基于知识图谱的商品数据规范管理研究：面向多源异构数据的动态建模和智能应用

陈佳佳 庞恒莉 胡秋霞 邵静雯 黄羽飞 李寒蕾(通讯作者)

广西职业师范学院工商管理学院，广西南宁，530007；

摘要：当前的全球电商市场规模已经突破 5 万亿美元、商品 SKU 达 10 亿级，但是海量商品数据却存在多源异构、语义模糊、更新滞后等问题。传统关系型数据库具有刚性模式限制，难以适配商品复杂关联关系与高效管理需求，为此本研究构建以知识图谱的商品数据规范管理体系为基础。这个体系通过商品领域本体模型明确数据关联、动态实体对齐算法解决多源匹配、数据清洗规则库提升质量，最终形成支持多维度关联推理的架构，实现商品数据的知识化表示、结构化存储与高效服务。

本次实验主要以京东 200 万商品数据、淘宝 150 万爬取数据及 6000 万条用户评论为基础（覆盖 6 大品类），对比传统 MySQL 与 ElasticSearch 方案验证有效性。结果显示，本方法核心性能指标显著优于现有的方案。

该体系可为电商智能搜索、供应链优化等上层应用提供知识支撑，同时丰富垂直领域知识图谱理论，为商品数据规范管理提供可行技术方案，具有重要理论与实践价值。

关键词：知识图谱；商品数据治理；多源异构数据融合；动态实体对齐

DOI：10.69979/3041-0673.26.02.001

1 研究背景

随着数字经济发展，电商行业成为全球贸易重要支撑，当前全球市场规模已突破 5 万亿美元，商品数据呈现爆炸式增长。面对现有问题，解决方案局限性非常明显。

传统的关系型数据库受到刚性模式约束，难以适配商品数据多源异构、动态变化特点，处理商品与品牌、品类等复杂关联时需要频繁表连接，查询效率低，无法满足高并发需求；规则引擎虽然可以基础校验数据，但是维护成本随品类增长而增加，面对新增商品与格式，规则迭代跟不上需求；联邦学习虽护隐私，却因为多参与方频繁数据交互导致效率下降，难以平衡隐私与效率。另外，现有的知识图谱多聚焦客观事实，忽视了用户主观观点整合，而用户观点对消费决策与平台优化至关重要，这会导致知识服务不全面。

相比之下，知识图谱技术为解决这些难题提供了新路径。它的属性图与超图模型能清晰刻画商品多维度关联，既描述 SKU、品牌等基础属性，又关联用户观点、竞品等，能够实现深度语义整合；而基于开放世界假设支持增量学习，新增数据或品类无需重构模型，仅需增量更新，从而降低管理复杂度；还能整合文本、图像、时空等多模态数据，为商品数据全面分析与智能应用打下良好基础。

2 研究方法

2.1 商品领域本体模型构建

在商品数据的标准化表示工作中，为了让数据能够更规范、统一地被描述和使用，同时为后续知识图谱的搭建打下扎实基础，本研究专门设计了一套商品领域本体模型，并通过这套模型明确了商品知识体系中的核心类、关键属性以及核心关系。

这套本体模型始终围绕“商品”这一核心类展开构建——所有属性与关系的设计都以“精准刻画商品特征、理顺商品间逻辑”为目标，在属性层面具体分为基础属性和扩展属性两大模块，两者各有侧重又相互补充，共同完善商品的信息维度。

其中，基础属性的设计重点在于保障商品信息的“基础性”和“准确性”，主要涵盖三个关键维度：一是 SKU 编码，作为每一个商品实体的唯一标识，它就像商品的“身份证”，能够精准区分不同规格、不同批次的独立商品，避免出现信息混淆；二是三级分类树，

通过“大类—中类—小类”的层层细化逻辑（例如从“电子产品”这个大类，到“手机”这个中类，再到“手机”下属的“智能手机”小类），让每一件商品都能清晰归位到对应的分类节点下，大幅提升分类结果的准确性；三是品牌实体链接，通过与模型中专门设计的“品牌类”建立关联，将商品与所属品牌的信息直接挂钩，既完善了商品的基础信息，也为后续品牌维度的分析提供了支持。

这三个基础属性从标识、分类、品牌三个层面相互配合，共同保障了商品实体的唯一性界定和分类结果的准确性。

而在扩展属性方面，考虑到不同品类商品的特性差异较大，且用户对商品的实际感知也很重要，本研究特意引入了“动态属性槽”和“UGC 标签云”两个设计，以弥补基础属性在灵活性和用户视角上的不足：动态属性槽的核心作用是“按需存储”不同品类商品的个性化规格参数——比如针对手机这类数码产品，会重点存储“屏幕尺寸”“处理器型号”“电池容量”等与产品性能强相关的参数；针对冰箱、空调等家电产品，则会存储“额定功率”“能效等级”“容量大小”等符合家电品类特性的参数，这种动态适配的设计让模型能更好地覆盖多品类商品；UGC 标签云则聚焦于“用户视角”，通过汇聚普通用户在使用商品后标注的标签（像用户评价中常提到的“续航强”“拍照清晰”“噪音小”“材质好”等），将用户对商品的直观感受和实际使用体验融入模型中，让商品的特征描述更贴近真实使用场景。

这两个扩展属性的加入，有效提升了模型对商品特性的全面刻画能力，也让商品信息更具实用性。

2.2 多源商品数据治理流程

在商品知识图谱的构建中，多源数据的治理是衔接“数据采集”与“知识生成”的关键环节——既要解决不同渠道数据的“杂乱性”，也要保证数据的“准确性”和“时效性”。我们围绕“采集-清洗-更新”三个核心步骤，设计了一套完整的多源商品数据治理流程，具体如下：

（1）多源数据采集：多渠道保障全面性与实时性

针对不同来源数据的特性，可以采用差异化采集策略：

品牌商数据：通过 API 对接官方数据库，获取标准化基础信息（如规格参数、生产厂家），为后续治理提供基准。

电商平台数据：用 Scrapy-Redis 分布式爬虫集群，爬取淘宝、京东等平台的商品详情页、价格及竞品数据，应对千万级商品量级需求。

（2）数据清洗与冲突消解：分类处理数据矛盾

针对多源数据冲突，按类型设计消解策略：

数值型冲突（如价格、重量差异）：采用加权平均法，按来源可信度分配权重（厂商 0.7、平台 0.2、用户 0.1），平衡准确与客观。

文本型冲突（如商品描述差异）：基于知识投票机制，统计多源文本中相同表述的频率，以多数表述为最终结果。

（3）动态更新机制设计：事件驱动实现实时迭代

为避免固定周期更新的滞后性，设计事件驱动型机制：

价格波动事件：设定 5%变动阈值，价格超阈值时触发实时更新，同步知识图谱中的价格属性。

舆情事件：通过情感分析评估评论影响力，高影响力评论（如点赞超 100 的负面评论）优先更新，及时把握用户反馈。

2.3 知识融合与分布式存储优化

（1）改进的动态语义哈希融合算法

在多源商品知识融合时，最关键的问题是怎么准确判断不同来源的“商品实体”是不是同一个——如果只看语义或者只看属性，很容易出现匹配偏差，而且旧数据的参考价值会随时间慢慢降低。所以我们提出了改进版的 DSH（DynamicSemantic Hashing）算法，把语义相似度、属性相似度和时间衰减这三个关键因素结合起来，综合计算实体间的相似度。

具体计算时，算法公式里的各项各有作用：$\alpha \cdot \cos(v_1,v_2)$对应的是基于词向量的语义相似度，比如“智能手机”和“智能手機”（不同平台的表述差异），通过词向量余弦值能判断它们的语义关联度，这里的 α 是这部分的权重系数；$\beta \cdot Jaccard(c_1,c_2)$则是基于属性集合的 Jaccard 相似度，比如两款手机的“屏幕尺寸”“处理器型号”等属性重合度，β 就是属性维度的权重系数；

还有γ·temporal_decay(t)这个时间衰减项，主要是为了弱化旧数据的影响——它的计算公式是 temporal_decay(t)=0.2e^(-0.1Δt)，其中Δt是不同来源数据的更新时间差，比如 A 渠道数据是昨天更的，B 渠道是上个月更的，Δt越大，这一项的数值就越小，γ则是时间维度的权重系数。这里要注意，α、β、γ三个系数加起来等于 1（α+β+γ=1），保证整体相似度计算的合理性。

（2）查询工作量敏感的分布式存储

我们用 RDF（资源描述框架）把商品知识拆成<主体，谓词，客体>的三元组形式，比如“<智能手机，属于，电子产品>”“<某手机，屏幕尺寸，6.7 英寸>”，这种结构清晰，也方便后续做查询和推理。但遇到一些涉及多个实体依赖的事实时，常规三元组就不够用了——比如“用户 A 对商品 B 的屏幕给出好评”，这里同时牵扯到用户、商品、商品部件三个实体，直接用普通三元组没法完整表达关系。所以我们引入了“超实体”的概念，把刚才这个例子重新表示成<超实体（用户 A,商品 B,屏幕），opinion_type，positive>，这样就能准确描述多实体间的关联逻辑。

为了让查询效率更高，我们采用了“垂直软划分”策略：把满足同一频繁查询模式的子图存在同一个数据分块里。比如所有和“手机用户评价”相关的三元组都放一个分块，这样用户查这类信息时，不用跨好几个分块找数据，能减少分块间的连接操作，速度自然快。

另外，我们还根据查询模式的“相关紧密度”设计了图聚类算法，将关联度高的两类子图对应的分块分配到同一个存储节点。这样一来，节点之间不用频繁传递数据，通信开销少了，整个分布式查询的吞吐量也能提上去。

2.4 关键技术实现

在实体识别与链接上，我们采用 BERT-BiLSTM-CRF 联合抽取模型：BERT 负责挖掘文本深层语义特征，BiLSTM 捕捉序列依赖关系，CRF 通过上下文标注约束，实现商品实体与属性的精准抽取。针对新品类冷启动问题，引入少样本学习技术，用已有品类标注数据训练迁移模型，提升对未见过品类的识别能力。

知识推理结合符号推理与向量推理：符号推理基于 Datalog 规则引擎，定义“商品品类从属传递”、“属性值范围校验”等规则，完成确定性推理；向量推理采用 RotatE 算法，将商品实体与关系映射到复向量空间，通过向量旋转挖掘潜在关联，比如推荐与用户好评商品相似的竞品。

3 实验

3.1 数据资源与实验平台

本研究采用多源异构商品信息构建实验数据集，数据采集范围覆盖全球主流电商平台。通过分布式爬虫系统获取的商品数据总量达 150GB，涵盖 5 万余种商品品类及超过 6000 万条用户反馈信息。数据集包含多维特征：除基础商品属性外，还整合了用户评分数据、评论内容、消费行为轨迹及商品关联网络。这种多模态数据结构为构建商品知识图谱提供了多角度分析视角，有助于深入探究商品与用户间的复杂关联。

在数据预处理环节，我们建立了完整的数据质量控制流程，采用多级数据清洗管道对原始数据进行标准化处理。通过基于规则引擎与机器学习算法的混合去噪方法，有效识别并剔除无效记录与异常数据。随后将规整化数据存入分布式文件系统，采用列式存储格式优化数据访问模式。为提升处理效能，创新性地实现了动态数据分片机制与弹性计算资源调度策略，通过智能负载均衡技术将计算任务合理分配至集群节点。

实验平台建设特别注重系统可靠性与可维护性，通过容器编排技术实现服务的高可用部署。同时构建了全方位的监控告警体系，实时追踪集群资源利用率与服务健康状态，建立预测性维护机制确保系统持续稳定运行。

3.2 性能评估体系与基准对比

为系统评估知识图谱在商品数据管理中的综合性能，建立多维评估指标体系与对比分析框架：

知识融合效能评估采用融合精度与时延指标双重衡量。基准对比选取了当前主流的数据匹配与链接框架，通过控制变量实验验证知识融合方案的性能优势。评估过程充分考虑了电商领域的数据特性，设计了面向多源异构数据的测试用例。

查询响应性能测试以查询吞吐量与响应延迟为核心指标，构建了多场景查询负载模型。基准系统选取了

两种不同架构的存储引擎，分别代表图原生存储与优化关系存储技术路线。测试过程采用渐进加压策略，系统评估了各系统在不同数据规模下的性能表现。

观点挖掘准确度验证采用多维度评估策略，综合考量模型的精确度与鲁棒性。基准模型选取涵盖了传统神经网络、深度语义模型和基于词典的方法，通过对比实验验证观点识别模型在准确率、召回率及泛化能力方面的综合表现。

3.3 实验方案与验证过程

通过设计系统化实验方案，验证各模块在真实场景下的性能表现：

实体对齐验证实验针对知识融合核心环节，设计了基于多特征融合的匹配模型。实验采用分层抽样方法构建测试数据集，通过特征加权机制优化实体相似度计算。测试结果验证了混合特征模型在复杂实体匹配场景下的有效性，为知识图谱的构建质量奠定基础。

存储系统压力测试设计了多轮渐进式负载实验，模拟真实业务场景下的查询模式。通过对比不同数据分布策略下的系统性能指标，探索最优存储配置方案。实验结果表明，基于语义关联的数据分片策略可有效提升系统并发处理能力。

4 讨论

4.1 研究成果的创新与价值

本研究的创新点主要体现在三个方面：一是构建了“属性-关系-事件”三维商品知识表示范式，将客观性商品事实与主观性用户观点深度融合，突破了现有知识图谱仅聚焦事实知识的局限，丰富了商品知识的维度；二是提出查询工作量敏感的分布式存储策略，通过挖掘搜索引擎日志中的频繁查询模式，采用“垂直软划分”将同类查询相关的子图存储于同一数据分块，并结合图聚类算法将关联紧密的分块分配至同一节点，减少跨节点数据交互，解决了传统存储架构在高并发场景下的性能瓶颈；三是设计事件驱动的动态更新机制，结合阈值触发与情感影响力评估，实现商品知识的实时迭代，保障了数据的时效性，避免因固定周期更新导致的信息滞后。

在实践价值方面，研究成果可直接应用于电商平台的多个核心场景：智能搜索中，基于知识图谱的语义理解能提升搜索结果相关性；供应链优化中，通过商品品类关联度分析可预测爆品组合，指导库存备货；消费者洞察中，用户观点知识的深度挖掘能帮助平台与商家精准把握用户需求。在理论价值方面，本研究为垂直领域知识图谱的构建提供了可复用的方法论，尤其是多源异构数据融合、动态更新机制的设计，可为医疗、金融等领域的知识图谱研究提供参考。

4.2 研究的局限性

尽管本研究取得了一定成果，但仍存在三方面局限性：一是数据集覆盖范围有限，实验数据主要来源于京东、淘宝等国内主流电商平台，对跨境电商的多语言数据、直播电商的特殊数据类型覆盖不足。跨境场景中，不同语言的语义差异可能导致实体对齐准确率下降，而直播电商的实时数据若无法及时整合，会使知识图谱难以反映“直播专属优惠”“限时库存”等动态信息，可能影响研究成果在跨境、直播电商场景的适用性。二是自动化程度有待提升，知识图谱的本体初始构建需人工定义核心类、关键属性及关系，冲突消解规则的制定也需人工校准；面对新增的小众商品品类，本体扩展需人工补充“材质工艺”“定制周期”等特殊属性，规则适配也需人工调整，自动化能力不足导致知识图谱维护成本较高。三是超大规模数据处理能力有待优化，实验基于约 8000 万条商品相关数据验证了性能优势，但当商品数据规模达到百亿级 SKU 时，现有分布式存储架构的分块策略会导致跨分块查询的连接开销增加，查询延迟可能从毫秒级变为秒级，且高并发更新场景下，数据写入吞吐量不足，难以满足极端高并发场景的需求。

4.3 未来研究方向

针对上述局限性，结合电商行业发展趋势，未来可从三个方向深化研究：一是拓展多模态与多语言数据处理能力，研究基于多语言 BERT 的跨语言实体对齐算法，构建多语言商品属性词典，解决“手机内存”与“mobilephonestorage”等表述的语义映射问题；针对直播电商数据，开发实时话术解析模型，提取直播中的商品卖点、限时优惠等信息，同时引入 CLIP 等视觉-语言预训练模型，将商品主图、直播视频帧中的视觉特征转化为结构化知识，实现对跨境电商、直播电商数据的有效

管理，提升研究成果的场景适配性。二是提升自动化与智能化水平，引入 GPT-4 等大语言模型，输入小众品类的商品描述文本，自动生成该品类的核心属性与关联关系，实现本体模型的无人工干预扩展；同时构建基于大模型的智能冲突消解系统，通过输入冲突数据与业务规则，让模型自主判定优先级，并结合强化学习持续优化，减少人工干预，降低知识图谱的维护成本。三是优化超大规模数据处理架构，探索基于分布式图数据库的分层存储策略，将高频访问数据存储于内存节点，低频数据存储于冷存储，平衡查询速度与存储成本；同时研究联邦学习与知识图谱结合的隐私保护方案，在多平台数据不共享原始数据的前提下，通过联邦训练实现跨平台实体对齐与知识融合，既保障数据安全，又能整合多平台商品知识，提升知识图谱的完整性。

5 结语

本研究围绕电商领域商品数据规范管理的核心难题，以知识图谱技术为核心支撑，构建了涵盖本体建模、数据治理、知识融合与分布式存储的完整管理体系。通过真实电商数据的实验验证，这个体系在数据一致性、查询效率、动态更新能力及实际应用效果上都展现出明显优势，不仅有效解决了商品数据多源异构、语义模糊、更新滞后等行业痛点，还为电商平台的智能搜索、精准推荐、供应链优化等上层应用提供了高质量的知识支撑，同时进一步丰富了垂直领域知识图谱的理论研究与实践应用体系。

当然，本研究仍存在一定局限性，比如数据集覆盖范围有待拓展，知识图谱构建与维护的自动化程度也需要提升。不过，未来可通过多模态数据处理、智能化优化、超大规模存储架构优化等方向的深入探索，持续完善这一基于知识图谱的商品数据规范管理体系。相信随着技术的不断发展，该体系将会在电商行业数字化转型中发挥出更大价值，从而推动商品数据管理向更智能、更高效、更安全的方向发展，为数字经济时代电商行业的高质量发展注入动力。

参考文献

[1]方创新.面向大规模商品知识图谱的查询处理技术[D].桂林电子科技大学,2023.DOI:10.27049/d.cnki.ggldc.2023.000499.

[2]李涛,王元卓,靳小龙.面向大规模知识图谱的表示学习技术研究[J].计算机研究与发展,2019,56(8):1581-1596.

[3]刘玮,王国仁,杜小勇.分布式图数据存储与查询技术研究[J].软件学报,2020,31(6):1689-1706.

[4]李翠平,王珊,刘胤.知识图谱实体对齐技术研究综述[J].计算机科学与探索,2022,16(5):985-1002.

[5]王宏志,李建中,高宏.多源数据融合技术研究进展[J].计算机学报,2018,41(3):587-610.

[6]张晓军,陈群,李战怀.海量数据存储系统架构研究综述[J].计算机研究与发展,2017,54(2):229-248.

[7]朱金沛.针对大宗商品知识图谱的实体识别算法[D].北京邮电大学,2023.DOI:10.26969/d.cnki.gbydu.2023.003188.

[8]钱浩东.基于社交媒体数据知识图谱的商品推荐算法研究[D].西安电子科技大学,2024.DOI:10.27389/d.cnki.gxadu.2024.003959.

[9]铁永正.基于知识图谱的产品概念设计研究与应用[D].西安工业大学,2025.DOI:10.27391/d.cnki.gxagu.2025.000043.

[10]程梦清.基于用户画像和知识图谱的混合推荐模型研究[D].江西财经大学,2024.DOI:10.27175/d.cnki.gjxcu.2024.001852.

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：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。