

基于预训练模型适配的烟草营销领域智能查询增强技术研究

李一平¹, 舒刚², 张滔¹, 祝凯¹, 程康²

1. 湖北省烟草公司黄石市公司, 湖北 黄石 435000

2. 湖北省烟草公司鄂州市公司, 湖北 鄂州 436000

DOI:10.61369/ERA.2026060042

摘要 : 针对 XX 烟草企业营销数据孤岛严重、业务人员数据分析能力不强、需求响应迟缓等问题, 本文结合业务需求及现有技术, 提出融合湖仓一体架构与改进型 NL2SQL 技术的查询增强方案, 以满足企业技术需求。

关键词 : 查询增强; NL2SQL; 维度建模; 烟草营销

Research on Intelligent Query Enhancement Technology in the Tobacco Marketing Domain Based on Pre-trained Model Adaptation

Li Yiping¹, Shu Gang², Zhang Tao¹, Zhu Kai¹, Cheng Kang²

1. Huangshi Company, Hubei Tobacco Corporation, Huangshi, Hubei 435000

2. Ezhou Company, Hubei Tobacco Corporation, Ezhou, Hubei 436000

Abstract : Addressing the issues of severe data silos, limited data analysis capabilities among business personnel, and sluggish demand response in XX tobacco enterprises' marketing operations, this paper proposes a query enhancement solution that integrates a lakehouse architecture with an improved NL2SQL technology, tailored to meet the enterprise's technological needs by combining business requirements with existing technologies.

Keywords : query enhancement; NL2SQL; dimensional modeling; tobacco marketing

引言与研究背景

某地烟草企业营销数据呈指数级累积, 分散在多个异构系统中, 形成“数据孤岛”。传统固定报表模式存在三大局限: 技术门槛高(需掌握 SQL)、需求响应慢(平均3-5工作日)、灵活性差(预定义报表无法满足动态需求)。

一、技术体系总体架构设计

(一) 分层架构设计原则

1. 数据编织 (Data Fabric) 理念引入

数据编织最早由 Gartner 提出, 是一种数据管理的设计思路, 其核心思想是: 不强制把所有数据都放入数据仓库中, 而是通过元数据来把分布在不同系统、不同环境中的数据“编织”到一起, 让调用像访问一个集中、统一、完整的数据平台一样去调用数据。

数据编织架构主要具备四方面的能力。第一, 数据集成可以做得更自动化。第二, 元数据管理要更智能。第三, 数据交付要能弹性编排。第四, 安全、质量和合规这类治理要求。

2. 组装式数据分析架构 (Composable D&A)

组装式数据分析架构是未来数据分析应用形态。该架构底层

依赖于数据编织设计模式, 上层将报表平台、分析工作站、自助式分析平台、数据科学平台等技术能力微服务化、容器化, 形成高度开放的能力平台。用户不再面对单一技术系统, 而是根据业务需求灵活组装所需能力, 快速构建领域专属的数据产品或分析应用。

3. 湖仓一体 (Lakehouse) 技术选型

湖仓一体架构底层逻辑是基于低成本的对象存储, 配上像 Parquet、ORC、Delta Lake、Apache Iceberg 这类开放的数据格式。不论数据是结构化的表格、半结构的 JSON 日志, 还是完全非结构的图片、文档, 均可处理。更关键的是, 它补充了数据湖短板——ACID 事务支持、数据版本管理、还有模式强制校验, 这些本来都是数据仓库才有的能力。总的来说, 是既有数据湖的“装得下”, 又有数据仓库的“管得住”。^[4]。

本文设计营销数据分析平台, 底层选择湖仓一体架构, 应用

Apache Iceberg。存储直接调用对象存储当底座，不设计专用存储。计算引擎选取 Spark 和 Flink 搭配，能够处理批处理和流处理。为了提升交互式分析效率，前面增加一层 ClickHouse（开源版本）做加速。以上各组件集成一起，构成了一个统一的数据底座。

（二）五层系统架构

1. 数据源层：多源接入与实时采集

数据源层负责对接湖北烟草营销相关的各类数据产生系统，实现数据的全面采集和高效接入。该层采用差异化的技术策略适配不同类型的数据源：

核心业务系统（省级营销管理系统、ERP、CRM 等）采用 CDC（Change Data Capture）技术捕获增量变更，结合 Debezium、Canal 等工具实现准实时数据同步；物联网设备数据（智能烟柜、RFID 标签、GPS 终端等）通过 MQTT 协议接入 Kafka 消息队列，实现高并发流式采集；外部数据服务通过 API 调用和文件传输方式获取，配置灵活的数据质量校验规则；日志与行为数据通过 Flume/Filebeat 等轻量级 Agent 收集，统一汇聚至数据湖原始区。

2. 数据存储层：ODS-DWD-DWS-ADS 分层模型

数据存储层采用经典的分层架构设计，将数据按照加工程度和应用场景划分为四个层次：ODS，主要功能是原始数据镜像与缓冲；DWD，主要功能是明细数据标准化与主题建模；DWS，主要功能是主题聚合与指标预计算；ADS，主要功能是应用专属数据服务。分层架构的价值体现在三个方面：数据能重复用；口径统一；扩展性强。

3. 数据处理层：批流一体 ETL/ELT 引擎

数据处理层主要负责：数据转换、清洗、整合，以及数据增强（enriched）。本文研究采用采用批流一体的混合模式，不选择固定单一模式。对于复杂的数据清洗和关联转换，采用 ETL 模式，在加载前完成数据清洗；对于探索性分析和灵活变换需求，采用 ELT 模式，先将原始数据加载至数据湖，再根据查询需求即时转换。

4. 数据查询增强层：NL2SQL 智能转换核心

数据查询增强层主要是负责把自然语言查询智能转换成 SQL 语句。该层采用微服务架构，其核心组件及对应功能定位、关键技术如下表所示。

组件	功能定位	关键技术
语义理解引擎	解析自然语言查询的意图和要素	BiLSTM-Attention 编码、意图分类、实体识别
Schema 感知模块	获取并编码数据库模式信息	图神经网络、元数据缓存、增量同步
NL2SQL 引擎	自然语言→SQL 核心转换	双编码器架构、语法树解码、执行反馈优化
查询优化器	SQL 重写与执行计划优化	代价模型、索引推荐、并行执行
智能缓存管理器	多级缓存与失效策略管理	语义缓存、增量更新、预热机制

在设计查询增强城时，主要考虑了烟草营销的特殊性。系统内一共收入了专业词汇共计 5000 余条，包括了组织机构名称，专业术语等等。同时可以自动识别专业术语。需要注意多表关联推

理模块主要是利用模型的外键关系和统计特点，可以自动推断出查询所要涉及的路径，而在处理数值范围、时间表达等等一些约束，可通过条件值智能匹配模块实现，提供了转换和解析能力，让用户使用起来更方便、简单。

5. 数据应用层：可视化与决策支持

数据应用层面向终端用户提供多样化的数据消费方式和决策支持工具，体现“千人千面”的设计理念：

用户群体	核心需求	功能形态	交互方式
高层管理者	宏观指标监控、战略执行跟踪	管理驾驶舱、预警推送	语音查询、移动端概览
中层管理人员	业务过程管控、团队绩效分析	自助分析工作台、固定报表	拖拽分析、自然语言查询
一线业务人员	客户信息查询、订单状态跟踪	智能问答助手、移动应用	自然语言、扫码识别
数据分析师	深度数据挖掘、模型开发	数据探索平台、API 服务	SQL/Python、Jupyter Notebook

历史问题推荐和语音输入等方便的功能可由智能问答界面提供支持。拖拽式分析，过滤筛选等一些交互式的功能，由自主分析工作台所提供。运用响应式设计管理驾驶舱可以适配不同终端，还能实现异常的报警以及实时刷新等服务功能。

（三）架构协同机制

1. 元数据驱动的数据血缘追踪

本次研究所构建的是一种统一的元数据库，利用三类元数据进行分析，包括技术元数据，操作元数据以及业务元数据，如技术元方面可提供 etl 作业等，操作元数据涉及时间更新，访问频率等，业务元数据则涉及了维度说明和指标定义等。该仓库实现的是全链路的数据血缘式追踪，简单来说就能从报表的指标向上追溯，一直到原系统的字段，能更好的对影响分析进行处理。

2. 智能查询路由与缓存优化

利用智能查询路由以及多级缓存可有效提高查询响应的性能，查询路由可根据查询特点进行执行路径的动态选择：

查询特征	路由策略	执行引擎	预期响应
预聚合指标查询	命中物化视图	ClickHouse	<100ms
明细数据探索	下推至 DWD 层	Presto/Trino	1-10s
复杂关联分析	分解子查询并行执行	Spark SQL	10-60s
实时流数据查询	直接读取 Kafka	Flink SQL	<1s

将缓存优化设计成三级结构，结果缓存可以用 Redis 纯热点查询的结果可以实现毫秒级的返回。子查询缓存是缓存公共的中间结果，查询复杂时可重复使用，并不需要过度计算。预计算缓存是利用物化视图支持常用的分析模型。

3. 安全治理与权限控制体系

数据安全治理与架构全程融会贯通，做好事前事中和事后的安全体系控制：

安全层级	核心机制	技术实现
身份认证	统一身份认证、多因素认证	LDAP/AD 集成、OAuth2.0、SSO
权限控制	RBAC+ABAC 混合模型	角色权限矩阵、动态属性授权
数据脱敏	行级过滤、列级脱敏、动态掩码	格式保留加密（FPE）、正则替换
操作审计	全量日志记录、异常行为检测	日志集中采集、ML 异常检测、合规报告

针对 NL2SQL 场景，在权限控制方面进行了如下优化。第一，查询解析阶段以用户权限约束模式的范围，为基础保障，用户只可访问已授权的数据。第二结果返回阶段，对于查询结果的敏感字段进行脱敏处理。第三操作审核阶段，完整记录自然语言查询的原文，执行结果等一系列的信息，支持合规报告和安全审计报告提供。

二、业务数据模型设计创新

（一）维度建模方法论应用

基于 Kimball 维度建模理论^[3]，主要构建营销区域、供应链条区域。再利用对于时间、产品、客户等一致维度，与总线架构可以实现跨主题的连接。

（二）动态维度建模机制

1. 缓慢变化维度（SCD）处理策略

在维度数据管理方面是整个设计的难题，本研究以烟草营销数据作为特点系统并创新使用了该策略。本研究的创新主要是提出该策略的自动化匹配制度，基于元数据标准维度的属性对策略类型进行变更，ETL 引擎可以识别变更事件。同时支持对应的版本管理，有效减少了操作的复杂程度和人工操作可能造成的风险。

2. 实时维度更新机制

突破传统 T+1 批量更新模式，构建主备双存储结构实现分钟级维度更新：

主存储（关系型数据库）：事务性维度维护，实时生效

热缓存（内存数据库）：秒级高并发读取

分析存储（分析型数据库）：分钟级异步同步

针对客户档位（核心分析维度）等关键属性，采用 SCD Type 2 策略，通过代理键 + 生效时间戳（effective_date/expiry_date）保留完整历史版本，支撑“当时有效”的历史分析，预计资源利用率提升 40%。

3. 自适应粒度聚合模型

对查询力度需求多样性进行分析，范围由明细及到高度汇总及跨越。本研究提出了适应性的力度聚合模型，以查询模式学习动态优化预计算的策略，该核心是通过识别高频查询的组合以及力度层次，可以优先去计算这些聚合点。对于一些低频的查询内容，则利用实计算或是从更粗力度上卷的一种模式开展处理，这样可以有效减少预计算的开销度，数学建模则是把聚合选择的建模分为约束优化的问题。

设候选聚合集合为 $A = \{a_1, a_2, \dots, a_n\}$ ，每个聚合 a_i 的存储成本为 c_i ，对查询负载的加速收益为 b_i ，存储预算为 C ，则：

$$\max_{S \subseteq A} \sum_{i \in S} b_i \quad \text{s.t.} \quad \sum_{i \in S} c_i \leq C$$

该问题为 NP 难的背包问题变体，利用贪心算法可求解。按收益成本比 b_i/c_i 降序选择聚合，直至预算耗尽。对于查询日志进行持续性的监控，周期性重新计算最优聚合的集合，可以有效调整预计算的策略。

三、改进型 NL2SQL 模型设计

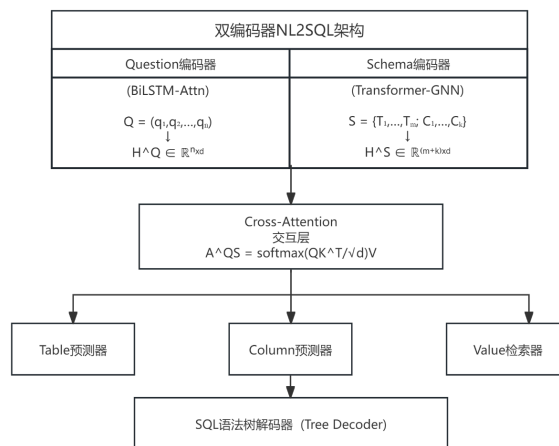
（一）领域挑战分析

烟草行业的自然语言到 SQL（NL2SQL）转换有以下三个难点：一是识别专业术语和分析比较困难，二是实际查询需要，联系多张数据报表，三是查询条件模糊或相对表述处理时较为复杂。

（二）双编码器架构

本研究提出双编码器架构，分别对数据库模式和自然语言问题进行深度编码，通过交叉注意力实现精准语义对齐^[4]。

架构示意图：



Question 编码器采用 BiLSTM-Attention 结构：

$$\vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t-1}), \vec{h}_t = \text{LSTM}(x_t, \vec{h}_{t+1})$$

$$h_t = [\vec{h}_t; \vec{h}_t] \in \mathbb{R}^{2d_h}$$

$$\alpha_i = \frac{\exp(w^T \tanh(W_h h_i + b))}{\sum_{j=1}^n \exp(w^T \tanh(W_h h_j + b))}$$

$$c_Q = \sum_{i=1}^n \alpha_i h_i$$

Schema 编码器采用 Transformer-GNN 混合结构，显式建模表 - 列层次结构和外键关系：

$$h_i^{(l+1)} = \text{TransformerLayer}(h_i^{(l)}, \{h_j^{(l)}\}_{j \in N(i)})$$

（三）核心模块设计

表名预测。利用层次化注意力制度，第一层基于相似度筛选 Top-K 候选表，第二层则利用 GNN 推理达到最优表组合条件。

条件值的检索工作。利用三层融合策略解决数值，时间等识别问题。

第三，语法树解码：自顶向下递归生成 SQL AST，结合约束感知束搜索（语法约束 + 模式约束 + 业务约束）与执行反馈优化生成策略。

（四）领域知识注入

本文构建包含 10,000+ 实体、50,000+ 三元组的烟草营销知

识图谱，使用 RotatE 嵌入^[5]增强语义表示。同时，采用 LoRA 技术^[6]微调中文预训练语言模型，可训练参数降至0.1%，训练速度提升10倍，同时保持领域适配效果。

四、实验评估与应用效果

（一）实验设置

构建 Tobacco-NL2SQL 数据集（15,500样本），定义执行准确率（EX，结果集匹配）与逻辑准确率（EM，SQL 字符串匹配）。测试集分布：单表查询30%、2-3表关联45%、4+ 表关联20%、嵌套子查询5%。

（二）性能评估结果

模型	执行准确率 (EX)	复杂查询准确率 (4+ 表)	平均响应时间
RAT-SQL (基线)	82.3%	68.7%	180ms
本研究模型	92.4%	85.2%	110ms

较基线提升10.1个百分点，复杂查询优势更明显（+16.5%）。消融实验表明：知识图谱嵌入贡献最大（+4.8%），执行反馈优化提升2.9%，LoRA 在训练成本降低10倍前提下性能损失仅0.6%。

（三）某地烟草企业应用成效

指标	实施前	实施后	提升幅度
平均数据获取时间	3-5工作日	5分钟	99.7%
业务自主分析比例	<10%	65%	+55个百分点
IT 开发工作量占比	100%	40%	-60%
决策响应周期	1-2周	1-2天	85.7%

五、结论与展望

（一）研究贡献

本研究所提出的是一种技术方案，通过整合湖仓一体架构以及优化后的自然语言 SQL（NL2SQL）方法，增强数据的查询功能。贡献主要包括了可动态维度建模，实现了分钟级维度的优化和自适应性。二是双编码器与领域知识注入结合在一起，使得 NL2SQL 执行准确率高达92.4%。三是五层组装式架构，可支持业务人员利用自然语言开展自助式分析工作。

（二）未来工作

在未来的工作中，需要从如下三个方面进行。第一，多轮对话式分析，引入对话状态跟踪支持用户在多方面交互中持续追问以及细化分析的需求。第二，拓展查询服务，探索跨预购数据库的查询能力，可以实现在数据不搬迁的情况下实现跨源的统一访问。第三，协同大小模型研究大语言模型以及领域专用小模型的合作架构，考虑理解力与推理的效率。

参考文献

[1] Gartner. Hype Cycle for Data Management, 2022[R]. Stamford: Gartner, 2022.

[2] Zaharia M, Ghodsi A, Xin R, et al. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics[C]// Proceedings of the 11th Conference on Innovative Data Systems Research (CIDR). Online: www.cidrdb.org, 2021.

[3] Elizabeth M. Pierce, Song I Y. Dimensional modeling: Identification, classification, and evaluation of patterns[J]. Decision Support Systems, 2008, 45(1): 59-76.

[4] Wang B, Shin R, Liu X, et al. RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). Online: ACL, 2020: 7567-7578.

[5] Sun Z, Deng Z H, Nie J Y, Tang J. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space[C]// Proceedings of the 7th International Conference on Learning Representations (ICLR). New Orleans: OpenReview.net, 2019.

[6] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of Large Language Models[C]// Proceedings of the 10th International Conference on Learning Representations (ICLR). Online: OpenReview.net, 2022.