

**本科生毕业设计（论文）开题报告**

题 目：基于图结构数据的协同推荐算法实现与优化

|  |  |
| --- | --- |
| 院 系 | 计算机科学与技术 |
| 专业班级 | 1501班 |
| 姓 名 | 莫海 |
| 学 号 | U201514489 |
| 指导教师 | 邵志远 |

2019年 3月 8日

**开题报告填写要求**

1. 开题报告主要内容：

1.课题来源、目的、意义。

2.国内外研究现况及发展趋势。

3.预计达到的目标、关键理论和技术、主要研究内容、完成课题的方案及主要措施。

4.课题研究进度安排。

5.主要参考文献。

1. 报告内容用小四号宋体字编辑，采用A4号纸双面打印，封面与封底采用浅蓝色封面纸（卡纸）打印。要求内容明确，语句通顺。
2. 指导教师评语、教研室（系、所）或开题报告答辩小组审核意见用蓝、黑钢笔手写或小四号宋体字编辑，签名必须手写。
3. 理、工、医类要求字数在3000字左右，文、管类要求字数在2000 字左右。
4. 开题报告应在第八学期第二周之前完成。

# 课题简介

## 研究课题

推荐系统是人们当今生活中最重要的技术之一，随着或联网的发展，人们每天接触的信息日益增加，再整个信息空间上用户评分数据及其稀疏，如何帮助人们有效的获取需要的信息日益重要，传统的基于评分的推荐算法各自存在自己的弊端，导致系统推荐的质量急剧下降。争对用户评分数据极端稀疏的的情况下传统度量方法的不住，提出了一种基于排名预测的协同过滤算法，用户之间项目的相似性进行排名，再次基础上，采用一种新颖的相似性度量方法计算目标用户的最临近邻居。

## 研究目的

实现并采用GRank算法提高音乐广告，音乐，电影的推荐效率。

## 研究意义

基于邻居的协同过滤（协同过滤的主要类别之一）根据相似用户的行为估计目标用户的排名/评级。尽管在这类算法中进行了多次研究，但它们仍无法精确计算用户的相似性。克服这个问题的一种方法是基于图形的推荐，其利用异构信息网络，即包含不同类型的节点和边缘的信息网络，以细化相似性度量。

基于图形的推荐算法将用户和项目之间的关系表示为二分图，其中用户与他评定的每个项目之间存在加权或未加权的链接。不幸的是，这种方法基本上是为评级/二进制反馈而设计的，并且对于基于邻居的协同过滤的排序导向类具有至关重要的不足。

GRank的新框架，它使用新的三方首选图（TPG）结构捕获用户的偏好，该结构演示了用户，项目和成对偏好之间的关系。 GRank还提供了一种新的排名算法，该算法扩展了个性化PageRank以获得top-k推荐。 据我们所知，该算法是第一种基于图的方法，能够捕获隐式反馈提供的偏好信息。

# 国内外研究现况及发展趋势

## 协同过滤算法（collaborative filtering）

CF算法的原理是汇总所有<user,item>的行为对, 利用集体智慧做推荐。其原理很像朋友推荐, 比如通过对用户喜欢的item进行分析, 发现用户A和用户B很像(他们都喜欢差不多的东西), 用户B喜欢了某个item, 而用户A没有喜欢, 那么就把这个item推荐给用户A。(User-Based CF)

当然, 还有另外一个维度的协同推荐。即对比所有数据, 发现itemA和itemB很像(他们被差不多的人喜欢), 那么就把用户A喜欢的所有item, 将这些item类似的item列表拉出来, 作为被推荐候选推荐给用户A。(Item-Based CF)

如上说的都是个性化推荐, 如果是相关推荐, 就直接拿Item-Based CF的中间结果就好啦。

该算法的好处是:

1.能起到意想不到的推荐效果, 经常能推荐出来一些惊喜结果

2.进行有效的长尾item

3.只依赖用户行为, 不需要对内容进行深入了解, 使用范围广

该算法的坏处是:

1.一开始需要大量的<user,item>行为数据, 即需要大量冷启动数据

2.很难给出合理的推荐解释

## 矩阵分解（MFCR）和基于邻居的算法（NCR）

1. 矩阵分解技术。

协作排名中的矩阵分解技术，尝试学习代表性潜在特征，以准确预测每个用户的项目排名。 CofiRank是第一个使用矩阵分解技术优化排名指标的算法。另一种技术ListRank，通过估计前1的概率来推断项目的排名。URM是另一种结合了ListRank和概率矩阵分解的模型，以提高系统在排名和评级方面的准确性。BoostMF是另一种矩阵分解方法，它基于偏好数据顺序学习一组弱矩阵分解模型。贝叶斯个性化排名及其变体，尝试优化贝叶斯预测模型的曲线下面积（AUC），该模型基于相关和不相关项目之间成对比较生成一组预测。

1. 基于邻居的算法

EigenRank是最著名的NCR技术，它基于类似于目标用户的用户的成对偏好来推断总排名。 EigenRank使用Kendall相关性计算用户的相似性，该相关性考虑了用户对成对比较的一致性和不一致性。在估计相似性之后，EigenRank估计偏好矩阵，其元素是邻居偏好的加权线性组合。最后，它使用贪婪或基于马尔可夫的方法来推断项目的总排名。据我们所知，所有NCR技术都遵循EigenRank提出的主要方法，稍作修改。然而，这种方法仍然存在稀疏问题，因为它仍然依赖于共同的成对比较来计算相似性。

## 3 基于图形的推荐算法

尽管没有基于图形的方法设计用于协作排名，但是在推荐系统的其他领域中已经进行了许多最近的研究。

基于图形的推荐算法由两个步骤组成：构建表示数据的图形并通过分析图形来提出建议。 这些推荐算法已经利用了不同类型的图。 但是，在所有这些中，图表的主要组成部分是用户与被他们评级的项目之间的关系。因此，最常见的方法是构建一个二分网络，其中连接是从一部分网络，用户到另一部分。一旦构建了二分图，就可以使用几种方法使用来自目标用户的邻居的信息对项进行排名。 在这个领域中使用了诸如使用普通邻居，Katz相似度，扩散分数和个性化PageRank等方法。

最近的方法通过向其添加一些层来扩展双向网络。一些研究人员（考虑使用会话层来考虑用户的长期和短期偏好，以便在特定时间内提出建议。其他人在多层结构中使用了不同类型的节点，以通过图中的随机游走来进行上下文感知推荐。三层图用于通过考虑用户使用基于扩散的分数分配给项目的标签来改进推荐。在一些作品中，网络结构已经修改。他们考虑星形异构网络，其中用户和项目可以连接到不同类型的节点。他们使用此图结构来改进基于模型的建议，或通过改进异构网络中的个性化PageRank算法来提出建议。这些算法都不是为了捕获用户的选择上下文和偏好而设计的。此外，它们中的大多数依赖于在所有应用中系统不存在或不可用的上下文信息（例如，时间，内容等），并且收集可能是昂贵的。

# 研究内容与总体研究方案

## 推荐算法学习

1. PageRank算法学习。

佩奇排名（PageRank），又称网页排名、谷歌左侧排名，是一种由搜索引擎根据网页之间相互的超链接计算的技术，而作为网页排名的要素之一，以Google公司创办人拉里·佩奇（Larry Page）之姓来命名。Google用它来体现网页的相关性和重要性，在搜索引擎优化操作中是经常被用来评估网页优化的成效因素之一。这个算法把互联网上的网页集合看成一个有向图，每个网页是一个节点，如果网页A有指向网页B的链接，则这两个点存在一条有向边A->B，那么如果A这个网页上有k个网页的链接，那么他会以1/k的概率从A这个网页跳到他所指的网页，所以可以用一个转移矩阵来表示一对网页集合的转移关系。

1. 协同过滤算法学习。

协同过滤(Collaborative Filtering)作为推荐算法中最经典的类型，包括在线的协同和离线的过滤两部分。所谓在线协同，就是通过在线数据找到用户可能喜欢的物品，而离线过滤，则是过滤掉一些不值得推荐的数据，比比如推荐值评分低的数据，或者虽然推荐值高但是用户已经购买的数据。

协同过滤的模型一般为m个物品，m个用户的数据，只有部分用户和部分数据之间是有评分数据的，其它部分评分是空白，此时我们要用已有的部分稀疏数据来预测那些空白的物品和数据之间的评分关系，找到最高评分的物品推荐给用户。

一般来说，协同过滤推荐分为三种类型。第一种是基于用户(user-based)的协同过滤，第二种是基于项目(item-based)的协同过滤，第三种是基于模型(model based)的协同过滤。

基于用户(user-based)的协同过滤主要考虑的是用户和用户之间的相似度，只要找出相似用户喜欢的物品，并预测目标用户对对应物品的评分，就可以找到评分最高的若干个物品推荐给用户。而基于项目(item-based)的协同过滤和基于用户的协同过滤类似，只不过这时我们转向找到物品和物品之间的相似度，只有找到了目标用户对某些物品的评分，那么我们就可以对相似度高的类似物品进行预测，将评分最高的若干个相似物品推荐给用户。

1. GRank算法学习。

GRank是通过一种新颖的异构图结构，称为三方偏好图（TPG），它在用户，偏好和聚合结构中的项之间嵌入不同类型的关系。 最后，用一种有效的算法来利用TPG来为每个目标用户排序项目。由于国内相关文献还比较缺乏，难度相对来说比较大。

## 实现GRank算法并对其做出一定改进

目标是实现GRank算法并对其做出一定的改进，提高算法的时间以及空间的利用率，通过一些公开训练集的训练，以及其他的一些公开代码的对比实验，实现广告，音乐，电影等内容跟准确高效的推荐。

## 测试GRank

运用相同的数据集训练GRank与互联网上公开的推荐算法，通过对比得出算法性能上的差距。

# 核心技术路线

1. 由于数据量大，用矩阵进行存储不能满足现有数据量的需求，需要用到MapReduce技术。
2. 使用GRank的算法思想，设计TGP三层存储结构。
3. 底层算法可能采用personal PageRank，EigenRank算法等。

# 课题进度安排

表 1 课题进度安排表

|  |  |  |
| --- | --- | --- |
| 学期 | 周次 | 工作任务 |
| 2018-2019  第一学期 | 18周——19周 | 拿到课题，作整体规划 |
| 19周——20周 | 翻译相关文献。了解和学习Grank相关背景知识 |
| 2018-2019  第二学期 | 1周——2周 | 学习文献上提到的重要算法，了解推荐算法的相关背景知识。 |
| 3周——4周 | 学习相关算法。OpenAI Gym上下载environment,利用Q-Learning算法对其训练 |
| 5周——7周 | 实现Grank算法，并完成文献中提及的工作。 |
| 8周——10周 | 进一步对GRank算法进行优化，并与其他算法对比得出结论。 |
| 11周——14周 | 完成毕业论文 |

# 参考文献

1. 1604.03147，Graph-based Collaborative Ranking，《Computer Science》 2016
2. TP391.1，基于CR-PageRank算法的个人事件自动摘要研究，《计算机工程》2016
3. TP391.3，基于长尾理论的物品协同过滤Top-N推荐算法，《齐齐哈尔大学报》2019
4. TP393,09，PageRank算法研究，《计算机工程》2006
5. 基于隐式反馈LDA模型的协同推荐算法研究，《计算机技术与发展》2019
6. TP399，基于项目评分预测的协同过滤推荐算法，《软件学报》
7. TP18，不确定近邻的协同过滤推荐算法，《计算机学报》

**华中科技大学本科生毕业设计（论文）开题报告评审表**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **姓名** |  | **学号** |  | **指导教师** |  |
| **院（系）专业** | |  | | | |
| **指导教师评语**   1. 学生前期表现情况。 2. 是否具备开始设计（论文）条件？是否同意开始设计（论文）？ 3. 不足及建议。 | | | | | |
| 指导教师（签名）：  年 月 日 | | | | | |
| **教研室（系、所）或开题报告答辩小组审核意见** | | | | | |
| 教研室（系、所）或开题报告答辩小组负责人（签名）：    年 月 日 | | | | | |