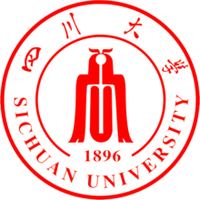
****

**本科生课程论文**

****

**题 目 基于用户协同过滤的音乐推荐**

**课 程 名 信息分析与决策技术**

**任课教师 周良**

**学 院 公共管理学院**

**专 业 信息管理与信息系统**

**小组成员 雷文瑶、郭亚琪、夏馨宇**

**年 级 2016级**

基于用户协同过滤的音乐推荐

【摘要】互联网与多媒体技术的普及，使音乐推荐系统成为了推荐领域的一大研究热点。个性化的音乐推荐是音乐检索系统的核心，因而研究音乐是如何进行推荐算法的实现是很重要的一部分。本实验选择MSD作为数据集，从用户协同过滤的角度设计了两个模型，以期对比预测效果，其一是基于物品间相似度对用户偏好进行计算，以反映用户的偏好行为；其二采用SVD方法对用户降维并进行特征分析，准确地对用户进行歌曲推荐。最近最后通过对比两种推荐算法模型，对方法进行评估。

【关键词】用户协同过滤；音乐推荐；相似度计算；矩阵分解

目录

[基于用户协同过滤的音乐推荐 1](#_Toc11315675)

[1. 研究意义 2](#_Toc11315676)

[2. 文献综述 3](#_Toc11315677)

[3. 研究思路 3](#_Toc11315678)

[4. 数据获取与预处理 4](#_Toc11315679)

[4.1数据源选取 4](#_Toc11315680)

[4.2数据预处理 4](#_Toc11315681)

[5. 模型建立与结果展示 6](#_Toc11315682)

[5.1基于歌曲相似度的推荐模型 6](#_Toc11315683)

[5.1.1构建相似度模型 6](#_Toc11315684)

[5.1.2分析用户数据并进行歌曲相似度模型的建立 7](#_Toc11315685)

[5.2基于矩阵分解的音乐推荐 12](#_Toc11315686)

[5.2.1构建潜在因子模型 12](#_Toc11315687)

[5.2.2使用SVD方法进行降维 13](#_Toc11315688)

[5.2.3分析用户数据并进行用户歌曲偏好模型的建立 14](#_Toc11315689)

[6. 模型评估 16](#_Toc11315690)

[7. 未来展望 17](#_Toc11315691)

# 研究意义

近年来，随着网络服务的迅速发展，用户可以方便的使用手机终端在线收听音乐。由于网络上存在海量的音乐数据，用户检索自己喜欢的音乐时，面临信息过载的问题。因此,为用户提供有效的个性化推荐是一项必要的音乐服务。但在面对实际数据时，由于数据集过大，计算任务过重，非常不利于个人学习。我们希望能从用户数据的角度突破，减少计算任务的同时训练出音乐推荐模型，为有志于音乐推荐算法研究的同学提供参考。

# 文献综述

通过对现有文献和研究的调研，我们将机器学习在音乐推荐领域的应用归纳如下：

在基本研究思路中，我们总结出以下三种主流方法：

**（1）基于内容过滤的推荐。**该方法通过获取音乐的音频特征或社会化网络标签，为用户推荐最近播放的相似音乐，如虾米、SongTaste等热门音乐网站。这种方法推荐结果直观且容易理解，几乎不需要特别的用户历史数据，缺点是推荐结果缺乏新颖性、易受音乐特征提取能力的影响。

**（2）基于用户协同过滤的推荐。**该方法基于用户对音乐的评分数据，找到具有与该用户最相似偏好的用户，进而推荐相似用户评分最高的音乐，如Ringo、Last.fm等音乐服务提供商。这种方法善于发掘用户新的兴趣点，可处理复杂的非结构化对象，缺点是冷启动问题和数据稀疏性。

**（3）基于语境推荐。**该方法获取用户历史的音乐评论、博客百科等文化背景知识、播放次数等数据，记录用户偏好，为用户推荐个性化的音乐。比如豆瓣电台。这种方法的优点是更具个性化、准确率高，答案却点事语境信息数据量庞大，受语境提取率影响。

在研究过程中，大部分的研究学习都是基于Last.fm与Million Songs Database的数据集。由于数据异常值会影响到模型质量，学者们普遍采用简单统计量分析、36原则、箱型图分析等方法进行数据异常值的分析，[[1]](#footnote-0)然后去重、去噪、去缺失、筛选相关数据集成符合自己研究状况的数据集进行规范[[2]](#footnote-1)。

在算法选择中，CNN（卷积神经网络）、CRNN（卷积-循环神经网络）、KNN（最近邻算法）、决策树、VSM（支持向量机）等方法普遍用于音乐推荐系统的研究。

在模型评价中，模型评估的方法主要采用真值表，利用准确率（Precision）、召回率（Recall）、F值、覆盖率这四项数值进行。

# 研究思路

我们倾向于从**基于用户协同过滤推荐**的角度进行研究。该方法基于用户对音乐的评分数据，找到具有与该用户最相似偏好的用户，进而推荐相似用户评分最高的音乐。这种方法善于发掘用户新的兴趣点，优点是处理的信息和对象简单，易分析，只需实时搜集即可，无需像基于内容的推荐，需要计算大量复杂的音频数据特征；缺点是冷启动问题和数据稀疏性，对用户过往产生的使用数据要求比较高。

在算法上，我们将主要使用用户行为数据，以研究聚类问题为出发点，从两个角度构建推荐模型，第一种是计算歌曲与歌曲之间的相似度，第二种是计算“音乐-元素”的潜在因子矩阵。之后我们将对这两个模型进行评估并进行准确率对比，查看在同一个数据集的应用中，哪种模型（及对模型选择何种参数）更优。

在数据集上，我们将使用MSD（Million Songs Dataset）。这是一个包含七个数据平台（包括SecondHandSongs dataset、Last.fm dataset等）真实数据集的资源整合平台，该平台在原始数据的基础上进行了必要的分析和提取，所包含的数据量和信息量都非常之多，全部数据量达到280G，但我们可以下载其中1%的数据，其数据也能较好地反应整体数据集的面貌。

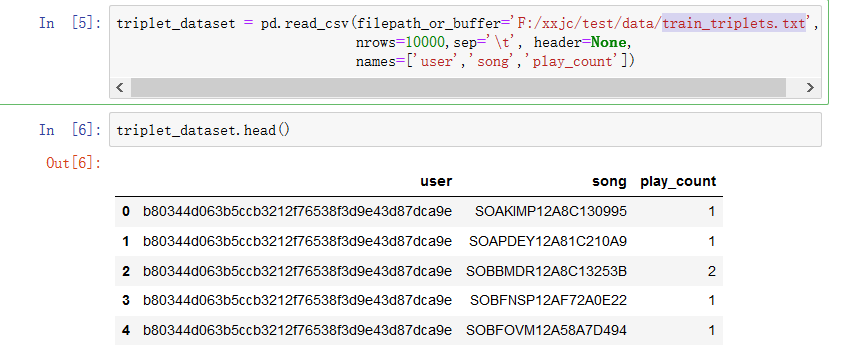
# 数据获取与预处理

## 4.1数据源选取

我们选取了百万歌曲数据集Million Song Dataset（MSD）中的train\_triplet.txt。这个是The Echonest喜好画像子集，包含了匿名用户的歌曲播放次数的记录。即使只是百万歌曲数据库的一个子集，但它的数据量也非常庞大，因为它包含了130,000,000行左右的三元组信息：（user，song，play\_count）。

## 4.2数据预处理

由于数据量对于单机工作而言太大（2.79GB），所以我们需要对数据进行一定裁剪。首先，我们对数据的前五行进行读取，大致了解一下数据样例：

图1 数据样例

对于这样规模大小的数据集，我们首先要做的是有多少用户（或者歌曲）是我们应该要考虑的。在原始数据集中，有大约100万的用户，但是这里面是不是所有用户我们都需要纳入考虑的。例如：如果20%的用户的歌曲播放了占了80%的总体播放量，那么其实我们只需要考虑这20%用户就差不多了。所以我们需要统计一下播放量的累积求和来探知多少用户占了80%的总体播放量，并将统计结果保存为song\_playcount\_df.csv。



图2 用户播放数量裁剪

预读一下新生成的这份数据：

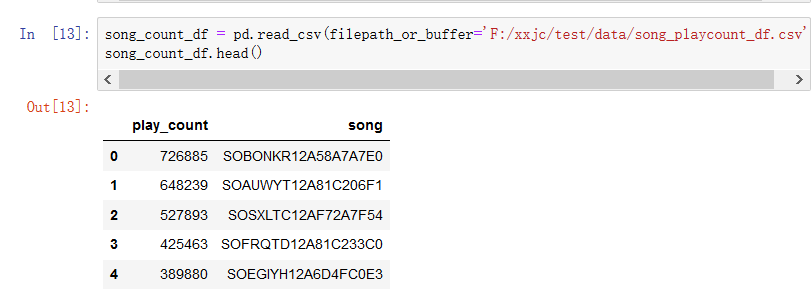


图3 歌曲播放数量裁剪

接下来就可以通过找到前多少用户占了40%的总体播放量，来人为控制数据集的大小。可以看到在这个数据集中，大约前100,000用户的播放量占据了总体的40%。

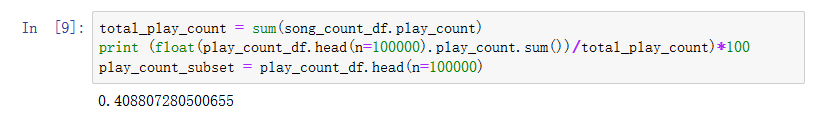
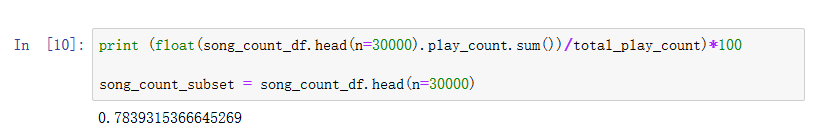


图4 用户播放占比

同样的，我们可以发现大约有30,000首歌占据了总体80%的播放量，即：10%的歌曲占据了80%的播放量。

图5 歌曲播放占比

通过这样一些条件，我们就可以从原始的数据集中抽取出最具代表性的数据出来，从而使得需要处理的数据量在一个可控的范围内。新建了一个表头为（user，song，play\_count）的表，抽取目标用户，过滤非目标歌曲。这个数据集的概况如下：

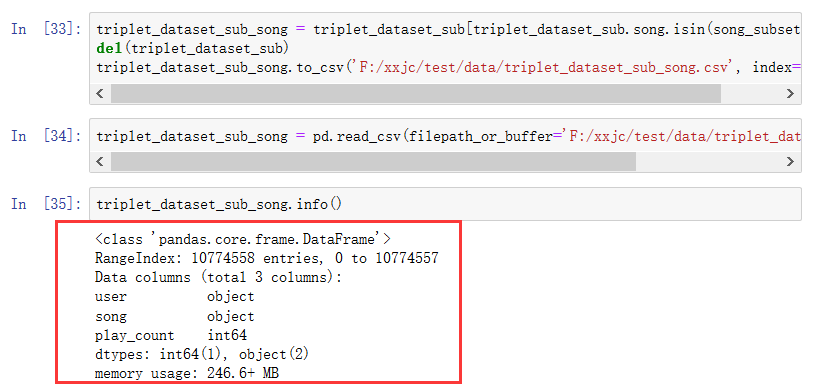


图5 过滤非目标歌曲结果

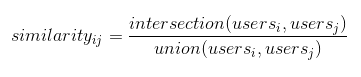
数据准备完毕，可以看到这个文件仍然很大，有246.6+MB，共包含10774588条记录。

# 模型建立与结果展示

## 5.1基于歌曲相似度的推荐模型

### 5.1.1构建相似度模型

能够判断两首歌相似的方法和依据有很多，如歌曲的流派、音乐风格和音轨的相似度。如果从用户的协同过滤的角度来考虑歌曲的相似度的话，我们不妨可以这样设想：对不同的两首歌曲，我们可以分别获得两个听过此首歌的用户列表，两个列表重复度越高，说明两首歌曲相似度越高。我们可以依据此构造如下歌曲相似度的算法：

图6 歌曲相似度算法similarity

先找出用户i听过的歌曲，针对他听过的每首歌计算跟所有歌曲的相似度，通过用户的听歌记录对其音乐偏好进行统计，以相似度为准，找出相似度最高的歌曲对其进行推荐。由于本音乐数据集一共有约一万条歌曲的详细信息，三百多万条用户信息，计算歌曲的相似度是一个计算密集型任务，其时间复杂度为**O(m×n)**，所以我们需要先对数据集进行缩减，将所有歌曲控制在5000首左右。步骤如下：

1. 选择最流行的5000首左右的歌，得出用户累计听歌统计数据和歌曲累计数据，重新对数据进行筛选；
2. 将5000首歌的用户和播放记录作为训练集，导入推荐算法对其进行训练；
3. 指定目标用户并对其进行推荐。

图7 热度最高的5000首歌曲信息

### 5.1.2分析用户数据并进行歌曲相似度模型的建立

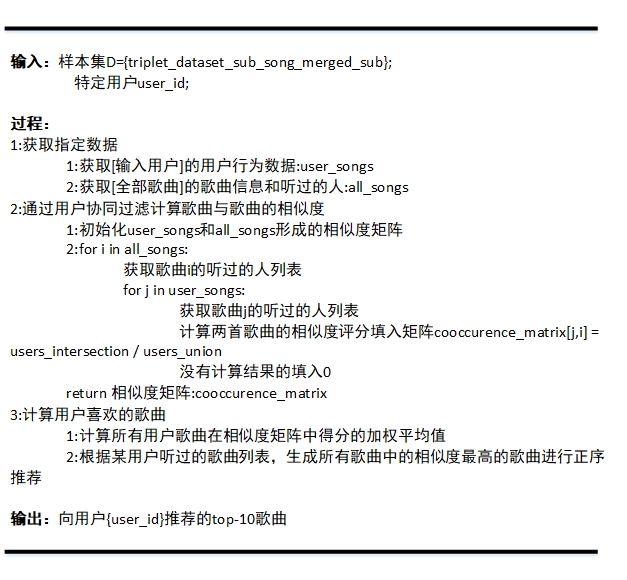
本算法首先对每个用户听过的所有歌曲与全部歌曲建立成矩阵，再统计每首歌所听过的用户id和次数，通过对歌曲与歌曲之间的用户交集进行歌曲相似度的计算。再对某一特定用户的所有歌曲与全部歌曲进行相似度分析，得出该用户的相似度排名推荐。

表1 歌曲相似度计算矩阵

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 用户\所有歌曲 | 歌曲A | 歌曲B | 歌曲C | … |
| 用户1 | 1 | 0 | 1 | … |
| 用户2 | 0 | 0 | 1 | … |

\*1为听过，0为未听过

算法整理如下：

图8 基于歌曲相似度的推荐算法

即使数据缩减到五千条，计算量仍然很大，计算时间大约在一个小时左右。

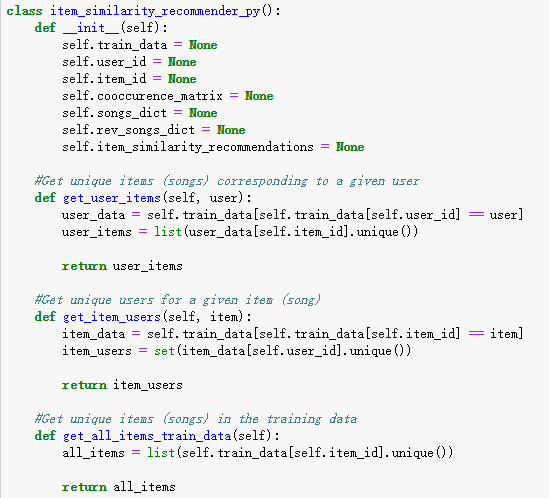


图9 基于相似度的推荐算法-1

声明用户、歌曲、矩阵等参数后，获取所有用户的歌曲列表、所有用户列表和全部歌曲列表。

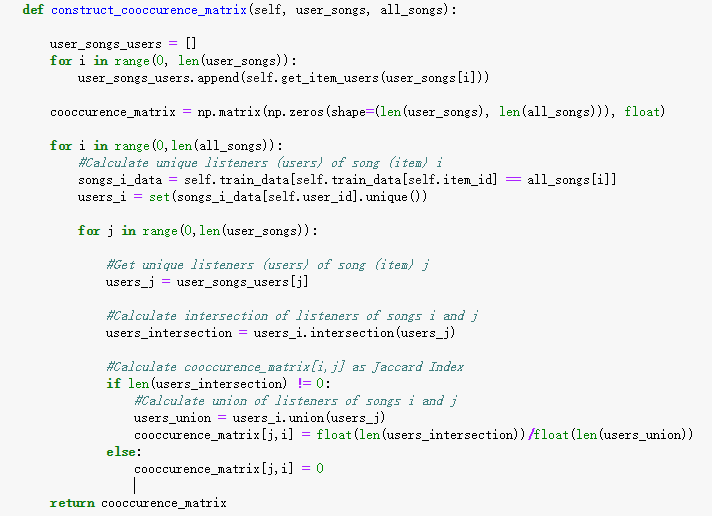


图10 基于相似度的推荐算法-2

建立矩阵，先将用户的特定歌曲循环，形成矩阵的行；将其在所有歌曲列表中做循环，得出一首歌i中的特定用户；得到歌曲j中的特定用户数据，将i和j的用户交集做计算，将交集除以并集的商作为矩阵的特定指数，最后得出所有歌彼此之间的相似度矩阵。

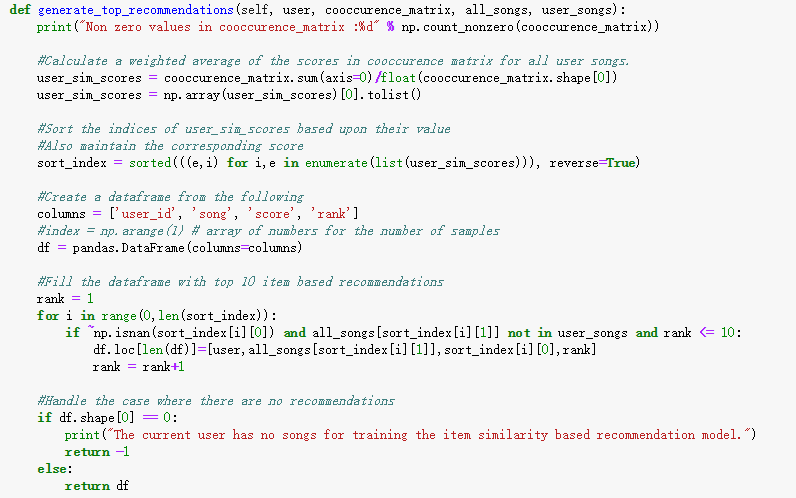


图11 基于相似度的推荐算法-3

计算所有用户歌曲的共生矩阵中得分的加权平均值，根据加权平均值对其索引进行排序，同时保持相应的分数；从数组中创建一个数据帧，以获取样本数，使用前10个基于项目的建议填充数据框，即进行十首歌曲的推荐并对没有推荐的数据进行处理。

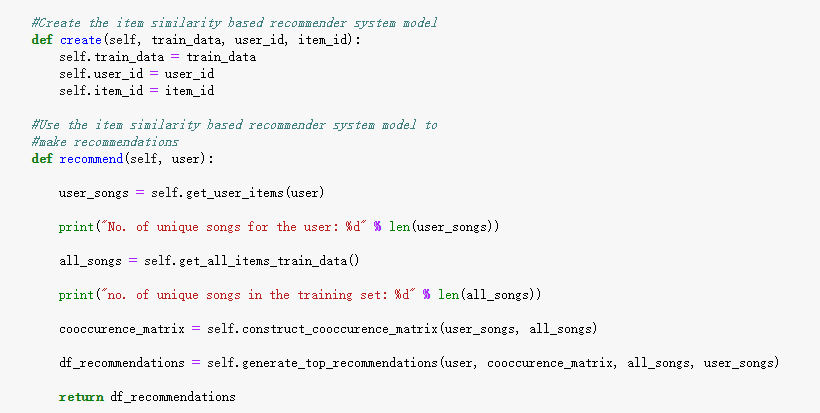


图12 基于相似度的推荐算法-4

创建基于歌曲相似度的推荐系统模型。

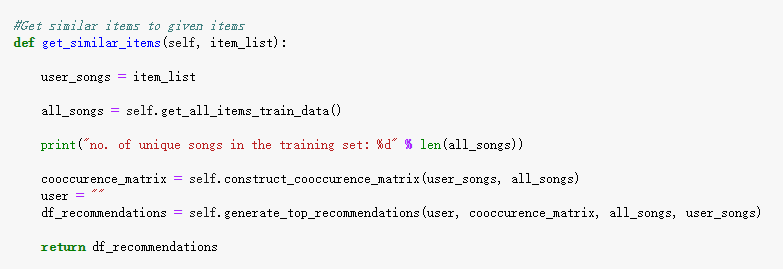
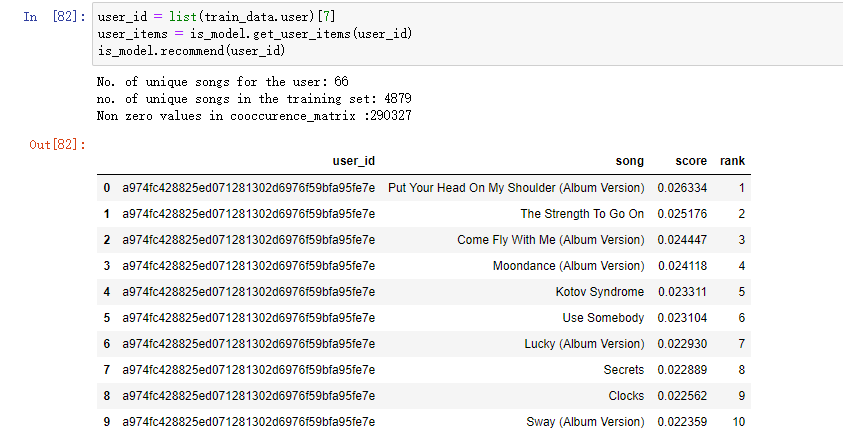


图13 基于相似度的推荐算法-5

对给定的用户进行推荐算法处理，得出推荐结果。

测试用户id为[7]的用户，推荐结果如下：

图14 针对用户7的推荐结果

结果显示，这名用户听过的歌曲有66首，整个训练集的歌曲有4879首，上列十首歌则是基于其歌曲相似度得出的推荐结果。

## 5.2基于矩阵分解的音乐推荐

### 5.2.1构建潜在因子模型

在向用户推荐其可能会喜欢的歌曲时，我们的另一个思路是通过一个“潜在因子”将用户与音乐联系起来，即构建“用户-潜在因子-音乐”模型。我们假设用户A喜欢的歌曲都有一定的风格，拥有一些元素，而如果一首歌具备这些元素，我们就将它推荐给用户A。每个人对音乐的喜好都不尽相同，每首歌所包含的元素也完全不同，所以我们希望能得到如下两个矩阵：用户喜好矩阵U，音乐内容特征矩阵V，将其相乘，可以得到“用户-潜在因子-音乐”模型：



图15 用户喜好矩阵



图16 音乐内容特征



图17 “用户-潜在因子-音乐”模型

事实上，用户喜好矩阵和音乐内容特征矩阵获得难度较大：在音乐系统中少有用户为自己或歌曲打标签的行为，即使有这种行为，由于自定义标签过多，矩阵维度太高，分析起来也太过复杂。但我们有用户的播放数据，就可以对“用户-歌曲”构建一个“用户歌曲偏好”矩阵，通过一定的行为数据来定量描述各个用户的喜好和各个物品的属性，即直接建立“用户-潜在因子-音乐”模型。

但问题是，该模型是一个m(用户)×n(歌曲)的矩阵，维度仍然较高，且由于歌曲库庞杂，用户只听过里面一部分的歌，造成最后矩阵较稀疏，计算存在一定困难。

### 5.2.2使用SVD方法进行降维

主成分分析（Principal Component Analysis，简称PCA）是最常用的一种降维方法，它有许多名字，例如线性代数中的散度矩阵奇异值分解（SVD）、统计学中的因子分析（factor analysis）等。在如上模型中，SVD可以起到很大作用。

SVD又称奇异值分解，是线性代数中一种矩阵分解的技术，它能够将任意一个m×n的矩阵A分解成为U、S、V，U是m×m的正交矩阵，V是n×n的正交矩阵，S是m×n的矩阵，且A=U×S×V。

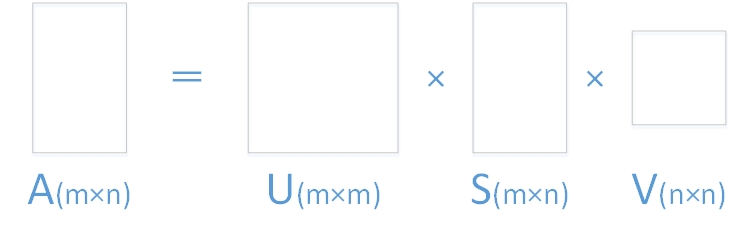


图18 SVD分解-1

其中S矩阵又叫奇异值矩阵，其有一个特点：奇异值矩阵是由个沿对角线从大到小排列的奇异值组成的方阵，且其数值减小速度特别快，大部分情况下，前10%甚至前1%的奇异值的和就占了全部奇异值和的99%以上。这使得我们可以只摘取前K个维度的奇异值，使用“瘦身”后的新矩阵来做原矩阵的近似，维度大大减小，存储与计算成本降低：

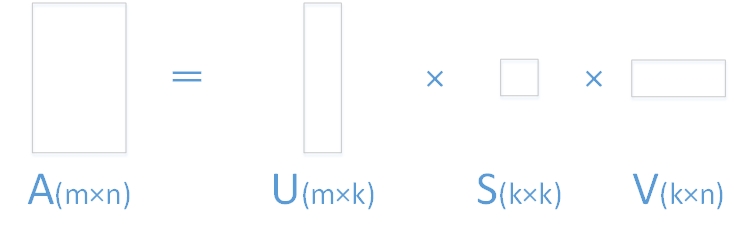


图19 SVD分解-2

使用如下具体实例来说明：

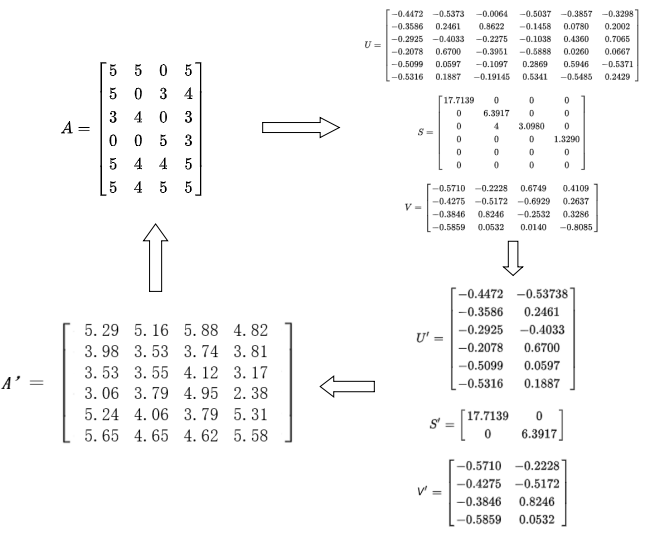
1. 假设存在以下行为user和列为item的数据矩阵A，0表示跳过此首歌，5表示单曲循环，从小到大用户喜好程度依次加大；
2. 对矩阵A进行SVD分解，可获得右侧三个矩阵；
3. 选取k=2来对U，S，V进行降维，k=2即表示我们默认该数据集含有两个潜在因子；
4. 通过降维后的U’、S’、V’相乘来得到A’（保留位小数），可以较直观地看出这两个矩阵比较类似，所以我们可以将SVD方法看作一种数据有损的压缩

图20 音乐偏好矩阵SVD分解实例

### 5.2.3分析用户数据并进行用户歌曲偏好模型的建立

算法整理如下：

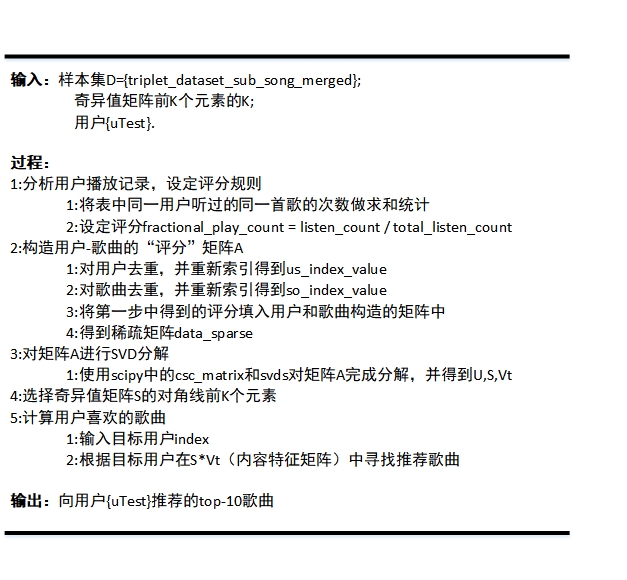
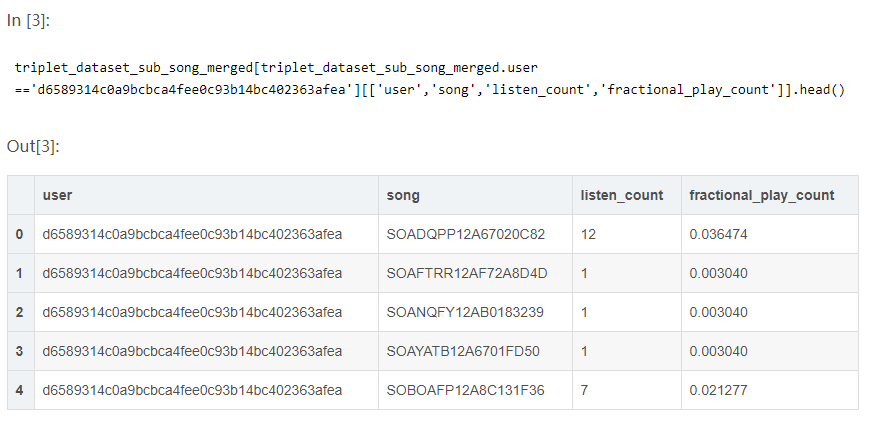


图21 基于矩阵的音乐推荐算法

过程1.1中所作的求和统计如图：

图22 过程1.1中对用户A所作的求和统计情况

测试用户index为{ 4,5}的用户，推荐结果如下：

图23 针对用户4，5的推荐结果

# 模型评估

推荐系统是机器学习技术在企业中最成功和应用最广泛的应用之一。想要找到最有效的模型，只有通过持续调整和调优才能适应不断变化的数据和目标。模型的应用是一个迭代循环的过程。

我们的两个模型都属于聚类问题，是一种无监督学习。通过歌曲相似度和潜在因子模型进行歌曲的推荐，从而将相似的歌曲组成一个类，向用户进行推荐。

聚类评估估计在数据集上进行聚类的可行性和被聚类方法产生的结果的质量。聚类评估主要包括：估计聚类趋势、确定数据集中的簇数、测定聚类质量。

**估计聚类趋势**：对于给定的数据集，评估该数据集是否存在非随机结构。盲目地在数据集上使用聚类方法将返回一些簇，所挖掘的簇可能是误导。霍普金斯统计量(Hopkins Statistic)是一种空间统计量，检验空间分布的变量的空间随机性。计算步骤：

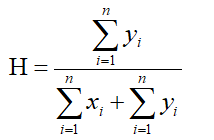
1. 均匀地从D的空间中抽取n个点p1,p2,...pn,对每个点pi(1≤i≤n),找出pi在D中的最近邻，并令xi为pi与它在D中的最近邻之间的距离，即

IMG_256

1. 均匀地从D的空间中抽取n个点q1,q2,...qn,对每个点qi(1≤i≤n),找出qi在D-{qi}中的最近邻，并令yi为qi与它在D-{qi}中的最近邻之间的距离,即

IMG_256

1. 计算霍普金斯统计量H



**确定数据集中的簇数：**K-均值这样的算法需要数据集的簇数作为参数，簇数也可以看作是数据集的有趣并且重要的概括统计量。因此，在使用聚类算法导出详细的簇之前，估计簇数是可取的。

经验方法：IMG_256

肘方法(elbow method)：给定k>0,使用像K-均值这样的算法对数据集聚类，并计算簇内方差和var(k)。然后，绘制var关于k的曲线。曲线的第一个（或最显著的）拐点暗示“正确的”簇数。

交叉验证法：将数据分为m部分；用m-1部分获得聚类模型，余下部分评估聚类质量（测试样本与类中心的距离和）；对k>0重复m次，比较总体质量，选择能获得最好聚类质量的k。

**测定聚类质量：**在数据集上使用聚类方法之后，需要评估结果簇的质量。两类方法：外在方法和内在方法。

外在方法：有监督的方法，需要基准数据。用一定的度量评判聚类结果与基准数据的符合程度。（基准是一种理想的聚类，通常由专家构建）

内在方法：无监督的方法，无需基准数据。类内聚集程度和类间离散程度。

# 未来展望

对于建立的两个模型，我们认为也有许多不足之处，因此在此提出问题并希望能在将来进行改进。

## 7.1.数据使用单一

本次研究涉及的两个模型均只涉及到了用户行为数据，且只包含了user,song,play\_count三个字段，数据维度较为单一，数据使用不够丰富，使训练结果具有一定的片面性。我们可以尝试使用更多的字段来分析歌曲与歌曲之间其自身属性之间的差异和相似度计算，以此更为准确地进行用户推荐。

## 7.2评分设计比较单一

在模型二中，评分的设计比较单一，计算结果值偏小，歌曲与歌曲之间区分度不明显。在完善过程中可以使用其他评分方式（如5分制推荐模型矩阵）对其进行改进。

## 7.3模型评价不完善

在本次研究中，由于算力不足导致无法正常进行聚类问题的评估，不能很好地通过用户反馈评估模型的准确率。所以后面应该引入监督学习，为数据进行结果标记，再做模型评估。

# 参考文献

[1]于润杰. 基于AP聚类算法的推荐系统研究[D].河北大学,2017.

[2]吕倩倩. 基于机器学习的音乐流行趋势预测[D].兰州大学,2017

[3]刘雨亭. 基于内容的音乐特征提取及分类技术研究[D].北京邮电大学,2016.

[4]琚春华,黄治移,鲍福光.融入音乐子人格特质和社交网络行为分析的音乐推荐方法[J].电信科学,2015,31(10):122-130.

[5]李碧. 基于卷积神经网络的钢琴音乐推荐算法研究[D].武汉工程大学,2017.

[6]刘奕. 基于音乐内容语义的推荐算法的研究与实现[D].北京邮电大学,2018.

[7]邓腾飞. 个性化音乐推荐系统的研究[D].华南理工大学,2018.

[8]谭学清,何珊.音乐个性化推荐系统研究综述[J].现代图书情报技术,2014(09):22-32.

[9]李瑞敏,林鸿飞,闫俊.基于用户-标签-项目语义挖掘的个性化音乐推荐[J].计算机研究与发展,2014,51(10):2270-2276.

[10]叶展鹏. 基于CNN自动标注用户偏好的音乐推荐算法[D].华南理工大学,2018.

[11]李博,陈志刚,黄瑞,郑祥云.基于LDA模型的音乐推荐算法[J].计算机工程,2016,42(06):175-179+184.

[12]刘亚林. 基于FP-growth算法的音乐推荐应用研究[D].北京交通大学,2015.

1. 吕倩倩. 基于机器学习的音乐流行趋势预测[D].兰州大学,2017. [↑](#footnote-ref-0)
2. 叶展鹏. 基于CNN自动标注用户偏好的音乐推荐算法[D].华南理工大学,2018. [↑](#footnote-ref-1)