融合知识图谱与大语言模型的
广告点击率预估优化研究

开题答辩

答辩人：李芮

指导教师：王小宁 吴殿义

专业：2021级广告学
(计算广告双学士学位复合型人才培养项目)

目录

CONTENTS

第一部分

PART ONE

选题背景和意义

01 选题背景

商业价值

在当前的数字经济时代，互联网广告已经成为全球商业活动中不可或缺的一部分。作为衡量广告系统性能的核心指标，**点击率（Click-Through Rate，CTR）预估**在广告投放的精准性、用户体验的优化以及平台收益的提升中起着**关键作用**。

全球数字广告市场规模正在以稳健的速度增长，预计到2025年，全球数字广告支出将突破7000亿美元。

这一发展趋势对广告平台提出了更高的技术要求，其中CTR预估模型的性能直接影响了广告投放的效果与商业价值。

传统方法缺陷

传统CTR预估方法有三大主要挑战：**数据稀疏性、语义理解不足和知识关联缺失**。

- 数据稀疏性表现为用户与广告交互的**长尾分布**、新用户**冷启动**问题，以及对**高维稀疏特征**的建模能力不足。
- 语义理解不足体现在对广告内容与用户兴趣的**深层语义关联**、**上下文信息**、多模态融合及用户**意图建模**能力的**欠**缺。
- 知识关联缺失则包括**商品与广告间关系的表达不足**、**隐含特征关联未被挖掘**、领域知识利用有限以及**跨场景迁移能力较弱**。

近年来，**大语言模型（Large Language Models, LLMs）**在自然语言处理领域取得了革命性突破，展现出**强大的语义理解和知识推理能力**。例如，GPT和BERT等模型通过预训练和微调技术，可以高效处理复杂的自然语言任务。此外，**知识图谱（Knowledge Graph, KG）**技术日益成熟，为结构化知识的表示和推理**提供了全新途径**。将大语言模型与知识图谱相结合，为CTR预估模型提供了全新的研究方向，有望克服传统方法的瓶颈。

01 选题意义

学术价值

提出**融合知识图谱与大语言模型**的新范式，探索**知识图谱**在**特征工程**中的**深度应用**，弥补传统特征表达方式的不足；

利用知识图谱提供**结构化的解释能力**，**增强模型的透明性**，结合大语言模型，提供自然语言形式的解释，降低模型复杂性对用户理解的影响，构建多层次的解释体系，满足不同场景下的解释需求。

实践应用

提高广告投放的**精准性**，优化广告收益分配机制，降低获客成本，支持**广告系统向智能化方向发展**，促进广告生态的可持续增长；

提升个性化**推荐质量**，**减少**无关广告对用户的**干扰**，**冷启动**下实现**快速适应**，为新用户提供良好的初始体验；

提供融合知识图谱和大语言模型的工程实践方案，降低实际应用门槛。

第二部分

PART TWO

研究现状与方法

02 现存CTR预估方法

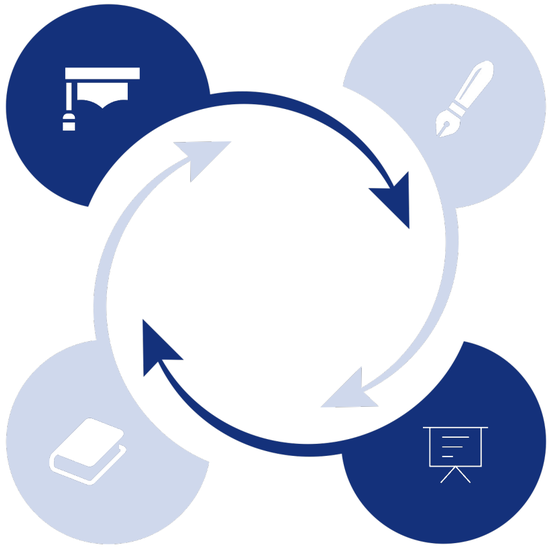
逻辑回归

逻辑回归通过线性组合输入特征（如用户和广告属性），将结果映射到点击概率上，**核心在于直接学习每个特征的权重**，但无法捕捉特征之间的交互。

Wide & Deep与DeepFM

Wide&Deep 将**线性模型**和**深度神经网络**相结合，通过同时建模**记忆能力**和**泛化能力**来提升效果。

DeepFM 在 **Wide&Deep** 基础上改进，通过**因子分解机** 替代 Wide 部分，从而去掉人工特征设计需求



因子分解机

因子分解机通过隐向量分解捕捉特征之间的二阶交互关系，**不仅学习单个特征的权重，还建模特征对之间的隐含影响**。

DIN

DIN (Deep Interest Network)通过动态兴趣建模，引入用户行为序列信息，用**注意力机制**动态加权不同历史行为，捕捉用户在特定场景下的兴趣，提升推荐的实时性和个性化程度。

02

知识图谱结合大模型方法

GACE(基于图的跨页面广告嵌入)

通过构建包含语义知识、页面知识和交互知识的三元组图结构，建立广告节点之间的关联关系，并利用内容相似度和页面相似度的乘积作为边权重，从而预测新广告效果。

LLM-KERec(通过推理知识图谱在工业推荐系统中利用大语言模型)

LLM-KERec 系统通过实体提取、大语言模型构建互补知识图谱和 E-E-I 权重决策模型三个核心模块的协同，实现了更优的互补商品推荐效果。

K-RagRec(基于知识图谱检索增强生成的LLM推荐系统)

K-RagRec通过知识图谱检索增强生成(RAG)来提升LLM推荐性能，包括知识子图索引、自适应检索、重排序和GNN编码四个核心模块，以解决LLM推荐中的幻觉和知识缺失问题。

第三部分

PART THREE

研究方法

02

数据来源

淘宝展示广告点击率预估数据集

raw_sample

用户ID，广告ID，时间，资源位，是否点击
114万用户8天内的广告展示/点击日志（2600万条记录）

ad_feature

广告ID，广告计划ID，类目ID，品牌ID

user_profile

用户ID，年龄层，性别等

raw_behavior_log

用户ID，行为类型，时间，商品类目ID，品牌ID
22天内的购物行为(共七亿条记录)

补充：Amazon Review Dataset

包括评分数据（rating），产品元数据（descriptions, category information, price, brand 和 image features）以及链接数据（共同查看/共同购买的关系图）

03 研究过程

第一章
引言

1.1 研究背景
1.2 研究意义
1.3文献综述
1.4 创新点

1

第二章
知识图谱构造

2.1数据集介绍
2.2 数据处理方案

2

第三章
模型介绍

3.1 知识增强的CTR预估框架

3

第四章
实验应用

4.1实验过程
4.2 实验结果

4

第五章
总结与展望

5.1研究总结
5.2局限与未来工作

5

03 创新点

LOGO HERE

原始特征的深层语义高维表示

将原始特征转化为具有深层语义理解的高维表示，在现有研究中未被充分探索。
如几点进行什么行为，向相似场景/商品拓展。

可解释的推荐结果

关注推荐结果的可解释性，在当前的研究中尚未得到充分重视

综合利用知识图谱和大型语言模型

知识图谱与LLM结合，旨在提升点击率预估的准确性，实现更全面的用户理解、更深入的商品关联分析和更准确的兴趣匹配。

03 研究思路概述

特征提取和增强

通过用户行为数据、商品信息、上下文数据的图谱提取，使用LLM处理语义理解，并用知识图谱特征计算相似性。

特征融合和预测

通过融合用户特征（历史行为、语义理解、图谱特征）与广告特征（基础信息、上下文）进行预测并给出解释和建议。

效果评估

通过评估AUC、基准提升、准确率、召回率等判断效果。

第四部分

PART FOUR

进度安排

04 进度安排

1月1日 之前
确定选题，查找大量资料进行文献阅读与整理
完成文献综述和开题报告

1月2日 - 3月1日
完成模拟研究和实验
完成论文初稿的撰写

3月2日- 4月1日
修改完善初稿
形成第二稿

4月2日- 5月1日
修改第二稿
完成论文撰写

5月2日之后
准备答辩

请老师批评指正

答辩人：李芮

指导教师：王小宁 吴殿义