**北 京 邮 电 大 学**

**本科毕业设计（论文）中期进展情况检查表**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 学院 | | 计算机学院 | 专业 | | 计算机科学与技术 | | 班级 | 2015211307 |
| 学生姓名 | | 王睿嘉 | 学号 | | 2015211906 | | 班内序号 | 31 |
| 指导教师姓名 | | 石川 | 所在单位 | | 计算机学院 | | 职称 | 教授 |
| 设计（论文）题目 | | （中文）异质信息网络中的相似性推荐算法研究 | | | | | | |
| （英文）Research on Similarity Recommendation Algorithm in Heterogeneous Information Network | | | | | | |
| 目前已完成任务 | 1. **调研研究现状，深入理解选题背景和意义**   推荐是解决信息过载的有效方法，被广泛应用于电子商务和互联网服务中。而融合更多信息进行混合推荐是解决数据稀疏性的有效技术。  近年来，异质信息网络（Heterogeneous Information Network），即多种类型节点和链路的建模方式，在数据挖掘领域大放异彩。作为重要的数据挖掘任务，推荐系统包括许多对象类型（如电影推荐中的用户、电影、演员和导演等）及对象间的丰富交互关系（如观看信息、社交信息和属性信息等），可以自然而然地利用异质信息网络建模。而异质信息网络优越的信息综合能力及丰富的语义信息使得更准确的推荐成为可能。  元路径（meta-path），即连接两类对象的关系组合，被广泛用于异质信息网络中的语义信息建模和关系抽取。Shi等人实现了一个基于语义的推荐系统HeteRecom，利用元路径的语义信息来评估电影间的相似性。此外，考虑属性值，如链接上的评分，他们进一步将推荐系统建模为加权异质信息网络，并提出基于语义路径的个性化推荐方法SemRec。为充分利用关系异质性，Yu等人引入基于元路径的潜在特征来表示沿不同类型路径的用户和产品间的连接，然后利用贝叶斯排名优化技术在全局和个性化水平上定义推荐模型。同样基于元路径，Burke等人提出在加权混合中包含多个关系的推荐方法。  但是，现有异质信息网络的推荐算法大多聚焦于用户和物品间的交互信息，即购买历史，而这在实际应用中仅能反映用户口味和物品特性的一个方面。因此，为深入挖掘不同方面的语义关系相似性，**Neural network based Aspect-level Collaborative Filtering model (NeuACF)**进一步设计基于元路径的推荐算法，使得推荐模型融合多方面语义信息，得到更加全面准确的结果。  但NeuACF中，用户和物品的交互可能性是利用embedding间的点积来度量的。由于点积本身所固有的缺陷——不满足三角不等式，可能导致相似性的传递特征被破坏，从而限制模型效果的提升。与此同时，NeuACF使用point-wise损失函数，关注于交互可能性的绝对数值。但对于Top-N推荐而言，相对交互可能性更为重要，因而通过修改损失函数显式捕捉排序信息，可能产生更为精准的推荐列表。   1. **确定具体改进方案，并根据实际需求选择数据集，编码实现模型**   模型分为三个模块，即首先抽取aspect-level特征，其次进行表示学习并融合，最后利用适当打分和损失函数完成优化，总体框架如下图所示。接下来，分模块阐述算法原理、与NeuACF的异同及相应实现进度。     1. **基于PathSim的aspect-level相似性矩阵计算（NeuACF）**     给定元路径，需要选择合适的算法抽取aspect-level特征。本工作遵循NeuACF，利用PathSim计算相似性矩阵，并作为特征输入后续模型。原因有二，其一，基于相似性可以缓解噪音；其二，利用相似性矩阵可以抽取高阶特征。  为此，查找PathSim源码，并在MATLAB运行，完成该模块的编程实现。   1. **基于Multi-Layer Perception（MLP）的aspect-level表示学习（NeuACF）**     相对于浅层模型而言，深层模型优越的非线性关系捕捉能力使其应用范围越来越广。因此，利用MLP将相似性矩阵作为输入，并输出相应aspect-level embedding：    其中，是相似性矩阵B中用户i的对应向量，是MLP第l层的隐向量，和是相应第l层的权重和偏置矩阵，是MLP输出的用户embedding。与此同时，所有层的激活函数均选择ReLU。  与NeuACF完整模型不同，本阶段暂时以average方式融合aspect-level embedding，完成模型简化版本的编码实现，方便调试与正确性检验。后续阶段再加入attention机制，考虑不同aspect的重要性，进一步提升模型效果。   1. **引入metric learning和pair-wise损失函数（改进）**   首先，NeuACF利用点积度量用户和物品间的交互可能性：    其中是用户embedding，是物品embedding。因为点积不符合三角不等式的约束，因此可能无法反映更细粒度的用户口味，从而降低模型准确度。以一个简单的例子来说明，如下所示：    表格是用户对物品的偏好信息，左图展示基于点积的矩阵分解方法的一组稳定解，即若用户喜爱物品，则他们间点积为2，否则为0。此时，可以观察到，同时喜欢和，但和间点积为0，即用户对物品的偏好虽然得以体现，但用户和用户、物品和物品间的相似性可能遭到忽视，而这一点对于推荐至关重要。  为此，引入距离的概念。距离，是许多机器学习算法的核心概念，如K-nearest neighbor（kNN）、K-means等。Metric learning技术利用距离定义的metric捕捉数据间的关系，从而使得相似对间距离较近，不相似对间距离较远。而metric所需满足的重要性质之一——三角不等式，自然地蕴含相似性的传递过程，即将已知的相似性信息通过距离传递给未知的点对，这与协同过滤的本质思想不谋而合。因此，**Collaborative Metric Learning（CML）**将metric learning应用至协同过滤，通过用户-物品metric，在反映用户偏好的同时，捕捉用户-用户及物品-物品间的相似性。因此，本工作考虑以metric替换NeuACF中的点积估计形式，实现更全方位的信息传递。具体地，参考CML，修改NeuACF打分函数：    其中是用户embedding，是物品embedding。  其次，NeuACF利用point-wise损失函数进行模型优化：    其中，和分别是正负实例集合，是预测分数，是ground truth。在此种情况下，选取top-ranked结果作为最终的推荐集合。但point-wise损失函数作为预测结果与ground truth的交叉熵，更为关注分数的绝对数值，较适合基于显式反馈的rating推荐系统。而基于隐式反馈的Top-N推荐系统注重物品列表的相对分数。因此，本工作进一步考虑使用基于负采样的pair-wise损失函数替换NeuACF中的point-wise损失函数，产生排序更合理的推荐列表。具体地，在metric基础之上，与不喜欢的物品相比，用户应与喜欢的物品有更小的距离。为此，修改损失函数为：      其中，是正样本集合，是超参数margin。  以上两点即为模型的优化部分。至此为止，编码实现了NeuACF改进模型的简化版本（无attention）。   1. **数据集及编程环境说明**   为验证模型有效性，选择公开的MovieLens和Amazon数据集。其中，前者广泛用于电影推荐系统，后者用于商品推荐。除电影和用户外，NeuACF还需导演和演员信息来学习基于元路径的aspect-level节点表示，而这些信息不包含在原数据集中。因此，利用爬虫工具，从IMDb上抓取相应电影的导演和演员列表补充至MovieLens。  进一步，基于Python 3.6搭建TensorFlow 1.2.1环境，并参考NeuACF源码和官方指导手册，熟悉TensorFlow的函数使用方法。与此同时，实现部分官方样例，在动手操作中，加深对框架的理解和体会。 | | | | | | | |
| 是否符合任务书要求进度 是 | | | | | | | |
| 尚需完成的任务 | 1. 进一步优化模型，实现以attention方式融合aspect-level embedding； 2. 在不同数据集下验证模型有效性，并与baseline（NeuACF）进行比较； 3. 撰写毕业设计论文，进行答辩。 | | | | | | | |
| 能否按期完成设计（论文） 能 | | | | | | | |
| 存在问题和解决办法 | 存  在  问  题 | 1. 对深度学习框架TensorFlow的了解不够深入，模型搭建过程难点颇多； 2. 参数较多，难以确定最佳参数组合。 | | | | | | |
| 拟  采  取  的  办  法 | 1. 根据遇到的实现难点，反复阅读TensorFlow官方指导手册、博客或GitHub中相关源码，理解并熟悉函数接口； 2. 采用控制变量法或网格搜索确定最优参数组合。 | | | | | | |
| 指导教师  签字 | |  | | 日期 | | 2019 年 月 日 | | |
| 检查小组  意见 | | 负责人签字： 年 月 日 | | | | | | |

注：可根据长度加页。