智能商品数据管理体系：基于知识图谱的分布式存储优化与用户观点深度挖掘在电商领域的应用升级

陈佳佳 庞恒莉 胡秋霞 邵静雯 黄羽飞

广西职业师范学院工商管理学院，广西南宁，530007；

摘要：电商蓬勃发展，Web 商品数据量暴增。但是数据多源、异构、异质且稀疏，阻碍智能化管理。为解决多源异构数据融合难题，实现高效知识服务，本文构建基于知识图谱的商品数据规范管理框架，旨在将商品数据知识化、结构化存储并高效服务。

构建融合客观商品信息与主观用户评价的大规模异质知识图谱；设计高性能分布式存储及增量更新机制；研发人机协同的混合查询技术，应对高并发。通过从五大电商平台采集 150GB 数据（含 4000 万商品、6000 万评论）实验，结果表明该框架在知识融合效率、查询响应速度和用户观点识别准确率方面优势显著。分布式存储策略使查询延迟降低 40%，观点识别模型 F1 值达 89.7%，优于现有方法。本研究为商品知识服务智能化、实时化提供理论与实践支持。

关键词：知识图谱；商品数据管理；分布式存储；用户观点挖掘

DOI：10.69979/3041-0673.25.07.008

研究背景

在当下数字化浪潮中，电商行业一路狂飙，平台上的商品数据规模冲破 PB 级大关了。但随之而来的问题也不少，数据冗余严重，存储效率低；数据格式杂乱，难以统一管理；关键信息易缺失，使得数据很难被充分理解和利用。

传统数据库秉持“封闭世界假设”，在面对商品领域里不断变化的用户观点知识时，显得力不从心。在电商购物场景中，用户的主观评价对大家的购买决策影响极大，所以怎么把这些主观观点有效整合起来，就成了亟待解决的问题。

从研究现状来看，国内外在知识抽取、分布式存储和观点挖掘方面都取得了一定成果。但是，多源知识融合困难重重，为攻克这些难题，实现多源异构商品数据的高效融合，提升知识服务质量，我们开展了基于知识图谱的商品数据规范管理研究。打算构建能融合多源异构数据的知识图谱，提出高效又可扩展的管理框架，助力电商知识服务朝着智能化、精细化方向大步迈进。

1 方法

1.1 整体框架设计

本研究提出一种分层式系统架构，其核心框架共四个功能模块：

（1）多源异构数据采集层：

我们团队采用 Web 爬虫自动化抓取与规范化 API 接口接入相结合的技术范式，实现对结构化、半结构化及非结构化数据的全面获取。该层集成数据清洗、去重及格式标准化等预处理流程，为后续知识处理提供高质量数据基础。

（2）知识图谱构建层：

我们通过本体论指导下的语义消歧算法完成实体对齐，并运用基于规则推理与深度学习结合的属性融合策略。该层特别设计了异构数据源的动态映射机制，通过知识拓扑结构优化解决跨域语义冲突问题，构建具有强语义关联性的多维度知识网络。

（3）分布式存储与智能查询层：

采用支持横向扩展的多模态图数据库架构进行数据持久化存储，同时引入人机协同的查询优化机制。通过将专家规则引擎与群体智能算法相结合，建立动态负载均衡策略，有效协调分布式计算资源与众包人工校验的协作关系。

（4）多模态服务接口层：

提供标准化的 SPARQL 语义查询接口与基于深度学习的自然语言问答系统双重访问通道。该层特别集成上下文感知与意图识别模块，通过服务组合引擎实现查询请求的智能路由，支持从结构化查询到多轮对话的多样

化服务范式。

1.2 关键技术实现

（1）多源知识融合

第一个是实体对齐机制：在跨源知识融合过程中，通过构建三重相似度度量体系实现实体匹配。其一，在结构相似度计算层面，采用基于集合论的 Jaccard 系数量化属性集合的重合程度；其二，在文本语义相似度分析方面，运用预训练语言模型 BERT 生成深度上下文嵌入向量；与此同时，在主题分布相似度评估维度，借助潜在狄利克雷分布（LDA）模型提取主题概率特征。

第二个是众包知识协同整合机制：本研究提出一种基于群体智能的质量控制方法，通过构建多维度的标注者可靠性评估模型（包含历史标注准确率、任务响应一致性及专家信任度等指标），建立动态筛选机制。在知识融合阶段，采用基于 k/n 共识阈值的投票聚合算法，当且仅当超过预设比例（k/n）的独立标注者在语义空间达成收敛性共识时，才采纳该知识条目。

（2）分布式存储优化

第一个是 RDF 编码与分布式存储优化，我们团队为实现大规模知识图谱的高效存储与查询，提出基于哈希函数的编码压缩方法。该编码机制通过将 RDF 三元组的主客体进行位映射转换，实现三元组存储空间的优化。同时结合频繁子图挖掘技术（FP-Growth 算法），构建基于数据访问模式的垂直分块策略，通过识别高频关联子图结构，形成语义紧密的存储单元。

第二个是动态负载均衡机制，在数据分布层面，提出基于复杂网络分析的图聚类算法（Louvain 算法）驱动的分片策略。通过计算节点间拓扑结构的模块度指标，将强关联的知识子图聚合为同构数据块。

（3）目标依赖观点识别

第一个是联合语义表征学习框架，本研究提出一种基于深度神经网络的多模态联合嵌入方法，其核心架构采用双向长短期记忆网络（BiLSTM）作为编码器基座。该模型通过并行编码观点目标实体及其上下文语境的词向量表示，并设计目标导向的注意力机制（Target-oriented Attention），在隐藏状态序列中动态计算注意力权重分布，从而精准捕获目标词与语境特征间的语义关联模式。特别地，我们通过引入多头自注意力（Multi-head Self-attention）子模块，实现不同抽象层级特征交互的显式建模，有效增强模型对长距离依赖关系的捕捉能力。

第二个是跨领域情感分析模型，在情感分类任务中，构建具有领域适应能力的深度神经网络架构。我们团队输出层采用 Softmax 函数将高维特征映射至情感极性概率空间，同时创新性地融合领域对抗训练（Domain Adversarial Training）与最大均值差异（MMD）损失函数。该复合损失函数通过特征分布对齐策略，约束源域与目标域在共享隐空间中的表示相似性，从而缓解领域偏移（Domain Shift）导致的语义迁移障碍。除此之外，我们引入梯度反转层（Gradient Reversal Layer）实现领域判别器的对抗训练，使特征提取器生成领域无关的泛化表征，显著提升模型在跨领域场景下的分类鲁棒性。

2 实验

2.1 数据集与实验环境

在本研究中，我们构建了一个大规模的商品数据集，数据来源于亚马逊、沃尔玛和 eBay 等主流电子商务平台。这些平台是全球范围内用户广泛使用的电商网站，涵盖了丰富的商品类别和大量的用户评论。通过爬取和整合，数据集总量达到 150GB，涵盖了 5 万个商品概念及 6000 万条用户评论。这些多样化的数据为知识图谱的构建和商品数据管理提供了坚实的基础，使我们能够从多个维度分析和理解商品与用户之间的复杂关系。

为了高效处理和分析如此大规模的数据，我们设计并部署了一个高性能的分布式集群环境。该集群由 8 个节点组成，每个节点配备 256GB 内存和 NVMe SSD 存储。NVMe SSD 存储提供了极高的 I/O 性能，能够有效支持大规模数据的快速读写操作，而分布式架构则确保了计算任务的高效并行处理。

这一实验环境将能够支持更加复杂和智能的商品数据管理应用，为用户提供更加个性化和高效的购物体验。我们的数据集和实验环境为基于知识图谱的商品数据管理研究提供了强大的支持。

2.2 评估指标与对比方法

（1）知识融合效率：

评估指标：知识融合的效率主要通过 F1 值和对齐耗时来评估。F1 值是准确率和召回率的调和平均数，能够综合反映知识融合的精确性和完整性。对齐耗时则衡量了知识融合过程中所需的时间，时间越短，效率越高。

对比方法：为了验证知识融合的效率，研究采用了

Dedupe 和 Silk 框架作为对比方法。Dedupe 是一种用于数据去重和记录链接的工具，能够有效处理重复数据。Silk 框架则是一个专门用于数据集成和链接的框架，支持复杂的数据匹配和融合任务。通过对比这些方法，可以更好地评估知识融合的效率。

（2）查询性能：

评估指标：查询性能主要通过响应时间（QPS，每秒查询数）来评估。QPS 反映了系统在单位时间内能够处理的查询请求数量，是衡量系统查询性能的重要指标。响应时间越短，QPS 越高，系统的查询性能越好。

对比方法：为了评估查询性能，研究采用了 gStore 和 RDF-3X 作为对比方法。gStore 是一个基于图结构的 RDF 数据存储和查询系统，能够高效处理复杂的图查询。RDF-3X 则是一个高性能的 RDF 存储和查询引擎，以其快速的查询响应时间著称。通过对比这些系统，可以全面评估查询性能。

（3）观点识别准确率:

评估指标：观点识别的准确率主要通过 F1 值和 AUC 来评估。F1 值综合了准确率和召回率，能够全面反映观点识别的精确性和完整性。AUC（Area Under Curve）则是衡量分类模型整体性能的指标，AUC 值越高，模型的分类性能越好。

对比方法：为了验证观点识别的准确率，研究采用了 LSTM、RNTN 和 SentiWordNet 作为对比方法。LSTM（长短期记忆网络）是一种常用的深度学习模型，擅长处理序列数据。RNTN（递归神经张量网络）则是一种用于处理树结构数据的深度学习模型。SentiWordNet 是一个基于词汇的情感分析工具，能够提供词汇的情感极性信息。

2.3 实验设计

实验设计部分通过一系列精心设计的实验，全面验证了系统在实体对齐、分布式存储和观点识别等方面的有效性和性能。这些实验不仅为系统的优化提供了数据支持，还为实际应用中的性能表现提供了可靠的评估依据。

首先，实体对齐实验是研究中的重要环节。实体对齐是知识图谱构建和知识融合的核心任务之一，其目标是将不同来源的实体进行匹配和链接，以形成统一的知识表示。为了验证混合特征模型的有效性，研究从数据集中随机采样了 10 万对实体进行实验。实验结果表明，混合特征模型在处理大规模实体对齐任务时表现出色，能够有效提升对齐的准确率和召回率，为知识融合提供了坚实的基础。

其次，分布式存储压力测试是评估系统性能的重要实验之一。随着数据规模的不断增长，传统的集中式存储系统已经无法满足高效查询的需求，分布式存储系统因其高扩展性和高并发处理能力而成为主流选择。结果表明，合理的分块策略能够显著降低查询延迟，提高系统的并发处理能力。

最后，观点识别跨领域验证实验旨在评估模型在不同领域中的泛化能力。实验采用了多种评估指标，包括 F1 值和 AUC（Area Under Curve），以衡量模型在不同领域中的表现。结果表明，模型在多个领域中均能保持较高的识别准确率，证明了其在实际应用中的广泛适用性。这一实验不仅验证了模型的鲁棒性，还为跨领域应用提供了有力支持。

3 讨论

本研究构建的商品数据规范管理框架在多方面取得显著成果，但仍存在可优化与拓展的空间。在知识融合环节，尽管提出的融合策略有效整合了多源数据，但电商数据持续增长，来源更加多样复杂，现有的实体对齐和属性融合方法面临更高挑战。

分布式存储优化虽提升了查询效率，但在数据规模急剧膨胀时，网络传输瓶颈可能凸显。目前的负载均衡策略在极端高并发情况下，对资源分配的动态调整还不够精细，可能影响部分查询的响应速度，需要进一步优化负载均衡算法，提高系统在超大规模数据和超高并发场景下的稳定性。

目标依赖观点识别技术在多领域表现良好，但面对语义模糊、情感隐晦的评论，仍存在误判可能。此外，在实际应用中，如何将观点识别结果更精准地转化为营销策略和产品优化建议，还需深入探索。

4 结语

本研究围绕“基于知识图谱的商品数据规范管理研究--多源异构数据融合与高效知识服务实现”展开。针对电商商品数据多源、异构的特点及对高效知识服务的需求，构建了创新的商品数据知识图谱管理框架。

在多源异构数据融合方面，设计了融合策略，整合电商平台商品基础信息、产品说明书专业数据及用户评价反馈，打破数据源壁垒，丰富了知识图谱内容，拓展

覆盖范围，为知识服务奠定数据基础。

为了实现高效知识服务，优化分布式存储。通过改进存储结构与算法，合理分块、分布式部署数据，提升存储与读取效率，快速响应海量数据查询，降低延迟，有力支撑电商平台高并发请求。

通过真实电商数据集验证，本管理框架在知识融合效率、查询响应速度、观点识别准确率等关键指标上都有优于现有技术，为电商智能化管理提供了可行且高效的方案。

展望未来，鉴于数据安全与隐私保护的重要性，计划在学习框架下探索隐私保护知识共享机制，实现不同平台或机构间的安全知识共享与协作，拓展商品数据知识图谱跨平台应用，推动电商行业智能化发展。

综上，本研究在多源异构数据融合与高效知识服务实现方面成果显著，为电商商品数据规范管理提供了新思路和方法，对推动电商行业智能化发展具有理论与实践意义。

参考文献

[1]张芳,吕鑫.我国电商供应链研究现状与热点分析——基于 Citespace 的知识图谱分析[J].中国物流与采购,2025,(01):119-120.DOI:10.16079/j.cnki.issn1671-6663.2025.01.068.

[2]唐瑞锋.大数据在供应链管理领域的研究热点及其演化——基于 CiteSpace 的知识图谱分析[J].物流工程与管理,2023,45(11):79-82.

[3]谢佳洋,谢瑜,靖富营.中国企业高质量发展概念内涵、研究热点与前沿——基于 CiteSpace 可视化知识图谱分析[J].电子科技大学学报(社科版),2024,26(04):63-72.DOI:10.14071/j.1008-8105(2024)-4006.

作者简介：陈佳佳.2004,女，汉族,河南淮滨,在读管理学学士,本科在读，研究方向为大数据管理与应用

基金资助：国家级大学生创新创业训练计划项目资助,项目号 202414684006。