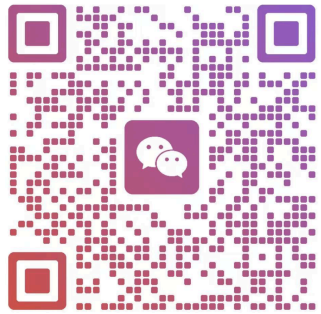
面向可控人工智能的推荐系统

打破信息茧房的因果反事实探索

王佳
西交利物浦大学
2025 年 9 月 17 日
国家自然科学基金青年项目（C类）交流报告



CONTENT

教育背景（北京交通大学–瑞典皇家理工学院–香港理工大学）

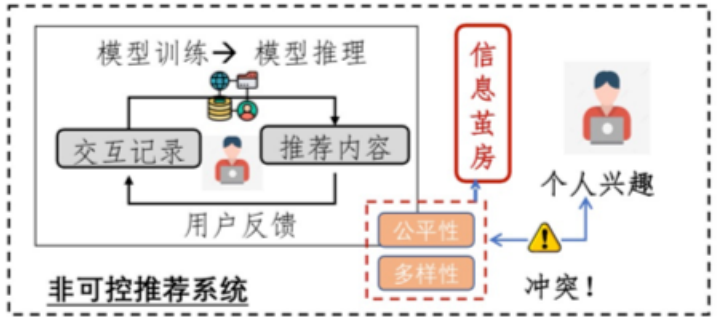
发表AAAI, IJCAI, ICDM, TBD, TAI等期刊会议论文； 参与开发新浪微博投票、香港机场调度项目。

1. 研究背景：个性化推荐的“悖论”

1.1 个性化推荐的“悖论”：越“懂”你，越“困”住你

- 现状: 智能推荐已成为信息获取的主要入口，追求极致的个性化。
- 问题: 算法过度迎合，导致信息获取的“窄化”和“极化”。
- 引出概念: “信息茧房” ——用户在无意识中被同质化信息所包围。

> **政策导向**: 网信办《发力“破茧”》，强调抗拒“信息茧房”。
> **学术前沿**: 张钹院士强调，安全可控是第三代人工智能的核心。



1.2 问题的本质：为什么“破茧”这么难?

- 核心挑战：当前解决方案的“三不”困境

困境	表现	关键词
“控不到”	茧房程度难以实时、在线地衡量。 用户和平台都处于“无意识”状态。	**理论建模空白**
“看不清”	推荐机制“黑盒”，茧房成因的因果路径不明。 用户无法针对性调整。	**可解释性缺失**
“控不准”	简单增加多样性损害体验。 用户的“新偏好”与“旧表征”产生冲突。	**表征冲突**

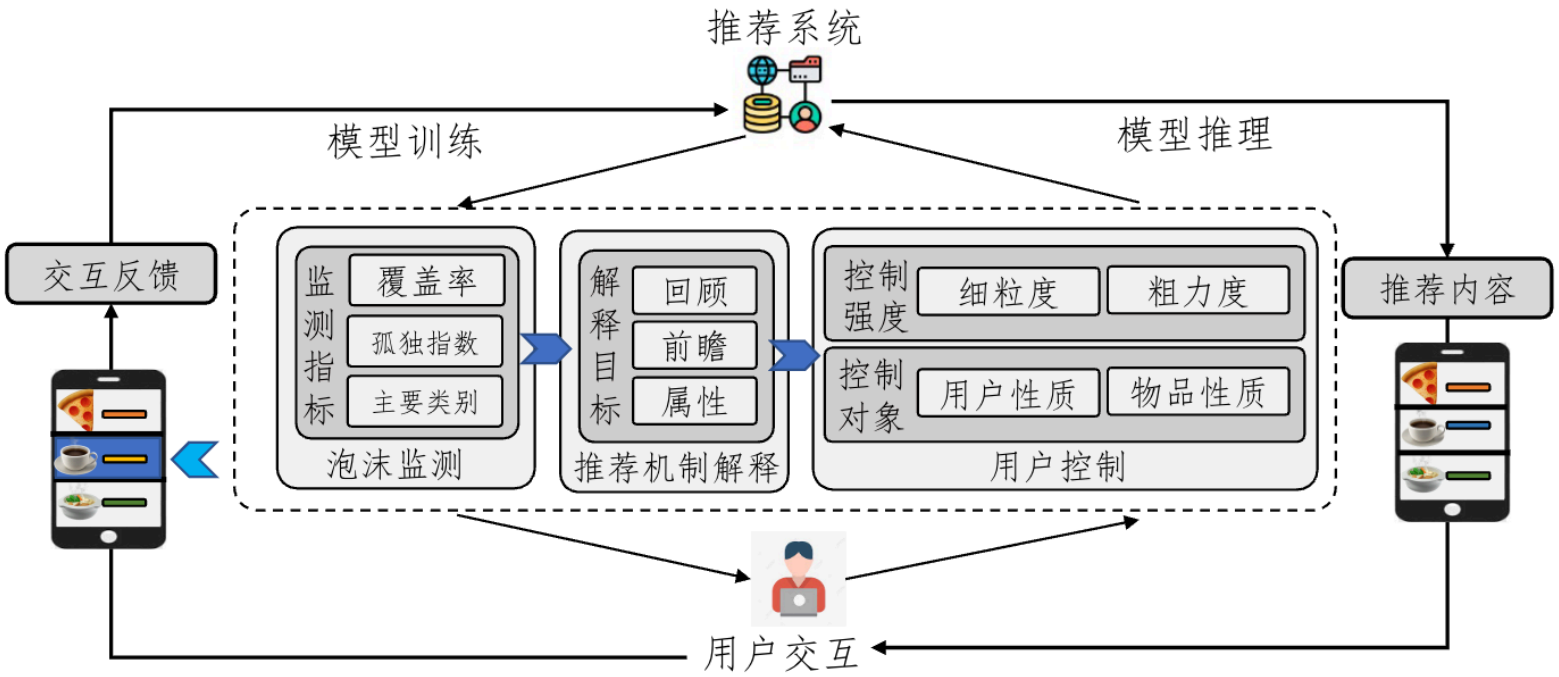
2. 核心思路: 从“被动接受”到“主动可控”的范式转变

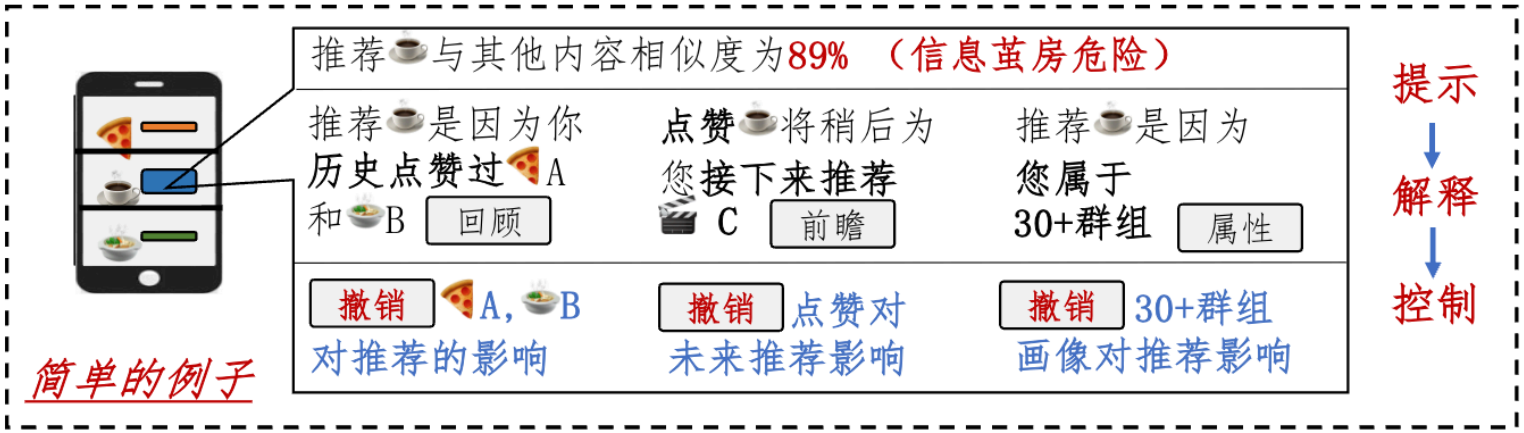
2.1 我们的解决方法：赋予用户“反事实”的控制权

三大核心支柱

- 控得到 (Quantifiable):
  - 实时感知茧房程度
- 控得清 (Explainable):
  - 洞悉茧房形成因果
- 控得准 (Controllable):
  - 精准干预推荐结果

1. **感知 (Perceive)**: 通过多维度指标建模，量化并预警信息茧房。
2. **理解 (Understand)**: 利用因果反事实推理，为用户生成可理解的“破茧”路径解释。
3. **行动 (Act)**: 基于因果干预，实现用户对推荐内容的精准控制。


推荐系统
模型训练
模型推理
交互反馈
监测指标
覆盖率
孤独指数
主要类别
泡沫监测
解释目标
回顾
前瞻
属性
推荐机制解释
控制强度
细粒度
粗力度
控制对象
用户性质
物品性质
用户控制
推荐内容
用户交互


推荐☕与其他内容相似度为89% （信息茧房危险）
推荐☕是因为你历史点赞过🍕A和☕B
回顾
点赞☕将稍后为您接下来推荐🎬C
前瞻
推荐☕是因为您属于30+群组
属性
撤销 🍕A, ☕B 对推荐的影响
撤销 点赞对未来推荐影响
撤销 30+群组画像对推荐影响
提示
解释
控制
简单的例子

3. 研究方案: 技术框架

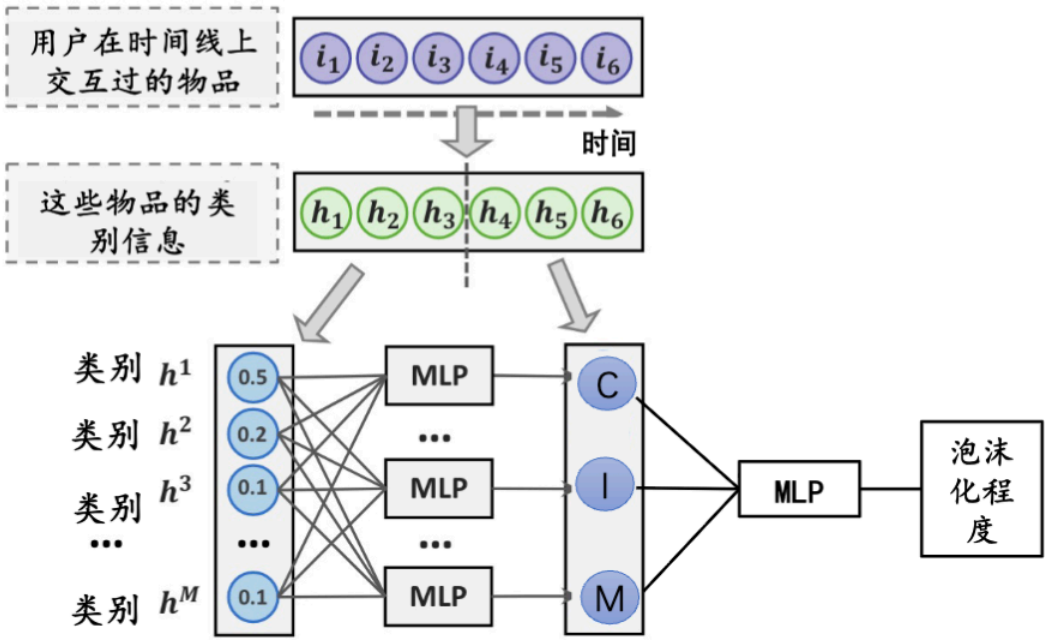
3.1 关键技术一：为“信息茧房”建立“仪表盘”

- **多维度量化指标:**

  i. **覆盖率（Coverage)**: 衡量推荐内容的**广度**。

  ii. **孤立指数（Isolation Index)**: 衡量用户群体间的**隔离度**（社会学视角）。

  iii. **主要类别支配指数（MCD)**: 衡量内容类别的**集中度**（物品性质视角）。



3.2 关键技术二：让“黑盒”开口说话”

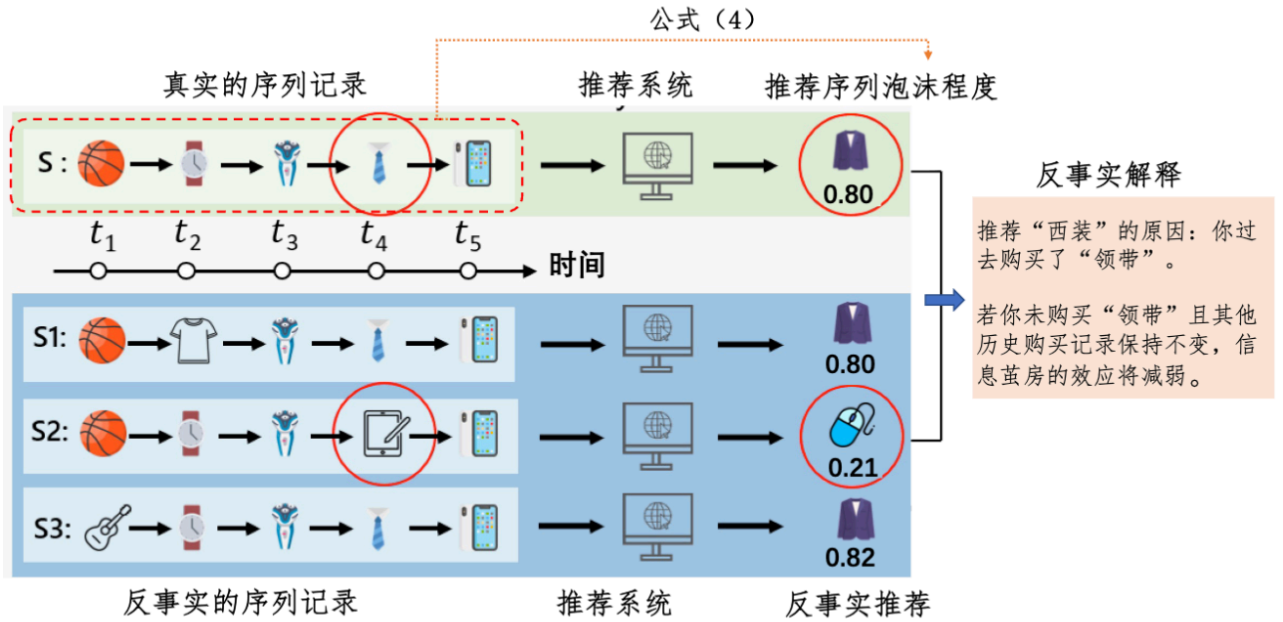
- **核心思想：回答 "What-if" 问题**

> “**如果**你过去没有点击那些视频，现在还会给你推荐这些吗？”

- **方法：基于因果反事实生成解释**
  - **用户可理解的解释模板**

i. **回顾性**：“系统内容同质化，是因为您之前喜欢了[项目A]。要停止，请撤销该行为。”

ii. **前瞻性**：“由于您现在的点击，系统同质化正在上升。要阻止，请撤销该行为。”



3.3 关键技术三：“控得准”

- **核心挑战**:
  - 特征的修改，会因**混杂因子**导致“控不准”。
  - 解决用户控制引入的“**新旧表征冲突**”难题。
- **方法**

  i. **因果干预** 调整预测结果里，减去“旧偏好”带来的影响。

  ii. 冻结基座模型的 **参数高效微调** 的Prompt tuning调控。

  iii. 基于**大模型**的“自然语言” Prompt Engineering调控/知识编辑调控。

控制类型	用户指令示例
细粒度	“我想看‘Z世代’喜欢的内容。”
粗粒度	“别总根据我的‘年龄’给我推荐。”
细粒度	“多给我推荐一些‘科幻’电影。”
粗粒度	“最近‘喜剧’看腻了，少推点。”

3.3 关键技术三：“控得准”

路径一：因果干预	路径二：大模型Prompt+RAG	路径三：小模型Prompt Tuning
“结构化”调控	**“零样本”调控**	**“领域自适应”调控**
核心思想: 解耦混杂，精准干预。将控制视为对因果图的 do 操作，从根本上切断不希望的关联。	**核心思想: 理解开放意图，无需训练。**利用固定LLM的强大NLU能力，将用户自然语言指令转化为推荐策略。	**核心思想: 高效微调，精准适配。**冻结LLM主体，仅训练少量“软提示”参数，使模型高效适应推荐控制任务。
实现方式: 通过反事实推断，计算移除特定特征影响后的推荐得分。 最终分 = 调整后预测 – α * (原预测 – 反事实预测)	**实现方式:** 设计精巧的文本提示（Prompt），引导LLM直接生成控制信号或重排物品列表。	**实现方式:** 将用户指令与历史作为输入，通过微调后的Soft Prompt，让LLM直接输出最优的控制向量。

三路径协同策略：LLM as an Interpreter, Causal Model as an Executor

4. 近期成果: 本项目研究的进展

ReLAX: 训练一个AI智能体来解释“黑箱”模型

核心思想: 通过训练一个智能体，以最少的“动作”（修改特征）来“通关”（改变模型预测结果），从而生成高效、简洁且模型无关的反事实解释。

> ReLAX 将寻找反事实解释的过程建模为一个**序贯决策任务**:

- **玩家 (Agent)**: 一个深度强化学习智能体。
- **游戏目标 (Goal)**: 改变输入样本预测结果（如"不推荐→推荐"）。
- **动作 (Action)**:
  1. 特征选择（离散动作）
  2. 修改幅度（连续动作）
- **奖励 (Reward)**:
  * 预测翻转 → 高回报 ✓
  * 特征修改 → 惩罚机制 ✗

Chen, Z., et al. Explain the Explainer: Interpreting Model–Agnostic Counterfactual Explanations of a Deep Reinforcement Learning Agent. IEEE Transactions on Artificial Intelligence.

基于检索增强生成的推荐遗忘框架 （Prompt+RAG）

核心思想: 推荐遗忘要解决的是如何在不扰动整体系统的情况下**精准擦除**特定用户数据影响，同时避免传统参数更新方式引发连锁反应的难题。

> 推荐系统遗忘困境：
>
> - 传统遗忘方法导致**全局参数扰动**
> - 参数级更新引发**非目标用户性能衰退**
> - 动态偏好捕捉与**偏差传播控制**难以兼顾

> 核心组件：
>
> - 偏好感知检索器（长期兴趣建模） → 类目多样性采样
> - 时序注意力过滤 –原子化遗忘检索 → 用户粒度操作
> - 动态提示生成器 –检索后商品重排 → LLM可控生成

Haichao, Z., et al. Customized Retrieval–Augmented Generation with LLM for Debiasing Recommendation Unlearning. The 25th International Conference on Data Mining November 12–15, 2025 Washington DC, USA

(Prompt+Tuning）推荐控制的工作 AAAI在投

LLM-Based Recommendation 的语意解码

现状: 基于大语言模型（LLM）的推荐系统，在生成推荐结果时，仍然沿用为自然语言处理(NLP)设计的解码策略（如Greedy Search, Beam Search）。

- **矛盾**:
  1. **目标不同**: NLP生成流畅的**文本序列**，而推荐系统只需要一个准确的**物品ID**。
  2. **语义冗余**: 用户可能对多个**功能相似**的物品（如不同品牌的蓝牙耳机）有同等兴趣。传统解码策略将这些物品视为独立选项，导致模型的“置信度”被分散。

  ✕ **简单来说：用写作文的方法来做选择题，既低效又抓不住重点。**

> 核心组件：
>
> - 聚合相似物品: 将那些能满足同一需求、在语义上等价的物品**聚类**成一个“意图簇”。
> - 衡量意图不确定性: 不再计算模型对单个物品，而是计算对这些 **意图簇** 的困惑度（Semantic Entropy）。
> - 不确定性指导的自适应解码，优化LLM Decoder解码策略。

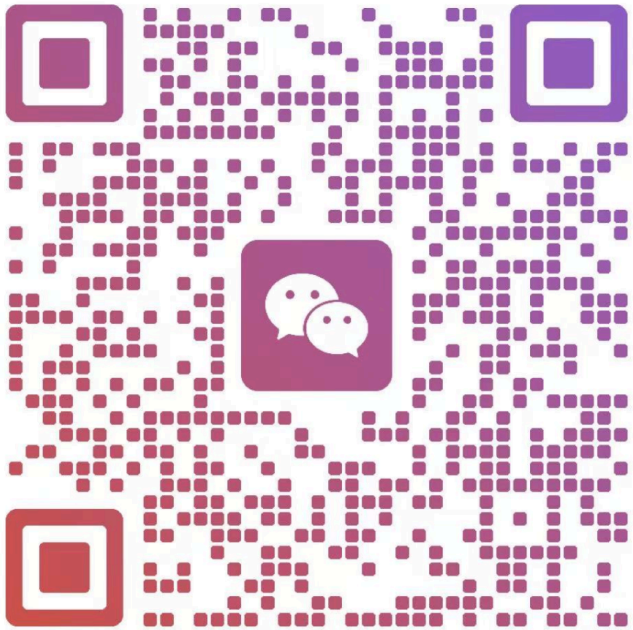
Chenke, Y., et al. Uncertainty-Aware Semantic Decoding for LLM-Based Sequential Recommendation." The 9th APWeb-WAIM joint International Conference on Web and Big Data. 2025.

5. 总结与展望

预期成果与创新贡献

- **理论创新**
  - 建立首个信息茧房**量化建模与预测理论**。
  - 提出**因果–大模型融合**的推荐系统解释与控制新框架。
- **技术突破**
  - 开发一套**用户可控、实时响应、支持自然语言**的个性化推荐算法。
  - 解决用户控制引入的“**新旧表征冲突**”与**意图理解**难题。
- **社会价值**
  - 推动形成更健康、多元、包容的信息生态。
  - 为**可控、可信、负责任的第三代人工智能**发展提供技术支撑。
- **近期工作**
  - **更智能的交互**: 探索基于多轮对话的持续性用户控制。
  - **更广阔的场景**: 将框架应用于多模态推荐（如短视频、新闻）。

欢迎交流 ~



- 邮箱：jia.wang02@xjtlu.edu.cn
- 欢迎借用算力
- 欢迎老师学术交流
- 欢迎学生RA访问