李涵宇

l-hy12@outlook.com · adamli12.github.io · Birthdate: 2002/01/15

教育背景

清华大学，北京 2020 – 至今

在读博士研究生 （导师：马少平）计算机科学与技术

纽约州立大学布法罗分校，美国 2019.7

访问学者 （导师：Junsong Yuan）

清华大学，北京 2016 – 2020

学士 数理基础科学

耶鲁大学，美国 2018.7

暑期学校 统计学导论 (A)，心理学导论 (B+)

研究兴趣

推荐系统，序列、重排序、跨域推荐模型，扩散模型，大语言模型等技术在推荐中的应用

主要论文

1. **Hanyu Li**, Sabrina Racine-Brzostek, Nan Xi, Jiwen Luo, Zhen Zhao, Junsong Yuan. **Learning to Detect Monoclonal Protein in Electrophoresis Images.** (VCIP 2021). *[论文链接]*
2. **Hanyu Li**, Hongyu Lu, Songhao Huang, Weizhi Ma, Min Zhang, Yiqun Liu, Shaoping Ma. **Privacy-Aware Remote Information Retrieval User Experiments Logging Tool.** (SIGIR 2021, CCF A). *[论文链接]*
3. **Hanyu Li**, Weizhi Ma, Peijie Sun, Jiayu Li, Cunxiang Yin, Yancheng He, Guoqiang Xu, Min Zhang, Shaoping Ma. **Aiming at the Target: Filter Collaborative Information for Cross-Domain Recommendation.** (SIGIR 2024, CCF A). *[论文链接]*
4. Jiayu Li*, **Hanyu Li***, Zhiyu He, Weizhi Ma, Peijie Sun, Min Zhang, Shaoping Ma. **ReChorus2.0: A Modular and Task-Flexible Recommendation Library.** (RecSys 2024, CCF B). *[论文链接]*

项目经历（以下成果均为一作或共一）

凝胶电泳图像中异常单克隆蛋白检测 *[论文1]* 纽约州立大学布法罗分校 2019.7-2021.3

- 电泳法检测单克隆蛋白对多种疾病的诊断至关重要。该工作用高斯混合模型表示电泳图像，并用峰值检测方法识别异常视觉特征。该分类器具有良好的可解释性，并且在小训练集上表现也较好。
- 该成果已发表在计算机科学会议视觉通信与图像处理（VCIP）*[代码链接]*。

考虑用户隐私的远程线上实验工具 *[论文2]* 清华 2020.9-2021.7

- 为了促进信息检索领域用户实地实验的开展，减少研究人员的开发工作量，消除平台设计与开发壁垒，并在数据的自然收集的过程中最大程度保护用户隐私，我们设计、开发了一个考虑用户隐私的开源远程用户行为实验工具 RUS-toolkit。
- 该成果已发表在顶级计算机科学会议国际信息检索大会（SIGIR demo paper, CCF A）*[代码链接]*。

考虑动态多样性需求的重排序 清华-腾讯 2021.7-2022.7

- 大多数多样性感知研究都认为，提供更多样化的结果总能提高用户满意度。然而，用户对多样性有不同程度的需求，这种需求在不同的交互会话中会发生动态变化。
- 我们通过对大规模真实推荐数据集的广泛分析验证了这一说法。然后，我们提出了一种新颖的重排序方法来满足用户的动态多样性需求，即考虑动态多样性的重排序（DDAR）模型。
- 该方法在多个数据集上取得显著提升，目前在投*[代码链接]*。

过滤协同信息的跨域推荐框架 *[论文3]* 清华-腾讯 2022.7-2024.1

- 尽管目前关于跨域推荐的研究已经相当丰富，但大部分工作主要集中在让重叠用户在不同领域的建模更具表现力，以传递领域间的信息，而在监督信号层面上解决负迁移问题的研究相对较少。此外，现有的方法大多是整体的跨域模型，无法跟上最新的单域推荐进展，整体架构不够灵活。
- 针对上述研究难题和现有模型的不足，我们提出了一种新的框架——协同信号正则化的用户转换(CUT)。该可扩展的跨域推荐框架通过过滤源域中无关信息，利用更多的有用知识，以提高目标领域的性能。同时，它还可以使用各种单领域推荐系统作为基础，并将其扩展到跨域推荐任务。
- 该成果已发表在顶级计算机科学会议国际信息检索大会（SIGIR full paper, CCF A）*[代码链接]*。

ReChorus2.0：适用于多种任务的模块化推荐系统开源代码库 *[论文4]* 清华 2024.1-2024.6

- 越来越多的研究关注于推荐系统中数据输入、模型和任务设置的各个方面。我们设计了一个灵活的开源研究工具来帮助研究人员高效地实现他们所需的实验策略。
- 现有的推荐场景开源库通常对输入数据有一定限制，很少支持同一模型的不同任务和输入格式，从而限制了用户进行定制化探索。为填补这一空白，我们提出了 ReChorus2.0，一个面向推荐系统研究人员的模块化、灵活、轻量化的开源库。基于 ReChorus，我们升级了支持的输入格式（包括丰富的上下文信息)、模型以及训练和评估策略，以帮助实现更多数据类型的推荐任务（包括重排序和 CTR 预测等)。ReChorus2.0 的实现和详细教程请关注*[仓库链接]*。
- 该成果已被推荐系统领域的国际顶级学术会议录用（RecSys reproducibility paper, CCF B)。

实习经历

利用神经网络不确定性处理带噪声数据集 京东 2019.12-2020.3

- 数据集标签的噪声问题广泛存在，应对噪声的办法包括正则化，重标签，噪声建模，样本加权，数据增强等。实习期间，广泛调研了回归分类等多种任务下量化神经网络不确定性的方法，以及其在处理数据集噪声场景下的应用。
- 该工作应用 MC dropout 方法量化神经网络的不确定性，并据此为样本分配权重进行训练，从而提升带噪声数据集的分类任务准确率。

基于大语言模型的冷启动跨域推荐（进行中） 快手 2024.6-至今

- 冷启动用户具有交互历史缺失，协同信息不足等特点，因此推荐系统需结合用户和物品的内容信息，包括用户画像和物品描述等来进行推荐。而在过往的文献中，大模型因为其丰富的世界知识，已经有工作将其应用在冷启动推荐任务上。我们计划利用大语言模型整合用户物品的内容信息，对电商场景下的冷启用户推荐表现进行增强。
- 目前需要解决的困难包括如何对物品进行编码和表征，监督微调过程中训练辅助任务和提示词的设计，推理过程任务和训练过程的对齐，生成目标商品的方式选择等。
- 电商场景下的冷启用户往往在视频域有交互历史，而商品和视频之间的语义信息具有较强的关联性，因此将视频域交互引入到用户画像中能够提升商品推荐效果。下一步我们计划对引入跨域语义信息的方式进行探索。

技术能力

- **编程语言：** Python, C, C++, Java, SQL
- **开发平台：** Pytorch, Tensorflow, Django
- **英语：** TOEFL: 109, GRE: 329

学术活动与个人荣誉

清华大学《信息检索的前沿研究》课程助教 2022 年 3 月-至今
M 奖 (前 7%), 美国大学生数学建模竞赛 2019 年 1 月