of Software and Microelectronics, Peking University, Beijng 102600, China

Abstract

With the advancement of national data element policies, how to achieve efficient discovery and convenient use of data resources in data spaces across various industries and fields has become a key prerequisite for the effective circulation of data elements. The work proposed a cross-data-space search system for the scenario-based data usage. In response to the security requirements that data must not leave its data space or not centrally aggregated, a distributed search framework based on the Digital Object Architecture (DOA) is proposed. Based on the pragmatic data relationships abstracted from the scenario-based data usage, a Digital Object ranking algorithm, named DO-Rank is proposed. It considers both query similarity and data pragmatic relationships. It provides upper-layer applications with the capability to search and discover data resources across different data spaces. Experiments show that the cross-space search system achieves the precision, recall, and nDCG scores close to the methods that aggregate data from data spaces.

Key words

data space, data circulation, Digital Object Architecture, search system, search ranking

1 引言

随着数字经济的快速发展，数据已成为驱动社会创新和产业变革的核心生产要素，国家数据局2024年11月21日印发《可信数据空间发展行动计划（2024—2028年）》（国数资源〔2024〕119号）明确提出“以推动数据要素畅通流动和数据资源高效配置为目标，以建设可信可管、互联互通、价值共创的数据空间为重点，分类施策推进企业、行业、城市、个人、跨境可信数据空间建设和应用，为充分释放数据要素价值，激发全社会内生动力和创新活力”。在这一政策背景下，各级政府、行业龙头企业及科研机构纷纷开展数据空间的规划与落地工作，政务数据空间、工业互联网数据空间、医疗健康数据空间等行业、领域、区域数据相继涌现，将会形成覆盖多领域、多层次的数据资源体系。然而，在数据确权、隐私保护和访问控制日益严格的趋势下，尽管数据空间的数量不断增多、单个数据空间内部数据可以流通利用，如何在“数据不出域”“可见不可得”的数据安全要求下实现不同数据空间之间的高效互联与协同使用仍是需要解决的一大难题。

数字对象架构（Digital Object Architecture，DOA）是由互联网发明人、图灵奖获得者Robert Kahn于2006年提出的一套分布式数据互操作架构[1]。将数据封装成包括标识、元数据、数据实体三部分的数字对象（Digital Object，DO），标识是数字对象的身份ID，全局唯一且持久，不以DO的所有者、存储位置、访问方式的改变而改变，元数据表示数字对象和应用、业务相关的描述信息，用于根据应用需求搜索所需数字对象，数字对象实体是实际被封装的数据内容。采用数字对象标识解析系统负责标识的创建、解析、修改和销毁，数字对象注册系统负责元数据的注册、发布、修改和删除，数字对象仓库系统负责实体的封装、存储、访问和销毁。采用数字对象标识解析协议（Digital Object Identifier/Resolution Protocol，DOIRP）和数字对象接口协议（Digital Object Interface Protocol，DOIP）分别规范对数字对象标识的管理与解析[2]和数字对象的检索与访问等操作。DOA将互联网中的数据资源通过统一的形式抽象和建模成DO，从而屏蔽了数据资源的异构性，通过标准协议实现数据的创建、获取、更新和删除等，从而实现互联网上异构数据的互联互通[3]。数据提供方将数据资源抽象成数字对象，并将元数据发布到注册系统上供公开检索查询，数据需求方检索所需元数据后根据元数据中的访问信息申请使用数据资源。如果数据需求方想要使用多个数据空间中的数据，则需要与各个数据提供方管理的注册系统分别建立连接后进行检索。这种方案存在两大问题，其一，由于各个数据提供方注册系统的数据库、软件实现机制都不同，需要数据需求方根据每个注册系统编写不同的访问接口和查询语言，复杂度极高；其二，各个注册系统返回的搜索结果依赖底层数据库的排序算法，数据需求方获取的数据缺少高效排序方法，查询结果的数量和质量无法保证，进一步影响数据发现效率。例如，医生在有权限访问电子病历系统、医保用药系统、卫健委诊疗指导系统的情况下查询某疾病信息，需要分别向三个系统发送查询请求，对于为病人开药的数据

使用场景，电子病历和医保用药数据较为重要，对于研判当前流行病风险与治疗方案制定场景，电子病历和诊疗指导数据较为重要。而现有数联网技术的实现方案仅根据各个系统运行和网络速度将查询结果直接返回，需要医生花费大量时间进行人工筛选。针对上述问题，研究的主要贡献如下：

（1）提出基于数字对象架构的分布式搜索框架，设计统一的查询生成、查询转换、查询执行方法，将数据需求方从与各个注册系统建立连接、面向异构注册系统的查询语言转换等繁琐操作中解放出来，满足对数据空间数据的可搜索可发现需求。

（2）提出基于数据场景使用关系的数字对象搜索结果排序算法，首先结合各数据空间数据数据可见不出域的要求，在各个数据空间内部构建搜索结果的数据使用关系视图；随后综合文本相似度和数据关系权重，对汇总的各数据空间搜索结果进行排序，确保面向数据场景化使用的数据搜索的查准率和用户满意度。

（3）使用不同领域数据集、模拟异构注册系统进行了实验，分析了系统跨空间发现数据的查全率、查准率和用户满意度。

2 相关工作

互联网上的搜索从基于人工分类的导航目录发展为基于相关性的检索系统，再到当前广泛使用的基于网页链接的搜索引擎，并不断向个性化、多样化、智能化、社会化的智能化搜索引擎发展[4]。当前主流搜索引擎信息源来自于互联网中网页的内容，网络爬虫爬取整个互联网的网页信息，经过网页去重、解析、建立倒排索引，进行集中或分布式存储[4][5]。搜索请求经过意图理解、搜索匹配后，根据网页之间的链接关系通过链接分析算法评估页面的相对重要性，为用户提供搜索结果。传统搜索引擎通常需要将整个网络的数据爬取下来，而数据空间的数据有不出域的安全要求，无法支撑数据的全量汇聚。元搜索引擎采用分布式架构，元搜索引擎通过统一的用户搜索界面接收用户搜索词，在多个独立的搜索引擎上执行搜索任务，并综合这些搜索引擎的结果，为用户提供更加全面和准确的信息[6]。Pandey K. K. 和 Shukla P. K. 等人研究表明元搜索引擎在检索互联网上相关信息方面比传统搜索引擎更有效[7]。元搜索引擎是为了扩大搜索数据来源的覆盖范围，本质上还是基于已有同质化的 Web 搜索引擎，无法有效应对异构数据空间的搜索问题。为了解决传统搜索引擎无法访问大量受权限或商业限制的孤立信息源等问题，联邦检索技术应运而生，是一种针对多源异构数据资源的分布式检索技术，支持用户通过一次查询实时检索多个独立数据资源，并将检索结果直观且统一的形式呈现给用户[8]，联邦检索的研究主要关注数据源描述、数据源选择和搜索结果合并三个关键问题。文章提出搜索系统基于联邦搜索技术，并重点解决面向数据使用场景的搜索结果合并问题。

对搜索结果排序的方法主要包括查询相似度、链接分析等。经典的相似度算法包括词袋模型、TF-IDF 算法、Word2vec 等。词袋模型通过提取文本特征建模，忽略文本结构与顺序，仅关注单词出现的情况[9]；TF-IDF 算法通过对文本特征进行提取来评估一个词在文档集中的重要性[9]；Word2Vec 是一种用于将词语转换为向量表示的工具，通过训练神经网络来学习词语之间的语义关系[10]。随着自然语言处理与深度学习技术的快速发展，预训练语言

模型也被广泛应用于长文本排序工作，DSSM 模型首次采用双塔结构对查询与文档进行语义向量编码，通过点击反馈学习相关性匹配，实现了从关键词匹配到语义匹配的初步跨越[11]；BERT 等模型进一步引入深层上下文建模能力，能够捕捉更细粒度的语言结构与推理关系，实现对候选结果的精细化重排序[12]，并已广泛应用于Google、Microsoft 等搜索引擎的精排模块；ReasonRank 采用自动化数据合成框架和两阶段训练策略（监督微调+强化学习）的先进段落重排器，实现了突破性的推理能力[13]；基于大模型的多智能体协同框架的发展，进一步提高了在复杂推理场景下的智能排序能力[14]。链接分析算法通过分析网页被其他页面所链接的数量、质量以及被链接页面的权重来评估网页重要性[15]。经典的链接分析算法包括PageRank、HITS 等，也有面向不同搜索需求的经典链接算法的各种改进算法，例如，Xing W. 根据页面的流行度分配排名分数，提出加权 PageRank 算法（Weighted PageRank Algorithm，WPR）[16]，MA Dayeh 等引入出版时间和引用时间作为新的指标对 PageRank 算法进行改进[17]。随着搜索引擎向着个性化和智能化发展，排序算法也将个性化特征、网页质量等引入评价指标。Mercy Paul Selvan 等通过考虑可达性、网页价值和用户反馈等多个因素来进行网页排名，以实现更有效的个性化搜索和排名机制[18]；Ayad Abdulrahman Saleem 等基于挖掘技术、方法学、输入参数和结果相似度等，分析不同的网页排名算法在返回与用户查询最相关的结果方面的表现[19]；Kang G. 等通过探索用户个性化特征，查询历史记录来推断潜在的用户，从而对网页进行排序[20]；Drivas I. C. 等通过大数据分析量化网站规模、网站安全状况、用户行为等影响搜索引擎的因素，并使用模糊认知映射对文化遗产机构网站进行排序[21]；Usta A. 等通过训练教育领域搜索的机器学习排名模型对教育领域垂直搜索引擎的结果进行排序[22]；以 GNNRank 为代表的研究使用图神经网络将节点属性的成对比较值构建有向图，通过图神经网络学习全局排序，在稀疏、噪声环境中表现出更强的鲁棒性、可解释性和迁移能力[23]。当前基于深度学习和预训练模型的排序算法在特定垂域搜索应用中表现良好，但是随着覆盖数据空间持续增多、涉及场景持续变化，模型需要不断更新迭代泛化，会造成持续开销。相比之下，基于链接分析的排序范式（如PageRank、HITS）等，不依赖于内容语义，仅通过结构关系建模重要性，在原理上更具有可扩展性与领域无关性，更适合作为跨数据空间搜索的基础排序框架。然而，传统链接分析需要获取整个网络的全局视图，而数据空间的数据有不出域的安全要求，搜索系统无法获取全局视图进行分析排序。此外，搜索结果排序往往依赖于搜索需求，现有搜索排序主要面向为用户找到更受欢迎网页、兴趣推荐、分析推理等需求，而在跨数据空间数据流通利用中，使用场景是一个重要的影响因素。为此，文章利用数据使用场景和语用关系构建局部链接关系视图，进行搜索结果排序。

为了在“数据不出域”“可见不可得”的数据安全要求下实现不同数据空间数据的高效发现与便捷使用，学术界和产业界提出了多种解决方案。联邦学习、多方安全计算等技术在数据提供方执行算法、采用加密机制交换计算结果，从而在不暴露原始数据的前提获取所需计算结果[24]；可搜索加密技术构建各数据源的本地安全索引，通过为用户生成搜索令牌并分发至各

数据提供方，数据提供方在本地密文中执行匹配并将加密结果返回，全程无需解密原始数据，保证了数据的隐私[25]。上述方法均适用于有限个数据提供方在提前沟通算法、模型、机制的前提下，开展数据跨域搜索与共享，但是随着接入数联网的数据空间数量变多、异构性增加，算法、模型、秘钥的管理和沟通成本指数上升。孙金烨等从理论架构层面提出了一套面向权属治理的分布式数据空间架构模型，提升数据要素市场交易的有序性与生态协同水平[26]。黄罡、罗超然等提出了一种协议化的解决方案，基于DOA将各个数据空间的数据封装成数字对象，可以通过标准的访问协议进行元数据检索和数据实体访问，同时，融合分布式账本和智能合约等技术将数据搜索、传输等过程进行可信存证，保证各数据源的行为可被验证、搜索过程可审计、结果可解释[27]。

研究以DOA为基础，采用联邦检索技术，设计了一个面向多数据空间异构注册系统新型数据搜索系统架构，并重点解决异构注册系统下的统一查询和面向数据使用场景的结果排序问题，为上层应用提供了跨数据空间数据资源搜索发现的能力。

3 跨数据空间搜索系统架构

在数据空间中，数据实体并非是孤立存在的，它们之间因为面向不同场景的使用产生了丰富关联关系，称为数据语用（data pragmatic，简称DP），即“数据在应用特定的语境中的含义”，指的是数据与应用的具体结合方式[28]。例如，在一个电商平台的数据集中，用户的购买记录数据不仅包含了商品购买信息，在数据语用层面，这些数据还可以用于推断用户的消费偏好、购买习惯等，并且这些数据的使用方式会受到隐私政策、商业目的等语境因素的影响。数据语义侧重于数据本身的定义和逻辑关系，例如数据库中某个字段代表的含义。而数据语用更关注数据在实际场景中的用途和效果。例如，在医疗数据库中，“血压值” 这个数据的语义是明确的，但它的数据语用可能包括用于诊断疾病、评估治疗效果等多种用途。数据语用重点描述了数据的实际使用情况中数据之间的关系，它提供了关于数据如何在特定上下文中被应用和理解的一种描述方式。在数联网技术中，将数据空间中的数字对象通过DOI来唯一标识，数字对象的使用关系使用统一的DPML格式来描述[26]。

研究针对各个数据空间数据不出域的需求，设计了基于DOA的跨数据空间搜索系统架构，根据数据场景化使用形成的语用关系，提出了基于语用的搜索排序算法，如图1所示，将用户从与各个数据空间注册系统建立连接、撰写面向各个数据空间查询语言等繁琐操作中解放出来，全部交由搜索系统完成。

1. 查询生成模块

查询生成模块负责接收用户输入的查询请求，支持自然语言形式输入查询关键字和关键语句，并将原始输入转化为标准化的JSON格式查询语句。

2. 查询转换模块

查询转换模块将查询请求转换成可由不同注册系统直接执行的查询语句，包括查询解析、查询校验、查询执行等过程。在查询解析阶段对标准化JSON格式查询请求进行深度解析，将其转换为抽象语法树，提取其中的关键信息，如查询条件、排序规则、过滤选项等。在查询校验阶段对解析结果进行严格校验，确保其符合目标注册系统的语法和语义要求，避免因语

法错误或语义不一致导致的查询失败。在查询执行阶段生成针对各个注册系统的具体查询语句，如SQL、Cypher等，并通过DOIP协议建立与各个注册系统的连接，发送查询请求。查询转换模块通过统一的查询解析、校验和转换机制，实现了对多个注册系统的高效访问和集成查询，大大提升了数据需求方的查询体验和系统的整体易用性。

3. 结果排序模块

结果排序模块负责对从各个注册系统返回的元数据进行综合评估和排序，以确保数据需求方获得最相关、最有价值的查询结果。主要包括结果集视图构建、查询相似度计算、语用重要性计算和权重值计算与排序阶段。结果集视图构建汇聚来自不同注册系统的检索结果及其语用关联对象集合，形成具备上下文关联性的元数据集合视图；查询相似度采用TF-IDF算法量化评估结果集中的数据与查询请求之间的文本相似度；语用重要性计算基于数字对象之间的语用权重计算结果集中每个数字对象的语用重要度；结果计算与排序阶段，计算每个数字对象的综合评分并排序。查询结果排序模块基于数据场景化使用形成的语用关系，为数据需求方提供全面且符合需求的查询结果列表，有效提升了查询效率和满意度。

4. 展示与获取模块

系统分别展示元数据检索列表、元数据详情，并支持通过元数据申请符合权限的数字对象。元数据结果已列表形式展示排序的元数据；用户选择某个元数据后，可展示元数据详细信息，并可对该元数据所属数字对象提交访问申请；数据需求方具备该数字对象的访问权限，系统会与对应的仓库系统建立DOIP链接，并调用DOIP服务的Retrieve方法发送DOIP请求，获取数字对象的具体内容并展示给数据需求方。DOIRP协议和DOIP协议内置了数据需求方进行身份认证和权限验证功能，只有通过验证的数据需求方才能获取数字对象。

4 基于数据语用的数字对象排序

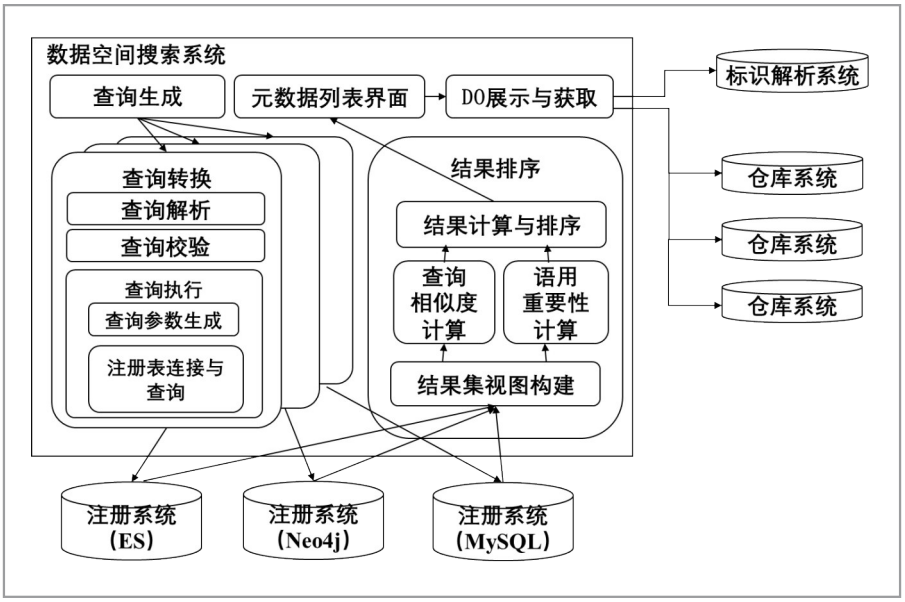


图1 跨数据空间搜索系统总体架构

算法

搜索结果排序模块接收各个数据空间注册系统返回的元数据，并基于数字对象之间的语用关系进行排序和展示，确保越精准、越相关、越有用的数字对象的排名越靠前。数字对象分散的存储在各个数据空间的注册系统和仓库系统上，不能汇聚到集中的第三方存储设施上，因此，基于集中式全局视图的链接分析的PageRank和HITS算法不再适用。研究提出了基于数据语用的数字对象排序算法DO-Rank，在各个注册系统上利用数字对象语用关系，扩展了元数据关键字检索结果，构建了元数据集合的语用关系视图；随后计算元数据的相似度和语用重要性，以此对元数据进行排序。数字对象排序模块使用的DO-Rank算法框架如图2所示，包括结果集视图构建、查询相似度计算、语用重要性计算和结果集排序四个部分。

1. 搜索结果集合视图构建

各个数据空间注册系统接收查询执行模块通过DOIP协议的Search操作发送的查询请求，在底层存储系统中基于关键字检索结果构成基础元数据集合（Basic DO Set，BDO）。仅获取命中关键字的元数据往往是不够的，例如在科研论文检索场景中，搜索某论文后通常也非常需要该论文引用的论文、使用的数据集、相关代码等数据。因此，将基于数据场景使用形成的关系抽象成数据语用，对基础元数据集合进行扩展，根据语用关系和关系深度D获取D跳语用关系链接到的元数据，称为扩展元数据集合（Extend DO Set，EDO）。BDO和EDO及数字对象之间的语用关系可以构建形成搜索结果集合视图，如图3所示为关系深度D=2时的BDO和EDO形成的元数据搜索结果集视图。

2. 查询相似度计算

算法使用经典的TF-IDF算法计算数字对象元数据与查询请求的文本相似度，计算公式如公式1所示。

$$Sim(do_k, Q) = \sum_{i=0}^{n} TF(t_i, do_k) * IDF(t_i, DOs) \#(1)$$

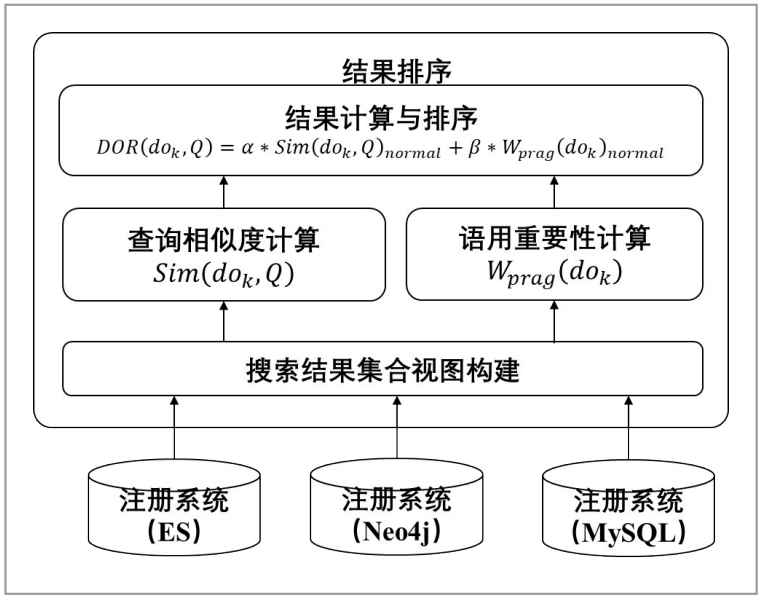


图2 DO-Rank算法框架

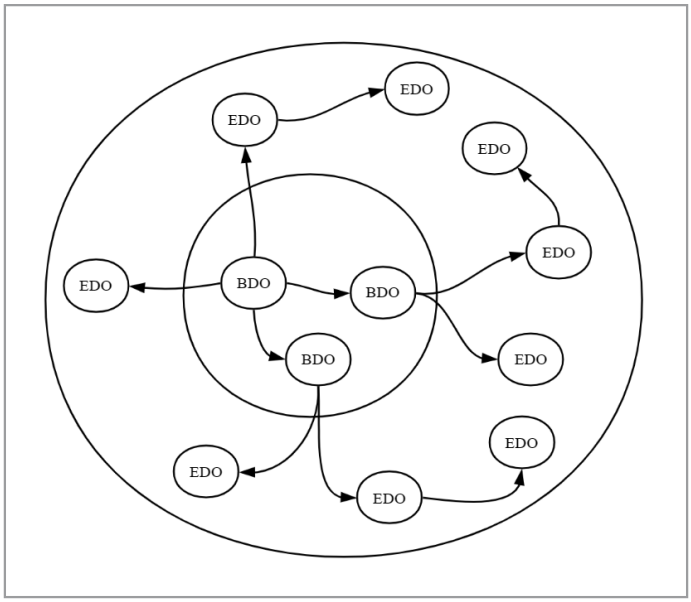


图3 元数据搜索结果集合视图(D=2)

其中 $Sim(do_k, Q)$表示编号为 k 的元数据 do_k和查询 Q之间的相似度；$TF(t_i, do_k)$表示查询词 t_i在 do_k的元数据中出现的次数；$IDF(t_i, DOs)$表示在整个元数据集合中包含查询词 t_i的个数。对计算出的查询相似度需要进行归一化处理，避免不同特征数据范围差距过大对最终排序结果的影响。对于查询相似度 $Sim(do_k, Q)$，其归一化计算方式如公式2所示。

$$Sim(do_k, Q)_{normal} = \frac{Sim(do_k, Q) - Sim_{min}}{Sim_{max} - Sim_{min}} \#(2)$$

其中 $Sim(do_k, Q)$表示归一化后的查询相似度；Sim_{min}表示所有查询相似度中的最小值；Sim_{max}表示所有查询相似度中的最大值。

3. 语用重要性计算

定义1. 数字对象语用关系，是指数据的场景化使用中数字对象之间的关系，形式化定义如下：

$$r = (do_m, do_k, t) \quad (3)$$

其中，do_m 为语用关系的起点数字对象，do_k 为语用关系的终点数字对象，$do_m, do_k \in \mathbb{I}$，$\mathbb{I}$ 表示各个数据空间中全体数字对象集合。t为语用关系描述，$t \in \mathrm{T}$，T表示各个数据空间中全体语用关系描述。

定义2. 数字对象之间的语用权重，是面向特定数据场景化使用中，数字对象之间特定语用关系的重要程度，计算公式如下：

$$prag_{mk} = \mathrm{Sim}(s, t) \quad (4)$$

其中，s为数据使用场景的自然语言描述，t为数字对象语用关系描述，$Sim(s,t)$为数据使用场景描述与数字对象语用关系描述的归一化语义相似度。该公式表示，一对数字对象之间的语用关系与当前数据使用场景越相似，语用权重越高，反之越低。

定义3. 数字对象语用重要性，是基于数字对象之间的语用关系计算出的语用重要程度。对于编号为 k 的数字对象 do_k，其语用重要性的计算方式如公式5所示。

算法1 单个数字对象的语用重要性计算

```
function calculatePrgmaticWeights(graph, DF)
    weights ← initializeWeights(graph) //初始化数字对象语用重要性
    for iteration in range(1, maxIterations + 1) do
        for do in graph.nodes do
            weight ← 0
            neighbors ← getNeighbors(do)
            for each neighbor in neighbors do
                pragmaticWeight ← getPragmaticWeight(graph, neighbor, do) //获取语用权重
                neighborWeight ← getWeight(graph, neighbor)//获取相邻数字对象语用
                totalPragmaticWeight ← getTotalPragmticWeight(graph, neighb //获取邻居数字对象指向其他数字对象的语用权重之和
                weight ← weight + (pragmaticWeight × neighborWeight) /totalPragmaticWeight
            end for
            weight ←(1 − DF)/graph.size + DF × weight
            setWeight(graph, do, weight)
        end for
    end for
    return weights
end function
```

$$W_{prag}(do_k)=\frac{(1-DF)}{N}+DF*\sum_{do_m\in M(do_k)}\frac{prag_{mk}*W_{prag}(do_m)}{P(do_m)}\#(5)$$

其中，$W_{prag}(do_k)$表示数字对象 do_k 的语用重要性；DF（Damping Factor）表示阻尼因子，用于控制数字对象之间根据语用关系跳转的概率；N表示当前有向图中全部数字对象的个数；$M(do_k)$表示语用关系直接指向 do_k 的数字对象的集合；$prag_{mk}$表示数字对象 do_m 和 do_k 之间的语用权重；$P(do_m)$表示 do_m 指向其它数字对象的语用权重之和。依据使用场景、数据类型、搜索需求不同数字对象的语用重要性不同。数字对象的语用重要性越高，越可能符合用户搜索需求，即在搜索结果列表中排序应该越靠前。计算语用重要性的伪代码如算法1所示。

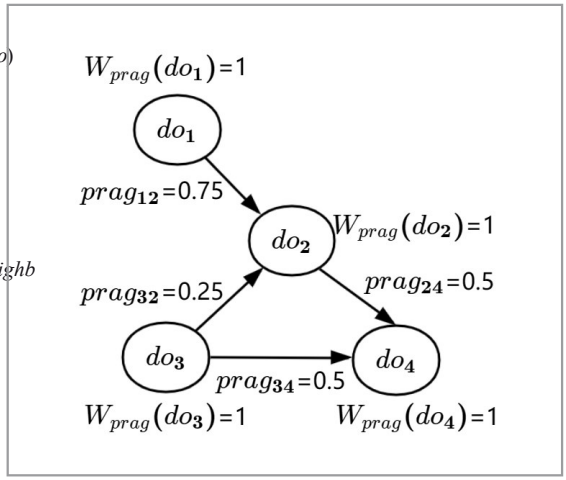


图4 DO语用重要性计算示例

如图4所示，初始状态下 do_1、do_2、do_3、do_4 的语用重要性均设置为1，四个DO之间的语用权重分别为：$prag_{12}=0.75$、$prag_{24}=0.5$、$prag_{32}=0.25$，阻尼因子 $DF=0.85$，则经过一轮迭代后，do_2 的语用重要性 $W_{prag}(do_2)$ 计算结果如下：

$$W_{prag}(do_2)=\frac{(1-DF)}{N}+DF*\left(\frac{prag_{12}*W_{prag}(do_1)}{P(do_1)}+\frac{prag_{32}*W_{prag}(do_3)}{P(do_3)}\right)$$

$$=\frac{(1-0.85)}{4}+0.85\times\left(\frac{0.75}{0.75}\times1+\frac{0.25}{0.25+0.5}\times1\right)=1.17\#(6)$$

数字对象的语用重要性也需要进行归一化处理，如公示7所示。

$$W_{prag}(do_k)_{normal} = \frac{W_{prag}(do_k) - W_{min}}{W_{max} - W_{min}} \#(7)$$

其中 $W_{prag}(do_k)_{normal}$ 表示 do_k 的归一化语用重要性；W_{min} 表示所有数字对象的语用重要性的最小值；W_{max} 表示所有数字对象的语用重要性最大值。

4. DOR 值计算与排序

一个数字对象元数据与查询请求相似度越高、数字对象语用重要性越高，在最终的排序结果中就越靠前，使用 DOR 值进行计量。对于一个查询 Q，编号为 k 的元数据 do_k 的 DOR 值的计算方式如公式 8 所示。

$$DOR(do_k, Q) = \alpha * Sim(do_k, Q)_{normal} + \beta * W_{prag}(do_k)_{normal} \#(8)$$

其中 $Sim(do_k, Q)_{\text{normal}}$ 表示归一化处理后的元数据 do_k 与查询 Q 之间的查询相似度；$W_{prag}(do_k)_{\text{normal}}$ 表示归一化处理后的元数据 do_k 的语用重要性；α 和 β 表示权重参数，用于调整查询相似度和语用重要性的权重。

5 跨数据空间搜索系统测试

实验测试了搜索系统排序算法，使用开源医学数据集，具体包括 1）诊疗指南数据，由 OpenKG 的公开的结构化临床诊疗指南数据集，涵盖 7,7000 余条覆盖多学科的临床路径与诊疗标准，具备明确的疾病分类、诊断依据与治疗推荐；2）媒体健康信息，通过爬虫获取主流媒体健康专栏的公众科普内容数据集，包含约 80,000 篇非结构化健康文章；3）教学病例数据，通过大模型模拟的多家三甲教学医院联合贡献的教学病历数据集，包含约 18,000 例脱敏后的真实病例，每例均标注主诉、检查、诊断过程及所依据的指南条目。

实验平台使用 4 台阿里云 4 核 16GB 服务器 SSD 云盘，其中 3 台模拟安装 3 个数据空间的注册系统，1 台用于部署运行数据空间搜索系统。将数据封装成数字对象后，临床诊疗指南数字对象的元数据结构化程度高、支持精确元数据查询，使用 MySQL 关系型数据库存储；公众健康科普数字对象的元数据非结构化文本密集、强调全文检索性能，使用 Elasticsearch 数据库存储；教学病历数字对象的元数据蕴含丰富语义关系的教学病历及其与指南、疾病、药物之间的映射关系，使用 Neo4j 图数据库存储。

实验对比了以下三种情况：①分别与各个数据空间注册系统建立连接使用 DOIP.Search 服务进行查询，并根据查询结果返回的先后顺序排序，简称“DO-Search”；②采用集中汇聚式搜索架构，并使用基于访问链接的加权 PageRank 算法(Weighted PageRank based on Visits of Links, W-PageRank)[17] 进行结果排序；③基于分布式搜索架构并使用 DO-Rank 算法进行结果排序。

（1）算法查全查准效果分析

由于人工标注完整检索结果相关集合在跨多个领域数据空间搜索场景下不可行，实验基于领域经验设定 R = 30，使用不同检索词进行重复试验，分别计算了前 k 个结果（top-k）的精确率(Precision)、召回率(Recall)、F1 值(F1-Score)和归一化折损累计增益(normalized Discounted Cumulative Gain, nDCG)的平均值，结果如图 5-8 所示。

精确率、召回率、F1 值用于衡量搜索系统的查准率、查全率，随着 k 值增大，搜索系统的精确率降低，表明随着搜索结果总数的增多，命中结果占所有搜索结果

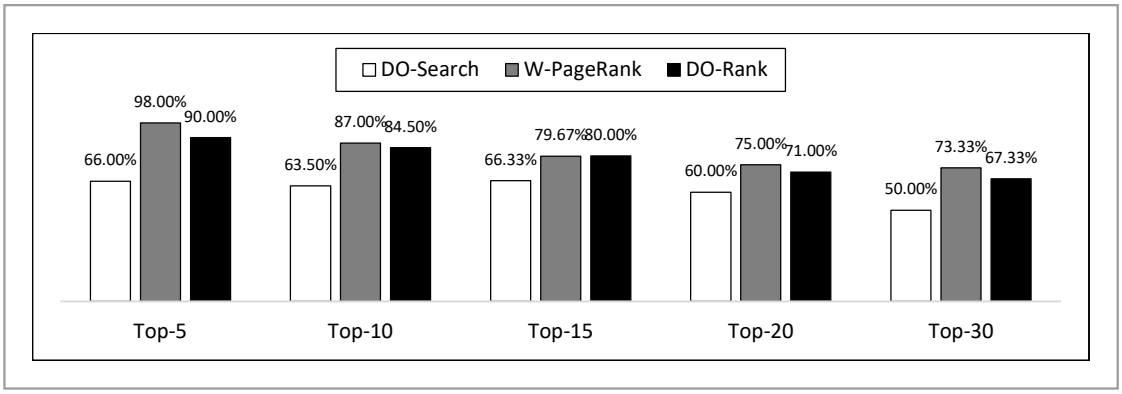


图5 各算法精确率比较

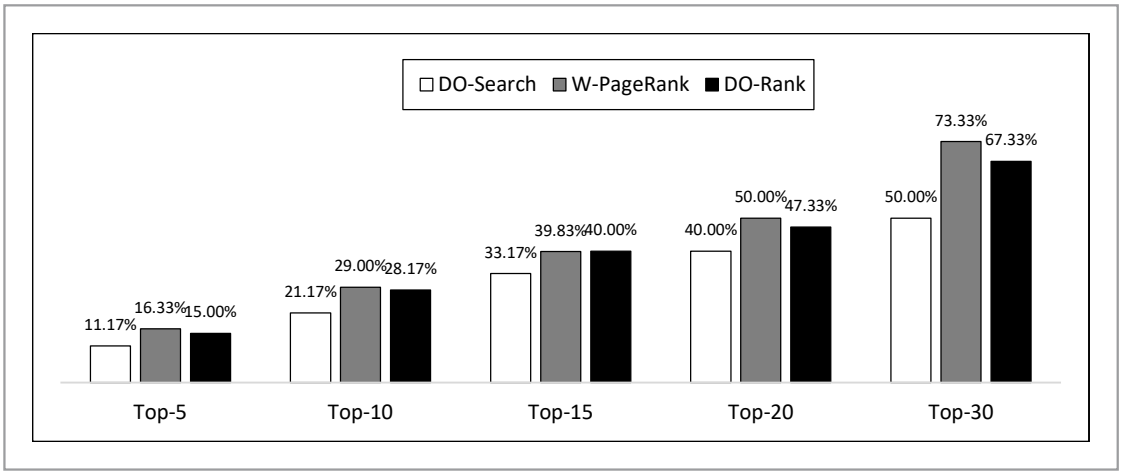


图6 各算法召回率比较

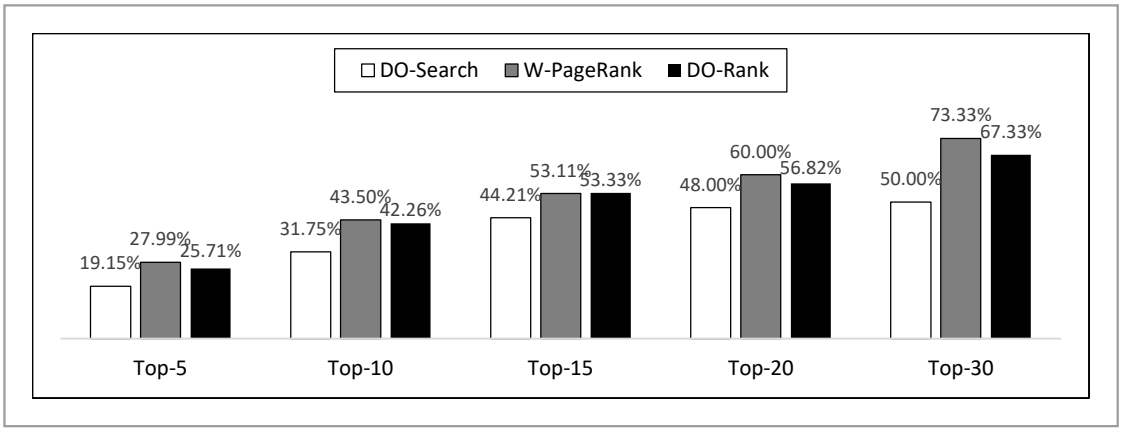


图7 各算法F1值比较

的占比降低；召回率、F1值升高，是因为 随着搜索结果数量增加，命中结果数量增

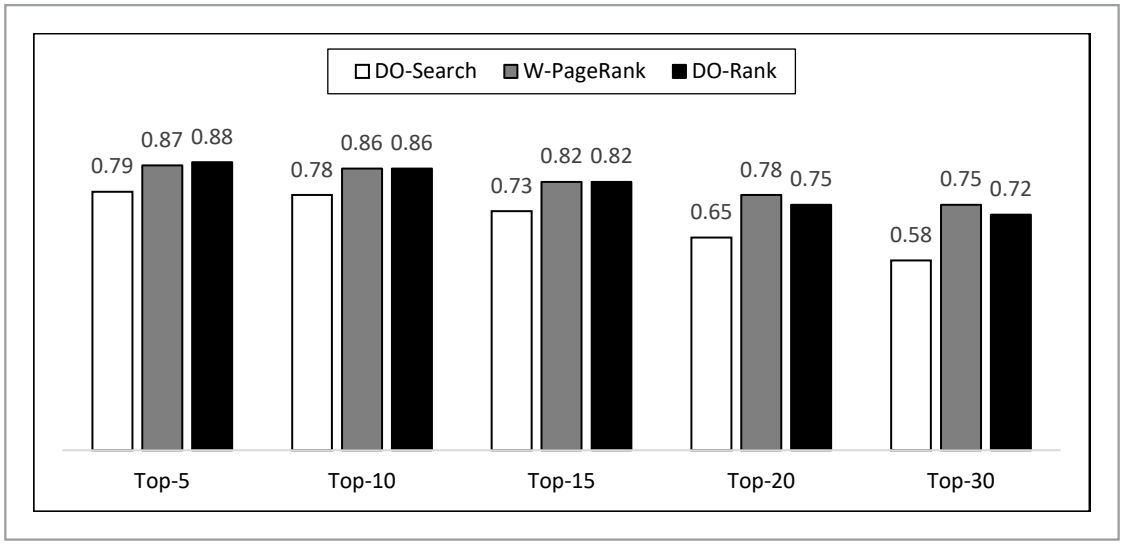


图8 各算法nDCG值比较

加，因此在全部数据集中的占比也增加。W-PageRank算法的查准率、查全率高于DO-Search和DO-Rank算法，这是因为W-PageRank算法采用集中汇聚式搜索架构将数据全量汇聚后排序，得到的是全局最优解，而DO-Search和DO-Rank采用分布式搜索架构，不集中汇聚各注册系统的全量数据，无法获得全量最优解，在top-20时，DO-Rank算法相比于W-PageRank算法精确率低5.33%、召回率低5.34%、F1值低5.33%，相比于DO-Search方法精确率高18.3%、召回率高18.32%、F1值高18.38%。结果表明，DO-Rank算法通过数据语用关系扩充了在各个注册系统上的基础搜索集合再进行排序，较直接使用注册系统查询服务的方法有较大的改进，且查全查准效果接近汇聚全量数据的W-PageRank算法。

结果显示，随着数据规模增加，W-PageRank算法需要遍历分析的全局数据量显著增加，执行时长明显增加，DO-Rank算法仅获取关键字匹配后数据的有限跳数数据，时间不会随着数据量的增加有大幅增长，因此，随着数据量的增长，DO-Rank算法的时间优势越来越明显。

nDCG用于衡量搜索系统的用户满意度，随着k值增大，搜索系统的nDCG值降低，W-PageRank算法和DO-Rank算法均优于DO-Search方法，DO-Rank算法的nDGC比DO-Search提高了10%。W-PageRank和DO-Rank的nDCG值差别不大，因为二者均基于语用权重进行排序，可以将更符合用户使用场景的数字对象排在前面。

（2）算法时间性能分析

在3个数据空间注册系统中分别按比例进行数据抽样，抽取1/10到10/10共10个梯度比例规模的数据，对算法进行重复试验，计算算法运行的平均响应时间，如图9所示。

综上所述，相比于传统的注册系统查询服务，DO-Rank算法引入了数字对象之间的语用关系，使用户获得了更好的搜索体验。同时也避免了传统基于链接的排序算法需要汇聚和全局数据关系视图的限制，

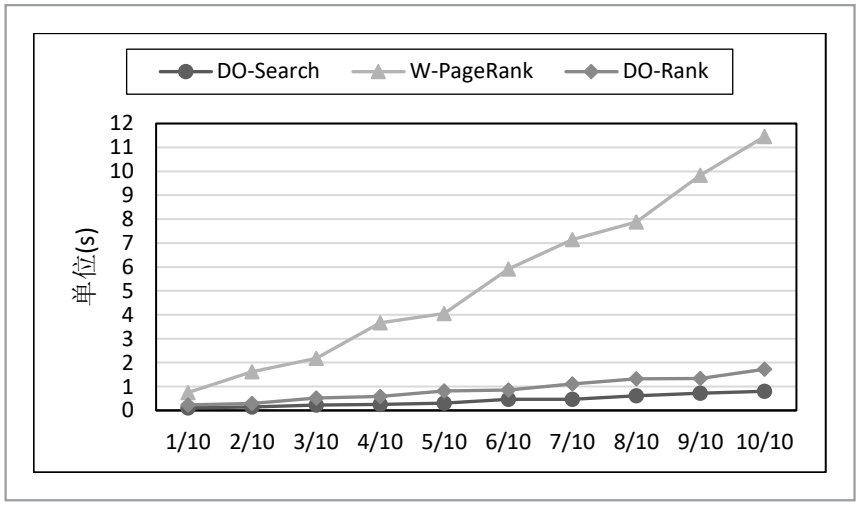


图9 各算法执行时长比较

使其能够更好的适用于跨数据空间的分布式数据搜索。

6 结束语

研究针对分散于数据空间各主体中数据资源的高效发现、便捷使用、有效流通的关键需求，提出了面向数据场景化使用的数据空间搜索系统架构，基于数据场景化使用形成的语用关系的设计了搜索排序算法DO-Rank，并进行了实验验证。研究为实现“数据可见不出域”前提下的数据跨空间高效、可信访问提供了关键技术支撑。

随着大语言模型技术的发展和广泛运用，下一步计划结合大语言模型与检索增强生成技术，增强用户查询理解与语义匹配能力，进一步提升系统的智能化水平与服务适应性。此外，还将进一步研究面向基于大模型的智能应用场景的跨数据空间搜索服务改进，如研究基于语义标签与场景特征的数据关联机制，支持智能应用按需获取高质量数据资源。

参考文献：

[1] Kahn R, Wilensky R. A framework for distributed digital object services[J]. International Journal on Digital Libraries, 2006, 6(2):115-123.

[2] CNRI, RFC 3650: Handle System Overview [EB/OJ]. https://www.ietf.org/rfc/rfc3650.txt.

[3] 黄罡.数联网:数字空间基础设施[J].中国计算机学会通讯,2021,17(12):60-62.
HUANG G. Internet of Data: The Infrastructure of Dataspace [J]. Communications of the China Computer Federation, 2021,17(12):60-62.

[4] 刘凡平.大数据搜索引擎原理分析及编程实现[M].北京:电子工业出版社,2016.
LIU F. Principles and Programming Implementation of Big Data Search Engines [M]. Beijing: Publishing House of Electronics Industry, 2016.

[5] 李晓明,闫宏飞,王继民. 搜索引擎——原理、技术与系统[M](第二版). 北京: 科学出版社, 2012.
LI X, YAN H, WANG J. Search Engines: Principles, Techniques, and Systems [M] (2nd ed.). Beijing: Corpo-

rate Profile of China Science Publishing & Media Ltd., 2012

[6] 钱立兵,季振洲.分布式搜索引擎的模型综述[J].智能计算机与应用,2015,5(05):133-116.
QIAN L, JI Z. A Survey on Models of Distributed Search Engines [J]. Intelligent Computer and Applications, 2015, 5 (05) :133-116.

[7] 张俭恭,陈定权,吴振新.关于搜索引擎与元搜索引擎的讨论[J].现代图书情报技术,2002,(02):36-38.
ZHANG J, CHEN D, WU Z. A Discussion on Search Engines and Meta-Search Engines [J]. New Technology of Library and Information Service, 2002, (02): 36-38.

[8] GARBA A, WU S, KHALID S. Federated search techniques: an overview of the trends and state of the art[J]. Knowledge and Information Systems, 2023, 65(12): 5065-5095.

[9] 吴平博,陈群秀,马亮.基于特征串的大规模中文网页快速去重算法研究[J].中文信息学报,2003,17(2):28-35.
WU P, CHEN Q, MA L. The Study on Large Scale Duplicated Web Pages of Chinese Fast Deletion Algorithm Based on String of Feature Code[J]. Journal of Chinese Information Processing, 2003, 17 (2):28-35.

[10] Robertson S. E.; Walker S. Some simple effective approximations to the 2 -poisson model for probabilisticweighted retrieval. In Proceedings of the International ACM Sigir Conference on Research and Developmentin Information Retrieval SIGIR'94, Dublin, Ireland, 3-6 July 1994; pp. 232-241.

[11] Po-Sen Huang, Xiaodong He, Jianfeng Gao, et al. Learning deep structured semantic models for web search using clickthrough data. CIKM '13: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, 27 October 2013, P. 2333-2338.

[12] Rodrigo Nogueira, Kyunghyun Cho. Passage Re-ranking with BERT. arXiv: 1901.04085. https://arxiv. org/abs/1901.04085

[13] Wenhan Liu, Xinyu Ma, Weiwei Sun, et al. ReasonRank: Empowering Passage Ranking with Strong Reasoning Ability. arXiv:2508.07050.

[14] YongqiFan, KuiXue, ZelinLi, et al. An LLM-based Framework for Biomedical Terminology Normalization in Social Media via Multi-Agent Collaboration[C]. Proceedings of the 31st International Conferenceon Computational Linguistics, 2025. pages10712-10726.

[15] Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine[J]. Computer Networks and ISDN Systems, 1998, 30(1-7):107-117.

[16] Xing W, Ghorbani A A. Weighted PageRank algorithm[C]//Conference on Communication Networks & Services Research. IEEE, 2004. DOI: 10.1109/DNSR.2004.1344743.

[17] Dayeh M A, Sartawi B, Salah S. A Bias-Free Time-Aware PageRank Algorithm for Paper Ranking in Dynamic Citation Networks[J]. 智能信息管理(英文), 2022(002):014.

[18] Selvan M P, Shekar A C, Babu D R, et al. Efficient ranking based on web page importance and personalized search[C]// 2015 International Conference on Communications and Signal Processing (ICCSP). IEEE, 2015. DOI: 10.1109/ICCSP.2015.7322671.

[19] Abdulrahman A. Web Pages Ranking Algorithms: A Survey[J]. 2021.

[20] Kang G, Liu J, Tang M, et al. An Ef-

fective Web Service Ranking Method via Exploring User Behavior[J]. IEEE, 2015 (4).DOI:10.1109/tnsm.2015.2499265.

[21] Drivas I C . Big Data Analytics for Search Engine Optimization[J]. Big Data and Cognitive Computing, 2020, 4(2). DOI:10.3390/bdcc4020005.

[22] Usta A , Altingovde I S , Ozcan R ,et al. Learning to Rank for Educational Search Engines[J].IEEE Transactions on Learning Technologies, 2021, 14(2): 211-225. DOI:10.1109/TLT.2021.3075196.

[23] Yixuan He, Quan Gan, DavidWipf, et al. GNNRank: Learning Global Rankings from Pairwise Comparisons via Directed Graph Neural Networks. ICML: 8581-8612.

[24] 窦路遥, 魏凤, 邓阿妹,等. 联邦学习赋能数智化科技情报工作中的多源数据融合[J]. 情报杂志, 2025,44 (05): 191-198.
DOU L, WEI F, DENG A. Empowering Intelligent Technology Information Work with Federated Learning for Multi-Source Data Integration[J]. Journal of Intelligence, 2025, 44 (05): 191-198

[25] 兰亚杰, 马自强, 陈嘉莉,等. 基于区块链的可搜索属性加密技术应用综述[J]. 计算机科学, 2024 ,51(6A): 882-895.
LAN Y, MA Z, CHEN J, et al. Survey on Application of Searchable Attribute-based Encryption Technology Based on Blockchain[J]. Computer Science, 2024 , 51(6A): 882-895.

[26] 孙金烨, 郭树行. 面向权属治理的分布式数据空间架构模型研究[J]. 大数据, 2025, 11(2): 140-151.
SUN J, GUO S. Research on Distributed Data Spatial Architecture Model for Ownership Governance[J]. Big Data Research, 2025, 11(2): 140-151.

[27] 罗超然,马郓,景翔等.数据空间基础设施的技术挑战及数联网解决方案[J].大数据,2023,9(02):110-121.
LUO C, MA Y, JING X, et al. Internet of Data: a Solution for Dataspace Infrastructure and its Technical Challenges [J]. Big Data Research, 2023, 9(02): 110-121.

[28] 黄小龙,杨婧如,柳熠,等. ReproLink: 面向可复现性的科研数据管理系统[J]. 软件学报, 2025,36(12):5801－5820.
HUANG X, YANG J, LIU Y, et al. ReproLink: Reproducibility-oriented Research Data Management System[J]. Journal of Software, 2025,36(12):5801－5820.

[29] 蔡华谦,刘逸豪,关天鹏,等.DPML: 一种面向科学数据语用的标记语言[J]. 数据与计算发展前沿, 2024, 6(4): 46-58.
CAI H, LIU Y, GUAN T, et al. DPML: A Markup Language for Scientific Data Pragmatics[J]. Frontiers of Data&Computing, 2024, 6(4): 46-58.

[30] 国家卫生健康委员会. WS/T305-2023. 卫生健康信息数据集元数据标准[S]. 2023.
National Health Commission of the People's Republic of China. WS/T305-2023. Metadata specification of health information dataset[S]. 2023.

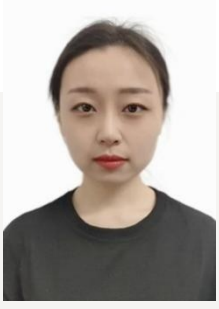
作者简介

姜海鸥（1987–），女，博士研究生，北京大数据先进技术研究院，副研究员，主要研究方向为云计算、大数据、软件工程等。

胡子逸（1997–），男，北京大学，硕士研究生，主要研究方向为软件工程、数据智能。

刘冉(1996–)，男，硕士研究生，北京大学软件与微电子学院，主要研究方向为数据智能。

罗超然（1995–），男，博士研究生，北京大数据先进技术研究院副研究员，主要研究方向为数据空间、数联网、系统软件等。



杨婧如（1995–），女，博士研究生，北京大数据先进技术研究院，副研究员，主要研究方向为大数据治理技术与系统

收稿日期: XXXX–XX–XX

通信作者: 姜海鸥，jiangho@aibd.ac.cn

基金项目: 数据空间技术与系统全国重点实验室基金项目（No. QZJZ2024KJ001–04–SJKJSS–KTBG）

Foundation Items: National Key Laboratory of Data Space Technology and Systems (No. QZJZ2024KJ001–04–SJKJSS–KTBG)