# 面向稀疏数据优化的协同过滤推荐算法

孟俊才1

1.安庆师范大学计算机与信息学院安庆 246011

**摘 要**:协同过滤算法在目前推荐系统中应用广泛，而如何寻找相似用户是协同过滤算法的核心，能否准确地找到相似用户决定了协同过滤算法的性能。由于评分矩阵的稀疏性问题，使得寻找相似用户十分困难或者不准确，严重制约着协同过滤算法的性能。为了解决稀疏性问题，提出了一种融合稀疏度进行加权的协同过滤算法。该算法首先重新定义了矩阵稀疏度计算为方法，融合矩阵稀疏度对用户相似度进行加权，并以此来改进协同过滤算法。改进后的算法由于有效地利用了稀疏矩阵中的相似信息来寻找相似用户，实验结果表明，与传统的协同过滤算法比较，改进算法使推荐结果的准确率和覆盖率都得到了提高。

**关键词:**协同过滤;稀疏度;相似性计算;推荐系统

**中图分类号:TP183**

**文献标识码:A**

**Collaborative Filtering Recommendation Algorithm for Sparse Data Optimization**

Meng Juncai1

1.School of Computer and Information, Anqing Normal University, Anqing 246011

**Abstract**: The collaborative filtering algorithm is widely used in the current recommendation system, and how to find similar users is the core of the collaborative filtering algorithm. Whether the similar users can be accurately found determines the performance of the collaborative filtering algorithm. Due to the sparsity of the scoring matrix, it is very difficult or inaccurate to find similar users, which seriously restricts the performance of the collaborative filtering algorithm. In order to solve the sparsity problem, a collaborative filtering algorithm based on sparsity is proposed. The algorithm firstly redefines the matrix sparsity calculation method, and the fusion matrix sparsity degree weights the user similarity, and improves the collaborative filtering algorithm. The improved algorithm effectively uses the similar information in the sparse matrix to find similar users. The experimental results show that compared with the traditional collaborative filtering algorithm, the improved algorithm improves the accuracy and coverage of the recommendation results.

**Key words:** collaborative filtering; sparseness; similarity calculation; recommendation system

## 1 引言

在互联网高速发展的当代，互联网为每个人带来的信息量也在高速增长，人们面对海量信息时很难从其中提取出对自己最有用的信息，就如在电商网站中用户面对海量商品，很难选择出最适合自己的商品，这既浪费了用户的时间，也使得网站的利益受损。为了使电商网站更好地销售物品，方便用户，推荐系统在各大电商网站得到了应用，其中协同过滤算法是广泛应用的一种推荐算法。

协同过滤算法由Goldberg等提出，研究者将该算法应用在一个文档过滤系统中，系统要求用户对文档评分，其他用户浏览时可以输人查询语句来过滤自己需要的文档。之后协同过滤算法被GroupLens用于新闻推荐，极大地促进了协同过滤算法的发展，也迅速推动了协同过滤算法在Internet上的应用及其相关研究。

协同过滤拥有如下优势:

(1)算法可以推荐任何资源，而不仅限于文本或者网站等。这是因为协同过滤算法需要考虑的只是用户间对物品的评价，而无需关注资源的特性，这使得协同过滤算法的适用范围变得十分宽广，可以为用户推荐任何资源。

（2）模型简单，易于实现，各大网站可以方便地使用协同过滤算法来实现推荐，而不需要为此付出过高的代价。

（3）对用户的推荐具有新颖性，不在是根据用户历史来重复推荐一个类别或几个类别的物品。协同过滤算法因为是从相似用户中寻找物品进行推荐，所以算法可能会为用户推荐该用户历史上从未接触到的物品，并且该物品也符合用户的喜好。这样可以挖掘出更多的推荐项，不仅对用户友好，而且有利于挖掘长尾物品。

协同过滤算法在现阶段的主要问题是稀疏性问题。由于网站特别是大型网站拥有大量的用户和物品数量，而绝大多数用户只会对极少的一些物品进行评价，因此评分矩阵非常稀疏，在这种稀疏矩阵中，协同过滤算法很难准确地找到相似用户或者找到的相似用户的评分稀少，最终使得推荐效率低下。

为了解决协同过滤算法由于稀疏性带来的性能低下问题，提出了融合稀疏度加权的协同过滤算法。首先，对稀疏矩阵的稀疏度进行计算评估;然后，在寻找相似用户阶段通过用户密度与矩阵稀疏度对两两用户间进行稀疏度加权，这样寻找的相似用户更加准确。

## 2 数据稀疏常用解决方法

目前，为了解决协同过滤算法中的稀疏度问题，研究人员提出了许多解决方法，大致分为三类:填充类、降维类和改进相似度计算类。

填充类方法-通过在稀疏矩阵中添加数据从而使得矩阵不再稀疏，降低稀疏性对协同过滤算法的影响。最基本的填充方法是在稀疏矩阵中添加均值或者中位数等等。例如，李灿等在IALM和填充可信度的协同过滤算法及其并行化研究中提出利用可扩充的拉格朗日乘法对评分矩阵进行填充;稀疏矩阵在经过填充之后的确可以在一定程度上解决稀疏性所带来的问题，但是填充算法大部分较为复杂，这样就使得改进后的协同过滤算法在效率上受到限制，不能为用户频繁更新推荐列表。(增加引用文献)

降维类方法-通过对稀疏矩阵进行降维达到缓解矩阵稀疏度的目的。例如，孙光福等引人用户与物品间的时序关系，然后将时序关系融人到概率矩阵分解算法中，最后将矩阵分解降维。降维类算法虽然降低了矩阵的稀疏度，但是在降维的过程中损失掉了更多相似信息，使算法更加难以找到相似用户。(增加引用文献)

改进相似度的方法-通过改进相似度计算使得稀疏性对协同过滤算法的影响降低。这类方法既避免了填充类方法带来的代价过大问题，也不像降维类方法那样降低推荐性能。例如， Luo等提出全局相似度和局部相似度，文中分别计算各个用户间的局部相似度与全局相似度，然后将两个相似度融合成最终的相似度来寻找相似用户。(增加引用文献)

在这些研究的基础上，本文中提出了融合稀疏度加权的协同过滤算法。算法通过矩阵稀疏度进行加权来改进相似度的计算，可以让稀疏矩阵中的相似信息得到更有效的利用，从而提高协同过滤算法的性能。

## 3 基于改进稀疏度权重的协同过滤算法

### 3.1 矩阵稀疏度提出

在协同过滤算法中，稀疏性问题严重制约着协同过滤算法的性能。由于稀疏矩阵中所包含的用户相似信息十分有限，因此更有效地利用这些信息成为了提高协同算法性能的关键。对此，文中提出在计算用户相似度时对稀疏度进行加权。

首先，对稀疏度加权需要更准确地计算稀疏度。传统计算矩阵稀疏度的方法是将矩阵中未评价的评分单元数比上矩阵总单元数。然而这样计算得到的矩阵稀疏度并不能准确表示矩阵真实的稀疏度。因为影响矩阵稀疏度的因素并不是仅仅包含数据本身的稀疏度，它还包括用户间的关系密度。关系密度是指用户在各个项目上评价的集中程度，如果用户在各个项目上的评价很分散，就会导致用户间的关系密度很低，即原始评分矩阵中所包含的有效相似信息很少，就认定矩阵十分稀疏。所以提出一个矩阵稀疏度的计算用户评价项目总数的比值，在计算出所有比值后取平均值，利用平均值来衡量矩阵的稀疏度。

（公式）

### 3.2 改进后的相似度计算方法

文中在协同过滤算法中融人上述稀疏度加权项，从而更有效地利用评分矩阵中的有限信息。

（1）余弦距离。

余弦距离通过计算用户之间的余弦值来衡量用户之间的相似性，在评分矩阵中每个用户被看作一个向量，向量的维度就是物品数，用户对各个物品的评分就是用户向量在相应维度上的值，计算公式如下:



（2）皮尔逊相关系数。

通过协方差与标准差来评判两个向量的相似性，计算公式如下：



本文在计算用户相似度时，利用融合稀疏度加权项来皮尔逊相关系数计算公式，因此改进后的皮尔逊相关系数相似度计算公式为:

*Wij*

具体算法步骤如下:

（1）根据式 计算矩阵稀疏度;

（2）根据式 计算用户间关系密度与整个矩阵的稀疏度的比值，并归一化得到稀疏度加权项;

( 3)使用式 计算各个用户间的相似度;

( 4)找到相似用户后，预测评分，进行推荐;

( 5)计算推荐结果的准确率和覆盖率。

## 4 实验步骤与结果分析

### 4.1 实验数据集

实验使用MovieLens公开数据集作为实验数据集，该数据集中包含1000个用户对3900部电影的10万条评分记录，其中每个用户的评价记录在20条以上，用户评分范围在1-5之间。

### 4.2 实验步骤

（1）将数据集分为两部分，一部分作为测试数据集，占比为五分之一，剩下的为训练数据集。

（2）应用文中算法，分别通过式 进行改进加权，以推荐准确率和推荐覆盖率作为评价标准。

（3）实验结果取相似用户的数量为变量，以10个相似用户为基础，以5个相似用户为梯度增加到50个。计算推荐准确率与覆盖率。

### 4.3 评价标准

准确率是指在推荐条目中，用户喜欢的条目数量与推荐的条目总数量的比值，其计算公式如下:

其中，为向用户推荐的条目总数;为推荐条目中用户喜欢的条目。

覆盖率可以反映出算法发现长尾物品的能力。长尾物品指在物品中用户较少评价的一些物品，覆盖率越高，算法发现长尾物品的能力越强，具体是指在推荐条目中用户喜欢的条目数量与所有用户喜欢的条目数量之比，计算公式如下:

其中，为向用户推荐的推荐条目总数;*I*为整个数据集中的物品数。

### 4.4 实验结果分析

## 5 结论

协同过滤算法的性能受到评分矩阵稀疏度问题的制约，稀疏的评分矩阵导致寻找相似用户困难或寻找到的相似用户不准确，最终影响协同过滤算法的性能。文中通过矩阵稀疏度加权来改进协同过滤算法，从而更有效地利用了用户评价信息，避免了稀疏性带来的问题，推荐结果可以更准确地体现用户的喜好程度。实验结果表明，该算法可以有效提高推荐的准确率与覆盖率。