姚亮

yaoliang@zju.edu.cn
深圳市南山区科兴科学园
https://yao8839836.github.io/

研究兴趣

自然语言处理　数据挖掘　医学信息学　大语言模型　图神经网络　概率图模型

工作经历

腾讯科技 (深圳) 有限公司 *2019 年 9 月- 至今*
高级研究员

教育经历

美国西北大学 (Northwestern University) *2018 年 1 月- 2019 年 9 月*
博士后研究员 (Postdoctoral Fellow)

浙江大学 *2012 年 9 月- 2017 年 9 月*
工学博士, 计算机科学与技术

四川大学 *2008 年 9 月- 2012 年 6 月*
工学学士，计算机科学与技术
GPA: 87/100
经济学双学士，金融学第二专业

发表论文　谷歌学术引用: 3120　H 指数: 21

Liang Yao, Chengsheng Mao, Yuan Luo. Graph Convolutional Networks for Text Classification. *33rd AAAI Conference on Artificial Intelligence* (**AAAI 2019**), pp. 7370-7377. (CCF A 类, 引用: **1702**, GitHub star: **1296**)

Liang Yao, Chengsheng Mao, Yuan Luo. KG-BERT: BERT for Knowledge Graph Completion. Submitted to *IEEE Transactions on Pattern Analysis and Machine Intelligence* (**TPAMI**, arXiv 引用: **449**, GitHub star: **589**)

Liang Yao, Jiazhen Peng, Chengsheng Mao, Yuan Luo. Exploring Large Language Models for Knowledge Graph Completion. arXiv preprint arXiv:2308.13916.

Liang Yao. K-LLaMA: Knowledgeable LLMs for Question Answering. Work in progress.

Liang Yao, Jiazhen Peng, Shenggong Ji, Qiang Liu, Hongyun Cai, Feng He, Xu Cheng. Friend Ranking in Online Games via Pre-training Edge Transformers. Accepted by **SIGIR 2023**. (CCF A 类)

Liang Yao, Yin Zhang, Baogang Wei, Wenjin Zhang, Zhe Jin. A Topic Modeling Approach for Traditional Chinese Medicine Prescriptions. *IEEE Transactions on Knowledge and Data Engineering* (**TKDE**) 30.6 (2018): 1007-1021. (CCF A 类, SCI, IF:8.9)

Liang Yao, Yin Zhang, Baogang Wei, Zhe Jin, Rui Zhang, Yangyang Zhang, Qinfei Chen. Incorporating Knowledge Graph Embeddings into Topic Modeling. In *31st AAAI Conference on Artificial Intelligence* (**AAAI 2017**) pp. 3119-3126. (CCF A 类)

Liang Yao*, Xu Cheng*, Feng He, Chenhui Zhang, Wenzheng Feng, Jie Tang. PB-GNN: Partition-based Billion-scale Graph Neural Networks Framework for Online Games. Submitted to **ICDE 2024**. *equal contribution

Liang Yao, Zhe Jin, Chengsheng Mao, Yin Zhang and Yuan Luo. Traditional Chinese Medicine Clinical Records Classification with BERT and Domain Specific Corpora. *Journal of the American Medical Informatics Association* (**JAMIA**). 26, no. 12 (2019): 1632-1636. (CCF B 类, SCI, IF: 6.4)

Liang Yao, Chengsheng Mao, Yuan Luo. Clinical Text Classification with Rule-based Features and Knowledge-guided Convolutional Neural Networks. *BMC Medical Informatics and Decision Making*, 9(3) (2019), p.71. SCI, IF: 3.5

Liang Yao, Yin Zhang, Baogang Wei, Zherong Li. Traditional Chinese Medicine Clinical Records Classification using Knowledge-Powered Document Embedding. In *2016 IEEE International Conference on Bioinformatics and Biomedicine* (**BIBM 2016**). (CCF B 类)

Liang Yao, Yin Zhang, Baogang Wei, Lei Li, Fei Wu, Peng Zhang, Yali Bian. Concept over time: the combination of probabilistic topic model with Wikipedia knowledge. *Expert Systems with Applications* 60 (2016): 27-38. (SCI, IF:8.5) (CCF C 类)

Liang Yao, Yin Zhang, Qinfei Chen, Hongze Qian, Baogang Wei, Zhifeng Hu. Mining Coherent Topics in Documents using Word Embeddings and Large-scale Text Data. *Engineering Applications of Artificial Intelligence* 64 (2017): 432-439. (SCI, IF:8.0) (CCF C 类)

Liang Yao, Yin Zhang, Baogang Wei, Wei Wang, Yuejiao Zhang, Xiaolin Ren, and Yali Bian. Discovering treatment patterns in Traditional Chinese Medicine clinical cases by exploiting supervised topic model and domain knowledge. *Journal of Biomedical Informatics* 58 (2015): 260-267. (CCF C 类, SCI, IF:4.5)

Yang Zhao, **Liang Yao**, Yin Zhang. Purchase prediction using Tmall-specific features. *Concurrency and Computation: Practice and Experience* 28.14 (2016): 3879-3894. (CCF C 类, SCI, IF:2.0)

Liang Yao, Yin Zhang, Baogang Wei, Hongze Qian, and Yibing Wang. Incorporating probabilistic knowledge into topic models. In *19th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (**PAKDD 2015**). (CCF C 类)

Shansong Yang, Weiming Lu, Dezhi Yang, **Liang Yao**, and Baogang Wei. Short text understanding by leveraging knowledge into topic model. In *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (**NAACL 2015**) (CCF B 类)

Xiangzhou Huang, Yin Zhang, Baogang Wei, and **Liang Yao**. A question-answering system over Traditional Chinese Medicine. In **BIBM 2015**. (CCF B 类)

Liang Yao, Yin Zhang, Baogang Wei, Wei Wang, Yuejiao Zhang, and Xiaolin Ren. Discovering treatment pattern in traditional Chinese medicine clinical cases using topic model and domain knowledge. In **BIBM 2014**. (CCF B 类)

Liang Yao, Yin Zhang. and Baogang Wei, 2014. An Evolution System for Traditional Chinese Medicine Prescription. In *Knowledge Engineering and Management* (pp. 95-106). Springer Berlin Heidelberg.

发明专利

一种基于图分割和掩码自监督的图表征学习算法以及在游戏未成人年识别的应用 (荣获腾讯 “鹅厂好专利”)
一种引入领域知识图谱的多语言多品类游戏玩家名字生成方法
一种基于多维数据的关键用户识别方法和智能游戏拉新方法
一种引入时序的异构网络表征学习推荐算法

专业活动

会议 PC/审稿人: AAAI, IJCAI, ACL, EMNLP, ECML-PKDD.
期刊审稿人: TKDE, Neural Networks, ACM Computing Surveys, JBI, etc

研究和工程经历

大规模图神经网络 *2019 年 9 月- 至今*

将图神经网络用于腾讯游戏异构网络

- 9 亿以上点, 1000 亿以上边
- 节点分类 + 链路预测
- 图分割 + 子图 GCN 优化
- 特征传播, 图增强 MLP
- 自创 Edge Transformer，并首次用 Edge MAE + 大规模边预训练, 发表于顶会 SIGIR 2023
- 广泛用于好友召回、拉新、未成年人识别、道具推荐等业务
- 多个业务 A/B 测试效果提升 10% 以上。训练提速 50 倍
- 图机器学习挑战赛 ogbl-collab 全球第一
- 知识图谱链路预测挑战赛 ogbl-wikikg2 全球第一

NLP 大模型 *2023 年 2 月- 至今*

NLP 大模型 + 知识图谱

- 在知识图谱补全任务微调 LLaMA, ChatGLM 等，性能超过 GPT-4 和 ChatGPT
- 知识增强的 NLP 大模型用于问答，获问答系统挑战赛 OpenBookQA 全球第一
- 在部门业务数据落地

知识引导的游戏玩家名字生成 *2021 年 12 月- 2022 年 07 月*

将知识图谱与 GPT2 等生成模型结合

- 训练游戏知识图谱 embedding
- prompt + 知识
- 生成符合业务预期的玩家名字

基于 BERT 的知识图谱补全 *2019 年 7 月 - 2019 年 9 月*

利用预训练 BERT 模型和实体、关系描述预测三元组的合理性。

- 将知识图谱补全问题转化为文本序列分类问题。
- 拼接三元组中的实体/关系的文字描述语句作为 BERT 输入。
- 微调 BERT 以预测三元组序列是否为真，或两个实体间的关系。
- 在三元组分类、链接预测、关系预测任务中取得最优结果。

基于 BERT 和领域语料的中医病例分类 *2019 年 4 月 - 2019 年 6 月*

利用中文预训练 BERT 模型和大规模无标注领域语料训练中医 BERT 模型

- 用无标注中医病例继续微调语言模型。
- 用微调后的中医 BERT 初始化文本分类器。
- 用有标签的病例微调文本分类器，取得最优结果。

基于图卷积网络 (GCN) 的文本和知识图谱建模 *2018 年 4 月 - 2019 年 3 月*

基于图卷积网络 (GCN) 的文本分类:

- 利用词的共生关系和文档-词关系为语料库构图。
- 利用图卷积网络 (GCN) 学习文档/词向量。
- 预测未标注文档标签。

联合学习文本和知识图谱向量:

- 变分自编码器 (VAE)
- 图卷积网络 (GCN)

临床文本分类 *2018 年 1 月 - 2018 年 4 月*

肥胖症挑战赛，根据病例文本预测病人是否患有肥胖及十五种并发症:

- 用正则表达式抽取疾病的各种表达及其疑问/否定形式
- 利用医疗文本得到医疗词向量和医疗实体 (UMLS CUIs) 向量
- 知识引导的卷积神经网络 (CNN), TensorFlow 实现

基于知识的主题模型 *2013 年 9 月 - 2017 年 9 月*

将外部知识引入主题模型 (LDA), 以提高其主题可解释性和特征表示能力，外部知识来源包括:

- 知识图谱 (例如微软 Probase, YAGO)。
- 维基百科。
- 大数据文本 (利用词向量编码知识)。

中医药数据中的知识发现 *2013 年 9 月 - 2017 年 9 月*

- 从中医处方中发现治疗模式, 以实现对不同的症状表现推荐处方，预测处方的功效等任务。

- 从中医病例中发现治疗模式。
- 中医病例的科属分类。

利用天猫历史数据进行购买预测 *2014 年 11 月 - 2014 年 12 月*

- 设计适合天猫数据的特征，包括点击特征，购物车特征，购买特征等。
- 探索了不同分类器的预测性能。

中草药专业知识服务系统 *2012 年 9 月 - 2015 年 9 月*

- http://zcy.ckcest.cn/tcm/
- 贡献了超过 8000 行的前端/后端代码，主要完成药方病证分析系统。
- 使用 Spring MVC + MyBatis 框架。

程序设计语言和工具

程序设计语言	Python, Java, Spark/Scala, C/C++, Matlab, R, LaTeX, Javascript
工具	SVN, GitHub, TensorFlow, PyTorch, Keras, Stanford CoreNLP, NLTK, Mallet

荣誉和获奖情况

全球前 2% 顶尖科学家 (2022, 2023; 斯坦福大学发布)
AI 2000 人工智能全球最具影响力学者 (2022, 2023; 清华大学 AMiner 团队发布)
问答系统挑战赛 OpenBookQA 全球第一 (2023)
知识图谱链路预测挑战赛 ogbl-wikikg2 全球第一 (2023)
图机器学习挑战赛 ogbl-collab 全球第一 (2022)
腾讯 Outstanding Contributor (五星员工)，2022
腾讯 IEG 公共数据平台部季度之星 (2022)
腾讯 IEG Tech Future 项目优秀个人 (2022)
深圳市海外高层次人才 (2020)
欧琳奖学金 (2016)
优秀博士生岗位助学金 (2015, 2016)
浙江大学优秀研究生 (2013, 2015, 2016)
国家励志奖学金 (2010)
四川大学优秀学生 (2009)