附件 1

中山大学
大学生创新训练计划项目申请书

项目名称：基于用户和项目的混合协同过滤推荐算法

相关学科： 计算机科学与技术

申 请 人： [redacted]

所在院系： [redacted]

指导教师： [redacted]

预期成果类别：

□社科类社会调查报告及学术论文

☑自然科学类学术论文

教 务 处 制

填表时间 二O一 八 年 5 月 31 日

一、 申请理由（包括自身具备的知识条件、自己的特长、兴趣及开展本项目的基础等）

我上过数据结构、算法和人工智能、数据挖掘等课程，学到了很多与计算机相关的知识，所以希望能够学以致用，将之前学习过的知识结合起来，增强自己对知识的理解，同时提高自己将理论知识用于实践的能力。

1. 在数据挖掘这门课程的学习过程中，我接触到了推荐系统，学习了关于推荐系统的一些基础知识以及了解了推荐系统目前面临的一些问题。课后我也阅读了一些关于推荐系统的论文，了解和掌握了一些关于推荐系统的经典算法。我对于基于协同过滤的推荐系统比较感兴趣，所以希望可以研究如何通过填充空缺值来减弱数据稀疏性问题对于推荐系统的影响，如何改进相似度计算方法，以及如何将基于用户的 TOP-N 算法和基于项目的 TOP-N 算法结合起来。
2. 推荐系统可以解决 Web 系统面临的信息过载和信息迷失现象，有效地提高用户使用系统的频率，高质量的推荐系统还可以让用户产生依赖性。在电子商务的领域，推荐系统得到了广泛的引用，例如 Amazon、淘宝都采用了推荐系统来为用户提供个性化的服务。电子商务推荐系统能够给商家带来长期稳定的客户关系和巨大的的经济效益。除此之外，在新闻、电源、微博等的个性化推荐中，推荐系统也得到了广泛的应用。协同过滤是目前最成功且应用最广泛的推荐技术。协同过滤算法的优势是：① 对推荐对象无特殊要求；② 只需要显式或隐式的用户历史评价数据，不需要有关用户本身的属性知识，且不会对用户的推荐体验带来任何负面影响。但是协同过滤算法还有很多问题没解决，例如数据稀疏性问题等。
3. 基于用户的推荐算法根据用户的相似性来产生推荐，基于项目的推荐算法根据项目的相似性来产生推荐。两种算法都只考虑了用户或项目单方面的影响，忽略了用户与项目之间的关系。如果能够将两种算法结合起来，达到提高预测准确率的目的，这样就可以获得更好的推荐质量。

二、 项目创新特色概述(50 字以内)

1. 通过用户兴趣偏好来预测用户项目评分矩阵中的空缺值，从而缓解数据稀疏性问题；
2. 改进现有的相似度计算方法，考虑用户特征和项目属性对于用户相似性和项目相似性的影响。
3. 融合基于用户和基于项目的 TOP-N 算法，提高推荐精度；

三、 计划项目实施思路（3000 字以内）

1、研究意义与目的，同类研究工作国内外研究现状与存在的问题等

随着互联网的普及和互联网用户数量的迅速猛增，互联网上的信息呈现爆炸式的增长。海量的信息为满足互联网用户纷繁复杂的信息需求带来了前所未有的机遇，但是也对信息处理技术提出了严峻的挑战——“信息过载问题”，即用户无法从海量的信息中快速准确定位到自己所需要的信息的问题。目前，解决信息过载问题的主要手段是信息检索技术和信息过滤技术。信息过滤技术的主要思想是通过挖掘海量的用户行为数据，分析出用户的需求，主动向用户推送他所感兴趣的信息。推荐系统作为信息过滤的一种重要应用，已经成为新一代 web 应用中的个性化信息服务形式。近年来，随着电子商务的兴起，推荐系统得到了广泛的应用，例如 Amazon、淘宝等都采用了智能推

荐系统来为用户提供个性化的推荐服务。高质量的推荐能够增加用户对网站的信任度，使用户产生依赖信息，提高用户的忠诚度。除了在电子商务领域的应用之外，随着个性化的信息服务逐渐成为 web 应用技术的热点，推荐系统在新闻、电影、书籍、微博等的个性化推荐中也取得了不同程度的成功。

目前常用的推荐算法主要包括：基于内容的推荐、基于协同过滤的推荐、基于关联规则的推荐以及混合推荐技术。其中，协同过滤是目前最成功且应用最广泛的推荐技术。协同过滤的基本假设是具有相同或相似兴趣偏好的用户的信息需求也是相似的。协同过滤算法通过挖掘用户的历史标注信息来发现相似用户或项目，然后利用相似用户和项目的评分信息来预测当前用户对项目的喜好程度。

目前协同过滤算法所面临的难点问题要包括：数据稀疏性问题、冷启动问题等。数据稀疏性问题是协同过滤算法面临的最大挑战。在实际的商品推荐系统中，在实际的商业推荐系统中，用户和项目的数量十分的庞大，而用户往往只在很少的项目上有评分记录，这就导致了评价矩阵是非常稀疏的，通常商业推荐系统的评价矩阵密度不会超过 1%。在数据稀疏的情况下，由于共同标注的数量过少，基于内存的方法无法准确地计算用户（项目）之间的相似度，从而导致邻居集合选择不准确，影响推荐精度。在有的推荐系统中，用户的数量远远少于项目的数量，这就导致了许多项目从未被标注过，从而无法对这些项目进行推荐。冷启动问题是数据稀疏问题的另外一种表现，也称为新用户问题（或者新项目问题），在实际的推荐系统中，用户和项目的数量的增长速度都较快，这就意味着不断的有新的用户和项目加入到系统中，当新用户或项目刚加入系统时，由于缺乏对应的标注信息，导致无法对这些用户或项目进行推荐。

相似度的计算是基于内存的协同过滤算法中最关键的一步，用户或项目之间的相似度估计准确与否直接影响推荐质量的高低。数据稀疏性问题对相似度估计有影响。协同过滤系统中，评价矩阵的信息一般都是十分稀疏的。对于许多用户（项目）而言，他们共同标注的项目数（标注该项目的用户数）非常少，有时甚至为 0，这就导致传统相似度计算方法不准确，从而降低了推荐精度。因此，如果能有效地改进现有的相似度计算方法，降低其对数据稀疏性的敏感程度，使得其对相似度的估计更为合理，就能够有效地提高推荐质量。

根据不同的假设，协同过滤算法可以分为基于用户的方法和基于项目的方法。这两种算法的有效性都已经被许多成功地商业推荐系统所证明，实际上不管是基于用户的方法还是基于项目的方法均能达到较高的推荐精度，然而这两种方法的推荐结果的覆盖范围却具有显著的差异。基于用户的推荐算法根据用户的相似性来产生推荐，基于项目的推荐算法根据项目的相似性来产生推荐。两种算法都只考虑了用户或项目单方面的影响，忽略了用户与项目之间的关系。因此通过将基于用户和基于项目的方法进行有效的融合，能达到提高预测准确率的目的，显著地提高推荐精度。

2、研究内容及工作方案

1）算法研究

算法研究主要分成三个部分，第一部分是通过用户兴趣偏好来预测用户项目评分矩阵中的空缺值，从而缓解数据稀疏性问题，第二部分是改进相似性的计算方法，通过结合用户特征和项目属性来改进用户相似性和项目相似

度的计算方法，第三部分是将基于用户的 TOP-N 算法和基于项目的 TOP-N 算法通过动态加权的方法结合起来。

（A）通过用户兴趣偏好来预测用户项目评分矩阵中的空缺值

用户规模和项目数目的增加， 使得评分矩阵的数据稀疏性逐渐增加， 从而导致预测评分精度低。目前解决办法就是对用户未评分项进行预填充。最简单且常用的预填充方法是固定缺省值，例如用平均值来填充，该方法可以有效的提高系统的推荐精度。但实际生活中评分矩阵中的缺省值不可能都一样， 因此填充固定缺省值的方法没有从根本上解决数据稀疏性问题。所以可以通过以用户兴趣分布评分矩阵求相似用户，对空缺值进行预测填充，减小用户评分数据的稀疏性。

（B）改进相似性的计算方法

传统的基于用户的协同过滤推荐技术在计算用户之间的相似性时只考虑用户单纯评分关系，没有考虑用户特征对用户相似性的影响, 导致相似性结果不精确,从而导致产生的最近邻居不准确,最终导致产生的推荐结果不准确。不同用户特征的人他们的兴趣爱好可能不一样，而具有相同类别用户特征的人，他们的兴趣爱好具有一定的相似性，例如：女大学生可能都比较喜欢浪漫爱情片，男孩子比较喜欢动画片，老年人喜欢纪录片等等 ，所以将用户特征的相似性与传统的基于用户的计算相似性方法相结合，能够提高推荐系统的推荐精度。传统的基于项目的协同过滤技术只考虑了用户-项目评分矩阵数据，没有考虑项目属性对项目相似性的影响, 从而得出不准确的邻居项目集合，最终导致推荐结果不准确。目标项目的邻居项目的属性与目标项目的属性应当具有一定的相似性，例如目标项目如果是动作片，则我们产生的邻居项目中是动作片的可能性更高。所以将项目属性的相似性与传统的基于项目的计算相似性方法相结合，能够提高推荐系统的推荐精度。

（C）结合基于用户和基于项目

为了结合基于用户和基于项目的算法，我引入了两个权重因子，分别代表基于用户的算法对预测结果的动态影响程度和基于项目的算法对预测结果的动态影响程度。利用用户相似性和项目相似性的比值来动态设定两个权重因子。

2）应用

我们可以将改进后的基于协同过滤的推荐系统应用到电子商务等需要推荐的领域。电子商务可以采用推荐系统来给用户推荐商品，有助于用户更快找到自己想要的商品，而且还可以促进用户的消费。在新闻、电源、微博等的个性化推荐中，推荐系统也能帮助用户获得更好地体验。

3）实验

首先通过运用利用用户兴趣偏好矩阵来进行预填充和不进行预填充的 TOP-N 模型来进行实验，对比实验结果，如果预填充的模型会得到更高的准确率，那么就说明预填充的做法是正确的。然后分别运用修改了的

相似性计算公式和原始的相似性计算公式进行实验，对比实验结果来判断修改了的相似性公式是否能够提高推荐精度。最后再分别运用基于用户的 TOP-N 模型、基于项目的 TOP-N 模型 以及 两者的结合模型进行实现，对比实验结果，来判断结合模型是否能够提高推荐精度。

3、拟解决的主要问题

（1）构建用户兴趣偏好矩阵，运用其来预测用户项目评分矩阵中的空缺值；

（2）改进相似性的计算方法，通过结合用户特征和项目属性来改进用户相似性和项目相似度的计算方法；

（3）将基于用户的 TOP-N 算法和基于项目的 TOP-N 算法通过动态加权的方法结合起来。

4、项目创新之处（原始创新：重大科学发现、技术发明；集成创新：融合多种相关技术，形成新产品、新产业；引进消化吸收再创新：在引进国内外先进技术的基础上，学习、分析、借鉴，形成具有自主知识产权的新技术）

1) 通过用户兴趣偏好来预测用户项目评分矩阵中的空缺值，从而缓解数据稀疏性问题。

2) 改进了相似性的计算方法，结合了用户特征和项目属性。

3) 将基于用户的 TOP-N 算法和基于项目的 TOP-N 算法通过动态加权的方法结合起来。基于用户的推荐算法根据用户的相似性来产生推荐，基于项目的推荐算法根据项目的相似性来产生推荐。两种算法都只考虑了用户或项目单方面的影响，忽略了用户与项目之间的关系。将两种算法结合起来可以达到提高预测准确率的目的，这样就可以获得更好的推荐质量。

4) 相关成果应用到电子商务系统等需要推荐的领域，用户可以获得更好的推荐质量。