 **JIANGXINORMALUNIVERSITY**

专业学位硕士研究生学位论文

**基于隐马尔可夫模型的音乐推荐系统设计与实现**

**Design and Implementation of Music Recommendation System Based on Hidden Markov Model**

**学 号 201641600018**

**姓 名 李春俊**

**学位类别 硕士**

**专业领域 数据挖掘**

**所在学院 计算机信息工程学院**

**指导教师 张光河**

## 声明

本人郑重声明：

所呈交的毕业设计（论文）是本人在指导教师指导下进行的研究工作及取得的研究成果。其中除加以标注和致谢的地方，以及法律规定允许的之外，不包含其他人已经发表或撰写完成并以某种方式公开过的研究成果，也不包含为获得其他教育机构的学位或证书而作的材料。其他同志对本研究所做的任何贡献均已在文中作了明确的说明并表示谢意。

本毕业设计（论文）成果是本人在江西师范大学读书期间在指导教师指导下取得的，成果归江西师范大学所有。

特此声明。

声明人（毕业设计（论文）作者）学号：

声明人（毕业设计（论文）作者）签名：

签名日期：　　　　年　　月　　日

## 摘要

随着信息技术的飞速发展，数据的种类和数量也与日俱增，借助现有的搜索系统，人们越来越难寻找到自己所需要的信息。因此推荐系统在处理日益繁多的信息的时候，就显得十分重要。

本文首先介绍推荐系统的发展历史和研究现状，然后介绍其中的一个应用——音乐推荐系统。在实现音乐推荐中，阐述了传统推荐算法——协同过滤算法、评价指标以及优劣势。在此基础上，采用聚类和按标签分类的方式对原有的推荐算法进行改进，提高其准确度和覆盖率。而在面对音乐推荐系统中的特殊问题——播放歌单延续性推荐时，本文采用隐马尔科夫模型，基于用户的行为来推测用户隐含歌曲类别喜好，以此实现歌曲的动态推荐功能。

关键词：音乐推荐系统；协同过滤；隐马尔科夫

## Abstract

With the development of information technology.

**Key words:** Music Recommender System; Collaborate Filter; Hidden Markov Model

目录

[声明 2](#_Toc526011038)

[摘要 i](#_Toc526011039)

[Abstract ii](#_Toc526011040)

[第一章：绪论 1](#_Toc526011041)

[1.1研究工作的背景及意义 1](#_Toc526011042)

[1.2推荐系统的发展历史和研究现状 2](#_Toc526011043)

[1.3音乐推荐系统所面临的问题和挑战 5](#_Toc526011044)

[1.4论文的研究内容 6](#_Toc526011045)

[1.5论文的组织结构 7](#_Toc526011046)

[1.6本章小结 7](#_Toc526011047)

[第二章：推荐技术相关简介 7](#_Toc526011048)

[2.1推荐系统概述 7](#_Toc526011049)

[2.2常用的推荐算法及其分类 7](#_Toc526011050)

[2.2.1基于内容的推荐算法 7](#_Toc526011051)

[2.2.2协同过滤推荐算法 7](#_Toc526011052)

[2.2.3混合推荐算法 7](#_Toc526011053)

[2.3推荐算法的评价 7](#_Toc526011054)

[2.3.1推荐算法的实验数据集获取 7](#_Toc526011055)

[2.3.2推荐算法的评价指标和性能分析 7](#_Toc526011056)

[2.4本章小结 7](#_Toc526011057)

[第三章：大数据 8](#_Toc526011058)

[3.1 大数据的技术介绍 8](#_Toc526011059)

[3.1.1大数据的原理与发展 8](#_Toc526011060)

[3.1.2大数据的主要特点 8](#_Toc526011061)

[3.2 MapReduce编程模型 8](#_Toc526011062)

[3.2.1MapReduce编程模型介绍 8](#_Toc526011063)

[3.2.2MapReduce编程模型原理和基本组成 8](#_Toc526011064)

[3.2.3MapReduce编程模型任务执行流程 8](#_Toc526011065)

[3.3 Hadoop开源框架 8](#_Toc526011066)

[3.3.1 Hadoop开源框架简介 8](#_Toc526011067)

[3.3.2 Hadoop开源框架基本组成 8](#_Toc526011068)

[3.3.3 MapReduce在Hadoop中的实现 8](#_Toc526011069)

[3.4本章小结 8](#_Toc526011070)

[第四章：个性化音乐推荐算法 8](#_Toc526011071)

[4.1常见算法所面临的问题 8](#_Toc526011072)

[4.2大数据平台下的音乐推荐算法 9](#_Toc526011073)

[4.3采用隐马尔可夫模型对用户播放列表预测和推荐歌曲 9](#_Toc526011074)

[4.4与普通推荐算法的实验结果对比 9](#_Toc526011075)

[4.5本章小结 9](#_Toc526011076)

[第五章：个性化音乐推荐系统实现 9](#_Toc526011077)

[5.1系统的分析 9](#_Toc526011078)

[5.1.1需求分析 9](#_Toc526011079)

[5.1.2数据库设计与分析 9](#_Toc526011080)

[5.2系统设计与实现 9](#_Toc526011081)

[5.2.1系统目标 9](#_Toc526011082)

[5.2.2系统功能结构图 9](#_Toc526011083)

[5.2.3系统流程图 9](#_Toc526011084)

[5.2.4系统各个模块实现 9](#_Toc526011085)

[5.3综合测试 9](#_Toc526011086)

[5.3.1为用户生成推荐列表测试 9](#_Toc526011087)

[5.3.2歌单列表延续性测试 10](#_Toc526011088)

[5.4系统数据的整理和对比 10](#_Toc526011089)

[5.5本章小结 10](#_Toc526011090)

[第六章：总结与展望 10](#_Toc526011091)

[6.1总结 10](#_Toc526011092)

[6.2展望 10](#_Toc526011093)

[参考文献 10](#_Toc526011094)

[致谢 10](#_Toc526011095)

[作者简介及读研期间主要科研成果 10](#_Toc526011096)

# 第一章：绪论

## 1.1研究工作的背景及意义

随着当今信息技术的飞速发展，数据的种类和数量也与日俱增，人们越来越感觉在海量数据面前束手无策。例如，面对互联网数以千万的歌曲，用户无法从中挑选自己喜欢的歌曲。正是为了解决信息过载的问题，人们提出了两种方案，搜索系统和推荐系统。

搜索系统更倾向于人们有明确的目的。它可以将人们对于信息的寻求转换为精确的关键字，然后交给搜索系统，最后返回给用户一系列列表，用户可以对这些返回结果进行反馈。但它会造成越流行的东西随着搜索过程的迭代会越流行，使得那些越不流行的东西石沉大海。

而推荐系统更倾向于人们没有明确的目的。或者说他们的目的是模糊的，通俗来讲，用户连自己都不知道他想要什么，这时候正是推荐系统的用武之地。推荐系统将用户的历史行为或者用户的兴趣偏好来输入推荐算法，然后推荐系统运用推荐算法来产生用户可能感兴趣的项目列表，同时用户对于推荐系统是被动的。

大数据环境下的推荐系统是传统推荐系统的延伸，所带来的影响也与一般的推荐系统有所不同。大数据环境下，音乐系统产生的数据以隐式反馈数据为主(比如用户对歌曲的收藏，分享，跳过等)，推荐系统必须能对数据进行快速实时处理，以满足用户的需求。

大数据平台下的个性化音乐推荐系统是基于分布式数据平台的推荐系统，它通过对音乐库以及用户产生的海量行为日志进行分析，依据相应的推荐算法挖掘出用户的行为偏好，从而向用户提供个性化的音乐推荐服务。传统的推荐算法当要处理海量的用户行为日志时会显得费时费力，因此借助大数据的分布式存储以及并行计算方式能够有效的处理海量数据。

在音乐推荐系统中，一个很重要的应用场景便是给用户播放列表推荐延续歌曲。但由于用户的兴趣是不断发生变化的，而这种变化是会随着场景的不同而不同，如果仅仅凭借用户之前听取的歌单列表为依据向用户推荐歌曲，这种推荐往往是不及时的。所以如果能够收集用户在收听播放列表时候的最近一段时间的操作行为，并以此为依据推测用户现阶段想要收听的歌曲类别，这样往往能够更为准确的向用户推荐歌曲。在这里我们假设用户对于歌曲类别的即时喜好只与用户此时刻兴趣有关，与之前所有的状态无关，即用户的喜好符合马尔科夫性或无后效性。

我们将用户的歌曲喜好类别看作是一个个隐含的状态变量，而用户的最近一段时间的操作行为（单曲循环、收藏、跳过等）看作是用户展现出来的可观测变量。用户的喜好只与之前一个状态类别有关，这样便可以很精准的向用户推荐隐含的歌曲喜好类别，进而推荐歌曲。

## 1.2推荐系统的发展历史和研究现状

现如今的音乐推荐在国内外都取得了相当大的进展。在国外，著名的在线音乐推荐服务有Pandora、Last.fm以及Spotify等。而国内也有相当多的在线音乐推荐产品，例如，网易云音乐的每日歌单推荐、酷狗音乐的猜你喜欢等功能。相比于传统的音乐搜索功能，音乐推荐可以从用户隐含的动作和习惯中挖掘用户的喜好信息，从而避免了用户搜索音乐的操作，也是实现音乐长尾推荐的方法。例如，Pandora网站是基于“音乐的基因”，通过提取音乐的特征属性来实现推荐。而Last.fm网站则是根据用户的收听行为和收藏记录，找到与其兴趣爱好最相似的若干个用户以及用户喜欢的音乐推荐该目标用户。网易云音乐根据用户的歌曲播放列表和对歌曲播放列表的分类以及相似度来寻找推荐歌曲的。

推荐系统在进行相关设置的时候主要分为两个阶段：数据预处理和推荐生成阶段。在数据预处理阶段，推荐系统需要不断收集和筛选用户与系统交互时候产生的数据。推荐生成阶段，推荐系统需要根据用户的行为信息，利用相关的算法，从数据集中产生用户推荐的项目。总的来说，音乐推荐算法主要包括三种，即基于内容的推荐算法，协同推荐算法以及混合推荐算法。

基于内容的推荐算法，其基本思想是：首先根据用户以往的行为信息，比如收藏，分享的歌曲等，分析出这些歌曲的特征（旋律、风格、歌手、歌词等）信息，以此来构成用户的特征向量，同时歌曲库中的每首歌曲也有着自己的特征向量。进行推荐时，系统分析音乐库中的歌曲特征向量与用户特征向量的相关程度。其中相关程度最大的歌曲即为推荐给用户的歌曲。基于内容的推荐利用项目的内在品质或者固有属性来进行推荐，不需要构建UI矩阵。它是建立在项目的内容信息上作出推荐的，而不需要依据用户对项目的评价意见。该算法还可以通过机器学习的思想来通过训练拟合用户的特征属性。此时需要一个效用函数来评价特定用户c对于特定项目s的评分，以及根据项目的内容属性来学习到跟项目一样维度的用户属性。音乐的属性特征可以从多个角度进行抽取，例如蒋盛益的基于歌词的歌曲高潮片段自动提取。考虑到用户对项目赋予的标签有很大用途，一些学者开始研究标签的推荐。Yang和Hu从机器学习的视角，提出一种为社会化推荐系统自动推荐标签的算法。Marcus认为用户提交的查询中所包含的标签，明确表明了用户的意图，并能确切描述一个项目。因此，他提出一种新的形成标签的方法，即抽取用户的意图，在社会化网络中形成面向目标项目的检索。

基于协同推荐算法，其基本思想是：找到与此用户有相似兴趣的其他用户，然后将他们感兴趣的内容推荐给此用户。一般采用最近邻技术，利用用户的历史喜好信息计算用户之间的距离，然后利用目标用户的最近邻居用户对物品评价的加权评价值来预测目标用户对特定物品的喜好程度，系统从而根据这一喜好程度来对目标用户进行推荐，通常需要用到UI矩阵的信息。

基于混合的算法，顾名思义，是对以上算法的融合，像淘宝既有基于内容的推荐也有协同过滤的推荐。具体怎么融合还是要结合具体的应用场景，包括是对特征的融合还是对算法层面的融合。依据混合算法的推荐系统中最具影响力的是Stanford大学推出的Fab。而在国内也有很多这方面的研究成果，例如，潘宇的个性化彩铃推荐系统的设计与实现；王卫平和王金辉的基于Tag和协同过滤的混合推荐方法。

## 1.3音乐推荐系统所面临的问题和挑战

在音乐推荐系统中主要有三大类问题：

1. 冷启动问题

一个新的用户注册或者当新的项目被加入到音乐库目录中的时候，推荐系统没有充分的数据与这些新数据相关联。这种情况下系统无法产生有效的推荐（即不能把已经存在的项目给新注册的用户，或者无法把刚加入的物品项目推荐给已经存在的用户）。

1. 数据稀疏性问题

已有的用户项目评分数量远低于所有的用户项目评分数量。特别是当所给的用户和项目数量很多的时候。因为大部分用户都倾向于对物品项目中的一小部分做出评分。造成的结果是推荐系统给出的推荐项目列表不可靠，无法令用户满意。

③自动播放列表的延续问题

指的是自动增加原有歌曲列表中的歌曲。自动播放列表的延续问题包含增加一首或者增加很多首歌曲。自动播放列表的延续问题的任务是去准确推断所给播放列表的目的，这个问题的难点在于所给出列表反映的用户意图很宽泛，需要从其中找出用户所要需求歌曲的特征。

虽然今天的音乐推荐系统可以帮助用户从众多的音乐类别中找到他们所喜欢的歌曲，但是关于音乐系统的研究还是遇到了许多的挑战。尤其是，当音乐推荐系统要去构建、整合、评估推荐算法的时候。因此，深入挖掘用户的需求、了解用户的喜好和意图应该成为现如今音乐推荐系统新的研究方向。

## 1.4论文的研究内容

面对音乐推荐系统所面临的问题和挑战，本论文分别采用不同的推荐算法来解决。对于数据稀疏性问题，采用将用户事先进行聚类，以评分数据更多的用户簇来替代单个用户，以此实现用户评分过少导致推荐不准确的问题。而面对项目冷启动的问题，推荐系统采用贴标签的方式，以此快速对刚注册的用户和物品进行分类，来达到精准推荐的目的。最后面对用户的播放列表延续歌曲推荐功能，本论文采用收集用户在最近播放歌曲时的操作记录，来推测用户本身的隐含喜好歌曲类别，以此实现歌单的动态推荐。

本论文采用对比论证的实验方法，首先用传统的协同过滤技术对含有大量评分记录的用户进行歌曲推荐，分析推荐的准确度与覆盖率。然后采用聚类和贴标签分类的方式进行推荐歌曲，同时得到推荐的准确度和覆盖率，以此对比两种方法的优劣。在播放列表延续歌曲推荐时，采用用户对推荐歌曲的评分误差来评定这些推荐是否是用户喜欢的歌曲。

## 1.5论文的组织结构

本论文分为六章，第一章首先对推荐系统和搜索系统做了一个比较，阐述了现在推荐系统的发展趋势，以及在大数据环境下，推荐系统与大数据平台的结合，同时也分析了音乐推荐系统所面临的问题和挑战。论文的第二章主要介绍一些传统的推荐技术和对应的评价算法指标。第三章讲述了大数据环境下，推荐系统和Hadoop框架的融合，主要介绍MapReduce技术。论文的第四章介绍个性化音乐推荐算法，主要介绍利用隐马尔科夫模型估计用户喜好歌曲的算法。第五章主要是个性化音乐推荐系统的设计、实现和测试。第六章是对全文的总结和展望。

## 1.6本章小结

本章是论文的绪论部分，主要介绍推荐系统以及推荐系统中的应用——音乐推荐系统。阐述推荐系统的发展历史和面临的挑战。

# 第二章：推荐技术相关简介

## 2.1推荐系统概述

## 2.2常用的推荐算法及其分类

### 2.2.1基于内容的推荐算法

### 2.2.2协同过滤推荐算法

#### 2.2.2.1基于模型的推荐算法：常见的机器方法(聚类、分类、矩阵分解、神经网络、图、贝叶斯网络等)

#### 2.2.2.2基于内存的推荐算法：基于用户推荐、基于物品推荐

### 2.2.3混合推荐算法

## 2.3推荐算法的评价

### 2.3.1推荐算法的实验数据集获取

### 2.3.2推荐算法的评价指标和性能分析

## 2.4本章小结

# 第三章：大数据

## 3.1 大数据的技术介绍

### 3.1.1大数据的原理与发展

### 3.1.2大数据的主要特点

## 3.2 MapReduce编程模型

### 3.2.1MapReduce编程模型介绍

### 3.2.2MapReduce编程模型原理和基本组成

### 3.2.3MapReduce编程模型任务执行流程

## 3.3 Hadoop开源框架

### 3.3.1 Hadoop开源框架简介

### 3.3.2 Hadoop开源框架基本组成

### 3.3.3 MapReduce在Hadoop中的实现

## 3.4本章小结

# 第四章：个性化音乐推荐算法

## 4.1常见算法所面临的问题

## 4.2大数据平台下的音乐推荐算法

## 4.3采用隐马尔可夫模型对用户播放列表预测和推荐歌曲

## 4.4与普通推荐算法的实验结果对比

## 4.5本章小结

# 第五章：个性化音乐推荐系统实现

## 5.1系统的分析

### 5.1.1需求分析

### 5.1.2数据库设计与分析

## 5.2系统设计与实现

### 5.2.1系统目标

### 5.2.2系统功能结构图

### 5.2.3系统流程图

### 5.2.4系统各个模块实现

## 5.3综合测试

### 5.3.1为用户生成推荐列表测试

### 5.3.2歌单列表延续性测试

## 5.4系统数据的整理和对比

## 5.5本章小结

# 第六章：总结与展望

## 6.1总结

## 6.2展望

# 参考文献

# 致谢

# 作者简介及读研期间主要科研成果