

基于语义Web的旅游数据语义集成与标注技术研究

古圣钰^{1,2}, 章牧²

¹惠州学院地理与旅游学院, 广东省 惠州市

²暨南大学深圳旅游学院, 广东省 深圳市

发布日期: 2025年12月1日

摘要

随着文旅融合的持续推进, 旅游行业面临数据来源多元、结构异构、语义割裂等复杂挑战。语义Web技术通过构建本体、RDF建模、语义映射与逻辑推理, 提供了一种面向知识层的数据集成与标注解决方案。本文系统回顾了旅游语义数据集成与标注的关键技术路径, 包括数据抽取、转换映射、实体对齐与冲突解决机制; 比较了规则驱动、机器学习与深度学习标注方法, 评估了典型标注工具在结构化与非结构化旅游数据中的适用性。结合行业实践, 分析了当前面临的性能瓶颈、数据质量波动与语义演化难题, 提出融合大语言模型、增量本体演化、语义隐私保护等未来研究趋势。本文旨在为智慧旅游系统建设与语义化数据治理提供技术支撑与参考框架。

关键词

语义Web; 旅游知识图谱; 本体映射; 语义标注; 大语言模型; 数据融合

Semantic Integration and Annotation Techniques for Tourism Data Based on the Semantic Web

Shengyu Gu^{1,2}, Mu Zhang²

¹School of Geography and Tourism, Huizhou University, Huizhou 516007, Guangdong, China

²Shenzhen Tourism College, Jinan University, Shenzhen 518053, Guangdong, China

Abstract

As culture and tourism integration accelerates, the tourism industry is increasingly challenged by heterogeneous data sources, inconsistent structures, and semantic fragmentation. Semantic Web technologies-through ontology construction, RDF modeling, semantic mapping, and logical reasoning-offer a knowledge-driven approach to data integration and annotation. This paper com-

文章引用: 古圣钰, 章牧. (2025). 基于语义 Web 的旅游数据语义集成与标注技术研究. 计算机科学辑要, 1(2), 11 - 20.

ISSN Print: 3079-5362; ISSN Online: 3079-5370.

DOI: 10.63313/CS.8018

prehensively surveys key techniques for semantic tourism data integration and annotation, including data extraction, transformation, ontology alignment, and conflict resolution. It contrasts rule-based, machine learning, and deep learning annotation paradigms, and evaluates mainstream tools for structured and unstructured tourism data processing. The study further identifies technical bottlenecks in scalability, data quality, and semantic drift, proposing future directions such as LLM-assisted semantic annotation, incremental ontology evolution, and privacy-preserving reasoning. This research aims to provide a technical foundation and strategic reference for intelligent tourism systems and semantic data governance.

Keywords

Semantic Web; Tourism Knowledge Graph; Ontology Mapping; Semantic Annotation; Large Language Models; Data Integration

Copyright © 2025 by author(s) and Erytis Publishing Limited.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



1. 引言

随着数字经济发展与文旅融合加速,旅游业成为典型的数据密集型行业。旅游相关数据呈现多源异构、结构复杂与语义分散等特征,既包括政府、OTA平台、地理系统与社交网络的结构化数据,也包含用户行为日志、网页文本等非结构化信息。这类数据在命名粒度、实体表达与语义描述方面差异显著,致使语义对齐困难、共享能力不足,严重制约了智慧旅游服务的落地[1]。语义Web技术以本体构建为核心,结合RDF、OWL、SPARQL等语言标准,具备良好的语义互操作性与逻辑可解释性。其可通过本体建立术语统一解释机制,在多系统之间实现语义桥接,有效支撑跨平台的数据融合与知识挖掘[2]。相较于传统数据接口或关键词匹配方式,语义Web集成强调从概念层进行关联映射,更适用于应对旅游数据的更新频繁与语义漂移问题。尽管国内外已开发多个旅游语义系统,如Hontology、ONTOUR本体框架与Karma、OpenRefine等标注工具,整体而言仍存在方法分散、缺乏统一规范、跨系统复用性弱等问题。特别是在语义标注、数据映射与实体对齐方面,缺乏系统化技术归纳与实证评估,限制了语义技术在旅游领域的广泛推广。基于此,本文拟围绕以下方面开展综述:(1)梳理旅游数据语义集成的理论模型与支撑体系,明确其与知识图谱和Linked Data的关系;(2)系统评析主流的抽取、映射与融合方法及其适用性;(3)比较规则驱动与学习驱动的语义标注路径,分析典型工具功能;(4)识别当前面临的核心挑战,探讨未来如大语言模型(LLMs)、语义演化与隐私保护等融合趋势。全文结构如下:第2章阐述旅游语义集成的理论基础;第3章综述关键技术路径;第4章聚焦语义标注方法与平台;第5章分析存在问题;第6章提出未来研究方向;第7章总结全文。本研究旨在为旅游数据语义治理提供系统化框架与理论支撑,亦期望为智能旅游系统构建提供参考依据与工程启发。

2. 旅游数据语义集成的理论基础

随着语义Web技术逐步在旅游信息系统中得到推广应用,如何构建一套可扩展、可解释、具有互操作性的旅游语义数据集成框架,成为当前研究与实践的关键课题。旅游数据的本质具有多源异构性和语义歧义性,其内容既包括结构化的酒店、景区信息数据库,也涵盖非结构化的用户评论、游记文本,语义一致性难以保障,传统数据集成方式难以满足智能化分析需求。因此,研究基于语义Web的数据集成

模型与机制, 对于实现多平台旅游知识协同具有重要现实意义[3]。

2.1. 语义数据集成模型概述

目前主流的语义数据集成模型主要包括 Global-as-View (GAV)、Local-as-View (LAV) 以及 Global-Local-as-View (GLAV) 三种。它们分别基于不同的建模思想来建立全局本体 (global schema) 与本地数据源之间的映射规则, 是旅游数据整合系统架构设计的基础。GAV (全局视图定义) 模式中, 数据源被直接映射为全局本体的组成部分, 其优势在于查询优化较为简洁, 适用于数据结构稳定、源系统更新频率低的旅游数据场景, 如国家级旅游资源库、标准化景点信息平台等。LAV (局部视图建模) 则反其道而行之, 由数据源自身定义其在全局本体中的映射, 强调数据源的自治性, 适用于数据更新频繁、异构性显著的旅游内容, 如社交媒体旅游数据、旅游电商用户行为等。GLAV (混合视图) 结合了前两者优点, 可实现灵活而高效的多源语义融合, 在多语种、多结构、多粒度旅游信息聚合任务中表现尤为突出[4]。通过上述模型选择, 旅游语义数据集成系统可根据实际数据环境进行有针对性地建模设计, 从而实现结构化语义映射与统一访问接口。

2.2. 本体映射与数据对齐机制

在数据集成过程中, 不同数据源对同一旅游概念可能存在表述差异、结构冲突或命名不一致等问题。为实现语义一致的集成, 必须引入本体映射与数据对齐技术。本体映射是指在两个或多个本体之间建立语义对等、包容、近似或从属关系的过程, 常用方法包括基于词汇层的词义相似度计算、基于结构层本体层次匹配, 以及引入上下文语境的语义相似推理[5]。而数据对齐 (Data Alignment) 更侧重于实例级的融合, 即将不同数据源中指代同一旅游对象 (如“广州塔”与“Canton Tower”) 的实体进行识别与链接, 进而构建统一知识表示。目前广泛应用的本体映射语言与机制包括: (1) RDFS/OWL 语义约束逻辑: 支持类层级定义与属性约束表达; (2) SPARQL 构造查询: 用于跨源规则抽取与语义转换; (3) R2RML、D2RQ 等映射语言: 面向结构化数据的语义转化工具, 在旅游关系型数据库向 RDF 图转化中广泛使用。本体映射与数据对齐协同工作, 是实现跨系统、跨结构旅游知识整合的核心基础。

2.3. Linked Data 与知识图谱在旅游领域的适用性

Linked Data 作为语义 Web 实践的重要实现路径, 其核心目标是通过统一资源标识 (URI)、RDF 三元组与标准语义链接, 实现开放数据的互联共享。Linked Data 在旅游场景中可用于连接多个异构数据源, 构建统一、互通、可查询的数据网络。例如, W3C 推荐的 Five-Star Open Data 标准体系已在多个国家级旅游开放平台中推广应用, 提升数据重用率与语义透明度[6]。基于 Linked Data 机制构建的旅游知识图谱 (Tourism Knowledge Graph, TKG), 能够在本体支撑下将景点、线路、评论、天气、交通等异构资源有机整合, 支持旅游问答系统、行程推荐系统、智慧导览服务等多种应用场景, 成为近年来研究与开发热点[7]。此外, 知识图谱还具备以下优势: (1) 利用语义规则推理用户偏好, 实现个性化旅游服务; (2) 支持语义检索与问答, 提升用户查询效率; (3) 提供数据可视化支持, 助力旅游知识管理与决策支持。

3. 旅游数据语义集成技术综述

3.1. 数据抽取: 从网页、数据库、文本中提取实体与属性

旅游数据的多源特性决定了数据抽取在语义集成中的基础性作用。所谓数据抽取, 是指从不同格式的数据源 (如网页 HTML、关系型数据库、XML 文件及非结构化文本) 中, 识别出与领域相关的实体 (如景点名称、地址、开放时间、用户评论) 及其属性, 并将其转化为结构化或半结构化的表示形式, 以便

后续的语义映射与融合。

3.1.1. 网页数据抽取

网页数据通常来源于在线旅游平台（OTA）、旅游局官网和景区门户等，这类数据多以 HTML 为主，嵌入内容多样且标签结构复杂。主流方法包括基于 DOM 树解析的模板抽取、基于 XPath 或 CSS 选择器的规则匹配，以及利用机器学习进行半自动抽取（如 Wrapper Induction）。例如，使用正则表达式与 XPath 规则结合，可以从携程、去哪儿等平台的网页中快速抽取酒店价格、评分等关键信息。然而，该类方法易受网页结构变化影响，需不断维护模板[8]。

3.1.2. 数据库与 API 抽取

对于存储于关系型数据库（如 MySQL、Oracle）的旅游数据，可通过 SQL 查询结合 R2RML 映射直接将表格数据转换为 RDF 三元组格式，实现结构化到语义化的转换。对于以 API 形式开放的数据源（如 Google Places API、TripAdvisor API），则可通过接口抓取数据后再进行本体化转换。此类方式结构清晰，抽取精度高，但依赖于数据源接口的可用性与访问权限。

3.1.3. 文本数据抽取

非结构化文本（如旅游评论、博客、游记）包含大量用户生成内容（UGC），蕴含丰富的情感与偏好信息。主流的文本数据抽取方法包括命名实体识别（NER）、依存句法分析与关系抽取，常借助深度学习模型（如 BERT、BiLSTM-CRF）进行语义识别。例如，通过 NER 技术可以从评论文本中提取景点名称、餐饮服务评价等语义单元，为后续知识图谱填充提供数据支撑[9]。

3.1.4. 主流数据抽取技术比较

表 1. 主流数据抽取技术比较表

技术类型	主要方法与工具	优势	局限性	典型应用场景
网页抽取	DOM 解析、XPath、Scrapy	规则可控、精度较高	模板需频繁更新，维护成本高	OTA 网页数据抽取
数据库/API抽取	SQL 查询、R2RML、D2RQ	结构化程度高、转换快捷	依赖接口权限与格式一致性	酒店/景区数据库抽取
文本抽取	NER、关系抽取、BERT、BiLSTM-CRF	语义能力强、适应复杂文本	模型训练成本高、需大量语料支持	用户评论、游记信息解析

3.1.5. 旅游语义数据抽取的综合策略

面对旅游数据多样性与动态性，单一抽取技术难以满足所有需求。实践中常采用多策略融合方法，即结合模板规则、机器学习与深度学习模型进行分层抽取。例如，可先用规则识别基本结构化数据，再用 NER 与上下文语义模型解析非结构化文本中的隐含信息。为提高抽取效率，部分研究还引入半自动标注与主动学习机制，通过小样本人工干预提升模型泛化能力。同时，随着大语言模型（LLM）的发展，利用 ChatGPT 等模型对旅游文本进行自动信息抽取成为新趋势。其优势在于可理解复杂语境、适应非标准化表达，然而其生成结果仍需结合本体规则进行语义校验，以确保抽取结果的精度与一致性。

3.2. 数据转换与本体映射：R2RML、D2RQ、Ontop 等

在完成旅游数据的抽取之后，为实现数据的语义互操作与语义统一表达，需进一步将原始数据转换为语义格式，并与领域本体进行映射对齐。该过程在语义 Web 体系中称为“RDF 化”或“本体映射”，其核心任务在于建立数据源与本体之间的映射规则，并生成具备语义解释能力的三元组数据。本体映射

不仅仅是字段对应或名称替换，而是涉及对数据结构、语义约束、类型系统等方面的深度理解与语义关系重建。当前主要的数据库转换与映射技术体系包括 R2RML、D2RQ 及 Ontop 等，它们在旅游数据知识化处理与图谱构建中被广泛使用[10]。

3.2.1. R2RML 映射语言

R2RML (RDB to RDF Mapping Language) 是 W3C 推荐的标准映射语言，用于将关系型数据库中的数据以声明式方式映射为 RDF 图结构。其基本单位为“Triples Map”，可灵活定义逻辑表、主键、外键与本体中的类和属性之间的对应关系，适合用于数据结构清晰、元数据定义良好的旅游数据库，如旅游景区名录库、酒店信息登记系统等。R2RML 支持 SQL 兼容表达式，便于复杂逻辑关系的建模，并可结合 SPARQL 查询语言实现高效语义查询。但其配置复杂度较高，对开发人员技术能力要求较强。

3.2.2. D2RQ 系统

D2RQ 是较早期的关系型数据库到 RDF 映射系统，支持通过自动分析数据库结构生成初步映射规则，并可作为虚拟 RDF 存储，允许在不复制原始数据的前提下进行 SPARQL 查询。其优势在于部署便捷、开箱即用，适用于旅游行业初期尝试语义转换的轻量场景。但 D2RQ 在复杂映射表达能力方面不如 R2RML，且支持的语义约束类型有限，不适用于高度规范化的语义数据发布系统。

3.2.3. Ontop 引擎

Ontop 是近年来在语义 Web 社区中广泛采用的本体映射平台，基于本体驱动的数据访问 (OBDA) 思想，将 SPARQL 查询动态转换为 SQL 语句，并结合 R2RML 扩展规则支持语义推理能力。Ontop 具有高性能查询优化能力，适用于大规模旅游数据库与知识图谱系统的实时访问与推理场景。此外，Ontop 兼容 RDF4J、OWLAPI 等语义处理框架，便于与现有旅游语义系统集成，支持本体一致性校验与多源映射融合，适合构建面向服务的旅游智能系统[11]。

表 2. 数据转换与映射方法概览

映射工具/语言	技术特点	优势	适用场景
R2RML	W3C 标准，基于 Triples Map 声明式映射	映射能力强，支持复杂结构转换	国家级旅游数据 RDF 化、知识图谱构建
D2RQ	自动生成映射规则，轻量级配置	快速部署，适合小规模数据接入	地方旅游局数据库 SPARQL 访问
Ontop	支持本体驱动访问、语义推理、实时查询	性能优、语义层集成能力强	智慧旅游平台、语义搜索服务系统

3.2.4. 旅游领域映射实践与融合趋势

在旅游实际应用中，通常需针对多个数据源构建一体化的本体映射机制。例如，将来自政务系统的景区名录、OTA 平台的用户评论与地图服务的地理坐标信息进行映射整合，必须综合使用 R2RML 进行结构化映射，同时辅以 Ontop 进行语义一致性校验。此外，部分研究尝试将映射自动化，如引入机器学习或语言模型（如 GPT）辅助生成映射规则，以缓解人工配置的繁琐性。此类智能映射系统正成为旅游语义数据管理未来的研究热点。

3.3. 数据融合与冲突解决方法：置信度模型、优先规则、实体链接等

在完成抽取与语义映射之后，旅游语义数据仍面临多源间信息冗余、冲突及不一致等问题。由于不同数据源可能在实体表述、属性值、时间戳等方面存在差异，如何对同一旅游对象进行统一识别与信息整合，成为实现高质量语义集成的关键步骤。本节将从置信度模型、优先级规则系统与实体链接技术三

个维度，对旅游数据融合中常用的方法与策略加以阐述，并通过表格方式进行典型方法对比。

3.3.1. 置信度模型（Confidence-Based Models）

置信度模型主要通过为不同数据源分配可信度权重，来判定多源数据冲突时的信息选择依据。此类模型基于数据来源的权威性、更新频率、历史一致性等指标计算“信任度分数”，从而在属性冲突时自动选择最可信值。其优势在于自动化程度高，适用于处理大量冗余值，如不同 OTA 平台对同一酒店评分、营业时间或电话的不一致情况[12]。但置信度模型也存在数据“来源偏见”问题，即高权重源并非始终准确，需结合实际领域规则进行人工校验或引入反馈机制加以修正。

3.3.2. 优先规则系统（Priority Rules）

优先规则是一种面向领域经验的冲突解决机制，通常由人工预设数据源或属性优先级。当同一实体出现多属性冲突时，系统优先采纳高优先级来源或属性。例如，在集成某景点信息时，若来自国家旅游局的官方数据与某社交平台数据发生冲突，可默认优先采用官方发布内容。优先规则系统实现简单、可控性强，适用于对数据质量要求较高的政务旅游平台或标准数据工程。但缺乏动态适应能力，面对新数据源或非常规冲突情形时较为局限。

3.3.3. 实体链接技术（Entity Linking）

实体链接旨在识别和关联多个来源中语义上相同但名称或表达方式不同的实体项，例如将“广州塔”与“Canton Tower”识别为同一地标。该技术依赖于上下文分析、语义相似度计算与外部知识库（如 DBpedia、GeoNames）的支持。常用方法包括基于编辑距离的字符串匹配、基于词向量的语义表示（如 Word2Vec、BERT）及基于图模型的连接路径分析等。在旅游数据融合中，实体链接是实现“统一实体标识符（URI）”生成的前提，是构建旅游知识图谱的核心环节之一[13]。

3.3.4. 综合融合机制与典型方法对比

在实践中，往往需结合置信度模型、优先规则与实体链接，构建多策略融合系统。例如，初步由实体链接判断候选实体集合，再根据置信度评分与优先规则确定最终保留值。下表列举了三种典型融合方法的对比：

表 3. 融合策略与典型方法对比

融合策略	技术核心	优势	局限性	应用场景
置信度模型	基于数据源权重的自动评分	自动化程度高，适应海量数据	易受评分机制偏差影响	OTA 评分整合、用户评论汇聚
优先规则系统	手工设定源/属性优先级规则	规则明确、控制力强	灵活性差，维护成本高	政务旅游平台、行业标准数据汇总
实体链接技术	实体名对齐与上下文语义识别	可处理异名、模糊表述等情况	算法复杂，计算资源消耗较大	跨平台旅游地标识别、地址标准化

3.3.5. 旅游语义融合中的挑战与趋势

旅游语义融合面临如下问题：一是数据更新频率高，融合机制须具备动态更新能力；二是源可信度难以量化，需引入人机协同机制；三是实体歧义性强，特别是在同名不同地等场景下，需结合上下文语义判别。当前研究趋势聚焦于以下方面：（1）融合置信网络与图神经网络，实现多源融合的深度语义理解；（2）引入 LLM（大语言模型），自动学习融合规则与实体对齐关系；（3）建立跨语种链接机制，服务于“一带一路”背景下的多语言旅游系统。

4. 旅游数据语义标注研究进展

在旅游大数据持续扩展的背景下，数据语义标注作为实现结构化、知识化表达的关键技术，已成为语义 Web 技术体系的重要组成部分。通过语义标注，非结构化或半结构化的旅游数据能够关联至标准本体，实现语义互操作与智能推理支撑，广泛应用于旅游推荐系统、知识问答平台与智慧旅游服务中[14]。语义标注技术发展至今，主要分为两大范式：规则驱动方法与学习驱动方法。此外，近年来深度学习在实体识别与语义关系抽取中的引入，使得旅游语义标注的自动化水平显著提升。

4.1. 语义标注技术类型综述

4.1.1. 基于规则的标注方法

规则驱动方法是最早期的语义标注手段，通常通过构建领域本体术语表+正则表达式+上下文模式匹配器，在数据中识别与本体概念相匹配的术语并完成标注。例如，在“广州白云山风景名胜区”的文本中，系统可根据地点类规则将“白云山”映射为景区类概念。该方法标注精度高、可解释性强，便于在政务类、标准化旅游数据场景中部署。但其不足在于扩展性差、维护成本高，尤其难以适应语言表达多样性与语义漂移的场景需求[15]。

4.1.2. 基于机器学习的标注方法

此类方法依托人工构建的训练数据集，利用统计特征（词性、上下文窗口、TF-IDF、句法结构等）进行模型训练，常用模型包括条件随机场（CRF）、支持向量机（SVM）等。其优势在于具有较好泛化能力，适应不同数据源。但机器学习方法面临两大挑战：一是训练样本依赖性强，难以在低资源领域推广；二是特征工程复杂，需要领域专家参与设计，成本高。

4.1.3. 基于深度学习的标注方法

近年来，基于 BERT、BiLSTM、Transformer 等模型的深度学习方法在命名实体识别（NER）、关系抽取等任务中展现出强大性能。其具备自动学习上下文语义、适应复杂句式结构的能力，适用于旅游评论、攻略文本等非结构化信息处理场景。以 BERT 为基础的微调模型，在识别复杂地名、事件描述方面效果优异。但其也存在一定风险，如语义漂移问题、推理不可解释性高，对语义标注结果的稳定性带来挑战[16]。

表 4. 标注方法类型对比表

方法类别	技术原理	优势	局限性
基于规则	规则库 + 匹配器	精度高，可解释性强	扩展困难，维护成本高
机器学习	特征工程 + 分类模型	泛化能力好	依赖训练样本，特征复杂
深度学习	BERT/LLMs 模型表示语义	理解复杂上下文	模型黑箱，语义漂移风险

4.2. 常用语义标注工具对比分析

为了支持语义标注任务的系统化实现，国内外研究者开发了多种工具与平台，常用于旅游数据语义化处理流程。以下简要介绍三类典型工具：OpenRefine、Karma 与 DBpedia Spotlight，并比较其功能侧重与适用性。

4.2.1. OpenRefine

OpenRefine 原为 Google 出品的数据清洗平台，支持 CSV、XLS、XML 等结构化数据的预处理与标准化。通过扩展插件，OpenRefine 可连接 SPARQL 端点，对字段值进行语义对齐，如将“长隆野生动物园”对接到标准旅游本体中的“景区类概念”。优点是操作直观、学习门槛低，适合用于旅游行业中的数据预处理环节。但 OpenRefine 不支持非结构化文本、缺乏对本体语义约束的原生支持，适用范围受限。

4.2.2. Karma

Karma 由南加州大学开发，是一套支持结构化与半结构化数据语义映射的平台。支持 JSON、XML、数据库等格式，通过图形化界面定义本体映射规则，并自动生成 R2RML 映射语句，适合本体驱动语义标注场景。Karma 具有较强的本体支持能力，并支持规则学习与推理，广泛用于旅游信息集成、智能问答系统构建等项目中。缺点是操作复杂、资源需求较高。

4.2.3. DBpedia Spotlight

DBpedia Spotlight 是一款开源文本实体识别与链接工具，可自动从自然语言文本中识别出与 DBpedia 实体匹配的短语。其适合处理用户生成内容，如评论、游记等，在旅游 UGC 数据挖掘中具有广泛应用。其不足在于识别依赖英文/大语料训练模型，在中文环境下识别率偏低；同时缺乏上下文关系处理与领域本体适配能力，适合作为初步语义链接工具。

表 5. 标注工具功能对比表

工具名称	支持数据类型	可视化能力	本体支持	典型应用
OpenRefine	表格/CSV/XML	中等	无	数据清洗、格式标准化
Karma	JSON/XML/数据库	高	是	本体映射、语义对齐
DBpedia Spotlight	非结构化文本	低	是	实体识别、UGC 链接处理

4.3. 标注能力面向未来的挑战

尽管现有方法与工具在旅游数据语义化中已取得显著进展，仍存在以下亟待突破的问题：（1）异构数据融合背景下的统一标注框架缺乏：当前方法大多专用于特定格式或领域，尚未形成适配多源多模态的统一标注体系。（2）中文语义处理能力偏弱：主流标注模型与工具多基于英文资源训练，对中文地名歧义、短语结构识别能力不足，限制了在中国旅游数据中的广泛应用。（3）标注结果可解释性与可控性不足：深度学习类标注系统虽性能优越，但缺乏过程可控性，影响在政府治理类项目中的信任度。（4）跨语言、跨文化标注需求上升：随着“一带一路”旅游数据流通加速，构建支持多语言语义标注与映射机制成为新趋势。为应对上述挑战，学界正探索融合大语言模型（LLMs）能力与可解释规则系统的混合标注架构，并开发基于图谱监督学习的自动标注引擎，为实现跨平台、多源语义融合奠定基础。

5. 面临的挑战与问题

5.1. 性能瓶颈：大规模推理与实时响应

在多源异构的旅游数据场景下，语义集成系统需处理来自政府数据、OTA 平台、社交网络与地图服务的大量三元组。传统基于 OWL 的描述逻辑推理在百万级数据环境中面临显著性能瓶颈，尤其在需对游客查询、景点状态等实时请求作出低延迟响应时更为突出。当前多数语义推理框架缺乏对动态数据的高效处理能力，难以满足旅游场景的实时性需求。为此，亟须引入图数据库支持的增量推理机制，或结合大语言模型（LLMs）进行语义近似计算，以实现快速、鲁棒的知识访问与标注服务。

5.2. 数据质量波动：稀疏性与错配问题

旅游数据在采集与整合阶段普遍存在结构不统一、字段缺失、语义不规范等问题。如同一景点在政府系统与 OTA 平台之间描述不一致，导致实体抽取与对齐错误频发。此外，用户生成内容中常见命名歧义、上下位混淆与标签不一致等情况，进一步影响语义标注与知识融合质量。解决此类问题需构建面向旅游知识的语义纠错机制，引入上下文一致性判别算法，并在数据融合过程中嵌入置信度计算与多源协

同校验模型, 以提升整体数据质量与语义准确性。

5.3. 语义漂移与知识更新难题

旅游知识具有高度时效性, 景区开放状态、交通方式、节庆活动等信息更新频繁, 导致原有本体与标注规则迅速过时。若系统未能及时演化, 将产生“语义漂移”现象, 影响推荐准确性与用户体验。同时, 术语演变、实体别名更新及表达习惯变化也对基于规则或模型驱动的标注系统提出更高适应性要求。因此, 有必要构建具备学习能力的本体演化机制, 并结合知识图谱动态更新策略, 支撑旅游语义系统的持续适配与自我优化。

6. 未来研究趋势

6.1. 基于大语言模型的自动语义标注机制

大型语言模型 (LLM) 如 GPT-4、ChatGLM 在语言理解与生成方面展现出显著优势, 为旅游语义标注提供了全新思路。借助 Prompt 工程技术, 可实现对评论文本、景点描述等语料的结构化实体与关系抽取, 从而弥补传统规则方法在覆盖面与扩展性方面的不足。然而, 现阶段 LLM 仍存在输出可控性弱、语义漂移明显、缺乏解释性的挑战。为此, 当前研究趋向于“规则 + LLM”的混合建模策略, 即以人工构建的语义约束规则为导向, 引导 LLM 在可控范围内完成标准化输出, 以实现准确性与灵活性的平衡。

6.2. 面向多源异构数据的实时语义集成架构

旅游数据高度异构, 来源广泛、更新频繁。未来语义集成系统需突破传统的离线处理范式, 构建面向流式数据的实时集成架构。该架构应具备高并发、高一致性与低延迟能力, 以适应动态旅游信息场景。关键研究方向包括: SPARQL 流式查询优化、增量式本体映射与语义合并机制、语义缓存更新策略等。通过构建多源实时语义接入总线, 可提升知识图谱时效性与语义服务性能, 为智能推荐与问答系统提供强有力支持。

6.3. 融合隐私保护机制的语义计算模型

随着数据合规与伦理要求的提升, 旅游语义计算必须重视隐私安全问题。游客偏好、行为轨迹、消费记录等高敏感信息在语义处理中的泄漏风险不容忽视。未来研究可引入联邦学习、同态加密与差分隐私技术, 在保障用户数据安全的同时实现跨平台语义融合与本体推理。同时, 构建可加密本体结构推理机制, 有助于在不暴露本体结构的前提下完成语义服务, 提升系统的可信性与合规性。

7. 结论

本文围绕“基于语义 Web 的旅游数据语义集成与标注技术”主题, 系统梳理了异构旅游数据的集成背景与技术路径, 回顾了语义数据模型、本体映射方法、标注策略与典型工具, 剖析了当前存在的性能瓶颈、数据质量波动与语义更新难题。通过文献综述与模型归类, 本文指出: 建设面向旅游场景的语义系统, 应坚持“标准化建模、智能化标注、动态化演化”的技术原则, 融合大语言模型、流式处理机制及语义隐私保障技术, 逐步实现高质量语义数据的全生命周期治理。本研究为旅游信息系统开发者、知识工程研究人员以及地方智慧旅游管理者提供了较为完整的语义支撑参考, 对推动语义 Web 技术在文旅融合发展中的深化应用具有一定借鉴意义。

致 谢

本研究工作得到了广东省教育厅计划国内访学项目的有力支持, 特此致以诚挚感谢。访学期间, 惠州学院(服务单位)和暨南大学(访学单位)为本人提供了良好的学术环境与周到的行政保障, 切实保障了课题研究的顺利开展与阶段目标的如期实现。在此, 谨向在本研究过程中给予悉心指导与无私帮助的各级领导、专家学者及相关工作人员表示衷心感谢。

基金项目

本研究获惠州学院横向科研项目【惠东非遗展示馆运营技术服务】(项目编号: 18-35900259094), 惠州学院高等教育教学改革项目【惠州学院旅游管理专业国际化人才培养路径研究】(项目编号: 17-1702242054), 惠州市哲学社会科学规划项目【从‘两创’到‘数创’: 人工智能驱动下岭南书院文化赋能惠州‘大湾区文化高地’建设路径研究】(项目编号: HZSK2025GJ151/2025ZX114)共同资助。

参考文献

- [1] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). Web semantic. *Scientific American*, 284(5), 34-43.
- [2] Chen, L., Cai, X., & Liu, Z. (2025). Multi-Source Data and Semantic Segmentation: Spatial Quality Assessment and Enhancement Strategies for Jinan Mingfu City from a Tourist Perception Perspective. *Buildings*, 15(13), 2298.
- [3] Bizer, C., Heath, T., & Berners-Lee, T. (2023). Linked data-the story so far. In *Linking the World's Information: Essays on Tim Berners-Lee's Invention of the World Wide Web* (pp. 115-143).
- [4] El Archi, Y., Benbba, B., Nizamatinova, Z., Issakov, Y., Vargóné, G. I., & Dávid, L. D. (2023). Systematic literature review analysing smart tourism destinations in context of sustainable development: Current applications and future directions. *Sustainability*, 15(6), 5086.
- [5] Euzenat, J., & Shvaiko, P. (2013). Overview of matching systems. In *Ontology Matching* (pp. 201-283). Berlin, Heidelberg: Springer Berlin Heidelberg.
- [6] Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 233-246).
- [7] Kejriwal, M. (2022). Knowledge graphs: Constructing, completing, and effectively applying knowledge graphs in tourism. In *Applied Data Science in Tourism: Interdisciplinary Approaches, Methodologies, and Applications* (pp. 423-449). Cham: Springer International Publishing.
- [8] Kushmerick, N. (2000). Wrapper induction: Efficiency and expressiveness. *Artificial intelligence*, 118(1-2), 15-68.
- [9] Huang, C. (2024). *Advancing Social Insights Through NLP: Social Media Reactions, Mental Health and Beyond* (Doctoral dissertation, The University of Iowa).
- [10] Arenas, M., Bertails, A., Prud'hommeaux, E., & Sequeda, J. (2012). A direct mapping of relational data to RDF. *W3C recommendation*, 27, 1-11.
- [11] Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., ... & Xiao, G. (2016). Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, 8(3), 471-487.
- [12] Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, 2(1), 550-561.
- [13] Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., & Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3), 527-570.
- [14] Heath, T., & Bizer, C. (2011). *Linked data: Evolving the web into a global data space*. Morgan & Claypool Publishers.
- [15] Cimiano, P., Handschuh, S., & Staab, S. (2004, May). Towards the self-annotating web. In *Proceedings of the 13th international conference on World Wide Web* (pp. 462-471).
- [16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).