

从元数据到本体论概述

韩 凌¹
HAN Ling

摘 要

从图书馆到互联网的信息组织与检索经历了资源数字化、万维网、语义网三个发展阶段,元数据技术、万维网、搜索引擎、语义网、关联数据、本体论等技术实现了信息与知识的组织与检索技术从图书馆到互联网,从语法关联到语义关联的跨越。文章系统梳理了上述各阶段的基本理念和关键技术,描述了始于图书馆机读目录的元数据技术和基于万维网的搜索引擎技术的诞生与发展;介绍了语义网、关联数据、本体论等技术理念以及基于它们的搜索引擎技术,总结了技术演进规律,呈现了信息组织与检索技术从封闭的图书馆系统向开放的互联网环境演进的历史轨迹和内在逻辑,突出强调了从“语法关联”向“语义关联”的跨越这一根本性进步,为理解当前语义搜索、知识图谱等前沿技术奠定了基础。

关键词

元数据; 语义网; 本体论

doi: 10.3969/j.issn.1672-9528.2025.11.035

0 引言

如何有效地组织和检索信息已成为信息时代的一个重大挑战。元数据(metadata)作为信息资源的“语义骨架”,自20世纪60年代起便为图书馆学、信息科学等领域提供了基础性描述框架。其标准化实践(如MARC^[1]、Dublin Core^[2])通过定义“数据的数据”,解决了异构资源的分类、索引与互操作问题,为大规模数据集成奠定了基础。

1990年,Emtage等人^[3]开发了第一个真正意义上的搜索引擎Archie。随后,由Gray开发的Wanderer进一步扩展了搜索引擎的功能,成为第一个网络爬虫(web crawler)。Fletcher开发的Jump Station首次集成了爬虫、索引、查找三大特征^[4],Pinkerton^[5]开发的Web Crawler首次实现了对网页全文的索引与检索,标志着现代搜索引擎的雏形初现。以PageRank算法^[6]为代表的搜索技术虽然实现了海量数据的快速查找,但本质上仍停留在“语法”层面,缺乏对用户意图与内容语义的深度理解。

这一矛盾促使Berners等人^[7]提出了语义网(semantic web)构想。借助资源描述框架(resource description framework, RDF)^[8]、本体论(ontology)^[9]所提供的形式化知识表示方法以及关联数据(linked data)^[10]等技术标准将离散的数据单元转化为具有明确语义指向的知识单元,语义网首次将数据的语义关联显式化,为进一步实现“数据互联”奠定了基础。

本文总结历史经验与技术演进规律,揭示了从数字资源描述到语义互联的内在驱动力,为探索下一代智能化、个性化的信息获取服务提供理论参考和实践启示。

1 元数据标准的建立与搜索引擎的诞生

1.1 元数据标准的建立

1969年, Bagley^[11]在报告中提出“metadata”的概念,但在1966年,美国国会图书馆便制定了最早的图书电子编目标标准MARC(machine-readable cataloguing),也称为机读目录,以编码形式按特定结构形成作品及文献指代的格式。MARC成为数字化环境下一种重要的元数据格式,主要被图书馆信息专业人员用于组织数字化资源。MARC的成功也推动了MARC II^[12]的开发。自诞生以后, MARC格式在国际上迅速传播,形成了UKMARC、INTERMARC、UNIMARC、CANMARC等变体,在美国本土则进一步标准化,于20世纪80年代形成了USMARC格式^[13],并成为美国的国家标准。后来CANMARC与USMARC进一步协调,产生了MARC21格式^[14]。

比MARC稍晚些时候, AACR(anglo american cataloging rules)元数据标准^[15]于1967年发布。该标准由英美两国的图书馆协会基于巴黎原则(paris principles)联合制定。该标准更倾向于英语形式,在一定程度上限制了其在全球范围内的适用性。1971年,国际图联出版了第一版ISBD(international standard bibliographic description)^[16],旨在提供一种全球化的书目描述格式,促进全球书目数据的一致性和可交换性。

1. 南京林业大学 江苏南京 210037

为了融合北美与英国的阶段性成果,也为了提升 AACR 的国际影响力,美英加等国的图书馆协会于 1978 年联合出版了 AACR2^[17]。与 AACR1 一样, AACR2 基于巴黎原则制定,并且在结构上引入了 ISBD 格式^[18],扩大了其全球适用范围。到 1988 年修订版发布时, AACR2 已在大多数英语国家得到普遍接受。 AACR2 主要面向卡片时代的图书馆而生,未能很好兼容后来的数字资源,这也导致了资源描述与检索(resource description & access, RDA)标准的诞生^[19]。

1989 年,英国 Berners-Lee 教授^[20]发明了万维网(world wide web, WWW)。随着网络信息量的不断膨胀,如何对网络资源进行表示、标识、定位和访问得到了学者们的高度重视和广泛讨论。1995 年,首届“OCLC/NCSA 元数据研讨会”在美国俄亥俄州的都柏林市举行,旨在讨论如何识别元数据以实现在 Web 上定位和描述对象。都柏林核心元数据标准(dublin core metadata standards, DC)由此诞生。1996 年,第二届研讨会围绕网络图像资源进行讨论,确立了 15 个核心元素。1998 年, IETF (the internet engineering task force) 发布 RFC 2413, 正式定义都柏林核心元数据集。2007 年, IETF 发布 RFC 5013, 规范了 Dublin Core 1.1 版本的元素集。构成了许多网站和数字图书馆的基础,在语义网(Semantic Net)、知识组织等方面均有重要应用。随后,各个行业的元数据标准也相继诞生,如表 1 所示。这些元数据标准对信息资源的结构、属性和上下文信息进行标注,对数据资源进行组织和分类,使之能够以一种标准化的方式被描述和查询。

表 1 常见的元数据标准及其应用领域

名称	应用领域
DDI(data documentation initiative) ^[21]	社会科学
Darwin Core ^[22]	生物领域
CDWA (categories for the description of works of art)	艺术设计领域
MIDAS Heritage ^[23]	文物保护领域
DataCite Metadata Schema ^[24]	科学数据管理
schema.org ^[25]	搜索引擎优化

1.2 搜索引擎的诞生

1990 年, Alan Emtage 创建了搜索引擎 Archie。1993 年, MIT 的 Gray 编写了网络爬虫(web crawls)程序 World Wide Web Wanderer。同年,日内瓦大学的 Nierstrasz 编写了一系列 Perl 脚本,这些脚本构成了搜索引擎 W3Catalog^[26]的基础。同年, Fletcher 开发了搜索引擎 Jump Station, Fletcher 也被 BBC 誉为“搜索引擎之父”。1994 年, Pinkerton 设计

了 WebCrawler 搜索引擎。1994 年 1 月,斯坦福大学的两位研究生杨致远和大卫·费罗创建了一个名为“杰瑞的网络指南”(jerry's guide to the world wide web)的网站。1994 年 3 月,“杰瑞的网络指南”更名为“Yahoo!”。1998 年, Brin 等人^[27]借鉴引文索引方法,提出 PageRank 算法,创建了 Google。2000 年 1 月 1 日,百度公司在中关村成立,其后台核心算法是超链分析算法并拥有技术专利^[28]。上述几个搜索引擎的特点如表 2 所示。

表 2 早期典型的搜索引擎及其特点

名称	特点
Archie	全世界第一款搜索引擎,使用 Telnet 协议搜索匿名文档传输协议站点列表,创建可通过 FTP 使用的索引文档。
World Wide Web Wanderer	全世界第一个网络爬虫(web crawls)程序
W3Catalog	定期镜像原先手工维护的目录下的页面,并将其重写为标准格式。
Jump Station	第一个集成了爬虫、索引、查找三大特征的搜索引擎。
WebCrawler	第一个对网页进行全文检索的搜索引擎。
Yahoo!	手工索引的门户网站
Google	PageRank 算法
百度	超链分析算法

从纸质时代到计算机时代,再从计算机时代到互联网时代,源自图书文献编目与索引的信息资源组织方法与技术出现了两次飞跃。元数据和统一资源标识符(uniform resource identifier, URI)共同构建了类似于图书馆书录的网络数据目录,实现了网络资源的定位、编目索引及其分类管理。在诞生之初,搜索引擎通过爬虫程序、手动方式,以及借鉴引文索引的方法对网页进行索引排序。到了世纪之交,搜索引擎初步具备了网络信息的导航、浏览和检索功能,开始与问答系统结合^[29-34]。

2 语义网和语义关联显式化

2.1 语义网和关联数据

早期的搜索引擎和网络问答系统远远不能理解和使用网络资源的实际语义内容来回答用户请求,也无法提供对网络信息的高效访问。2001 年, Berners-Lee 等人提出了语义网(semantic web)的概念。在其设想中,语义网并不独立于现有网络,而是现有网络的延伸。在语义网中,信息被赋予的计算机可以理解的明确定义,对网络信息和元数据进行编码并提供网络信息和元数据的元素之间的链接,记录这些链接的语义,以便将它们变成计算机可以处理的元数据类型。

为此, Berners-Lee 等人建议使用可扩展标记语言 (extensible markup language, XML) 为网页或网页上的文本部分添加标签, 同时使用 RDF 方法来表达语义链接。RDF 中的三元关系语句可由 XML 语言的标签进行描述。

2006 年, Berners-Lee 进一步提出“关联数据 (linked data)”的概念。在 Berners-Lee 的设想中, URI 用于唯一标识网络中的实体或数据资源, 并通过 HTTP 协议进行访问, 通过 SPARQL 查询语言进行跨数据集查询和整合 RDF 数据。URI、RDF、HTTP、SPARQL 构成语义网及关联数据的四要素, 将以往独立存在的数据库以数据图的形式呈现出来, 这些数据图可以与其他数据集相互链接和整合, 同时将数据的语义关联显式化, 从而构建一个全球范围的数据交互空间。

2.2 本体论及其应用

在语义网的早期发展阶段, 本体论起着非常重要的作用。按照 Gruber 的说法, 本体论是对概念化进行明确说明的规范 (an ontology is an explicit specification of a conceptualization)。在语义网的语境下, 本体论是一种计算机建模思想, 其内核是本体模型定义了概念及概念之间的关系, 是一个结构化的关系数据库, 该模型能够被他人重复使用^[35]。搜索引擎可以通过检索网页的元数据和本体模型中已定义的概念及其之间的关系进行信息检索和数据挖掘。2004 年, OWL (web ontology language) 和 RDF 成为 W3C 标准。2008 年, W3C 制定了 RDF 查询语言标准 SPARQL, 与 OWL 的兼容性也得到大幅度提升。

早在 2000 年, Gene Ontology Consortium、FlyBase、Berkeley Drosophila Genome Project、Saccharomyces Genome Database 以及 Mouse Genome Database and Gene Expression Database 等研究机构的学者^[36]在《Nature Genetics》发表标志性论文《Gene Ontology: tool for the unification of biology》, 正式推出基因本体论 (gene ontology, GO) 项目框架。该项目试图通过结构化、标准化的本体论 (分子功能、生物过程、细胞组分) 解决相关领域的术语混乱问题, 为基因及其产物提供统一的生物学注释, 促进跨物种数据整合与比较。2001 年, Consortium^[37]发布了 GO 项目的详细设计与实施方案, 明确了本体构建原则和协作流程。之后, GO 迅速发展并引发大量关注。2006 年, 已产生了 600 多篇相关文献^[38]。时至今日, GO 依旧在医疗生物领域发挥重要作用^[39-43], 并通过持续的更新和扩展^[44-45]保持自身的准确性和全面性。2004 年, Castells 等人^[46]与领域专家互动, 采用语义网技术开发了一个经济金融信息管理平台。该平台包括一个本体驱动的知识库, 并且提供了该知识库的内容发布、访问和管理的功能。除了改善最终用户的体验外, 平台的搜索、可视化、管理模块也易于移植到其他本体。2008 年, Kontopoulos 等人^[47]设

计了 PASER 系统。PASER 使用人工智能规划技术和语义网技术动态构建自主学习路径。在 PASER 中建有课程的能力本体 (competencies ontology), 以树形结构存储, 用于将学习目标分解为几个小目标。PASER 的核心是其规划引擎。它将教育资源的选择与合成问题视为一个规划问题, 利用斯坦福研究所规划系统 (stanford research institute planning system, STRIPS) 表示法对学习过程进行建模, 存储、管理和挑选组合电子教材, 满足学习者的概况、偏好、需求和能力。2012 年, Bertaud-Gounot 等人^[48]提出一种本体论表示, 对 30 个强直性脊柱炎患者病例进行分类。指出本体和通用本体工具可以在疾病分类任务中发挥支持作用, 前提是分类标准是不涉及不确定性的逻辑表达式。为了解决不同航空数据供应商之间的数据可交换性和互操作性问题, 航空业开始寻求标准化的数据交换模型, 本体论也被引入航空数据管理领域^[49]。美国国家航空航天局 (national aeronautics and space administration, NASA)、美国联邦航空局 (federal aviation administration, FAA) 等机构均开发了基于本体的航空数据管理系统。

2.3 基于语义网的搜索引擎技术

在语义明确的前提下, 利用本体论有助于在专业领域内构建完整规范的知识体系, 可以进行高效的信息组织和检索信息。与传统的面向非结构化数据的网络文献检索和面向结构化数据的数据库检索均不相同, 在语义网时代, 越来越多的网络检索开始面向实体, 其检索对象体现出半结构化 (semi structured)^[50]的特点。表 3 是部分典型的语义网搜索引擎及其技术特征。

基于本体论的实践本体论通过对概念、实体及其相互关系的明确定义与组织, 使得语义网能够以结构化、可被机器理解的方式呈现海量复杂信息, 为信息建模提供了方法论基础。同时, 基于本体论所开展的实践积累了丰富的经验和和技术成果, 也为后来的知识图谱铺就了道路。

3 总结

本文总结了信息和知识的组织与检索从图书馆编目索引到互联网信息检索的技术发展历程。元数据、搜索引擎、语义网、关联数据和本体论等技术展现了网络信息的分类、组织、索引与检索从语法逻辑推理到语义理解关联的技术演进。块数据技术为信息资源的描述和管理提供了标准化的框架, 确保了数据的规范性和可访问性, 是信息组织的基石。信息以 Web 的形式发布和利用时, 基于万维网诞生了以文献索引和检索理论为基础的搜索引擎技术。搜索引擎在网络信息的爬取、识别、组织和定位基础上, 创建信息的电子索引条目, 检索用户的目标信息, 其检索算法一直在不断适应网络技术的进步。语义网的出现, 标志着网络信息的检索与组织开始

表 3 语义网搜索引擎及其特点

名称	特点
Swoogle ^[51]	通过元搜索（meta-search）和跟踪链接的方式，运行多个爬虫程序查找语义网文献（semantic web document, SWD），对 SWD 的统一资源标识符引用（uniform resource identifier reference, URIfref）进行关键词匹配检索。通过适用于语义网的“理性随机冲浪”（rational random surfer, RRS）算法，计及 SWD 间不同类型的链接及其权重计算 SWD 的排名。
AquaLog ^[52]	AquaLog 通过使用 JAPE 语法识别术语、关系，以及问题的模式或类型。用户输入的自然语言查询被转换为查询三元组。这些查询三元组被转换为与被查询本体兼容的三元组。AquaLog 调用 RSS（relation similarity service）组件查看本体的结构和语义网上可用的信息，使用字符串相似性匹配，通用词汇资源以及通过使用学习获得的专业领域的词典来理解输入查询，但只能进行单一本体搜索。
PowerAqua ^[53]	复用了 AquaLog 的语言组件，并且查询中的每个术语或关系中的名词都通过 WordNet 中的同义词（synonyms）、上义词（hyponyms）和下义词（hyponyms）关系进行查询扩展，使得 PowerAqua 可以在多个本体中搜索答案。
Falcons ^[54]	索引实体的虚拟文档 ^[55] 中的所有术语，即使查询词未直接提及实体名称，系统也能找到相关实体。
Sindice ^[56]	通过关键词、URI、逆向功能属性（inverse functional property, IFP）以及倒排索引技术来索引定位语义网资源，为其他语义网客户端应用程序提供服务。
Semaphore ^[57]	Semaphore 使用混合查询（hybrid query）形式，将关键字搜索与结构化查询相结合。其核心思想是，将单个关键词或数个关键词的组合视为关键词概念（keyword concepts），当被检索资源中包含某个关键词概念的全部关键词时，该资源即被视为满足该关键词概念的实例。Semaphore 将被检索的网络资源视为文献，将关键词概念视为关键词，经过适当处理后，Semaphore 使用倒排索引来索引存储在 IR 引擎中的网络数据，使用基于位置的索引（position-based index, PosIdx）方法来索引、检索并最终连接关系三元组。
FREyA ^[58-59]	通过斯坦福解析器（stanford parser） ^[60] 从提问语句中推导潜在本体概念（potential ontology concepts, POCs）。POC 与问题中的术语相关，通过遍历语法树来识别最接近该 POC 的 OCs（ontology concepts），然后使用本体推理来生成建议。当 FREyA 无法生成答案时，会以对话框的形式与用户互动，要求用户消除歧义或给出建议选项让用户选择。
Watson ^[61]	使用 Google、Swoogle、手动提交的 URL 等不同的来源来发现本体和语义数据，搜索结果中显示的 URI 是指向页面的链接。该页面提供相应本体或特定实体的详细信息，用户可以在实体和本体之间导航，探索本体的内容，浏览实体之间的关系，以及查看本体及其元数据。
SIREn ^[62]	使用 AFOR（adaptive frame of reference）高性能压缩技术，用于优化倒排索引的存储和查询性能。AFOR 压缩技术在索引大小、索引时间和查询时间之间提供了良好的平衡，特别是在处理大规模数据集时。这使得 SIREn 能够在普通硬件上处理大量复杂的查询。

向语义层面深入。本体论和关联数据的引入，提供了在语义明确的前提下进行高准确性检索的基础，进一步强化了信息关联的构建，为信息的语义表达和整合提供了强大的理论和实践支持，标志着网络信息检索从“被动响应”向“主动理解”的跨越，也为后来的知识图谱技术奠定了基础。

参考文献：

[1] VOLLE A. Metadata[EB/OL].(2025-10-18)[2025-10-26].
https://www.britannica.com/technology/metadata.
[2] ARAKAKI F A, ALVES R C V, DA COSTA SANTOS P L V
A. Dublin core: state of art (1995 to 2015)[J]. Informação &
sociedade, 2018,28(2): 7-20.
[3] EMTAGE A. Archie[M]//Intelligent information retrieval: the

case of astronomy and related space sciences. Springer Dor-
drecht, 1993: 103-111.

[4] KATHURIA M, NAGPAL C K, DUHAN N. Journey of web
search engines: milestones, challenges & innovations[J]. In-
ternational journal of information technology and computer
science, 2016,8(12): 47-58.
[5] PINKERTON B. Finding what people want: experiences with
the WebCrawler[EB/OL].1994[2025-09-12].https://xueshu.
baidu.com/usercenter/paper/show?paperid=e030d875267623
14c64067fc60a9c92d&site=xueshu_se.
[6] PAGE L, BRIN S, MOTWANI R, et al. The PageRank cita-
tion ranking: bring order to the web[EB/OL].(1999-11-11)
[2025-09-23].https://www.semanticscholar.org/paper/The-

- PageRank-Citation-Ranking-%3A-Bringing-Order-to-Page-Brin/eb82d3035849cd23578096462ba419b53198a556.
- [7] BERNERS LT, HENDLER J, LASSILA O. The semantic web[J]. Scientific american, 2001,284(5): 34-43.
- [8] MILLER E. An introduction to the resource description framework[J]. Bulletin of the American society for information science, 1998,25(1): 15-19.
- [9] GRUBER T R. A translation approach to portable ontology specifications[J]. Knowledge acquisition, 1993,5(2): 199-220.
- [10] DENNIS M A. Tim Berners-Lee[EB/OL]. (2025-10-17) [2025-10-25]. <https://www.w3.org/DesignIssues/LinkedData.html>.
- [11] BAGLEY P R. Extension of programming language concepts[EB/OL]. (1968-01-01)[2025-10-25]. <https://www.amazon.com/EXTENSION-PROGRAMMING-LANGUAGE-CONCEPTS-Philip/dp/B00AOH1W92>.
- [12] AVRAM H D. The MARC II format: a communications format for bibliographic data[EB/OL]. 1968[2025-10-11]. <https://www.semanticscholar.org/paper/The-MARC-II-format-%3A-a-communications-format-for-Avram/8e990cd0db90ca55f812d21a79c563a586669050>.
- [13] DEAN S. From cataloguing to metadata creation: a cultural and methodological introduction [J]. Journal of the Australian library and information association, 2024,73(1): 128.
- [14] DAS D. MARC 21: the standard exchange format for the 21st century[EB/OL]. (2004-11-01)[2025-10-11]. <https://www.semanticscholar.org/paper/MARC-21-%3A-The-Standard-Exchange-Format-for-the-21-Das/025324f837552d21815a8c929a29a12e0f531c76>.
- [15] MAXWELL M F. The genesis of the anglo-american cataloging rules[J]. Anglo-American cataloging rules, 1977,27(3): 238-262.
- [16] BYRUM J D. The birth and rebirth of the ISBDs: process and procedures for creating and revising the international standard bibliographic descriptions[J]. IFLA journal, 2001,27(1): 34-37.
- [17] MANNING R W. The anglo-american cataloging rules and their future[J]. Library resources & technical services, 2011,44(3): 129-134.
- [18] COLE J E. AACR2 and ISBD(S): correspondence or divergence?[J]. Serials review, 1982,8(3): 67-69.
- [19] MILLER L. Resource description and access (RDA): an introduction for reference librarians[J]. Reference & user services quarterly, 2011,50(3): 216-222.
- [20] BERNERS-LEE T, CAILLIAU R, GROFF J F, et al. World-Wide Web: the information universe[J]. Internet research, 1992,2(1): 52-58.
- [21] VARDIGAN M, HEUS P, THOMAS W. Data documentation initiative: toward a standard for the social sciences[J]. International journal of digital curation, 2008,3(1): 107-113.
- [22] WIECZOREK J, BLOOM D, GURALNICK R, et al. Darwin core: an evolving community-developed biodiversity data standard[J]. Public library of science ONE, 2012,7(1): 29715.
- [23] AGATHOS M, KAPIDAKIS S. Harvesting metadata for immovable monuments from "local inventories" and mapping them to standards[C]//Metadata and Semantic Research. Berlin: Springer, 2011: 424-432.
- [24] STARR J, AMMANN N, ASHTON J, et al. Datacite metadata schema for publication and citation of research data[EB/OL]. 2015[2025-10-15]. https://www.researchgate.net/publication/283284243_DataCite_Metadata_Schema_for_Publication_and_Citation_of_Research_Data.
- [25] GUHA R V, BRICKLEY D, MACBETH S. Schema. org: evolution of structured data on the web[J]. Communications of the ACM, 2016,59(2): 44-51.
- [26] SEYMOUR T, FRANTSVOG D, KUMAR S. History of search engines[J]. International journal of management & information systems, 2011,15(4): 47-58.
- [27] 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.
- [28] LEONG S. Prophets of mass innovation: the gospel according to BAT[J]. Media industries journal, 2018,5(1): 69-87.
- [29] KWOK C, ETZIONI O, WELD D S. Scaling question answering to the web[J]. Scaling question answering to the web, 2001,19(3): 242-262.
- [30] BRILL E, DUMAIS S, BANKO M. An analysis of the AskMSR question-answering system[C]//EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing. New York: ACM, 2002: 257-264.
- [31] DUMAIS S, BANKO M, BRILL E, et al. Web question answering: is more always better?[C/OL]//Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM, 2002[2025-10-22]. <https://www.docin.com/p-927679784.html>.

- [32] RADEV D, FAN W G, QI H, et al. Probabilistic question answering on the web[C]//WWW '02: Proceedings of the 11th international conference on World Wide Web.NewYork: ACM, 2002: 408-419.
- [33] ZHANG D, LEE W S A Web-based question answering system[EB/OL]. 2003[2025-10-03].<https://www.semanticscholar.org/paper/A-Web-based-Question-Answering-System-Zhang-Lee/ad23647c895d57668fc202259dccb29edb9e683>.
- [34] STOYANCHEV S, SONG Y C, LAHTI W. Exact phrases in information retrieval for question answering[C]//IRQA'08: Coling 2008: Proceedings of the 2nd workshop on Information Retrieval for Question Answering.NewYork:ACM,2008:9-16.
- [35] HITZLER P. Semantic Web: a review of the field[J]. Communications of the ACM, 2020,64(2): 76-83.
- [36] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology[J]. Nature genetics,2000,25(1): 25-29.
- [37] CONSORTIUM G O. Creating the gene ontology resource: design and implementation[J]. Genome research, 2001,11(8): 1425-1433.
- [38] CONSORTIUM G O. The gene ontology (GO) project in 2006[J]. Nucleic acids research, 2006,34: 322-326.
- [39] YE Y, DAI Q J, QI H B. A novel defined pyroptosis-related gene signature for predicting the prognosis of ovarian cancer[J]. Cell death discovery, 2021,7(1): 71-82.
- [40] ZHANG N Y, BI Z, LIANG X Z, et al. OntoProtein: protein pretraining with gene ontology embedding[EB/OL]. (2022-06-03)[2025-10-25].<https://arxiv.org/abs/2201.11147>.
- [41] GODBOLD G, PROESCHER J, GAUDET P. New and revised gene ontology biological process terms describe multiorganism interactions critical for understanding microbial pathogenesis and sequences of concern[J]. Journal of biomedical semantics, 2025,16(1): 4-13.
- [42] CIHAN P, OZGER Z B, CAKABAY Z. Computational analysis of virus-host protein-protein interactions using gene ontology and natural language processing[J]. Applied intelligence, 2025,55(6): 393-415.
- [43] MIAO Y W, GUO Y Z, MA H H, et al. GoBERT: gene ontology graph informed BERT for universal gene function[EB/OL]. (2025-01-03)[2025-10-22].<https://doi.org/10.48550/arXiv.2501.01930h>.
- [44] CONSORTIUM G O, ALEKSANDER S A, BALHOFF J, et al. The gene ontology knowledgebase in 2023[J]. Genetics, 2023,224(1): 31.
- [45] CONSORTIUM G O. The gene ontology resource: enriching a Gold mine[J]. Nucleic acids research, 2021,49(D1): 325-334.
- [46] CASTELLS P, FONCILLAS B, LARA R, et al. Semantic web technologies for economic and financial information [C]//The Semantic Web: Research and Applications. Berlin:Springer, 2004:473-487.
- [47] KONTOPOULOS E, VRAKAS D, KOKKORAS F, et al. An ontology-based planning system for e-course generation[J]. Expert systems with applications, 2008,35(1-2): 398-406.
- [48] BERTAUD-GOUNOT V, DUVAUFERRIER R, BURGUN A. Ontology and medical diagnosis[J]. Informatics for health and social care, 2012,37(2): 51-61.
- [49] KELLER R M. Ontologies for aviation data management[C/OL]//Ontologies for aviation data management.Piscataway:IEEE,2016[2025-10-20].<https://ieeexplore.ieee.org/document/7777971>.DOI: 10.1109/DASC.2016.7777971.
- [50] ABITEBOUL S. Querying semi-structured data[C]//Database Theory : ICDT '97. Berlin:Springer,2005:1-18.
- [51] DING L, FININ T, JOSHI A, et al. Swoogle: a search and metadata engine for the semantic web[C]//Proceedings of the 2004 ACM CIKM International Conference on Information and Knowledge Management.NewYork:ACM,2004:8-13.
- [52] LOPEZ V, PASIN M, MOTTA E. AquaLog: an ontology-portable question answering system for the semantic web[C]//The Semantic Web: Research and Applications. Berlin:Springer,2005:546-562.
- [53] LOPEZ V, MOTTA E, UREN V. PowerAqua: fishing the semantic web[C]//The Semantic Web: Research and Applications. Berlin:Springer,2006:393-410.
- [54] CHENG G, GE W Y, QU Y Z. Falcons: searching and browsing entities on the semantic web[C]//Proceedings of the 17th International Conference on World Wide Web. NewYork:ACM,2008:21-25.
- [55] QU Y Z, HU W, CHEN G. Constructing virtual documents for ontology matching[C]//WWW '06: Proceedings of the 15th international conference on World Wide Web.NewYork: ACM, 2006:23-31.
- [56] OREN E, DELBRU R, CATASTA M, et al. Sindice.com: a document-oriented lookup index for open linked data[J]. International journal of metadata, semantics and ontologies, 2008,3(1): 37-52.

基于属性基加密的电力市场交易数据隐私保护算法

柳 茜¹
LIU Qian

摘 要

在电力市场交易数据的隐私保护中，单纯依赖添加数据噪声进行加密的方法受限于隐私预算，难以提供有效的隐私保护，且数据被成功攻击的风险较高。为此，文章提出了基于属性基加密的电力市场交易数据隐私保护算法。在基于电力市场交易数据敏感度添加噪声的全局设置基础上，采用属性基加密方法为数据生成独立的属性密钥，并通过动态更新机制为市场交易参与者分配访问权限，以实现数据的隐私保护。实验结果表明，该算法保护下，电力市场交易数据的被成功攻击概率仅为 0.46%，数据安全性较高，能够提供更加可靠的隐私保护。

关键词

数据隐私保护；交易数据；属性基加密；市场交易数据；电力市场；隐私保护

doi: 10.3969/j.issn.1672-9528.2025.11.036

0 引言

如今，电力市场的规模日益增加。如果市场交易过程产生的数据出现泄露，可能导致企业在庞大的电力市场竞争环境中呈现劣势，扰乱市场秩序，阻碍了电力市场的健康发展。因此，避免用户隐私数据泄露，保障数据安全这一问题成为电力市场交易的重要环节之一。然而，由于电力市场的交易数据往往涉及发电、售电、运营商等多方主体，这就意味着交易数据的隐私保护较为困难。在这样的情况下，为了保障电力市场交易数据的安全性，领域内针对数据隐私保护问题

展开了热烈讨论。

奚增辉等人^[1]提出了基于 Top-K 频繁挖掘的数据隐私保护方法。设置 Top-K，通过频繁挖掘分析数据的内在关联，确定数据的敏感度。结合差分隐私处理，生成数据的私有账本，实现数据的隐私保护。该方法处理的数据出现错误的概率较低。但在实际应用中，该方法容易出现隐私分配不合理的问题，导致数据噪声过多，数据信息出现失真，无法有效抵御外部攻击，隐私保护性能不理想。赵锋等人^[2]提出了基于本地差分隐私的数据隐私保护方法。解析原始数据流特征，以此去除数据冗余。通过计算隐私预算，在 Laplace 机制下，为原始数据流添加相应时间尺度的随机噪声，从而保护数据

1. 甘肃中网通讯技术有限公司 甘肃兰州 730000

[57] WANG H F, LIU Q L, PENIN T, et al. Semplore: a scalable IR approach to search the Web of data[J]. Journal of Web semantics, 2009,7(3): 177-188.

[58] DAMLJANOVIC D, AGATONOVIC M, CUNNINGHAM H. Natural language interfaces to ontologies: combining syntactic analysis and ontology-based lookup through the user interaction[C]//The Semantic Web: Research and Applications. Berlin:Springer,2010:106-120.

[59] DAMLJANOVIC D, AGATONOVIC M, CUNNINGHAM H. FREyA: an interactive way of querying linked data using natural language[C]//The Semantic Web: ESWC 2011 Workshops. Berlin: Springer, 2012:125-138.

[60] KLEIN D, MANNING C D. Fast exact inference with a factored model for natural language parsing[C]//NIPS'02:

Proceedings of the 16th International Conference on Neural Information Processing Systems. New York: ACM, 2003: 3-10.

[61] D'AQUIN M, MOTTA E. Watson, more than a semantic web search engine[J]. Semantic web, 2011,2(1): 55-63.

[62] DELBRU R, CAMPINAS S, TUMMARELLO G. Searching web data: an entity retrieval and high-performance indexing model[J]. Journal of web semantics, 2012,10: 33-58.

【作者简介】

韩凌(1981—)，男，江苏南京人，博士，研究方向：机械、信息检索、人工智能。

(收稿日期：2025-06-15 修回日期：2025-11-05)