**附录 4**

**北京航空航天大学软件学院**

工程实践开题报告

课题名称： 基于强化学习的文本推荐系统的设计与实现

姓 名： 张韬 赵庆国

学 号： ZF2021428 ZF2021440

专业方向： 人工智能

所属院系： 北京航空航天大学软件学院

指导教师： 原仓周

实践起止时间： 2021 年 9 月 至 2022 年 1 月

**目录**

[1课题背景与意义 2](#_Toc356991466)

[2 国内外相关研究现状 3](#_Toc191339322)

[3课题目标及内容 4](#_Toc1067176184)

[5. 预期成果(含功能/性能指标） 6](#_Toc26351069)

[6. 工作计划 7](#_Toc352352604)

[7. 参考文献 7](#_Toc1946672022)

### 1课题背景与意义

#### 1.1 课题背景

我们生活在信息爆炸的时代。信息过载的问题，使得人们难以快速做出正确的决策，比如当我们在网上的购物商城中遇到一长串物品列表时，便意识到列表中的物品越多，从中选择就越困难。推荐系统的出现，就是为了解决这类问题，推荐系统通过预测用户对物品的偏好或评分来帮助用户找到他们感兴趣的物品。实际上，该想法是基于用户过去对物品的反馈以及在某种程序上对用户的了解，来推荐与他们的偏好相匹配的物品，如今，推荐系统己经成为互联网公司的重要组成部分，并被广泛应用于电子商务，新闻，在线学习等领域，而且未来将继续发挥重要作用

#### 1.2 选题意义

推荐系统中最常用的方法有基于用户-物品的协同过滤和矩阵分解，但是由于用户与物品的直接交互的数据中，普遍存在数据稀疏性的问题，推荐质量分受到很大的限制，而且这些方法都忽略了联信息，所以有人提出了基于图的算法和链接预测算法，来探索用户与物品间的关联传递，推荐问题可以看做是异构信息网络中的链路预测问题。

深度强化学习技术DRL，在Atari和围棋游戏中，己经取得了非常好的效果，近年来，将强化学习RL应用于推荐系统己越来越多的被大型公司所采用。与传统的协同过滤和基于深度神经网络的方法相比，强化学习可以优化针对长期目标的一系列推荐决策，现在的大多数研究工作，都提出了训练强化学习Agent与用户进行交互的方法，对于这种方法，由于人类行为的复杂性，难以从线上用户中获取到大量样本来训练Agent。因此需要找到更好的方法来提高强化学习对推荐系统的适配效果。

### 2 国内外相关研究现状

传统的推荐系统主要是基于明确评分的推存方法，包括预测用户评分和向用户推荐物品。主要方法有协同过滤和因子分解等。而在异构信息网络中，基于元路径相似性度量的方法，用于在不同语义假设下，测量信息网络中未观测到的用户和物体之间交互的可能性。在网络模型中，链接预测是一个关键问题，社交网络，文献引用网络，和推荐系统等领域，都有重大进展。该类方法考虑了用户和物体在图中的交互，这是协同过滤等方法中所忽视的。然而在异构信息网络中，元路径都是事先同人给同，对如何生成高质量的元路径的研究少之又少。近年来，深度学习作为最有前景的方法，被广泛用于如目标检测，语言识别等任务中。在推荐系统中，有研究者利用深度网络进行评分预测，也有研究利用知识图谱，把用户偏好传播到未知物品上，最近Wang等人，提出一和种考虑用户和物品节点之间所有hop-n链接的图谱注意力网络。不过由于深度学习对算力和数据的巨大需求和难以解释，都对这些方法的落地构成挑战。

上述的推荐系统都是基于静态的数据进行预测，在预测时，损失了用户与系统交互时的时序性质，而强化学习可以有效的处理这种时序信息。与上述方法不同之处在于，强化学习通常为了一个长期目标优化一系列推荐决策，大多数基于强化学习的推荐系统都是无模型的方法，通常需要与环境进行大量交互才能学习到一个好的策略，例如，Heocharous等人讨论了一种个性化广告推荐系统，该系统具有非策略评估保证，可实现终身价值优化，Zhou等人提出了一种基于马尔科夫决策过程MDP的解决方案来跟踪用户的兴趣转移，并直接优化即时和延迟的用户参与度指标，由于线上用户在推荐系统看起来随机且不符合兴趣的情况下会很快放弃推荐服务，因此提出了基于模型的强化学习方法来避免无模型方法的大样本复杂性，例如Chen等人提出了一种基于模型的强化不官瘾方法，用于学习基于页面视图和交互模式的用户模型。

### 3课题目标及内容

#### 3.1 课题目标

推荐系统是帮助人们解决信息获取问题的有效工具，本项目旨在实现一个可以对文本网页周期性爬取，对文本数据进行清洗分类后，使用NLP技术生成摘要，再根据用户的浏览偏好，基于主流的推荐算法和强化学习，进行个性化推荐的推荐系统。

#### 3.2 课题内容

本项目实现内容包括以下推荐系统的核心模块：

1 数据收集模块，构建推荐模型需要收集很多数据，包括用户行为数据，用户相关数据及推荐文本的相关数据。

2 ETL模块，通过后台管理模块将数据存储到SQLite数据库中。

3 特征工程模块，对原始数据进行降维和特征提取，将数据转换为推荐算法可以学习的特征矩阵。

4 推荐算法模块，推荐系统尝试采用多种机器学习算法、以及强化学习算法来学习用户偏好，在比较不同算法的效果后，从中选取最优者，并基于用户偏好来为用户推荐文本。

5 推荐结果存储模块，周期性的更新用户推荐结果，保存在Redis等易于横向扩展的数据库。

6 Web服务模块，主要作用是当用户在UI上触达推荐系统时，触发推荐接口，为用户提供个性化推荐。

7 模拟环境模块，一个用于训练强化学习的环境。

#### 3.3 实现（或研究）方法与技术线路

主要使用BeautifulSoup库以及正则表达式来爬取文本信息并进行初步的过滤，将纯文本结果存入SQLite数据库。

使用抽取式（textrank）和生成式（seq2seq）方法分别进行摘要的自动生成，并进行效果比较。

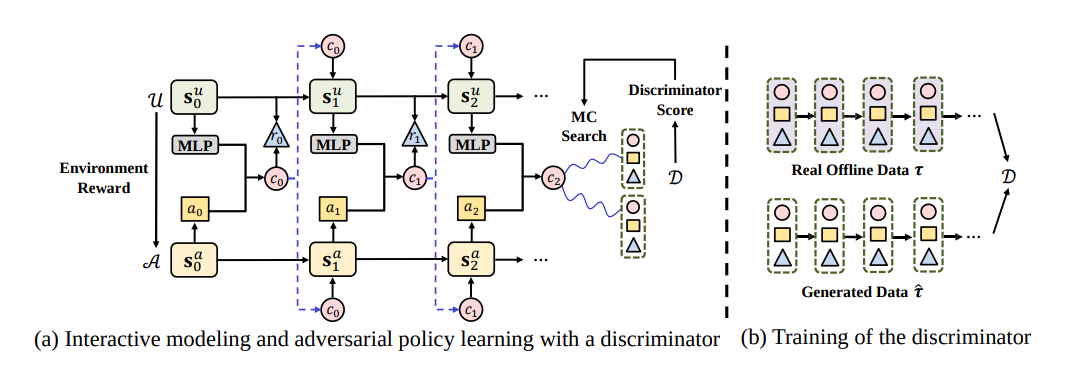
基于K-means等机器学习方法对文本进行分类，根据pagerank方法进行评分，用于在推荐系统冷启动或者探索时，在评分最高的类里选取topk。

使用DQN，A3C和IRecGAN的等强化学习方法对推荐Agent在模块环境中进行训练，以测试不同方法的效果。

使用Flask搭建web服务模块，与用户交互，留存日志做为下次推荐的依据。

#### 4. 关键技术难点与解决方案

基于模型的RL具有采样效率高、离线时有助于降低噪声等优点数据然而，由于其对真实环境的模型近似存在固有的偏差，这种优势很容易减弱。此外，在随后的政策更新中发生了巨大变化用户满意度降低的风险，即模型更新中的建议不一致。为了解决这些问题，我们在推荐人的策略学习中引入了对抗性训练，从脱机数据对鉴别器进行训练，以区分模拟的交互轨迹和真实性，以作为用户行为模型和改进策略学习。这是第一项在基于模型的RL框架上探索对抗性训练的工作，所参考的训练网络如下图所示：



### 5. 预期成果(含功能/性能指标）

实现全部主体功能模块，推荐系统在保证用户黏性和考虑长期收益的情况下，可以自动周期性探索新信息，并根据用户是否点击反馈进行调整。

用户界面响应时间在小于1秒。

### 6. 工作计划

10月 完成技术积累，并实现基本的框架代码。

11月 完成主体和模拟环境的编码，单元测试和系统集成。

12月 完成集成测试和技术文档写作。

### 7. 参考文献

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 22–31. JMLR. org, 2017.

1. Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. Top-k off-policy correction for a reinforce recommender system. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pages 456–464. ACM, 2019.
2. Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. Generative adversarial user model for reinforcement learning based recommendation system. In Proceedings of the 36th International Conference on Machine Learning, volume 97, pages 1052–1061, 2019.
3. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014.
4. Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In Proceedings of the 28th International Conference on machine learning (ICML-11), pages 465–472, 2011.
5. Marc Peter Deisenroth, Carl Edward Rasmussen, and Dieter Fox. Learning to control a low-cost manipulator using data-efficient reinforcement learning. 2011.
6. Marc Peter Deisenroth, Gerhard Neumann, Jan Peters, et al. A survey on policy search for robotics. Foundations and TrendsR in Robotics, 2(1–2):1–142, 2013.
7. Alexandre Gilotte, Clément Calauzènes, Thomas Nedelec, Alexandre Abraham, and Simon Dollé. Offline a/b testing for recommender systems. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, pages 198–206. ACM, 2018.
8. Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.
9. Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In International Conference on Machine Learning, pages 2829–2838, 2016.
10. Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pages 549–558. ACM, 2016.
11. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8): 1735–1780, 1997.
12. Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. Computer, (8):30–37, 2009.
13. Reinforcement Learning. An introduction, richard s. sutton and andrew g. barto, 1998.
14. Elad Liebman, Maytal Saar-Tsechansky, and Peter Stone. Dj-mc: A reinforcement-learning agent for music playlist recommendation. In Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, pages 591–599. International Foundation for Autonomous Agents and Multiagent Systems, 2015.
15. Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971, 2015.
16. Zhongqi Lu and Qiang Yang. Partially observable markov decision process for recommender systems. arXiv preprint arXiv:1608.07793, 2016.
17. David Meger, Juan Camilo Gamboa Higuera, Anqi Xu, Philippe Giguere, and Gregory Dudek. Learning legged swimming gaits from experience. In 2015 IEEE International Conference on Robotics and Automation (ICRA), pages 2332–2338. IEEE, 2015.
18. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013.
19. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529, 2015.
20. Jun Morimoto and Christopher G Atkeson. Minimax differential dynamic programming: An application to robust biped walking. In Advances in neural information processing systems, pages 1563–1570, 2003.
21. Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. In Advances in Neural Information Processing Systems, pages 1054–1062, 2016.
22. Junhyuk Oh, Satinder Singh, and Honglak Lee. Value prediction network. In Advances in Neural Information Processing Systems, pages 6118–6128, 2017.
23. Baolin Peng, Xiujun Li, Jianfeng Gao, Jingjing Liu, Kam-Fai Wong, and Shang-Yu Su. Deep dyna-q: Integrating planning for task-completion dialogue policy learning. arXiv preprint arXiv:1801.06176, 2018.
24. Doina Precup. Eligibility traces for off-policy policy evaluation. Computer Science Department Faculty Publication Series, page 80, 2000.
25. Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In ICML, pages 417–424, 2001.
26. John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International Conference on Machine Learning, pages 1889–1897, 2015.
27. Guy Shani, David Heckerman, and Ronen I Brafman. An mdp-based recommender system. Journal of Machine Learning Research, 6(Sep):1265–1295, 2005.
28. Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Machine Learning Proceedings 1990, pages 216–224. Elsevier, 1990.
29. Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pages 1057–1063, 2000.
30. Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. Journal of Machine Learning Research, 16(1): 1731–1755, 2015.
31. Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In advances in neural information processing systems, pages 3231–3239, 2015.
32. Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In International Conference on Machine Learning, pages 2139–2148, 2016.
33. Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256, 1992.
34. Qingyun Wu, Hongning Wang, Liangjie Hong, and Yue Shi. Returning is believing: Optimizing long-term user engagement in recommender systems. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 1927–1936. ACM, 2017.
35. Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In Thirty-First AAAI Conference on Artificial Intelligence, 2017.
36. Xiangyu Zhao, Long Xia, Yihong Zhao, Dawei Yin, and Jiliang Tang. Model-based reinforcement learning for whole-chain recommendations. arXiv preprint arXiv:1902.03987, 2019.
37. Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. Drn: A deep reinforcement learning framework for news recommendation. In Proceedings of the 2018 World Wide Web Conference on World Wide Web, pages 167–176. International World Wide Web Conferences Steering Committee, 2018.