

**学士学位论文（设计）**

**题目：** 基于HMM模型的在线视频个性化推荐研究

**姓 名**  冯运佳

**学 号**  2021234110701

**院 系**  信息科学与技术学院

**专 业**  信息科学与大数据技术

**指导教师**  张广辉

**2023年 11 月 16 日**

摘 要

基于隐马尔可夫模型的个性化推荐模型，通过对模型进行三个层级的推荐设计来实现个性化推荐目标，并且每个推荐层级针对不同的用户群体，对不同类别视频爬取的一些信息进行一定的加权处理，建立用户兴趣时间转移阈值，推荐给用户最可能感兴趣的视频内容的方法，完成对用户的个性化推荐。

**关键词**：个性化推荐；隐马尔科夫模型

ABSTRACT

THE personalized recommendation system based on Hidden Markov Medel was used on this paper. To achieve the personalized recommendation goal, We use the three-level recommendation design of the model and the layer-by-layer progressive approach. And each recommendation level is targeted at different user groups. We use the information which is obtained by crawling to weigh and build a user interest time shift threshold, recommending the online video content most likely to be of interest to the experienced user, and completing the personalized recommendation target of the senior user.

**Key words**: personalized recommendations; Hidden Markov model

目 录

[摘 要 III](#_Toc151129647)

[ABSTRACT III](#_Toc151129648)

[第一章 前言 1](#_Toc151129649)

[1.1 引言 1](#_Toc151129650)

[1.2 问题描述 1](#_Toc151129651)

[1.3 方法描述 1](#_Toc151129652)

[第二章 仿真实验及其分析 2](#_Toc151129653)

[2.1 建立隐马尔可夫模型 2](#_Toc151129654)

[2.2 隐马尔可夫模型应用 3](#_Toc151129655)

[第三章 仿真实验及分析 5](#_Toc151129656)

[3.1 模型参数准备 5](#_Toc151129657)

[3.2 隐马尔可夫模型训练 5](#_Toc151129658)

[3.3 获取观察值序列 2](#_Toc151129659)

[3.4 关键词提取 2](#_Toc151129660)

[3.5 三级推荐目标实现 4](#_Toc151129661)

[3.6 模型实验结果 4](#_Toc151129662)

[3.6.1 第一层级的个性化推荐结果（表1） 4](#_Toc151129663)

[3.6.2 第二层级的个性化推荐结果（表2） 5](#_Toc151129664)

[3.6.3 第三层级的个性化推荐结果（表3） 5](#_Toc151129665)

[参 考 文 献 5](#_Toc151129666)

# 前言

## 引言

随着互联网的普及和数字媒体内容的迅速增长，视频成为了人们获取信息、娱乐和交流的重要方式。然而，随之而来的问题是，用户在庞大的视频库中往往会感到困惑和不知所措。为了解决这一问题，视频推荐算法应运而生。

视频推荐算法旨在根据用户的兴趣和偏好，从海量的视频资源中选择最相关、最具吸引力的内容，以提供个性化的用户体验。通过分析用户的观看历史、搜索行为、社交互动和其他相关因素，视频推荐算法能够准确把握用户的兴趣爱好，并向其推荐符合其口味的视频内容。

## 1.2 问题描述

然而，视频推荐算法并非一蹴而就的成果。它背后涉及到多个复杂的技术和挑战，包括数据收集与处理、特征提取、相似度计算、推荐模型构建等。同时，还需要考虑到用户的多样性和时效性需求，以及平台的商业利益和内容的多样性。

本论文旨在通过使用隐马尔科夫模型（HMM）为用户进行个性化的视频推荐，进而提高平台的商业利益和内容多样性。

## 1.3 方法描述

隐马尔科夫模型（Hidden Markov Model，HMM）是一种统计模型，常用于建模时序数据和序列模式识别。该模型假设系统是一个隐藏的马尔科夫链，状态无法直接观测到，但会产生可观测的符号序列。在本文中，我们将使用隐马尔科夫模型来对视频推荐算法中的用户行为序列进行建模和分析。

首先，我们将介绍隐马尔科夫模型的基本原理和数学表达式，包括状态空间、观测空间、状态转移概率矩阵、观测概率矩阵等核心要素。隐马尔科夫模型的基本假设是当前时刻的状态仅依赖于前一时刻的状态，并且通过观测序列反映隐藏状态的变化。

其次，我们将阐述如何将视频推荐算法中的用户行为序列映射到隐马尔科夫模型的状态空间和观测空间。用户行为序列可以包括观看历史、搜索记录、点赞评论等多种类型的行为，我们将探讨如何将这些行为映射为隐马尔科夫模型的输入观测序列。

接着，我们将描述参数估计和模型训练的方法。针对隐马尔科夫模型的参数估计，我们将介绍基于极大似然估计或 Baum-Welch 算法的训练方法，以及如何处理模型的初始化和收敛性等问题。

最后，我们将说明如何利用训练好的隐马尔科夫模型进行用户行为序列的预测和推荐。通过隐马尔科夫模型，我们可以对用户下一步的行为进行预测，从而实现个性化的视频推荐。

# 仿真实验及其分析

## 建立隐马尔可夫模型

下面是以解码问题为例的隐马尔可夫模型建模过程：

1. 状态集：定义一个状态集合

G={G1,G2,…,Gn}；

1. 转移集：每个状态转移元素集合

(gi→gi+1)；

1. 观察集：输出的观察符号集：

∑=σ1，σ2，σ3,…σm

从开始，转移到新的状态，可以观测到一个 输出符号。经过转移到最终状态，会得到一个符号串：。每一个状态转移对应的概率为

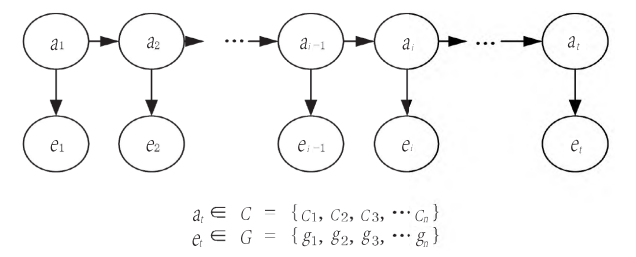
。在任意状态观测到特殊符号的概率 P (H |g )。在一个隐马尔可夫模型O上，序列H被观测到的概率是所有可能路径的概率和，如式（1）所示：

（1）

建立隐马尔科夫模型的目的就是找到能使观察概 率达到最大的状态序列Q (H|O)，如式（2）所示：



HMM用图形表示如图1所示。



图一 HMM图形

任意时刻状态序列的产生：序列在某时刻的状态是，在该状态下得到的观察值是，从转移到，在状态下得到观察值。按照此种方法，会得到一个观察值序 。因为在任意时刻产生的观察值e只依赖于状态值a，所以可以得到在a状态下出现e观察值的概率为所有时刻状态转移概率和观测概率的乘积，即。

## 隐马尔可夫模型应用

（1）个性化推荐：基于隐马尔可夫模型的偏好进化学习和高维数据稀疏问题处理。解决高维数据稀疏问题和因用户偏好学习困难而难以高效反应用户的个性化需求的问题，从而突破上下文感知推荐系统的应用瓶颈。利用评分分组，构造转移矩阵和观察矩阵。为防恶意刷分引入权重因子，完成隐马尔可夫的建模。

（2）文本分类：通过提取文本中的特征词、特征、词频、文档数量、特征词语义等信息，利用信息增益的方法进行特征词的合并和分类，通过分类分别建立隐马尔可夫分类模。新文本的分类则是通过分类向量的交集，作为观察序列输入隐马尔可夫模型，预测各分类输出的概率，以概率大小作为分类标准。

（3）舆情分析：以微博为平台，考虑微博用户的个人特征、感情走向，以具体的行为表现等为观察序列，以及舆情的发展、波动消亡等为状态序列，建立隐马尔可夫模型，预测和预警舆情的发展和一般规律，为最终决策提供科学高效的方法。

（4）智能学习：假设用户拥有良好的学习习惯和效果，需要从用户的学习序列中找到一个最优的学习序列引导学习的目的。建立隐马尔可夫模型，对模型进行训练，寻找知识点间的关联概率，利用模型预测知识点的迁移序列，找到最优的状态序列，智能引导用户学习，提升用户的学习效率和学习效果。

（5）时间序列预测：假设一系列的隐状态构成了一个时间序列，隐状态间的转换构成系统运行的根本。对时间序列分段后，根据先后发生的时间顺序和发生时间的相似性等进行聚类，以分类为隐含状态得到状态转移矩阵，建立隐马尔可夫模型，进行时间序 列的预测。

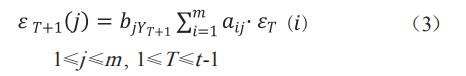
# 仿真实验及分析

## 模型参数准备

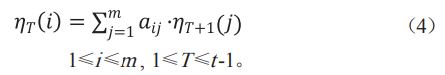
以在线视频的评分为状态序列，评分对应的视频内容即为观察序列。用户某一次对在线视频内容的评分只与前一次的评分有关，与下一次无关，所以用户的评分行为符合隐马尔可夫随机过程

## 隐马尔可夫模型训练

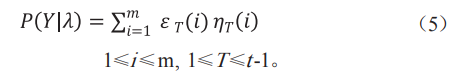
利用无监督的学习法（EM算法），训练集中t时刻使用并给出所得到评分的用户观察集，隐含的状态序列，通过对训练集参数的训练，得到基于隐马尔可夫最大概率，即。训练方法如下。

①根据向前-向后算法（训练HMM的标准算法是向前-向后算法（Forward-backward Algorithm）或者叫作鲍姆-韦尔奇算法（Baum-Welch algorithm），这是期望最大化（Expectation-Maximization algorithm，EM算法）的一种特殊情形)，定义一个在T时刻处于状态i 的前向变量概率是(i)，如果T+1时刻的状态为j，根据前向算法，则有公式（3）:

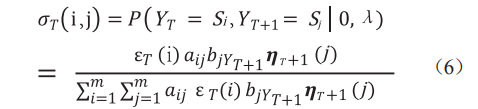
②根据向前-向后算法，定义一个在T时刻状态是Si的后向变量概率是ηT(i)，代表如果T+1时刻到t时刻观察到序列的概率，根据向后算法则有公式（4）：



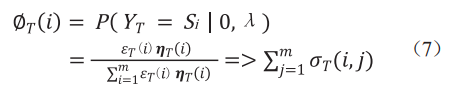
③采用向前-向后算法完成计算后，计算见公式（5）



④假如已知某刻T的在线视频概率变量。代表在这一时刻T对在线视频内容的的评分是，在下一时刻T+1时刻对视频 的评分是 的 概率，其中（1≤I, j≤m, 1≤T≤t）。把式（3）～（5）带入 式（6）可得



⑤假如已知用户在时刻的在线视频概率变 量是，代表在时刻在 线视频的评分是的概率。把式（3）～（5）代入式（7）可得：



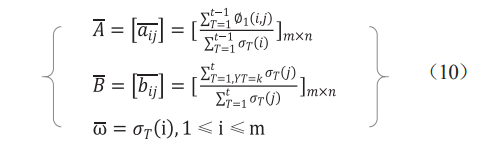
⑥根据步骤④和⑤，得到转移得到状态的期望值。评分值期望如式（8）所示：



由于转移到状态的期望可以计算的出，那么转移到其他状态的方法也可以得到，如式（9）所示：



⑦可以得到新的隐马尔科夫模型，各变量的计算公式如式（10）所示：



⑧迭代①～⑦步，直至三个参数的数值 收敛为止，得到最大的的概率值最大。

## 获取观察值序列

用户在视频网站上观看视频和评分的信息通过观看时间戳的方式记录，通过观看评分记录可以从中提取一个在线视频评分路径转移序列。定义该序列为。假如用户选择评分的在线视频资源都是用户喜欢的内容，那么用户选择评分的某在线视频中必然存在某些吸引用户的信息，而吸引用户选择观看并评分的信息就可以定义为该用户的兴趣点。通过一系列的用户浏览记录，就可以提取用户的兴趣集。这个兴趣集就是隐马尔可夫模型“隐”的内涵。通过挖掘用户的兴趣集当中兴趣点并进行加权，找出符合用户兴趣点权重的视频内容，推荐给该用户，那么就可以获得好的用户体验。

用户在评分转移的过程中，每一个评分对应一个视频内容。而用户评分对应的在线视频内容就是隐马尔可夫模型的观察值，获得观察序列，观察序列示意如图2所示。

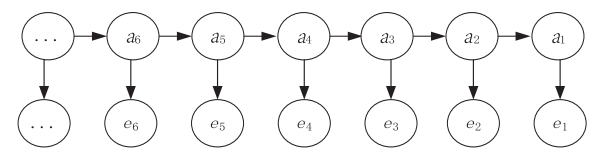


图2 观察值序列

## 关键词提取

在线视频网站上的某个视频内容简介，就是对该视频内容的简单概括，如图3所示。



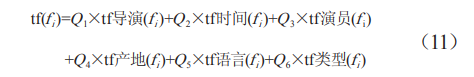
图3 豆瓣网站图片

图3中展示的是豆瓣网站上电影《流浪地球》的页面截图，从图中可以发现每部电影的简介信息结构相同，但页面的关键字不同，用户选择观看电影会根据电影简介的关键字进行选择，喜欢科幻或灾难题材电影的用户可能会选择观看《流浪地球》。假设不同用户因为自己的喜好不同而选择不同，那么用户最终选择观看影片的出发点可能就是网页中不同的关键字。

一部电影的简介页面中包含多个关键字，那么一部电影可以用一个关键字集合来表征，用户最终所选择观看的某部电影，决定因素可能就是因为被某个电影中的关键词吸引。为了找到用户的兴趣关键词，决定 采用Python定向爬取网站关键词信息的方式实现，具体步骤如下：

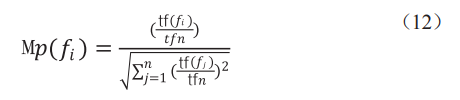
①确定爬取的目标信息。因电影简介页面的标准化程度较高，所以决定定向的爬取有用的特征词信息，建立兴趣关键词集。这里需要爬取的特征词主要包含：电影名，电影上映时间、评分、导演、演员、产地、语言、影片类型等信息。建立一个特征词集，F ={f 1,f 2,f 3,…,f n}，f i（i表示自然数1…n）则表示视频网站上爬取的一个特征词，n表示特征词的数量。

②特征词出现的频率计算：采用爬取后的信息建立起的特征词集即为当前影片的表征形式。由于电影的结构化程度较高，关键词的使用频率较固定，所以用户如果只存在观看了一部影片记录，那么并不能发现该用户的关键词兴趣。所以需要用户的一系列的浏览记录集，建立一个多事务（观看的影片）的关键词集。通过观看记录关键词的提取计算，才能发现用户真正感兴趣的关键词。计算公式如下：



式中，代表调整系数。

计算完特征词的出现频率，就可以算出每个特征词对该用户的权重值。计算公式如下：



通过公式（4）～（10），可以计算出每个特征词对该用户的兴趣权重，根据该权重就可以计算出该用户的对任何一部电影的感兴趣的概率，根据概率值得大小既可以预测该用户最可能感兴趣的视频资源。

豆瓣电影网的结构化的数据可以利用Python进行爬取。

## 三级推荐目标实现

基于隐马尔可夫模型的在线视频个性化推荐的最终目标是满足不同类型用户的使用需要。第一层级的个性化推荐：非注册用户的个性化推荐，主要解决的问题是非注册用户的冷启动问题。第二层级的个性化推荐：注册新会员用户的个性化推荐，解决的问题是新用户的用户稀疏问题。第三层级的个性化推荐：针对网站的资深会员用户的在线视频个性化推荐服务，提升用户的使用满意度。

## 模型实验结果

### 第一层级的个性化推荐结果（表1）

表1 第一层级的个性化推荐输出结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 概率P(%) | 63.9 | 61.7 | 61.2 | 60.8 | 56.9 |
| 电影id | 318 | 858 | 7502 | 77658 | 50 |
| 电影名 | Shawshank Redemption | Godfather | Band  of Brothers | Cosmos | Usual Suspects |

### 第二层级的个性化推荐结果（表2）

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 概率P(%) | 64.6 | 61.6 | 56.7 | 56 | 55.9 |
| 电影id | 318 | 858 | 527 | 2959 | 1221 |
| 电影名 | Shawshank Redemption | Godfather | Schindler's List | Fight Club | Godfather  Part II |

本层的数据结果本应是6个不同年龄段的输出结果，但由于数据集中的年龄数据是随机生成的，而随机生成的数据对于各年龄段的输出结果几乎没有差别，所以本输出结果只对其中一个输出结果进行展示。

### 第三层级的个性化推荐结果（表3）

表3是基于隐马尔可夫模型的个性化推荐结果展示，模型输出7120个结果，本文只截取前5个结果作为展示。由表3可以发现，前5个用户的兴趣关键词各不相同，所得到的推荐电影也各不相同，得到相应预测评分的概率也相差较大。但各自获得相应预测评分的概率都是在各自兴趣关键词集中概率最高的结果。第三级个性化推荐完成。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 用户ID | 预测评分 | 电影ID | 电影类型 | 电影名 | 概率P（%） |
| 1 | 4 | 356 | Drama | Forest Gump | 32.2 |
| 2 | 5 | 780 | Sci-Fi | Independence Day | 16.7 |
| 3 | 5 | 110 | Action | Braveheart | 42.4 |
| 4 | 4 | 296 | Comedy | Pulp fiction | 52.9 |
| 5 | 5 | 480 | Adventure | Jurassic Park | 41.1 |

基于对大型视频网站用户的观看记录数据的采集，建立基于隐马尔可夫模型的个性化推荐系统。推荐分成三个推荐层级，分别针对不同的用户类型，解决了不同用户的实际需求，并提出用户兴趣时间阈值观念，适时地满足用户的个性化需求，为在线视频网站的发展奠定基础。

参 考 文 献

[1]王新武.基于隐马尔可夫模型的在线视频个性化推荐研究[J].数字通信世界,2023(09):1-4.

[2] Baum L E , Petrie T .Statistical Inference for Probabilistic Functions of Finite State Markov Chains[J].Annals of Mathematical Statistics, 1966, 37(6):1554-1563.

[3] 朱志国．基于隐马尔可夫链模型的电子商务用户兴趣导航模式发现[J]． 中国管理科学，2014,22(4):67-73.

[4] 张浩，何杰，李慧宗，等．基于隐马尔科夫模型(HMM)与张量分解的 两阶段个性化推荐模型与应用[J]．科学技术与工程，2016,16(19):84- 90+115.

[5] 高知新，徐林会．基于隐马尔科夫模型与语义融合的文本分类[J]．计算 机应用与软件，2017,34(07):303-307.

[6] 何建民，李雪．面向微博舆情演化分析的隐马尔科夫模型研究[J]．情报 科学，2016,34(04):7-12.

[7] 翟琳琳，陈仪香．隐马尔科夫模型在智能学习系统中的应用[J]．计算机 工程与应用，2007(06):178-180.

[8] 杨安驹，杨云，周嫒嫒，等．基于隐马尔科夫模型的融合推荐算法[J]． 计算机与现代化，2015(09):60-65.

[9] 金敏，鲁华祥．一种遗传算法与粒子群优化的多子群分层混合算法[J]． 控制理论与应用，2013,30(10):1231-1238