

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

随着国家数据要素相关政策的推进, 在分散于各行业、各领域的数据空间之间, 实现数据资源高效发现和便捷使用, 已成为数据要素有效流通的关键。设计了一种面向数据场景化使用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, 构建了基于数字对象架构(digital object architecture, DOA)的分布式搜索框架; 基于数据场景化使用所形成的数据语用关系, 提出融合查询相似度与数据语用关系的数字对象排序算法——DO-Rank, 为上层应用提供了跨数据空间数据资源搜索发现的能力。实验结果表明, 该跨空间搜索系统在不汇聚数据的情况下, 其精确率、召回率、归一化折损累计增益(normalized discounted cumulative gain, nDCG)等指标接近采用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 scenario-based data usage

Jiang Haiou¹, Hu Ziyi², Liu Ran², Luo Chaoran¹, Yang Jingru¹

1. National Key Laboratory of Data Space Technology and System, Beijing 100195, China

2. School of Software and Microelectronics, Peking University, Beijing 102600, China

Abstract

With the advancement of national policies on data factors, efficient discovery and convenient utilization of data resources across data spaces scattered in various industries and fields have become a key for the effective circulation of data elements. A data space search system oriented to scenario-based data usage was designed and implemented. In response to the security requirements that data shall not leave its domain or be aggregated in each, a distributed search framework based on the digital object architecture (DOA) was constructed. Based on the pragmatic data relationships abstracted from the scenario-based data usage, a digital object ranking algorithm, named DO-Rank was proposed. It considered both query similarity and data pragmatic relationships. It provided upper-layer applications with the capability to search and discover data resources across different data spaces. Experiments results demonstrated that without data aggregation, the system achieved performance indicators including precision, recall and normalized discounted cumulative gain (nDCG) comparable to those of centralized data retrieval methods.

Key words
data space, data circulation, digital object architecture, search system, search ranking

0 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。国家数据局于2024年11月21日印发的《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，构建全国一体化数据市场提供有力支撑。”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、区域数据空间相继涌现，逐步形成覆盖多领域、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实现流通利用，但如何在“数据不出域”“可见不可得”等数据安全要求下，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）是由互联网奠基人、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]。该架构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）。其中，标识是数字对象的身份ID，

全局唯一且持久，不因DO的所有者、存储位置、访问方式的改变而改变；元数据表示与数字对象及其应用、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。在实现机制上，DOA中的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。同时，DOA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）分别规范数字对象标识的管理与解析^[2]、数字对象的检索与访问等操作。DOA通过统一的形式将互联网中的数据资源抽象和建模成DO，进而屏蔽了数据资源的异构性；通过标准协议进行数据的创建、获取、更新和删除等操作，实现了互联网上异构数据的互联互通^[3]。在DOA框架下，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。然而，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。这种方式存在两大问题：其一，由于各个数据提供方注册系统的数据库、软件实现机制各不相同，数据需求方需要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，以致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,进而影响数据发现效率。例如,医生在有权限访问电子病历系统、医保用药系统、国家卫生健康委员会下属各诊疗指导系统的情况下,若想查询某疾病信息,需要分别向3个系统发送查询请求,在为病人开药的数据使用场景下,电子病历和医保用药数据较为重要;而在研判当前流行病风险与治疗方案这类特定场景下,电子病历和诊疗指导数据则较为重要。现有数联网技术的实现方案仅根据各个系统的运行速度和网络速度将查询结果直接返回,导致医生需要花费大量时间进行人工筛选。针对上述问题,本文围绕跨空间数据搜索与发现展开研究,主要研究内容如下。

(1) 提出基于 DOA 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烦琐操作中解放出来,满足其对数据空间内数据的可搜索、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“数据不出域”“可见不可得”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具有较高的查准率和用户满意度。

(3) 使用不同领域数据集,并在模拟异构注册系统的环境下进行实验,分析了本文所提系统在跨空间数据发现方面的查全率、查准率和用户满意度。

1 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,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于互联网中的网页内容,网络爬虫抓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,再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。传统搜索引擎需要将整个网络的数据爬取下来,而数据空间中的数据有“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,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 和 Shukla 等^[7]的研究表明,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。元搜索引擎旨在扩大搜索数据来源的覆盖范围,但其本质上仍基于已有同质化的 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。联邦检索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,并将检索结果以直观且统一的形式呈现给用户^[8]。目前,有关联邦检索的研究主要聚焦于数据源描述、数据源选择和搜索结果合并这3个方面。本文提出基于联邦搜索技术的搜索系统,并重点解决面向数据使用场景的搜索结果合并问题。

搜索结果常用的排序方法主要有查询相似度、链接分析两类。其中,经典的查询相似度算法包括词袋模型、TF-IDF、

Word2Vec 等。词袋模型以文本特征提取为基础完成建模, 仅关注单词出现的情况, 忽略文本结构与顺序^[9]; TF-IDF 算法通过对文本特征进行提取来评估单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 模型率先采用双塔网络结构对查询与文档进行语义向量编码, 结合用户点击反馈学习语义相关性匹配规则, 实现了检索方式从关键词匹配向语义匹配的初步跨越^[11]; BERT 等预训练模型具备深层次上下文语义建模能力, 能够捕捉细粒度语言结构与逻辑关系, 实现对候选结果的精细化重排序^[12], 目前已广泛应用于 Google、Microsoft 等搜索引擎的结果精排模块; ReasonRank 基于自动化数据合成框架, 结合两阶段训练策略 (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]; 基于大语言模型的多智能体协同框架, 进一步提升了复杂推理场景下的智能排序能力^[14]。链接分析算法主要依据网页被其他页面链接的数量、质量以及被链接页面的权重来评估网页重要性^[15], 代表性算法包括 PageRank、HITS 等, 也有面向不同搜索需求的改进算法, 例如, Xing 等^[16]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); Dayeh 等^[17]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。随着搜索引擎向个性化和智能化发展, 排序算法也将用户个性化特征、网页质量等要素纳入排序评价指标。Selvan 等^[18]综合网页可达性、网页价值和用户反馈等多个因素进行网页

排名, 以探索更有效的个性化搜索和排名机制; Abdulrahman 等^[19]从挖掘技术、方法学、输入参数和结果相似度等维度, 对比不同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; Kang 等^[20]通过探索用户个性化特征, 并查询历史记录来推断潜在用户, 从而对网页进行排序; Drivas 等^[21]借助大数据量化分析网站规模、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; Usta 等^[22]构建教育领域搜索的机器学习排名模型, 实现了对教育领域垂直搜索引擎的结果排序; 以 GNNRank 为代表的排序方法借助图神经网络, 将节点属性的成对比较关系构建为有向图, 通过图神经网络学习全局排序, 在数据稀疏、存在噪声的复杂场景中, 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直搜索应用中表现良好, 但随着覆盖数据空间持续增多、应用场景不断变化, 模型需要不断更新迭代以提升泛化能力, 这将会带来持续的开销。相比之下, 基于链接分析的排序范式 (如 PageRank、HITS) 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, 因此, 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方可完成运算, 而数据空间中的数据普遍遵循“数据不出域”“可见不可得”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场景中, 使用场景是一个重要的影

响因素。为此，本文依托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允许在数据提供方本地执行算法，并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下获取所需计算结果^[24]；可搜索加密技术为各数据源构建本地安全索引，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、密钥的管理和沟通成本呈指数级上升。孙金桦等^[26]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，以提升数据要素市场交易的有序性与生态协同水平。黄罡等^[27]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，通过标准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验证、搜索过程可审计、结果可解释。

本文以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的新型数据搜索系统架构，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。

2 跨数据空间搜索系统架构

在数据空间中，数据实体并非孤立存在，它们之间因面向不同场景的使用而产生了丰富的关联关系，这种关系称为数据语用（data pragmatic，DP），即“数据在应用特定的语境中的含义”，也指数据与应用的具体结合方式^[28]。例如，在电商平台的数据集中，用户的购买记录数据不仅包含商品购买信息，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，如数据库中某个字段代表的含义，而数据语用则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，但它的数字语用可能包括诊断疾病、评估治疗效果等多种用途。数据语用重点描述了数据实际使用过程中所体现的关联关系，提供了一种关于数据如何在特定上下文中的应用和理解的方式。在数联网技术中，数据空间中的数字对象通过数字对象标识符（digital object identifier，DOI）进行唯一标识，数字对象的使用关系则采用统一的数据语用标记语言（data pragmatics markup language，DPML）格式进行描述^[29]。

针对各个数据空间“数据不出域”的需求，本文设计了基于DOA的跨数据空间搜索系统架构，其总体架构如图1所示。同时，根据数据场景化使用形成的语用关系，提出了基于数据语用的搜索排序算法。该系统架构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

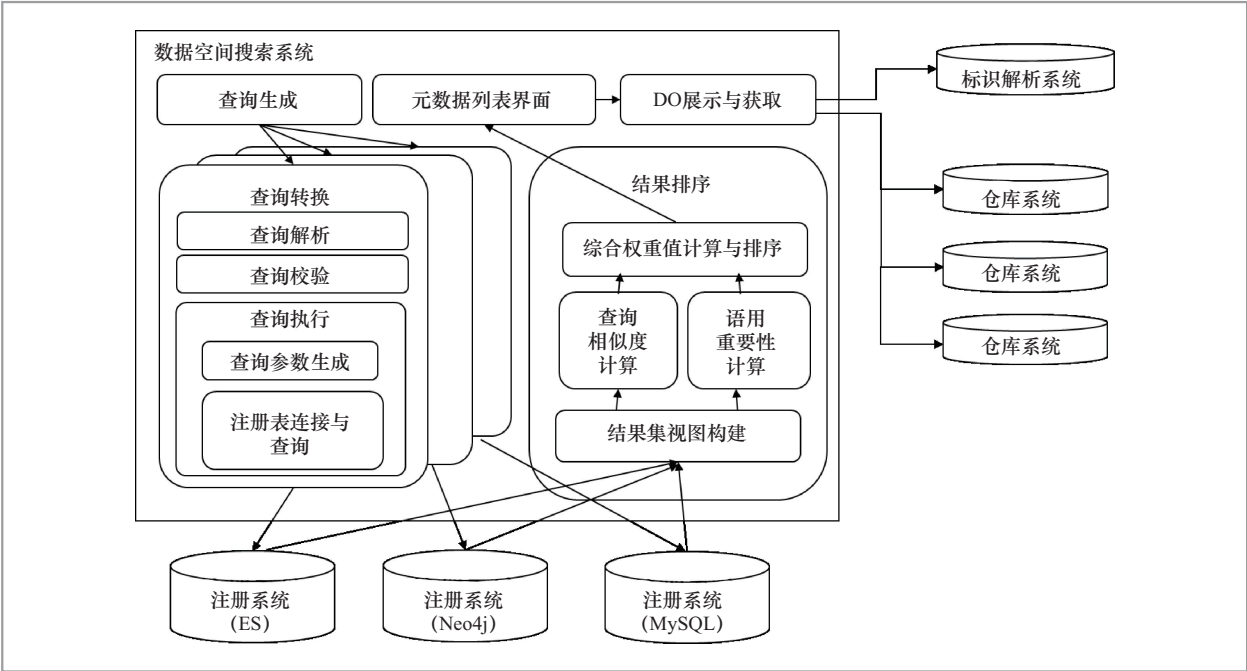


图1 跨数据空间搜索系统总体架构

询语言等烦琐操作中解放出来，全部交由搜索系统完成。

2.1 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，支持以自然语言形式输入查询关键字和关键词句，并将原始输入转化为标准化的JSON格式查询语句。

2.2 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，主要对标准化JSON格式查询请求进行深度解析，将其转换为抽象语法树，以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而导致查询失败。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，大幅提升了数据需求方的查询体验和系统的整体易用性。

2.3 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。该模块主要包括结果集视图构建、查询相似度计算、语用重要性计算和综合权重值计算与排序4个阶段。在结果集视

图构建阶段，系统汇聚来自不同注册系统的检索结果及其语用关联对象，形成具备上下文关联性的元数据集合视图；在查询相似度计算阶段，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；在语用重要性计算阶段，基于数字对象之间的语用权重计算结果，计算每个数字对象在结果集中的语用重要度；在综合权重值计算与排序阶段，计算每个数字对象的综合评分并进行排序。查询结果排序模块基于数据场景化使用所形成的语用关系，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用户满意度。

2.4 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。元数据结果以列表形式展示排序后的元数据；用户选择某个元数据后，系统可展示其详细信息，并允许用户对该元数据所属的数字对象提交访问申请。如果数据需求方具备该数字对象的访问权限，系统将与对应的仓库系统建立 DOIP 连接，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得数字对象。

3 基于数据语用的数字对象排序算法

搜索结果排序模块接收来自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，不能汇聚到第三方存储设施上，因此，基于集中式全局视图链接分析的 PageRank 和 HITS 算法不再适用。为此，本文提出了基于数据语用的数字对象排序算法 DO-Rank。该算法在各个注册系统上利用数字对象语用关系，扩展元数据关键字检索结果，构建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。数字对象排序模块采用的 DO-Rank 算法框架如图 2 所示，包括搜索结果集合视图构建、查询相似度计算、语用重要性计算和综合权重计算与排序 4 个部分。

3.1 搜索结果集合视图构建

各个数据空间的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，并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)。然而，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，搜索某篇论文后，用户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，本文基于数据场景使用形成的关系抽象为数据语用，并据此对 BDO 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、EDO 以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。关系深度 $D=2$ 时，BDO 和 EDO 形成的元数据搜索结果集视图如图 3 所示。

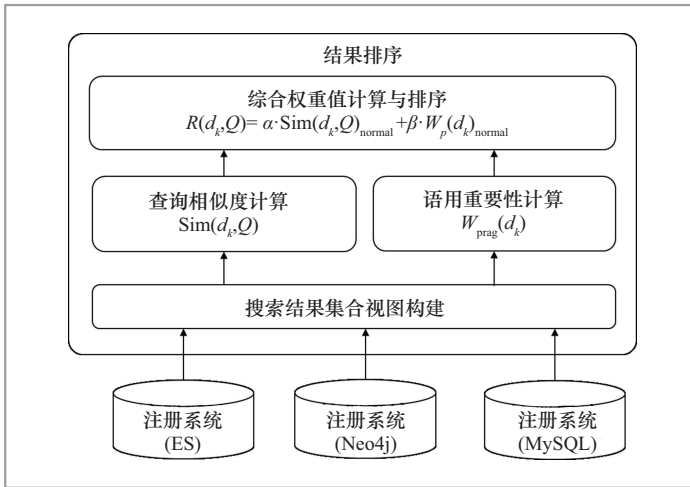


图2 DO-Rank算法框架

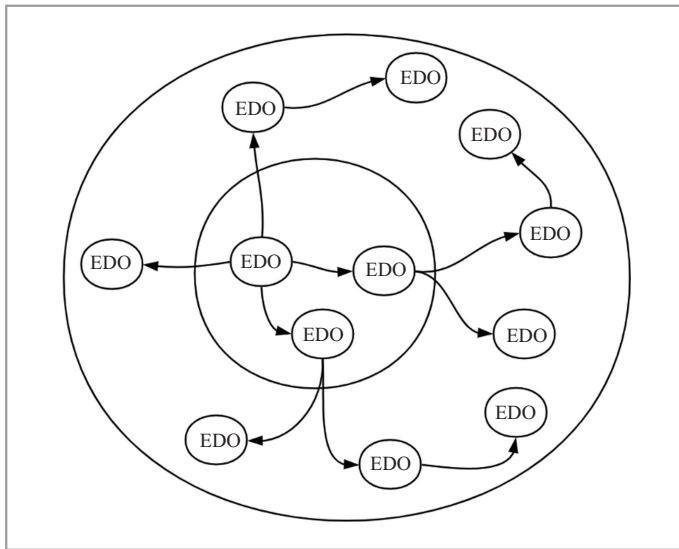


图3 元数据搜索结果集合视图(D=2)

3.2 查询相似度计算

DO-Rank 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, 可表示为:

$$\text{Sim}(d_k, Q) = \sum_{i=0}^n \text{TF}(t_i, d_k) \cdot \text{IDF}(t_i) \quad (1)$$

其中, $\text{Sim}(d_k, Q)$ 表示编号为 k 的元数据 d_k 和查询 Q 之间的相似度; $\text{TF}(t_i, d_k)$ 表示查

询词 t_i 在元数据 d_k 中出现的次数; $\text{IDF}(t_i)$ 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进行归一化处理, 以避免不同特征数据范围差距过大而对最终排序结果产生影响。对于查询相似度 $\text{Sim}(d_k, Q)$, 其归一化后的数学表达式如下:

$$\text{Sim}(d_k, Q)_{\text{normal}} = \frac{\text{Sim}(d_k, Q) - S_{\min}}{S_{\max} - S_{\min}} \quad (2)$$

其中, $\text{Sim}(d_k, Q)_{\text{normal}}$ 表示归一化后的查询相似度, $S_{\min}$ 表示所有查询相似度中的最小值, $S_{\max}$ 表示所有查询相似度中的最大值。

3.3 语用重要性计算

(1) 数字对象语用关系, 指数据场景化使用中数字对象之间的关系, 可形式化定义为:

$$r = (d_{\text{start}}, d_{\text{end}}, t) \quad (3)$$

其中, d_{start} 表示语用关系的起点数字对象, d_{end} 表示语用关系的终点数字对象, $d_{\text{start}}, d_{\text{end}} \in I$, I 表示各个数据空间中全体数字对象集合; t 表示语用关系描述, $t \in T$, T 表示各个数据空间中全体语用关系描述。

(2) 数字对象之间的语用权重是面向特定数据场景化场景, 数字对象之间特定语用关系的重要程度, 可表示为:

$$p = \text{Sim}(s, t) \quad (4)$$

其中, p 表示数字对象 d_{start} 和 d_{end} 之间的语用权重, s 表示数据使用场景的自然语言描述, t 表示数字对象语用关系描述, $\text{Sim}(s, t)$ 表示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。

式(4)表示,一组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,则越低。

(3) 数字对象语用重要性的计算流程如下:首先基于数字对象之间的语用关系构建数字对象语用关系图,随后迭代计算图中各个数字对象的语用重要程度。对于编号为 k 的数字对象 d_k ,其语用重要性计算如下。

$$W_{\text{prag}}(d_k) = \frac{(1-D_F)}{N} + D_F \cdot \frac{p \cdot W_{\text{prag}}(d_{\text{start}})}{P(d_{\text{start}})} \quad (5)$$

其中, $W_{\text{prag}}(d_k)$ 表示数字对象 d_k 的语用重要性; D_F (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; $P(d_{\text{start}})$ 表示 d_{start} 指向其他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也有不同。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越靠前。计算语用重要性的伪代码如算法1所示。

算法1: 单个数字对象的语用重要性算法

输入: 数字对象语用关系图 graph, 阻尼因子 D_F

输出: 数字对象语用重要性 W_{prag}

function calculatePragmaticWeights (graph, D_F)

$W_{\text{prag}} \leftarrow \text{initializeWeights}(\text{graph})$ // 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**

```

 $W_{\text{prag}} \leftarrow 0$ 
neighbors  $\leftarrow$  getNeighbors( $do$ )
for each neighbor in neighbors do
    pragmaticWeight  $\leftarrow$  getPragmaticWeight
    (graph, neighbor,  $do$ ) // 获取语用权重
    neighborWeight  $\leftarrow$  getWeight(graph,
    neighbor) // 获取相邻数字对象语用
    totalPragmaticWeight  $\leftarrow$ 
    getTotalPragmaticWeight(graph, neighbor) //
    获取邻居数字对象指向其他数字对象的语
    用权重之和
     $W_{\text{prag}} \leftarrow W_{\text{prag}} + (\text{pragmaticWeight} \times$ 
    neighborWeight)
    /totalPragmaticWeight
end for
 $W_{\text{prag}} \leftarrow (1 - D_F) / \text{graph.size} + D_F \times W_{\text{prag}}$ 
setWeight(graph,  $do$ ,  $W_{\text{prag}}$ )
end for
end for
return  $W_{\text{prag}}$ 
end function

```

DO 语用重要性计算示例如图4所示。初始状态下, d_1 、 d_2 、 d_3 、 d_4 的语用重要性均设置为1,4个DO之间的语用权重分别为 $p_{12}=0.75$ 、 $p_{24}=0.5$ 、 $p_{32}=0.25$, 阻尼因子 $D_F=0.85$, 经过一轮迭代后, d_2 的语用重要性 $W_{\text{prag}}(d_2)$ 计算结果如下:

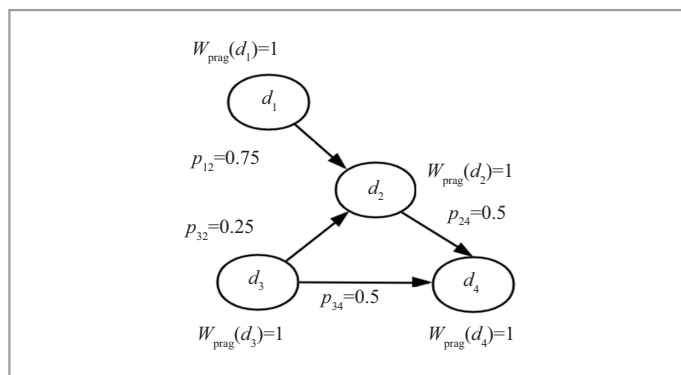


图4 DO语用重要性计算示例

$$W_{\text{prag}}(d_2) = \frac{(1-D_F)}{N} + \text{DF} \cdot \left(\frac{p_{12} \cdot W_{\text{prag}}(d_1)}{P(d_1)} + \frac{p_{32} \cdot W_{\text{prag}}(d_3)}{P(d_3)} \right) = \frac{(1-0.85)}{4} + 0.85 \times \left(\frac{0.75}{0.75} \times 1 + \frac{0.25}{0.25+0.5} \times 1 \right) = 1.17 \quad (6)$$

数字对象的语用重要性需要进行归一化处理, 可表示为:

$$W_{\text{prag}}(d_k)_{\text{normal}} = \frac{W_{\text{prag}}(d_k) - W_{\min}}{W_{\max} - W_{\min}} \quad (7)$$

其中, $W_{\text{prag}}(d_k)_{\text{normal}}$ 表示 d_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。

3.4 DOR 值计算

一个数字对象的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, 这可以通过综合权重值, 即数字对象排序 (data object rank, DOR) 值表示。对于一个查询 Q , 编号为 k 的元数据 d_k 的 DOR 值的计算如下。

$$R(d_k, Q) = \alpha \cdot \text{Sim}(d_k, Q)_{\text{normal}} + \beta \cdot W_{\text{prag}}(d_k)_{\text{normal}} \quad (8)$$

其中, $\text{Sim}(d_k, Q)_{\text{normal}}$ 表示归一化处理后的元数据 d_k 与查询 Q 之间的查询相似度; $W_{\text{prag}}(d_k)_{\text{normal}}$ 表示归一化处理后的元数据 d_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。

4 跨数据空间搜索系统测试

实验测试了搜索系统的排序算法, 使用开源医学数据集, 具体包括: (1) 诊疗指南数据, 来自 OpenKG 公开的结构化临床诊疗

指南数据集, 涵盖 77 000 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) 媒体健康科普信息, 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, 借助大模型模拟多家三甲教学医院联合贡献的教学病例数据集, 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 16 GB 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; 媒体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; 教学病例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 3 种排序方案: (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, 简称 “W-PageRank”; (3) 基于分布式搜索架构并使用 DO-Rank 算法进行结果排序, 简称 “DO-Rank”。

4.1 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

在跨多领域数据空间搜索场景下不可行，将检索深度设定为 $R=30$ ，即系统返回前 30 个候选结果作为搜索系统的检索结果，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 分数 (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 ~ 图 8 所示。

精确率、召回率、F1 分数用于衡量搜索系统的查准率和查全率。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，命中结果在所有搜索结果中的占比下降；召回率和 F1 分数升高，是因为随着搜索结果数量增加，命中结果数量增多，从而在全部数据集中的占比也增加。W-PageRank 算法的查准率、查全率均高于 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，将数据全量汇聚后再排序，得到的是全局最优解；而 DO-Search 算法和 DO-Rank 算法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，相比于 W-PageRank 算法，DO-Rank 算法的精确率

低 5.33%、召回率低 5.34%、F1 分数低 5.33%；相比于 DO-Search 算法，DO-Rank 算法的精确率高 18.3%、召回率高 18.32%、F1 分数高 18.38%。上述结果表明，DO-Rank 算法通过数据语用关系扩充各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。

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，原因是二者均基于语用权重进行排序，能够将更符合用户使用场景的数字对象排在前面。

4.2 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 至 10/10 共 10 个梯度比例规模的数据，对算法进行重复实验，计算各算法运行的平均响应时间。各算法执行时长比较如图 9 所示。

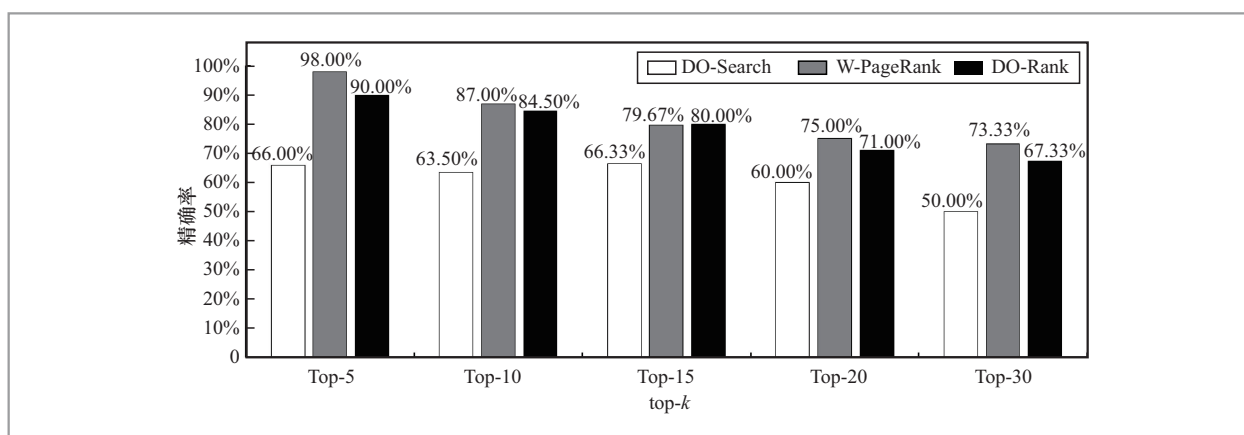


图5 各算法精确率比较

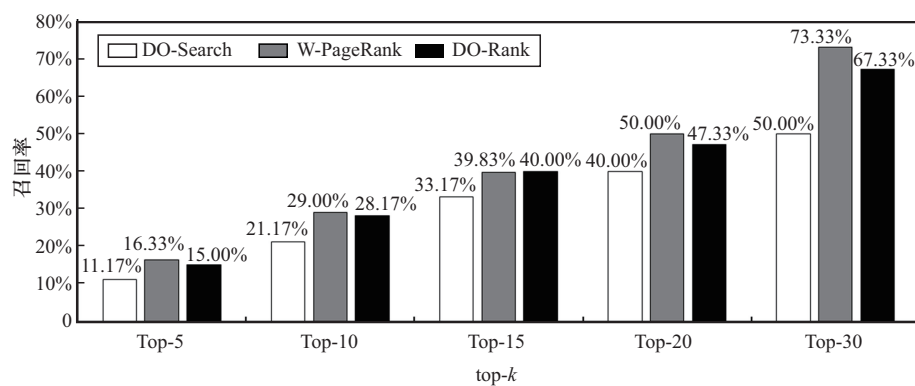


图6 各算法召回率比较

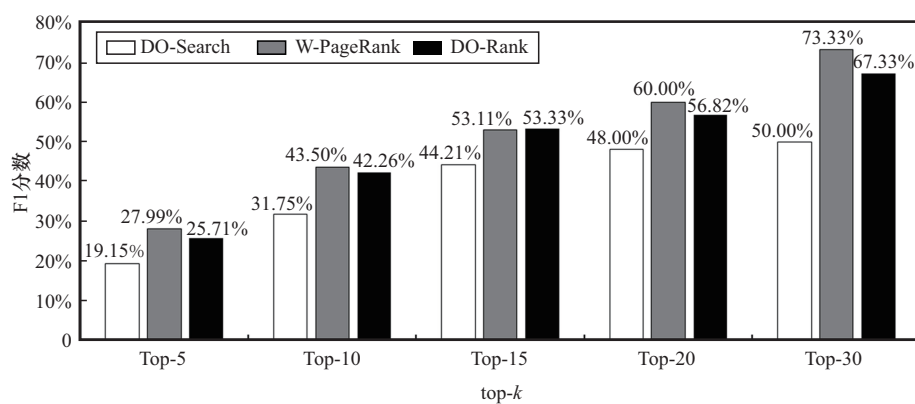


图7 各算法F1分数比较

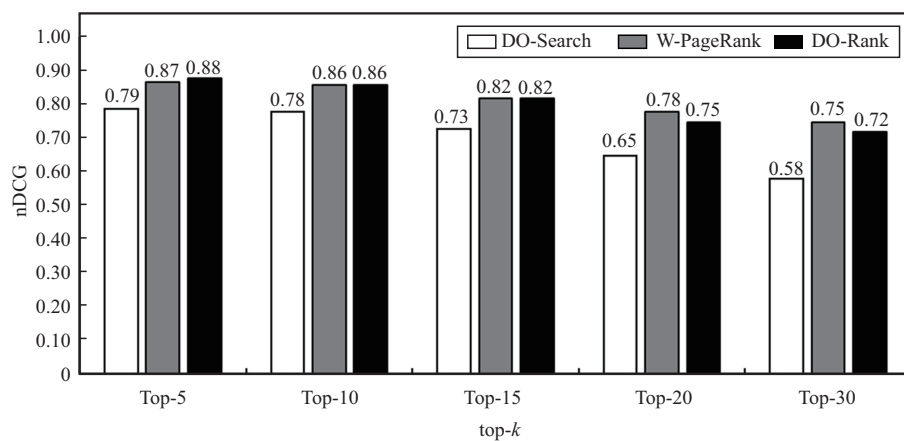


图8 各算法nDCG值比较

图9的结果显示,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的有限跳数数据,时间不会随着数据量的增加而出现大幅扩大。因此,随着数据规模的扩大,DO-Rank 算法的时间优势越来越明显。

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,使其能够更好地适用于跨数据空间的分布式数据搜索。

5 结束语

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,本文提出了面向数据场景化使用的数据空间搜索系统架构,同时基于数据场景化使用形成的语用关系设计了搜索排序算法DO-Rank,并进行了实验验证。实验结果表明,跨空间搜索系统在不汇聚数据的情况下,其精确率、召回率、归一化折损累计增益等指标接近采用数据集中汇聚的检索方法,为实现“数据不出域”“可见不可得”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应用,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,还将进一步研究基于大语言模型智能应用场景的跨数据空间搜索服务改进,如基于语义标签与场景特征

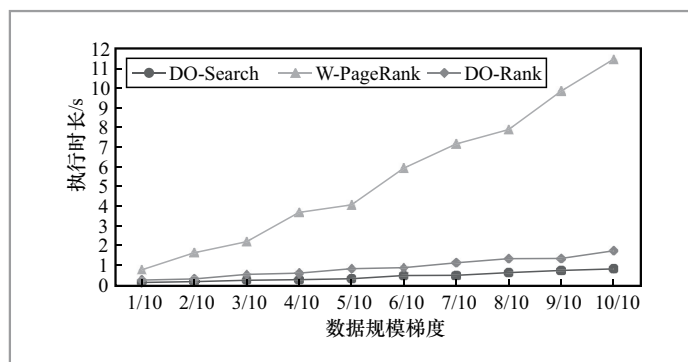


图9 各算法执行时长比较

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]. In - ternational Journal on Digital Libraries, 2006, 6(2): 115-123.
- [2] CNRI. RFC 3650: Handle System Over - view [EB].2003.
- [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 infra - structure of dataspace [J]. Communica - tions 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 P. Principles and programming implementation of big data search en - gines [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]. 2版. 北京: 科学出版社, 2012.
Li X M, Yan H F, Wang J M. Search Engines: principles, techniques, and systems [M]. 2nd. Beijing: Corporate

- 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(5): 113–116, 119.
- Qian L B, Ji Z Z. Review on distributed search engine model[J]. Intelligent Computer and Applications, 2015, 5(5): 113–116, 119.
- [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(2): 36–38.
- Zhang J G, Chen D Q, Wu Z X. Research on search engine and meta search engine [J]. New Technology of Library and Information Service, 2002(2): 36–38.
- [8] Garba A, Wu S L, 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 B, Chen Q X, Ma L. The study on large scale duplicated Web pages of ChineseFast 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 probabilistic weighted retrieval[C]//Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University. London: Springer, 1994: 232–241.
- [11] Huang P S, He X D, Gao J F, et al. Learning deep structured semantic models for Web search using click-through data[C]//Proceedings of the 22nd ACM International Conference on Information & Knowledge Management. New York: ACM, 2013: 2333–2338.
- [12] Nogueira R, Cho K. Passage re-ranking with bert[PP]. V5. arXiv (2020-04-14) [2026-05-27]. arXiv: 1901.04085.
- [13] Liu W H, Ma X Y, Sun W W, et al. ReasonRank: empowering passage ranking with strong reasoning ability[PP]. V3. arXiv (2026-04-22)[2026-05-27]. arXiv: 2508.07050.
- [14] Fan Y Q, Xue K, Li Z L, et al. An LLM-based framework for biomedical terminology normalization in social media via multi-agent collaboration[C]//Proceedings of the 31st International Conference on Computational Linguistics. Kerriville: Association for Computational Linguistics, 2025: 10712–10726.
- [15] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks and ISDN Systems, 1998, 30(1/2/3/4/5/6/7): 107–117.
- [16] Xing W, Ghorbani A. Weighted PageRank algorithm[C]//Proceedings of the Second Annual Conference on Communication Networks and Services Research, 2004. Piscataway: IEEE Press, 2004: 305–314.
- [17] Dayeh M A, Sartawi B, Salah S. A bias-free time-aware pagerank algorithm for paper ranking in dynamic citation networks[J]. Intelligent Information Management, 2022, 14(2): 53–70.
- [18] Selvan M P, Shekar A C, Babu D R, et al. Efficient ranking based on Web page importance and personalized search[C]//Proceedings of the 2015 International Conference on Communications and Signal Processing (ICCSP). Piscataway: IEEE Press, 2015: 1093–1097.
- [19] Abdulrahman A. Web pages ranking algorithms: a survey[J]. Qubahan Academic

- demic Journal, 2021, 1(3): 29–34.
- [20] Kang G S, Liu J X, Tang M D, et al. An effective web service ranking method via exploring user behavior[J]. IEEE Transactions on Network and Service Management, 2015, 12(4): 554–564.
- [21] Drivas I C, Sakas D P, Giannakopoulos G A, et al. Big data analytics for search engine optimization[J]. Big Data and Cognitive Computing, 2020, 4(2): 5.
- [22] Usta A, Altıngöve İ S, Özcan R, et al. Learning to rank for educational search engines[J]. IEEE Transactions on Learning Technologies, 2021, 14(2): 211–225.
- [23] He Y X, Gan Q, Wipf D, et al. GN-NRank: learning global rankings from pairwise comparisons via directed graph neural networks[C]//Proceedings of the 39th International Conference on Machine Learning. PMLR, 2022:8581–8612.
- [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(5): 191–198.
- Dou L Y, Wei F, Deng A, et al. Empowering intelligent technology information work with federated learning for multi-source data integration[J]. Journal of Intelligence, 2025, 44(5): 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(S1): 870–883.
- Lan Y J, Ma Z Q, Chen J L, et al. Survey on application of searchable attribute-based encryption technology based on blockchain[J]. Computer Science, 2024, 51(S1): 870–883.
- [26] 孙金焯, 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 Y, Guo S X. Research on distributed data spatial architecture model for ownership governance[J]. Big Data Research, 2025, 11(2): 140–151.
- [27] 罗超然, 马郅, 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(2): 110–121.
- Luo C R, Ma Y, Jing X, et al. Internet of data: a solution for dataspace infrastructure and its technical challenges[J]. Big Data Research, 2023, 9(2): 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 L, Yang J R, 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 Q, Liu Y H, Guan T P, et al. DPML: a markup language for scientific data pragmatics[J]. Frontiers of Data & Computing, 2024, 6(4): 46–58.

作者简介



姜海鸥 (1987–), 女, 博士, 数据空间技术与系统全国重点实验室副研究员, 主要研究方向为云计算、大数据、软件工程。



胡子逸（1997-），男，北京大学软件与微电子学院硕士生，主要研究方向为软件工程、数据智能。



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



罗超然（1995-），男，博士，数据空间技术与系统全国重点实验室副研究员，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，数据空间技术与系统全国重点实验室副研究员，主要研究方向为大数据治理技术与系统。

录用日期: 2026-01-07
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 Item: National Key Laboratory of Data Space Technology and Systems (No. QZJZ2024KJ001-04-SJKJSS-KTBG)