电 子 科 技 大 学

专业学位研究生学位论文开题报告表

攻读学位级别： □博士 ☑硕士

培养方式： □全日制 ☑非全日制

专业学位类别及领域： 计算机技术 云计算与大数据处理

学 院： 计算机科学与工程学院

学 号： 201952080901

姓 名： 刘邦健

论文题目： 基于用户行为特征的智能

推荐系统研究与应用

校内指导教师： 蔡洪斌

校外指导教师：

填表日期： 2021 年 03 月 01 日

电子科技大学研究生院

1. 学位论文研究内容

|  |  |  |
| --- | --- | --- |
| 课题类型 | | □应用基础研究 ☑应用研究 |
| 课题来源 | | □纵向 □横向 ☑自拟 |
| 学  位  论  文  研  究  内  容 | 学位论文的研究目标、研究内容及拟解决的关键性问题（可续页）  **研究目标**：  本文的研究目标是探索传统冷启动时协同过滤，以及大数据环境下机器学习等相关算法在推荐系统中的使用，并以此设计提出一套适用于教育、经济领域，包含视频，文章等内容推荐的算法方案及其应用系统。其中，推荐系统的算法部分研究包括选用合适的数据预处理技术、提升算法模型的训练、运行效率，以及提高推荐准确度三方面内容。而推荐系统的平台构建部分则采用java相关主流web技术，通过结合算法的研究开发出客户群体与目标资料基本匹配、平台使用体验基本满意的推荐系统。  **研究内容**：  随着互联网经济的蓬勃发展，互联网广告、在线教育、网上购物因为其覆盖范围广、信息量大、互动性强等优势已经在现代社会大面积铺展开来。而推荐系统的相关算法则是构成互联网实用性和方便性的重要基石。本文研究的主要内容包括：1)在推荐系统中对样本数据进行合理的特征设计2)选用合适的机器学习模型结构3)如何通过设置合适的超参数提升算法准确率。通过查阅相关文献，从数据预处理，基础系统的构建到算法调试，多方面分析及验证现阶段主流推荐算法模型的适用性及优缺点。  **拟解决的关键问题：**   1. 对于一般推荐系统平台初期，经常会存在初始数据不足的冷启动问题。需要解决训练样本数据初始数据源的获取以及如何不断获取后续数据集的问题。 2. 数据特征描述对于推荐算法来说至关重要，样本数据的特征设计需要满足系统的需要，并且要便于收集。 3. 在实际应用场景下，采用Factorization Machine（FM），Wide&Deep learning（WDL）等推荐算法进行模型构建，完成相应超参数调试及解决期间遇见的问题。 4. 本文为了方便用户使用，采用B/S架构对平台进行设计，需要解决web平台相关功能的设计开发问题。 | |

1. 学位论文研究依据

|  |
| --- |
| 学位论文的选题依据和研究意义，国内外研究现状和发展态势；选题在理论研究或实际应用方面的意义和价值；主要参考文献，以及已有的工作积累和研究成果。（2000字）  **1研究背景及意义**  在大数据，互联网经济蓬勃发展的大背景下，大型互联网公司凭借其优越的推荐系统对用户进行精准刻画，准确的为客户推荐相关的商品，广告，视频等内容，带来了极大的便利性和舒适性。由中国互联网络信息中心(CNNIC)发布的第42次和第43次《中国互联网络发展状况统计报告》，2018年互联网广告收入比上年增长24．2％达到3694亿元，其中上半年收入的同比增长率达到32％，而报纸和户外广告则均呈现出了负增长态势。[1]广告收入已经成为互联网公司经济收入的重要支柱，而推荐算法相关领域的研究也早早成为了企业，科研工作者研究的热门方向。  在机器学习还没大规模兴起的初期，AMAZON、Google等互联网公司均采用了协同过滤的方式来构建其推荐系统。然而随着大数据时代的到来，单一的协同过滤模型在其准确性不够，性能不高，系统延伸性不强等缺点上表现越来越突出，逐渐被一些兴起算法所替代。其中对后续研究产生重要影响的有Rendle等人对于FM算法的提出和研究[2,3]以及机器学习神经网络算法的快速发展。根据相关文献的查阅，现阶段主流的WDL,DeepFM,xDeepFM等推荐算法均采用了上述两种算法的相关原理，来进行演化和实现。本文也将基于这个思路，对推荐系统相关算法根据实验室与北京某公司合作的金融大脑，速百满教育平台等项目，在实际相关环境中进行应用与研究。  **2 国内外研究现状及发展趋势**  对于推荐系统这一研究课题，目前国内外许多研究者已经做了大量的相关研究。在2007年时，Richardson[4] 等人使用了逻辑回归模型(Logistic Regression, 简称 LR)预估 CTR并取得了一定成功。但由于其训练数据具有稀疏性，为达到准确有效的推荐结果需要的训练成本非常的高昂，不利于普遍推广。Rendle[2]等人在2010针对训练数据稀疏性的特点，发表了FM算法相关论文并对LR，SVM[16]，SVD++[17]，PITF[18]等相关算法进行分析比较证明其优越性。FM算法主要针对稀疏数据下的准确性，模型训练速度，以及探索特征向量之间的关系以预测新数据（未在训练样本中出现的数据）三个方面并取得明显的效果。  对于数据预处理部分，一般的推荐算法针对离散数据特征都进行one-hot转换，但这样会导致数据特征维度较高从而具有稀疏性（one-hot数据展示如图2-1）。如对于user域包含Alice,Bob,Charlie等100万个用户，那么作为输入数据的user域描述则为（如Alice）{1,0,0,0,0...}一百万维的高维向量，而且在此基础上还得再加上item域等其他参数域数据。而为了将高维数据用算法模型进行模拟，以LR为例则需要大量的数据去feed并且花费大量的计算资源和时间去训练才能得到较为可靠的效果。而且由于LR算法对于二阶以上的特征关系维护起来计算复杂度O(n2)较高，此处也是导致其不具有高效性的原因之一。  IMG_256  图2-1 电影推荐one-hot数据  对于二阶的FM模型的公式如下：  （2-1）  Vi是k维的权值向量，而k是需要模型设计者设置的超参数，只要k设计的足够大就可以表示任意的正定矩阵W。再根据Rendle的论文可证明在梯度下降的训练过程中，可化简计算方程使得时间复杂度降为O(n)。其次，在FM中由于通过潜向量Vi和Vj的内积来表示数据之间的关系，这种灵活的设计可以使得每当i(或j)出现在数据记录中时，FM就可以训练潜在向量Vi (Vj)。因此，FM可以更好地学习到训练数据中从未或很少出现的特征交互。在之后的2016年，为了强化样本数据中每一个域之间的关系，使得训练结果更直观更有效Juan Y[14]继续以FM为基础提出了FFM算法。  与此同时国内Xu 等人[5]以及 Xiong 等人[6]也分别在2010 年和 2012 年均对用户行为对广告推荐的影响做了相关研究。从2014年开始后大量的神经网络相关的模型[10]被发掘加入推荐系统的研究之中，先是基于RNN[8]，CNN[7]对CTR预测模型进行扩展。但是CNN只能考虑相邻特征之间的相互关系，而RNN则有是连续且密集的点击行为的限制。而且CTR预测的神经网络会与CNN,RNN有非常大的不同，具体来说，用于CTR预测的原始特征输入向量通常是高度稀疏、超高维、分类连续混合和按领域分组(如:性别、地点、年龄)。所以需要使用一层嵌入层，将稀疏向量压缩成一个低维、密集的实值向量，然后再进一步输入到第一隐藏层，否则网络会难以训练。之后提出的FNN[9]，PNN[11]考虑到了高阶特征交互泛化性强，但是却对低阶特征交互缺少关注使得泛化过强对样本数据需求较大。于2016年，Google团队[12]设计了的具有可观性能和效率的Wide&Deep learning模型（相关模型结构如图2-2），其同时考虑到了低阶特征和高阶特征，使模型更具有generalization和memorization。其中wide部分是一个线性模型，共现频率较高的特征组合能达到一个不错的baseline且可解释性强具有memorization。而deep部分具有generalization，对高阶的特征关系进行泛化。值得注意的一点是，wide部分和deep不是ensemble而是一个joint training，其中的不同的是ensemble是各练各的，而joint training从joint一词中就可以看出，wide侧和deep侧在训练中是有联动，同时进行优化的。在训练时，论文中对wide侧使用了FTRL[15]算法带L1正则，对deep侧使用了adagrad进行优化。但WDL模型仍然需要进行特征工程处理，需要注意哪些特征应该送进wide侧，哪些特征送进deep侧。在论文的案例中，作者将交叉项以及那些binary项送入了wide侧，deep侧在wide侧有的特征基础上继续增加了continuous的特征。  图2-2 CTR模型结构:FNN,PNN,WDL  在2017年Guo的团队在WDL的基础上继续完善混合模型的思路，提出了DeepFM[13]的方案（模型结构如图2-3）。其中DeepFM也是由两部分组合而来，对于低阶特征部分采用FM模型来处理，如上文所说FM模型处理低阶特征时具有相当的优越性，而高阶特征交互部分则依然采用DNN的结构。其中DeepFM的嵌入层有两个特点。一是尽管不同field的输入长度不同，但是embedding之后向量的长度均为K。第二个的话是在FM里得到的隐变量Vik现在也作为了嵌入层网络的权重，在输入端使用统一的embedding vector。这样DeepFM就实现端到端的输入输出，避免了像WDL这样需要做特征工程的步骤，降低了了系统的复杂度。  图2-3 DeepFM结构图  其中DeepFM中输出的联合训练模型的计算公式如下：  （2-2）  根据Guo团队的实验数据，展现的FM,FNN,PNN,WDL和DeepFM等算法的性能分析对比如图2-4（第一行表示数据来源）：  图2-4 算法模型的性能比较  最近仍然有新的算法被不断发掘出来，如2018年中科大、北大与微软合作发表的XDeepFM。XDeepFM与DeepFM虽然都是从WDL延伸而来，却也不尽相同而是基于DCN[20]模型提出的新模型。其最主要的区别是在wide部分自动地构造有限高阶的特征叉乘，xDeepFM将基于Field的vector-wise思想引入Cross，并且保留了Cross的优势。模型结构非常的简练，实验效果也得到了明显的提升。如果说DeepFM理解为“Deep & FM”，那么xDeepFM就可以理解为真正做到了“Deep”Factorization Machine。  **参考文献：**  [1] 吴翌琳,南金伶.互联网企业广告收入预测研究——基于低频数据的神经网络和时间序列组合模型[J].统计研究,2020,37(5):94-103. DOI:10.19343/j.cnki.11-1302/c.2020.05.008.  [2] S. Rendle, “Factorization machines,” in Proceedings of IEEE International Conference on Data Mining (ICDM), pp. 995–1000, 2010.  [3] S. Rendle and L. Schmidt-Thieme, “Pairwise interaction tensor factorization for personalized tag recommendation,” in Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM), pp. 81–90, 2010.  [4] Richardson M, Dominowska E, Ragno R. Predicting clicks: estimating the click-through rate for new ads[C]//Proceedings of the 16th international conference on World Wide Web. ACM, 2007: 521-530.  [5] Xu W, Manavoglu E, Cantu-Paz E. Temporal click model for sponsored search[C]//Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010: 106-113.  [6] Xiong C, Wang T, Ding W, et al. Relational click prediction for sponsored search[C]//Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012: 493-502.  [7] Qiang Liu, Feng Yu, Shu Wu, and Liang Wang. A convolutional click prediction model. In CIKM,2015.  [8] Yuyu Zhang, Hanjun Dai, Chang Xu,Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and TieYan Liu. Sequential click prediction for sponsored search with recurrent neural networks. In AAAI, 2014.  [9] Weinan Zhang, Tianming Du, and Jun Wang. Deep learning over multi-field categorical data - -  A case study on user response prediction. In ECIR, 2016.  [10] Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. A deep and autoregressive approach for topic  modeling of multimodal data. IEEE Trans. Pattern Anal.Mach. Intell., 38(6):1056–1069, 2016.   1. Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. Product based neural networks for user response prediction. CoRR, abs/1611.00144, 2016. 2. [Cheng et al., 2016] Heng-Tze Cheng, Levent Koc,Jeremiah Harmsen, Tal Shaked, Tushar Chandra,Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. CoRR, abs/1606.07792, 2016. 3. Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17). AAAI Press, 1725–1731. 4. Juan Y , Zhuang Y , Chin W S , et al. Field-aware Factorization Machines for CTR Prediction[C]// the 10th ACM Conference. ACM, 2016. 5. H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. 2013. Ad click prediction: a view from the trenches. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '13). Association for Computing Machinery, New York, NY, USA, 1222–1230. DOI:https://doi.org/10.1145/2487575.2488200 6. Yin-Wen Chang, Cho-Jui Hsieh, KaiWei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear SVM. JMLR, 11:1471–1490, 2010. 7. Y. Koren, “Factorization meets the neighborhood: a multifaceted collaborative filtering model,” in KDD ’08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM, 2008, pp. 426–434. 8. S. Rendle and L. Schmidt-Thieme, “Pairwise interaction tensor factorization for personalized tag recommendation,” in WSDM ’10: Proceedings of the third ACM international conference on Web search and data mining. New York, NY, USA: ACM, 2010, pp. 81–90. 9. Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. XDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). Association for Computing Machinery, New York, NY, USA, 1754–1763. DOI:https://doi.org/10.1145/3219819.3220023 10. Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. arXiv preprint arXiv:1708.05123 (2017). |

1. 学位论文研究计划及预期目标

|  |
| --- |
| 1.拟采取的主要理论、研究方法、技术路线和实施方案（可续页）  本文主要以FM和WDL相关的算法模型为基础进行理论研究，最终用以完成实验室合作项目需求的推荐任务。FM算法主要负责描述样本数据的低阶特征交互，与注重高阶特征交互的神经网络算法相比具有解释性强的特点。FM算法的处理一般是先对样本数据进行one-hot转换使其变成结构简单，便于描述的稀疏数据。然后通过使用特征向量做内积，以描述不同样本数据之间的关系。m阶的FM算法的公式如下，可根据实际应用场景的需求进行阶数的选择。  （2-1）  FM算法在稀疏数据下的准确性，模型训练速度，以及探索特征向量之间的关系以预测新数据（未在训练样本中出现的数据）三个方面均表现出了优秀的特征。以WDL为基础的相关算法，在注重低阶特征交互的同时，混合DNN相关的神经网络模型，使算法同时关注了高阶特征更具有泛化性增强稳定性。如xDeepFM模型结构如下：  图3-1 xDeepFM模型结构  本论文拟打算先通过完善实验室已经搭建的好的金融大脑平台，速百满教育平台，开发出收集算法模型样本数据的有关模块，使其能完成初期训练数据的收集工作。同时再一边使用微软提供的开源数据集进行模型训练，对性能，适应场景等优缺点进行分析评估。然后选择最适用的算法以及相应的之前训练好的模型进行迁移学习，最后使用该模型对平台的数据进行正式训练并投入使用。在整个实践期间也将继续查阅相关文献，去优化算法模型相关内容。 |
| 2.研究计划可行性，研究条件落实情况，可能存在的问题及解决办法（可续页）  **研究计划可行性，研究条件落实情况**  本文探索使用的算法模型，已在谷歌，微软，华为等计算机，通讯领域的大公司得到充分验证。推荐算法应用的web平台已完成搭建，初始的训练数据源采用微软于2021年提供的开源数据集，训练平台可使用实验室已部署的Dell高性能服务器。  **可能存在的问题及解决办法**   1. 推荐系统平台初始数据不足导致的冷启动问题。   对于平台构建初期遇见的初始样本数据不足，而无法充分训练算法模型的问题。本文打算先使用基于物品的协同过滤等初始成本不高的推荐算法，用于前期数据收集过渡，待积累足够多用户行为数据后再采用FM，xDeepFM等机器学习相关算法来进行模型训练。同时使用微软于2021年2月份开源的推荐系统项目和数据源，做算法模型的训练和测试，为以后的迁移学习做准备。   1. 数据特征描述的设计及相关的问题。   本论文采用one-hot的方法对训练数据进行标识处理，但one-hot方法常常会由于其稀疏性而导致训练成本和数据需求成本过高。目前因式分解机（FM）算法由于其线性时间复杂度O(kn)、高度稀疏数据下表现良好、通用性高等特点，已经成为了解决这个问题的主流。现阶段的主流先进算法也大部分都借鉴了或者混合了FM算法的相关内容。   1. 在实际应用场景下，使用FM，WDL等推荐预测器进行的建模，完成超参数调试及解决期间遇见的问题。   此部分涉及到模型训练的结构设计，参数调试，是本论文研究的主要内容。可以通过继续查阅相关文献，以及使用数据集对模型进行试验研究，最终获取项目所需的算法模型。   1. web平台的设计与开发问题。   本文最后针对平台实际应用需求，拟采用java语言相关web技术设计一款简便、实用基于 B/S 架构的在线平台。平台将会收集用户相关的行为信息，通过推荐算法向目标客户提供相应的学习资料产品。 |

|  |  |  |
| --- | --- | --- |
| **3.研究计划及预期成果** | | |
| 研  究  计  划 | 起止年月 | 完成内容 |
| 2020.9-2021.3 | 阅读相关文献并了解推荐系统的相关研究方向和前沿内容。 |
| 2021.3-2021.6 | 完善搭建本论文内容实际应用的web平台，并开始进行样本数据的收集工作。 |
| 2021.6-2021.10 | 使用微软的数据集进行模型训练、实验，并完成模型迁移的相关准备工作。 |
| 2021.10-2022.9 | 将训练好的模型迁移至平台上进行实际应用，并从中继续完善模型相关内容。 |
| 2021.9-2022.3 | 整理平台数据，总结整体算法流程，撰写论文，并给出下一步研究方向 |
|  |  |
| 预  期  创  新  点  及  成  果  形  式 | 1. 预期创新点： 2. 采用因式分解机对低阶的特征交互进行处理，在稀疏样本数据的准确性，模型训练速度，以及探索特征向量之间的关系以预测新数据（未在训练样本中出现的数据）等三个方面均表现出了优秀的特征； 3. 采用同时关注低阶特征交互和高阶特征交互的混合算法模型，使得算法同时具有generalization和memorization的特征； 4. 避免复杂的特征工程处理，采用端到端的方式对算法模型进行建模，降低复杂度。 5. 预期成果形式： 6. 毕业论文 7. 算法验证程序 8. 在实验室合作的web项目中应用 | |

1. 开题报告审查意见

|  |  |  |  |
| --- | --- | --- | --- |
| 1.导师对学位论文选题和论文计划可行性意见，是否同意开题：  校内导师（组）签字： 年 月 日  校外导师签字： 年 月 日 | | | |
| **2.开题报告考评组意见** | | | |
| 开题日期 |  | 开题地点 |  |
| 考评专家 |  | | |
| 考评成绩 | 合格 票 基本合格 票 不合格 票 | | |
| 结 论 | □通过 □原则通过 □不通过  **通过：**表决票均为合格  **原则通过：**表决票中有1票为基本合格或不合格，其余为合格和基本合格  **不通过：**表决票中有2票及以上为不合格 | | |
| 考评组对学位论文的选题、研究计划及方案实施的可行性的意见和建议： | | | |
| 考评组签名：  年 月 日 | | | |
| **3.学院意见：** | | | |
| 负责人签名： 年 月 日 | | | |