DZKD
电子科技大学
UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

硕士学位论文

MASTER THESIS

论文题目 基于用户偏好的视频推荐技术研究

学科专业 通信与信息系统

学 号 201021010231

作者姓名 唐 明

指导教师 阳小龙 教授

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

作者签名：[illegible]　　日期：2013 年 5 月 24 日

论文使用授权

本学位论文作者完全了解电子科技大学有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权电子科技大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后应遵守此规定）

作者签名：[illegible]　　导师签名：[illegible]

日期：2013 年 5 月 24 日

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

分类号 密级

UDC [注1]

学 位 论 文

基于用户偏好的视频推荐技术研究

唐明

指导教师 阳小龙 教 授

电子科技大学 成 都

（姓名、职称、单位名称）

申请学位级别 硕士 学科专业 通信与信息系统

提交论文日期 2013.04 论文答辩日期 2013.05

学位授予单位和日期 电子科技大学 2013 年 6 月 日

答辩委员会主席

评阅人

注 1：注明《国际十进分类法 UDC》的类号。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

VIDEO RECOMMENDATION BASED ON USER PREFERENCE

A Thesis Submitted to

University of Electronic Science and Technology of China

Major: **Communication and Information System**

Author: **Tang Ming**

Advisor: **Xiaolong Yang Professor**

School: **School of Communication & Information Engineering**

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

摘　要

伴随信息技术和互联网应用的发展，人们逐渐从信息匮乏时代迈入了信息过载（information overload）时代。在视频业务系统为代表的互联网资源信息平台中，传统的搜索技术已满足不了用户对网络个性化的需求。个性化网络时代要求资源业务平台能主动分析用户行为，发现用户兴趣偏好，为用户发现符合其个性需求的信息资源。个性化推荐技术正是为了应对网络个性化的这一挑战而出现的。

视频资源系统中，用户对视频不同的观看行为决定视频之间流行程度的偏差，视频流行程度呈现出长尾分布，即热门的视频只是整个系统中的少数资源，系统大多数视频资源都是相对冷门的。另外，用户的兴趣偏好不是一层不变的，通过利用用户行为信息中的时间信息可以进一步分析用户兴趣偏好随时间变化的特征和规律。同时，累积的用户评分行为可以体现出用户和视频的某种评分趋向，如有些用户比较苛刻，或者视频品质较差，他们趋向于有较低的评分值。传统视频推荐技术并未考虑上述的客观事实对推荐结果的影响，推荐质量欠佳。

本文详细介绍了视频推荐系统及其重要技术。在此基础上，本文就推荐技术的两个核心问题，Top-N 推荐和评分预测推荐，针对传统推荐技术的上述缺陷设计了改进的推荐算法模型。

首先，本文提出了一种综合利用用户隐性反馈和显性反馈信息的用户偏好建模策略。然后，在此基础上，提出了应用于 Top-N 推荐的基于用户的协作过滤和基于项目的协作过滤算法，并在基于用户的协作过滤推荐中引入视频流行度权重和用户偏好变化权重因子，提出改进算法。在实验中对比了用户-视频二进制关联矩阵模型下三种算法的性能，实验表明了，考虑视频流行度和用户偏好变化的推荐策略能给推荐质量带来有效提升。

其次，针对推荐系统评分预测问题，以用户评分数据稀疏性为切入点，简单介绍了现有的解决方案，并针对其中的典型算法，基于矩阵分解模型的协作过滤算法提出了引入偏置项的改进模型，最后通过实验分析，验证了改进的矩阵分解模型能有效地提高评分预测准确度。

关键词：视频推荐，流行度，兴趣变化，矩阵分解，协作过滤

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

ABSTRACT

With the development of information technology and Internet applications, the human society is striding from the information lacking age to information overload one. Traditional search technologies, as in the Internet information platform which is represented by video related service, cannot fulfill the users' content for personalized service. Active analysis of users' behaviors, to locate users' preferences and unearth the information resources which conform to users' individual demands is needed by the resources platform in the personalized Internet age. The personalized recommendation technology rises no other than to this challenge.

In video resource system, the popularity of the videos, determined by users' different watching behaviors, presents as fat-tailed distribution, i.e., most-viewed videos are only a few. Majority of the videos are less favored. Moreover, users' preferences are not always fixed. By making advantage of users' behaviors information, further analysis the features and rules of users' preferences which change s over time can be carried out. Meanwhile, users' accumulated rating behaviors could manifest some kind of rating trend, like some demanding users or poor video quality; they tend to a higher rating video. Excluding the effects of above-mentioned objective facts to the recommendation results, the traditional video recommendation technology has not do well in this respect.

This article gave a detailed introduction of video recommendation system and its essential technologies. Against this background, two core issues of recommendation technology, Top-N recommendation and score predict recommendation, are mentioned. The improvements of recommendation algorithm modeling aim at the traditional recommendation technology have also been stated.

First, this article presented users' preferences modeling strategies which comprehensively make uses of the users' implicit and explicit feedback. Then it proposed a collaborative algorithm, relied on users and items' collaborative filtering, which can be applied to Top-N recommendation. The popularity weight of video and the change weight of users' preferences are inserted into users' collaborative filtering algorithm. The performance of three algorithms, on the ground of user-video binary

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

related matrix model, is compared in the experiments. The experiments showed that the strategy to take the popularity of the video and users' preferences into consideration will effectively enhance the quality of recommendation.

Second, briefed on the existing solutions to the score prediction of recommendation system that based on the sparseness of the users' rating data, especially dwells on the representative algorithm, collaborative algorithm which leans against matrix decomposed model, and proposed an improved model inserted with bias terms.

Finally, by the experimental analysis, verified that the improved matrix decomposed model can effectively improve the accuracy of the score prediction.

Key words: Video Recommendation, Popularity, Interest Drift, Matrix Factorization, Collaborative filtering

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

目　录

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第一章 绪论

1.1 课题研究背景及意义

伴随信息技术和互联网应用的发展，人们逐渐从信息匮乏的时代迈入了信息过载（information overload）的时代。在这个时代，无论信息消费者还是信息生产者都遇到了很大的挑战：对应信息消费者，从大量信息中找到自己需要或者感兴趣的信息资源是一件非常困难的事情；对于信息生产者，让自己生产的信息脱颖而出，受到广大用户的关注与接收，也是一件非常困难的事情。搜索技术从最初雅虎公司的分类目录发展到当下以谷歌和百度为代表的自然语言搜索技术，它可以根据用户提交的搜索关键词从网络信息海洋中找到用户需要的信息资源。虽然被广泛应用的搜索技术能够在一定程度上解决这样的矛盾。但它仍存在着几点明显的缺陷[1]：

1）信息获取不够灵活。搜索引擎执行搜索的唯一依据是用户主动输入的关键词，搜索结果完全取决于这些关键词，而无法依据很多意义上相近的词语，这样的信息获取方式显得笨拙。如果用户对自己的需求没有清楚的认知，语义偏差的关键词也可能导致通过搜索引擎也找不到想要的信息资源。

2）被动的信息提供方式。搜索引擎只会在有用户搜索行为发生时，才能为用户提供信息服务。即搜索技术只能被动地等待用户的搜索行为，事件才会发生。

3）不能支持个性化服务。搜索引擎对用户一无所知，只要关键词相同，哪怕是不同用户，得到的搜索结果都是一样的。它不能获取用户历史行为，进而分析用户兴趣偏好，发现用户潜在需求，为用户主动提供个性化服务。

搜索技术存在以上缺陷，原因在于网络个性化的发展。网络个性化要求信息提供方式从用户主动明确的查询、系统被动的提供转变为系统主动地感知用户需求, 进而提供个性化信息资源。在这样的需求推动下, 个性化推荐（Recommendation Technology）技术应运而生。和搜索引擎一样，推荐系统也是一种帮助用户快速发现有用信息的工具。和搜索引擎不同的是，推荐系统不需要用户主动地提供关键字等需求描述，而会利用用户的历史行为分析挖掘用户的兴趣偏好，最后形成最大化满足用户需求和兴趣的推荐提供给用户。实际上，推荐系统是搜索引擎的发展，也是它的互补工具。搜索引擎满足了用户有明确目的时的主动查找需求，而

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

推荐系统能够在用户没有明确目的时帮助他们发现感兴趣的新内容。

推荐技术最早应用于电子商务领域。著名的电子商务网站亚马逊（Amazon）是个性化推荐的积极应用者和推广者，被 RWW（读写网）[2]成为“ 推荐系统之王 ”。个性化推荐系统对亚马逊的意义，其 CEO Jeff Bezos 在文献[3]中说过，亚马逊相对于其他电子商务网站的最大优势就在于个性化推荐系统，该系统让每个用户都能拥有一个自己的在线商店，并且能在商店中找到自己感兴趣的商品。实际上，个性化推荐系统是否成功已成为电子商务网站经济效益产生的关键。亚马逊的科学家 Greg Linden 曾经说过, 该公司至少有 35%的销售来自于推荐算法, 同时, 据 2012 年第二财季数据，该季度亚马逊的营收达到了 128.3 亿美元，与去年同期的 99 亿美元大涨了 29%。毫无疑问，如此的经济效益离不开推荐系统。

据 CNNIC（中国互联网信息中心）最新发布的《中国互联网发展状况统计报告》显示，截止 2012 年 12 月底，中国网民规模达到 5.64 亿，网络视频用户达到 3.72 亿，较上年增加了 4653 万人；另外，MRG 研究公司称，到 2014 年全球 IPTV 用户数将达到 1.02 亿，保持 25%的年增长率。网络视频和 IPTV 等视频业务用户与日猛增的同时，视频内容和数量也在迅猛发展，每个国内知名的视频网站中都有极其庞大的视频资源库。互联网视频的繁荣景象使得人们对丰富多彩的视频节目感到欣喜的同时，也为难以从浩瀚的节目资源中挑出自己真正喜欢的节目而感动无奈。个性化视频推荐系统成了解决如此烦恼的良药。

在视频业务系统中应用个性化推荐技术的好处是明显的：用户通过个性化推荐系统能够得到个性化的视频推荐服务，系统通过对用户自身的历史行为分析向用户推荐用户感兴趣的视频能够有效的解决用户需求与浩瀚的视频资源这一矛盾；通过资源的推荐，增加了系统资源的调度和利用率，为视频业务系统带来更多的观看率和经济效益；好的推荐能增加用户视频体验乐趣，从而有助于提升用户对视频业务系统的粘滞性和忠诚度。

随着近年来对推荐系统研究的开展，推荐技术中逐步凸显出了一些问题与挑战[4]：推荐技术利用的是用户历史行为数据，当系统中出现新用户或者新物品，他们没有有效的历史行为数据，推荐系统很难对新用户做符合其兴趣的推荐或者把新资源推荐给合适的系统用户；大型的推荐系统中，用户有过行为的资源数目可能远少于系统总的资源量，加上用户之间选择的差异就带来用户行为数据的稀疏问题，面对过于稀疏的用户行为数据，推荐系统很难做出好的推荐；现实中的用户兴趣不是一层不变的，分析用户历史行为中可能会随时间变化的用户兴趣也是推荐系统设计的重要挑战。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

本课题伊始于杭州某电视运营商的业务需求，他们期望通过建立有效的用户行为分析机制，设计智能的个性化视频业务推荐系统，把用户被动地接受海量资源变成不同用户看到量身定制的系统资源推荐，为用户减少多余信息和优化业务体验，进而为系统提高资源点击率。

1.2 国内外研究现状

推荐系统技术作为搜索引擎技术的发展和补充，一直都是近十年来计算机和信息领域的研究热点。九十年代中期，ACM 智能用户接口会议（ACMIUI）、国际人工智能联合大会（IJCAI）和美国人工智能协会春季会议（AAAI）等国际会议上出现的多篇关于个性会推荐系统的论文[5-8]拉开了个性化推荐研究的序幕。从这以来，推荐系统在电子商务、网络经济学和人类社会学等领域保持着很高的研究热度并逐渐成为了一门独立的研究领域。学术界和产业界对推荐技术的热情呈现出了一片蓬勃的景象。

致力于该领域的会议和研讨会出现：ACM 于 2007 年专门设立了推荐系统年会（ACM Recommender Systems，RecSys），该年会已成为推荐技术研究和应用的最重要会议。另外，人机交互、数据挖掘和机器学习等相关传统领域的顶级会议，如国际信息检索大会（SIGIR）、数据挖掘与知识发现会议（SIGKDD）和用户个性化会议（UMAP）等，也逐步加入了对推荐技术的研究探讨。

相关领域的国际顶级期刊曾发表关于推荐系统技术的研究成果：Physics Reports（2012）；AI Communications（2008）；IEEE Intelligent Systems（2007）；International Journal of Electronic Commerce（2006）；International Journal of Computer Science and Applications（2006）；ACM Transactions on Computer-Human Interaction（2005）和 ACM Transaction on Information Systems（2004）等。

很多知名高校成立了领先的推荐系统研究小组：包括明尼苏达大学、美国密歇根大学和卡内基梅隆大学等。其中密歇根大学更是在 2006 年开授了由 Resnick 主讲的推荐系统课程。

由知名机构开展了推荐算法设计的相关比赛：Netflix Prize 是由美国在线电影租租赁网站 Netflix 从 2006 年开设的推荐系统比赛，并同时公开了供比赛和相关研究使用的数据集。该公司为比赛设立了百万大奖鼓励参赛者设计出在公司已有的推荐算法基础上精度提高 10%以上的推荐策略；由 ACM 的数据挖掘及知识发现专委会主办的国际知识发现和数据挖掘竞赛（KDD-CUP）在近几年也专门针对个

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

性化推荐技术设立过比赛项目，获胜队伍被邀请在当年的 SIGKDD 会议上提交论文及相关技术报告[9-12]。

推荐技术是网络个性化时代的重要产物，随着网络应用和业务数量的与日俱增，包括电子商务、视频、音乐等范畴的网络业务运营商开始越发依赖于推荐技术，由此催生了推荐技术在实际业务系统中的广泛应用。表 1-1 中列举了当前常见的推荐系统应用。

表 1-1 常见的推荐系统

领域	推荐系统
电子商务	Amazon.com, eBay, dangdang.com, douban.com
音乐	Pandora, Lastfm, Ringo, CDNOW, 豆瓣
阅读社交	Facebook，人人网，sina 微博
阅读	Zite, Digg, Google News
视频电影	Youtube, Netflix, MovieLens

1、电子商务。亚马逊是将推荐系统应用到电子商务网站的鼻祖和成功典型。该网站的推荐系统采用一种基于商品到商品的协作过滤推荐算法，从海量商品中为用户筛选出量身定制的个性化商品推荐[13]，促进商品的打包销售。

2、个性化音乐网络电台。个性化网络电台 Pandora 的音乐推荐算法主要基于内容，音乐家和研究人员亲自试听上万首来自不同歌手的歌曲，然后对歌曲的不同特性进行标注，这些标注被成为基因，最后 Pandora 会根据专家标注的基因计算歌曲的相似度，并给用户推荐和他之前喜欢的音乐在基因上相似的其他音乐。

3、社交网站。知名社交网站 Facebook 曾推出一个称为 Instant Personalization 的推荐 API，该工具根据用户好友喜欢的信息，给用户推荐他们的好友最喜欢的物品。除了根据好友做物品推荐，社交网站还可以做信息流的会话推荐，以及给用户推荐好友。

4、个性化阅读。个性化阅读工具 Zite 允许用户给出喜欢或不喜欢的反馈，然后通过分析用户的反馈数据不停地更新用户的个性化文章列表，其推出后获得巨大成功，后被 CNN 收购。新闻阅读网站 Digg 在使用推荐系统后，用户的 dig 行为明显更加活跃，dig 总数提高 40%，用户的好友数增加了 24%，评论数增加了 11%。

5、电影和视频网站。Netflix 是个性化电影推荐的积极倡导者，该公司通过推荐系统的应用，60%的用户找到了自己感兴趣的电影和视频。明尼苏达大学开发的

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

电影推荐系统 MovieLens[14]是一个协作过滤推荐系统, 其通过搜集用户评分数据为用户建立推荐模型，同时，MovieLens 开发小组还将系统搜集数据整理成若干大小的数据集供研究人员使用。美国最大的视频网站 YouTube 在个性化推荐领域也进行了深入研究，其最新的论文[15]中显示他们正在使用的是基于物品的推荐算法。

1.3 课题主要研究内容

本课题研究旨在针对视频业务领域，针对协作过滤推荐算法中未考虑视频流行度和用户偏好变化的问题，以及数据稀疏性问题，在对现有方法进行分析研究的基础上，从用户兴趣度预测的 Top-N 推荐和用户评分预测推荐两个方面设计适用于视频业务推荐的算法模型。研究内容主要包括：

1、 基于视频流行度和用户偏好变化的协作过滤 Top-N 推荐

在视频推荐系统中，用户行为是推荐算法的基本推荐依据，累积的用户行为数据除了可以表达用户对视频节目的喜欢程度外，还可以体现某些统计意义。针对视频，丰富而不同的用户观看行为能够体现视频不同的流行程度，系统视频的流行度分布往往呈现出所谓的长尾分布；针对用户，积累的操作行为则可以通过时间轴信息体现出用户偏好的变化特征。由此，在利用用户行为关系进行推荐的传统协作过滤算法中引入以上两个元素，构建用户-视频二进制关联下的 Top-N 推荐算法模型，其推荐结果的质量得以有效提升。

2、 基于改进矩阵分解模型的协作过滤评分预测推荐

评分预测是推荐系统领域的另一研究热点，其产生推荐依据的是对用户未观看视频的预测评分值。基于矩阵分解模型的协作过滤算法具有易于实现和相对高效的特征，是一种能够有效缓解评分数据稀疏性问题的评分预测推荐算法，本文将在分析矩阵分解模型基础上，引入评分偏置项对基础模型进行改进，以得到更精确的评分预测值和推荐结果。

1.4 论文的组织结构

本文主要工作在于从视频推荐中的 Top-N 推荐和评分预测推荐两个方面，对利用用户行为信息的协作过滤视频推荐技术进行深入探讨和实验分析。本文的组织结构如下：

第一章主要介绍了视频推荐系统的研究背景和意义，综述了该领域的国内外

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

研究现状，对本文主要研究内容和方向进行概括后，说明了本文的主要工作和组织结构。

第二章先简单说明了推荐系统的几种定义及结构，接着，详细介绍了推荐系统的关键技术和推荐原理。然后介绍了推荐系统的常用实验方法和评测指标。最后简单说明了学术界常用的两个视频推荐实验数据集。

第三章从分析视频业务系统中的显性用户行为和隐性用户行为出发，提出了用户-视频二进制关联模型。在此用户行为信息模型下，建立了两种基于领域的协作过滤算法，并在基于用户的协作过滤中引入视频流行度和用户偏好变化权重因子，提出改进算法。通过实验，对比分析了该用户行为信息模型下，这几种算法的推荐性能。

第四章由评分数据稀疏性引出现有的针对该问题的评分预测推荐方案，重点阐述了基于 SVD 分解模型的和根据其发展而来的基于矩阵分解模型的协作过滤评分预测推荐方法。通过在矩阵分解模型中引入三个评分偏置项，提出对原算法的改进策略，得到了更准确的评分预测结果。

第五章是全文最后一个章节，总结了全文研究的目的和内容，以及主要的创新和取得的效果，同时根据领域的研究现状展望了进一步研究的内容及方向。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第二章 视频推荐系统及其相关技术

2.1 推荐系统概述

Hal R. Varian 和 Paul Resnick 在文章 Recommender Systems 中给出了对推荐系统的文字化定义：“ 个人需要在缺乏主观能动性或者足够领域知识的情况下做出选择时，主要会参考相关物品的评价，或者依赖于相关人士口头或者书面的推荐。推荐系统正是为应对这种情况而设计的，它是能够为用户的选择做出个性化参考的软件系统[16] ”。

除了上述文字化定义，Alexander Tuzhilin 和 Gediminas Adomavicius 于 2005 年在 IEEE Computer Society 上发表的 Toward the next generation of recommender systems 一文中对推荐系统进行了公式化定义[17]：O为系统所有对象（object）的集合，U为系统所有用户（user）的集合，效用函数$p(\ )$用来预测用户u对对象o的兴趣值，即对象o对用户u的效用，如果用R表示推荐结果集合，则可以形式化表达为：$p:U\times O\to R$。推荐的任务就是针对系统用户$u\in U$，找到能够最大化效用函数的系统资源对象$o^{'}\in O$。公式化表达为：

$$\forall u\in U,\ \ o_u^{'}=\underset{o\in I}{\operatorname{argmax}}\ p(u,o) \tag{2-1}$$

那么，就视频推荐系统而言，视频节目资源v（video）成为了系统的推荐对象，根据上述推荐系统的公式化定义可以对视频推荐系统做出如下公式化定义：

$$\forall u\in U,\ \ v_u^{'}=\underset{v\in V}{\operatorname{argmax}}\ p(u,v) \tag{2-2}$$

视频推荐系统的总体任务就是通过观测用户对系统视频资源行为从而获取用户行为数据，并结合用户和资源特性分析用户偏好，以此为基础预测用户对未产生行为的视频资源的兴趣大小，最终形成推荐。

以视频推荐系统为代表的推荐技术在学术界和信息产业界都得到广泛的研究和应用，尤其是在网络平台上，各大互联网公司都纷纷开发了自己的推荐系统，例如 Amazon，Baidu，Douban 和 Youku 等。事实上，有很多理由可以让服务提供商们开发如上的推荐系统，这些缘由组成了推荐系统的主要作用：

- 增加物品的销售量/业务使用量

这是一个商业推荐系统最重要的功能，相对于在没有任何推荐的情况下，它

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

能有助于销售额外的物品。能够取得这样的效果是因为推荐的资源迎合用户的需要或兴趣，而对于用户来说，推荐是由系统自动完成，并不需要用户做过多的操作，当用户接受了一个好的推荐结果后，自然可能会购买或使用系统提供的业务。

- 有助于发掘系统长尾资源

推荐系统的另一个重要功能就是能够让用户能够选择到在没有优质推荐的情况下难以找到的资源，例如 Netflix 这样的视频推荐系统，通常系统资源的大多数都不是最流行的，即长尾资源，本文后续章节会进一步介绍。服务提供者致力于推销各种类型的 DVD，而不只限于最流行的资源，Netflix 通过推荐系统能够把非热门的电影推荐给恰当的用户[18]。

- 提升用户满意度和忠诚度

一个设计良好的推荐系统能够让用户感受到系统推荐给他们带来的便利和趣味，并且，准确有效的推荐和实用的人机交互能够提升用户对系统的满意度；推荐系统不断累积用户行为信息，分析用户兴趣偏好，从而为用户计算推荐，系统会记录下用户的每次访问从而更新用户行为数据，用户使用系统的时长越长，推荐结果也越能迎合用户偏好，由此提升用户对系统的忠诚度。

2.2 推荐系统结构

一般的，个性化推荐系统至少包括用户输入模块，用户数据处理模块，推荐生成模块和推荐输出模块。图 2-1 展示了一个典型的个性化推荐系统的框架[19]，该框架的主要功能模块包括：

1） 用户行为日志数据库：记录了系统通过用户-系统交互接口采集到用户历史行为信息，这些信息在电子商务系统中，包括用户商品访问，商品购买和页面点击等；在视频业务系统中，包括用户对视频的评分，用户观看视频时长和视频分享等。

2） 数据挖掘引擎：从大量的用户行为数据中提取用户偏好模型可能用到数据挖掘技术，例如基于关联规则的推荐技术需要用到关联规则挖掘技术处理原始用户行为数据得到用户偏好特征。

3） 资源数据库：记录了系统资源的内容等相关信息，推荐系统提取这些信息，按对应的策略建立资源特征模型。根据推荐系统的不同，资源可能为视频，文本，商品，音乐的等。

4） 个性化推荐引擎：推荐引擎为推荐系统的核心模块，之前的数据准备都会

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

作为推荐引擎的数据输入，然后不同的推荐引擎根据不同的推荐策略生成推荐结果，传送给用户-系统交互接口。

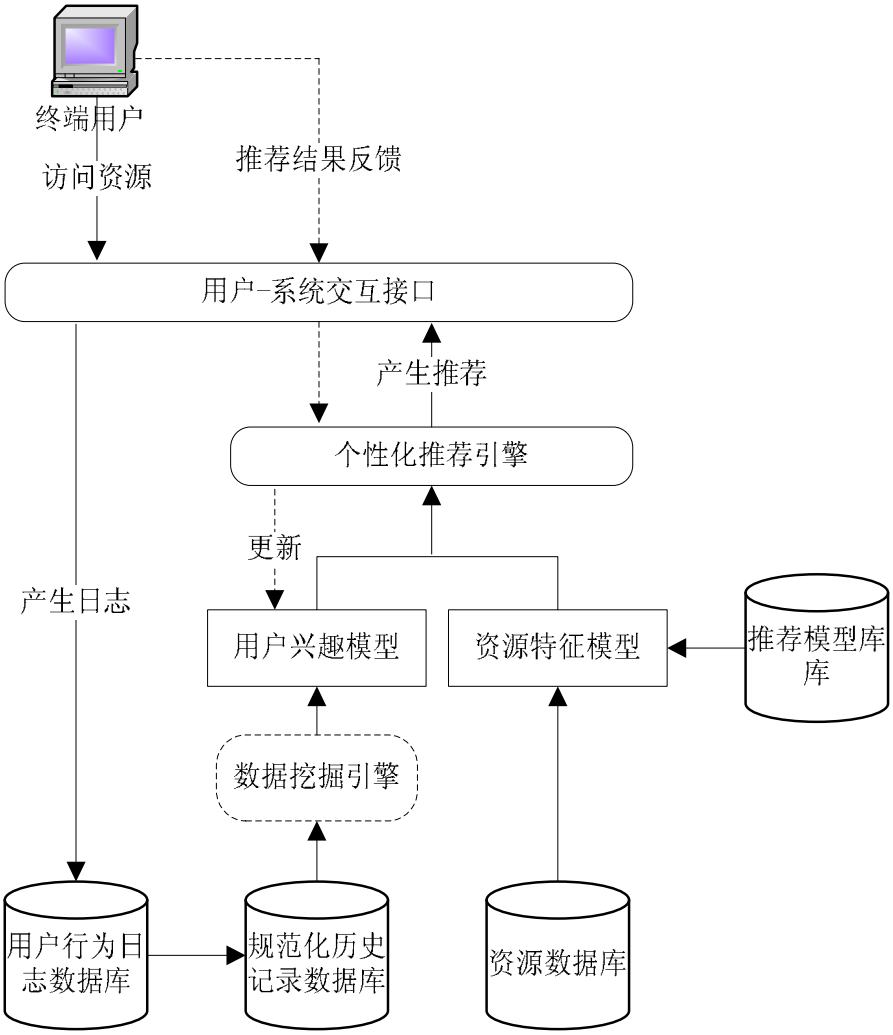

图 2-1 个性化推荐系统框架

该推荐系统推荐生成流程如图 2-1 中实线箭头所示：终端用户访问系统资源，系统通过用户-系统交互接口搜集用户行为信息后生成行为日志，并存储在用户行为日志数据库中，该数据库的行为经过统一的处理后形成规范化用户历史行为记录。系统以数据挖掘等技术从资源数据库和规范化历史记录数据库中生成用户兴趣模型和资源特征模型。然后，推荐引擎根据模型数据利用相关推荐算法生成对目标用户的推荐结果，最后，推荐结果传送到用户-系统交互接口，通过终端界面显示给用户。

图 2-1 中的虚线箭头则表示推荐反馈的流程：用户接收到推荐结果后，一般会对推荐结果有所反馈，以示用户对推荐结果的满意度，用户-系统交互接口采集用

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

户反馈信息并将其传递给推荐引擎，然后推荐引擎根据该信息和相关策略更新用户兴趣模型。

国内外学者通过近年来的研究，也提出了一些视频推荐系统的框架模型。比如图 2-2 展示的一种多模式带反馈的视频推荐模型[20-22]，该模型通过融合了文本、视觉和听觉信息，把视频节目资源 I 表示为：

$$I=(I_T,I_V,I_A,\omega_T,\omega_V,\omega_A)$$

其中，$I_i(i\in\{T,V,A\})$ 为各模式自身的相关性度量；$(\omega_T,\omega_V,\omega_A)$ 为模式间的相关性权重。

该模型中，把文本、视觉和听觉的模式相关性融合为单一相关性的线性加权和。并且，系统的用户反馈可以使相关性加权系数得到动态修正。

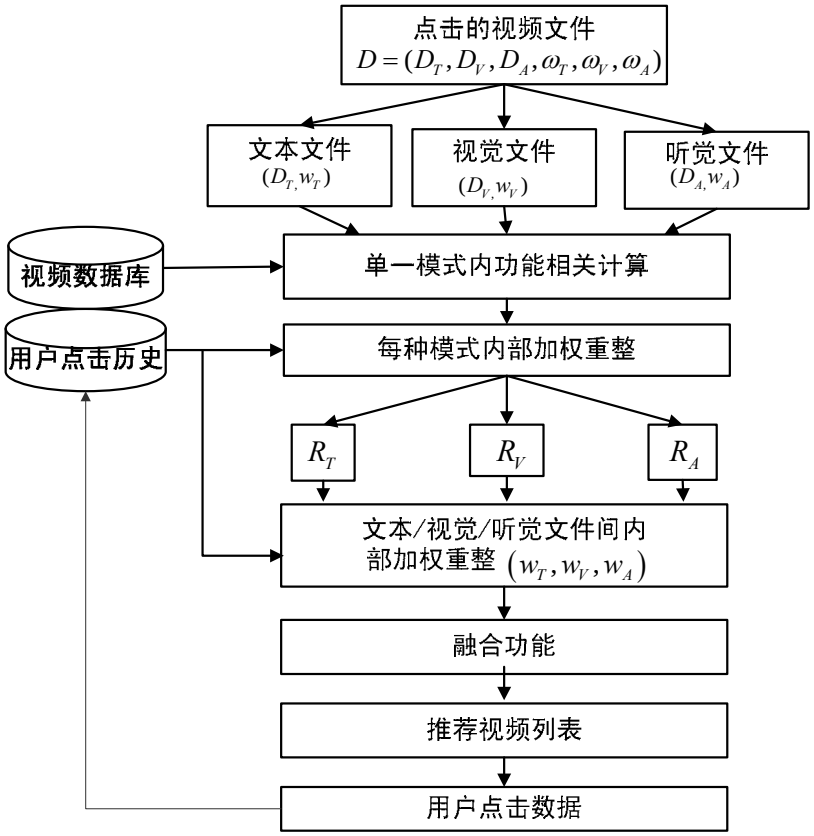

图 2-2 带反馈的多模式视频推荐模型

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.3 推荐系统常用技术

2.3.1 基于人口统计学的推荐

人口统计学（Demographic）特征包括了用户的性别、年龄、身高、学历等信息。通过这些每个人都具有的信息，可以发现用户间最基本的相似性，从而为预测用户兴趣提供依据。基于人口统计学的推荐技术单纯分析人口统计学信息，找到具有相似人口统计学特征的用户群体，形成用户分类，然后为用户推荐其所在类别中其他用户喜欢的资源，推荐原理如图 2-3 所示。

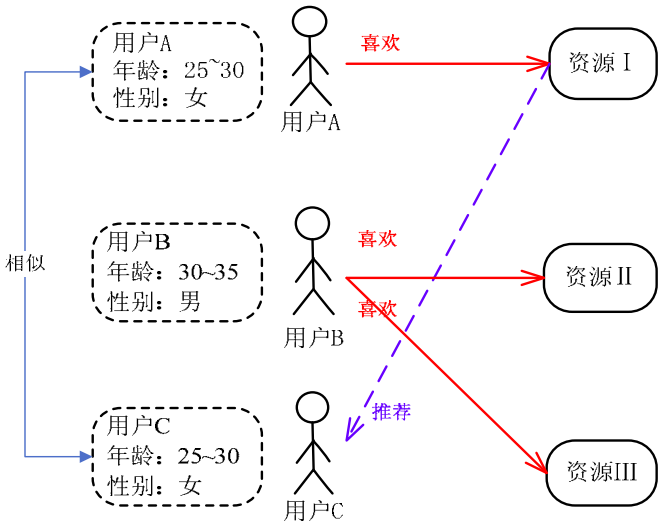

图 2-3 基于人口统计学的推荐原理图

上图中，系统为每个用户维护一个人口统计学信息样本，包括年龄、性别等。系统会根据该模型寻找信息相似用户，如上图，用户 A 和用户 C 具有相似的人口统计学信息，系统将把用户 A 和用户 C 视为相似用户，在推荐领域中也可以称之为“ 邻居 ”。系统根据邻居具有相似的兴趣偏好这一基本原则，为目标用户推荐其邻居喜欢的资源。正如上图中，系统将用户 C 的邻居用户 A 喜欢的资源 I 推荐给了用户 C。

基于人口统计学的推荐具有以下优点：①推荐算法思想简单，易于实现。②不需要用户对资源的评分等反馈数据，即不使用用户行为数据，因此，对于系统新用户不存在“ 冷启动 ”问题，当他们第一次访问系统时就能得到推荐。③不涉及资源本身的特征，因此该技术是领域无关的，即适用于各个领域的推荐系统。

但是该推荐技术也存在明显的缺点。最大的缺点就是，推荐基本不具有个性

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

化特征，基于人口统计学信息的用户相似粒度太粗，导致无法满足用户的个性化需求[23-24]，如图中具有人口统计学特征的用户将得到基本相同的推荐，推荐系统不具有明显的个性化特征。另外，搜集并统计用户的人口统计学信息，必然涉及到用户个人的隐私安全问题，正常情况下用户都不愿过多暴露真实的个人信息。

2.3.2 基于内容的推荐

基于内容的推荐技术是最早应用的推荐技术之一，该推荐技术的核心思想是为用户推荐那些与其之前喜欢的资源在内容上相似的资源。该技术针对资源类型选择相应的技术分析资源内容，然后根据用户已经观看过的资源的内容特征得到用户偏好概貌（Preference Profile），最后推荐给用户那些内容上与其偏好概貌相似的资源。图 2-4 展示了电影推荐系统中基于内容的推荐的基本原理。

系统先对影片特征进行建模，以一种统一的形式表达资源特征，该例子中以电影的类型为资源特征。接着，根据用户的对电影的观看历史分析取得用户偏好概貌，针对用户 A，系统检测到该用户对电影 I 的观看记录，以“爱情、动画”作为该用户的偏好概貌；同时，电影III的内容特征与用户 A 偏好概貌最为相似。由此，该系统将电影III推荐给用户 A。

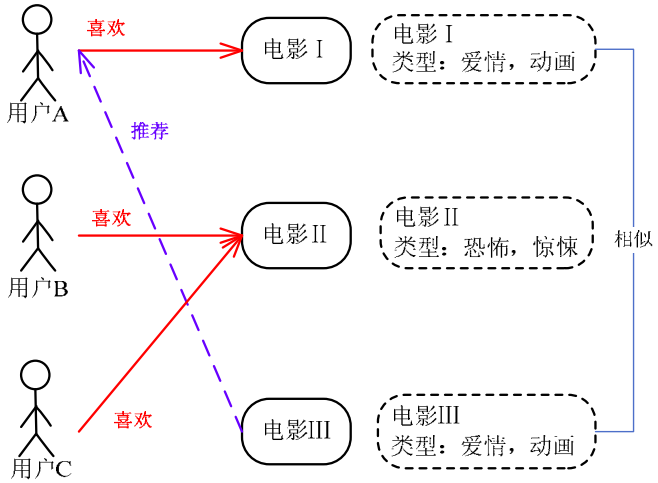

图 2-4 基于内容的推荐原理图

当然，实际的推荐系统中，为了达到理想的推荐准确度，电影的特征模型和用户偏好概貌设计会比较多样。根据上述电影推荐系统的例子，可以看出基于内容的推荐通常包括以下几个步骤：

1） 内容特征模型建立。每个新资源加入推荐系统时，系统都会按照统一的方

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

式建立资源的内容特征模型信息。

2） 用户偏好概貌获取。一般用户都不会只浏览某一个资源，推荐系统会分析用户历史浏览资源的内容特征，从而获取用户偏好信息，并以统一的形式组织成用户偏好概貌。除了系统分析用户历史行为记录外，系统也可以设计让用户主动提出个人偏好信息的辅助功能。

3） 相似度计算。推荐系统会计算目标用户未浏览资源的内容特征与该用户的偏好概貌之间的相似程度。

4） 生成推荐结果。按照上一步中得到的相似度对目标用户待浏览资源进行排序，然后将相似度最高的若干个资源推荐给用户。

5） 反馈处理。推荐系统还可以采集用户对推荐资源的反馈情况，从而分析用户对推荐资源的兴趣程度，在此基础上，调整优化用户偏好概貌信息，提高推荐的效率和质量。

2.3.3 基于关联规则的推荐

关联规则挖掘技术[25]用于发掘数据集中不同项目之间潜在的关联，它是数据挖掘领域中最为重要的分析模型[26]。该技术旨在发现数据集合中存在“ 并发关系 ”的资源项目，这种关系即成为关联。用关联规则挖掘理论的经典例子“ 尿布与啤酒 ”可以对该技术原理做初步说明：国外某大型超市多次在统计消费者购物清单后发现，看似没有任何关系的尿布和啤酒一起出现的频率却很高，于是超市售货员把它们摆放在一起，促进两种商品的销售。究其原因，可能是因为小孩的父亲在出门外婴儿买尿布的同时随便买点啤酒。本例中，超市利用“ 尿布与啤酒 ”这一关联规则，为父亲们提供了便利或者一种购买暗示，也为自己增加了盈利。

以集合为表达形式可以形式化地描述关联规则挖掘问题[27]。设定项目（Item）集合为$I=\{i_1,i_2.....i_m\}$，事务集合为$T=(t_1,t_2,...t_n)$，事务集合中的事务$t_i\left(i=[1,\cdots,n]\right)$是由不同项目组成的集合，满足$t_i\in I$。则关联规则可以描述为：

$A\to B$，其中$A\subset I$，$B\subset I$，且$A\bigcap B=\varnothing$。

A和B均为项目集合，简称项集，A为规则前件，B为规则后件，通过A可以推出B。同时，必须在满足一定的支持度[28]（Support）和置信度（confidence）的条件下，关联规则才成立。

如果项集X是事务t_i的一个子集，那么用$num(X)$代表事务集合T中包含了X的事务个数。假设n是T中事务的数目，则关联规则理论中的支持度和置信度可以

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

做如下定义：

$$\text{支持度}=\frac{num(A \cup B)}{n} \tag{2-3}$$

$$\text{置信度}=\frac{num(A \cup B)}{num(A)} \tag{2-4}$$

支持度是指“ T 中包含了 X 的事务个数占 T 中事务总数的比例 ”，可以认为是事务集合中 A 和 B 同时出现的概率。它是评价关联规则的重要指标，支持度很小，说明对应的规则只是偶然现象，失去现实参考意义。置信度是指“ 同时包含 A 和 B 的事务的数量占包含 A （规则前件）的事务的比例 ”，可以认为是在 A 出现的情况下，同时出现 B 的概率。假设关联规则 $A \rightarrow B$ 达到一定支持度情况下，置信度很低，说明事务集合中包含 A 的事务数目很大，所以在事务集合中同时出现 A 和 B 的概率并不可靠，关联规则难以让人信服。

总之，关联规则挖掘的任务是在设定最小支持度和置信度的要求后，从事务集合中，找出所有满足要求下，项集 A 推出项集 B 的关联规则。

把关联规则挖掘技术结合到推荐技术中，产生了基于关联规则的推荐技术。该技术[29]根据用户历史行为信息与关联规则库，向目标用户生成个性化推荐结果。基于关联规则的个性化推荐技术的具体实施通常包括以下几个步骤：

1） 构建资源事务集合。系统根据用户行为记录日志，为每个用户创建其使用资源事务记录，形成系统事务集合。

2） 提取关联规则，形成关联规则库。推荐系统使用某种关联规则挖掘算法[30-31]对系统事务集合进行关联规则挖掘，得到满足最小支持度和置信度的所有关联，形成关联规则库。

3） 初始化候选集。为目标用户初始化一个为空的推荐资源候选集合。

4） 匹配关联规则库。系统通过遍历关联规则库，找到属于用户历史使用资源的关联规则前件，即是说找出的关联规则中左边部分出现在目标用户已使用资源集合中。

5） 规则后件加入候选推荐集。将找出的所有关联规则的后件中全部资源加入到目标用户候选推荐集合中。

6） 候选推荐集过滤。首先，从候选推荐资源集合中删除目标用户已使用的资源项目。然后，根据关联规则的置信度对候选集中所有的候选资源项目排序，如果一个资源在多组关联规则后件中出现，选择这些关联规则中置信度最高的资源

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

项目参与排序。最后，从候选资源排序结果中选取置信度最高的 N 个资源作为该目标用户的推荐结果。

基于关联规则的推荐技术同利用人口统计学的推荐技术一样，推荐算法不涉及资源本身的内容信息，因此适用于各个领域的推荐系统。更重要的是，该基于关联规则的推荐是一种个性化推荐技术，它能够利用用户的历史偏好信息，发掘资源间的关联特征，这种关联并不是基于资源内容的，推荐结果可能与用户过去喜欢的资源内容相似，也可能不相似。所以，它能有效地为用户发掘新的兴趣点。

然后，基于关联规则的推荐技术也存在难以保障推荐质量的问题：该技术主要依赖于规则的质量，支持度和置信度等规则的评判标准若选取不恰当，都会影响推荐的准确性；规则的挖掘都是依赖与用户历史行为，若系统中某个资源没有用户使用过，或者是新加入的，则该资源无法推荐给任何用户。

2.3.4 基于协作过滤的推荐

现实生活中，人们往往会参照与自己兴趣相似的人的喜好来对商品或资源做选择，因为与自己兴趣相似的人喜欢的物品自己也很有可能会喜欢。协作过滤推荐技术（Collaborative Filtering Recommendation）就是基于这一思想的个性化推荐技术。所谓“ 协作 ”，就是利用用户与用户之间的行为异同，以相似关系为基础向系统用户进行推荐。

协作过滤技术作为推荐系统领域最为著名和最为重要的技术，不管是在学术界还是在信息产业界都得到了很多关注与研究。Paul Resnick 等人于 1994 年在 ACM Conference on Computer Supported Collaborative Work 上提出了最早的协作过滤推荐技术—基于用户的协作过滤算法[32]（User-CF），该技术被成功应用于 GroupLens 和 Tapestry 等推荐系统中。亚马逊公司团队于 2000 年在 IEEE Internet Computing 上发表的论文中首次提出了基于项目的协作过滤技术[33]（Item-CF）。两种算法的基本思想如图 2-5 所示, 基于用户的协作过滤推荐为目标用户寻找相似用户，进而为其推荐相似用户喜欢的资源；基于项目的协作过滤推荐为目标用户喜欢的资源寻找相似资源（基于用户行为的资源相似性，而不是资源本身属性的相似），进而为其推荐这些相似资源。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

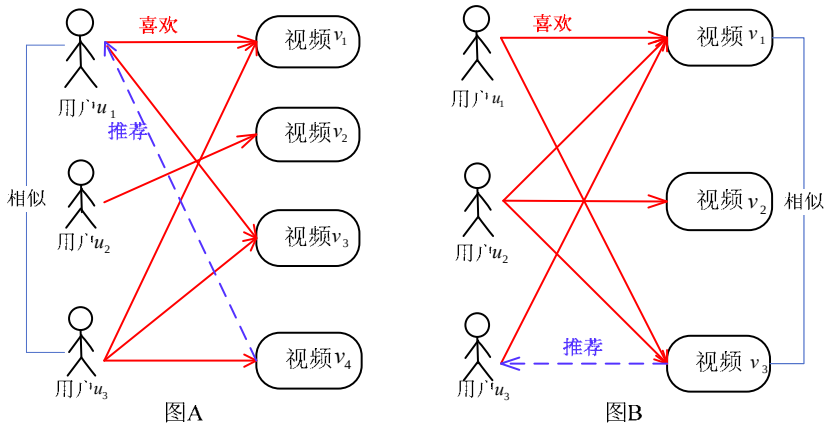

图 2-5 User-CF 和 Item-CF 的推荐原理示意图

本节将主要介绍三种传统的协作过滤推荐算法：User-CF、Item-CF 和基于聚类的协作过滤推荐。它们都是应用于评分预测推荐的算法。前两种算法思想类似，可统称为基于领域的协作过滤算法。

（1） 基于用户的协作过滤推荐技术

传统的基于用户的协作过滤推荐算法（User-CF）是以用户对项目的评分为数据基础建立用户偏好模型，以兴趣相似程度度量用户相似关系，找到目标用户的“邻居”用户，进而通过邻居的历史评分数据为目标用户生成推荐。可以看出，User-CF 主要有三个步骤：数据生成，邻居搜寻和推荐产生。

1、 数据生成

就视频推荐而言，输入数据通常为个 n 用户对 m 个资源的评分数据，系统整理这些评分数据到一个 $n \times m$ 的矩阵中，形成“用户-视频”评分矩阵 R。如表 2-1 所示。

表 2-1 协作过滤视频推荐评分矩阵示例

用户/视频	视频 v_1	视频 v_2	…	视频 v_m
用户 u_1	r_{11}	$\varnothing$	…	r_{1m}
用户 u_2	$\varnothing$	r_{22}	…	r_{2m}
…	…	…	…	…
用户 u_n	r_{n1}	r_{n2}	…	r_{nm}

2、 邻居搜寻

首先，利用相似度函数计算用户行为的相似性，然后，根据相似度大小进行

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

排序，选择相似度最大的 K 个用户作为目标用户的最近邻居集。相似度计算是邻居搜寻的关键环节，通常有两种方法度量用户相似性[34]：

- 余弦相似性（Cosine）

把用户评分矩阵看作 n 个 m 维的用户评分向量。假设两个用户 u_1 和 u_2 的评分向量用 a 和 b 表示，余弦相似性指的是两个用户评分向量的空间夹角的余弦值大小，函数值越大，说明两向量夹角越小，相似度就越大。余弦相似度计算函数如公式 2-5 所示。其中 r_{1i} 和 r_{2i} 表示用户 u_1 和用户 u_2 对视频 i 的评分值。

$$sim(u_1,u_2)=\cos(a,b)=\frac{a\cdot b}{\|a\|\times\|b\|}=\frac{\sum_i^m r_{1i}r_{2i}}{\sqrt{\sum_i^m r_{1i}^2}\sqrt{\sum_i^m r_{2i}^2}} \tag{2-5}$$

- 相关相似性

用户相似度还可以通过 Pearson 相关系数来度量。通过余弦相似度函数得到的相关性不会因为用户评分向量在空间中的距离长短而发生变化，Pearson 相关系数通过引入单个用户对所有项目评分的平均值来对克服这一缺陷。其计算公式如下。

$$sim(u_1,u_2)=\frac{\sum_i^m\left(r_{1i}-\overline{r_1}\right)\left(r_{2i}-\overline{r_2}\right)}{\sqrt{\sum_i^m\left(r_{1i}-\overline{r_1}\right)^2}\sqrt{\sum_i^m\left(r_{2i}-\overline{r_2}\right)^2}} \tag{2-6}$$

3、 推荐产生

确定目标用户的最近邻居集后，使用评分预测函数，根据邻居集中用户的历史评分和与目标用户的相似度大小预测目标用户对未评分视频的评分数值，最后选择预测评分值最高的 N 个项目推荐给用户。假设有某个未评分视频 j，传统协作过滤推荐算法的预测评分，是由目标用户 u_1 评分均值加上邻居集中用户评分与各自评分均值之差的相似度加权和而得到的。公式如下，其中 S 代表用户 u_1 的最近邻居集，c 代表 u_1 的邻居用户。

$$p_{1j}=\overline{r_1}+\frac{\sum_{c\in S}sim(u_1,c)\cdot\left(r_{cj}-\overline{r_c}\right)}{\sum_{c\in S}\left|sim(u_1,c)\right|} \tag{2-7}$$

（2） 基于项目的协作过滤推荐技术

Item-CF 技术存在如下假设：1）项目相似指的是很多用户对这些项目的评分相似；2）同一用户对相似项目会有相似的兴趣。

Item-CF 同样以用户-资源评分矩阵为基础，不同于 User-CF 推荐方法的是：它

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

并不是是寻找目标用户本身的最近邻居集合，而寻找的是目标项的最近邻居集合，并根据用户对相似项目的已有评分预测该用户对目标项的评分。它们的本质区别在于，预测某用户对一个项目的评分时，User-CF 参照的是与该用户相似的多个用户对这个项目的已有评分值；Item-CF 参照的是该用户对与这个项目相似的多个项目的已有评分值。

这里提到的项目之间的“相似”与前文中基于内容的推荐中提到的资源内容与用户偏好概貌“相似”有所不同。他们的相似性计算的数据来源就不同，前者是利用由用户行为数据得到的用户-资源评分矩阵，后者利用的是资源本身的内容特征和由此建立的用户偏好概貌信息；前者是基于用户的评分行为度量项目的相似程度，后者则是基于资源本身的属性度量用户偏好概貌与其他资源的相似程度。

由此，Item-CF 推荐技术的关键环节从 User-CF 的用户之间相似计算转变为了项目之间的相似性度量，但计算过程都利用了用户历史行为。其主要步骤类似于 User-CF，包括邻居寻找和推荐生成。

邻居寻找：遍历系统项目集合，先计算每一个项目于其他项目的相似度，然后对相似度进行排序，选择每个项目对应的相似度最大的 K 个项目作为该项目的最近邻居集合。与用户相似度计算类似，项目相似度也有两种常用的计算方法。以视频推荐为例，把用户评分矩阵看作 m 个 n 维列向量，即项目评分向量，令视频 v_1 和视频 v_2 的评分向量为 x 和 y。则两种计算公式为：

$$sim\left(v_1,v_2\right)=\cos\left(x,y\right)=\frac{\sum_i^n r_{i1}r_{i2}}{\sqrt{\sum_i^n r_{i1}^2}\sqrt{\sum_i^n r_{i2}^2}} \tag{2-8}$$

$$sim\left(v_1,v_2\right)=\frac{\sum_i^n\left(r_{i1}-\overline{r_1}\right)\left(r_{i2}-\overline{r_2}\right)}{\sqrt{\sum_i^n\left(r_{i1}-\overline{r_1}\right)^2}\sqrt{\sum_i^n\left(r_{i2}-\overline{r_2}\right)^2}} \tag{2-9}$$

推荐生成：同样的，针对某目标用户，预测起对为评分资源的评分数值，然后把评分数值最高的 N 个推荐给用户。设用户 u_1 的某个未评分视频为 j，其预测评分由用户 u_1 对视频 j 的邻居视频的评分的相似度加权和而得到。公式如下，其中 $S^{'}$ 代表视频 j 的最近邻居集，z 代表视频 j 的邻居视频。

$$p_{1j}=\frac{\sum_{z\in S^{'}}sim\left(j,z\right)\cdot r_{1z}}{\sum_{z\in S^{'}}\left|sim\left(j,z\right)\right|} \tag{2-10}$$

（3） 基于聚类的协作过滤推荐技术

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

聚类分析技术是数据挖掘和图像处理等相关领域的基础技术，聚类分析的任务是把大量的原始数据按照某种方法分成若干类\簇，类中的数据间相似度较大，类与类之间的数据相似度很小。即是把相似度较大的数据“聚集”在一起，使零散的数据变成多个类的数据。

用O表示一组样本数据集合，s表示样本之间的相似度，聚类分析的输入一般为有序对(O,s)，输出是对样本数据的聚类结果，即为：

$C=\{C_1,C_2,.....C_k\}$

其中，$C_i(i=1,2,.....,k)$包含于样本集O，且满足如下条件：

- $C_1 \cup C_2 \cup \cdots \cup C_k = X$
- $C_i \cap C_j = \varnothing, i \neq j$

聚类分析在很多领域都得到了应用，并积累了不少经典算法。文献[35]对聚类分析的思想进行了系统化分类，包括划分法、基于密度的方法、基于网格的方法和基于模型的方法[35]。基于聚类的协作过滤推荐技术是将聚类分析应用于协作过滤推荐流程中，具体方式是在搜索近邻之前对初始数据进行聚类，把邻居搜索范围缩小在各个聚类中，以提高近邻搜索的速度和准确性。

Chee 等人在 2000 年提出一种基于用户聚类的协作过滤推荐方法[36-37]。该方法首先按照用户历史行为信息把系统用户划分为若干不同的聚类，划分结果要求同一类中用户对各个项目的评分具有较高相似度，而不同类中用户对项目的评分相似度较低。然后，系统通过分析每个聚类中用户评分的总体特征，选取一个合适的用户代表该类中用户对项目的评分偏好。最后，系统在代表用户集合的评分数据基础上进行 User-CF 推荐。该方法缩小了用户的最近邻居搜索范围，提高了协作过滤的实时推荐速度。Chee 等人在上述基于用户聚类推荐中使用 K-means 聚类算法对用户进行聚类。该聚类算法的具体步骤如下：

1）随机选择 K 个用户作为初始用户节点，将聚类开始时的初始聚类中心定为这些初始用户节点。2）计算系统中其余所有用户与这 K 个中心用户的相似度，将其余每个用户各自聚合到相似度最高的某个中心用户对应的类中。3）对已生成的 K 个聚类，计算每个类中用户对项目的平均评分向量，并找出类中与该评分向量最接近用户评分向量对应的用户，然后将新的聚类中心定为这 K 个用户。4）重复步骤 2 和步骤 3，直到每个聚类不再发生变化为止。

该聚类算法能有效的进行用户聚类，但是算法也存在如下缺点：聚类数目需要事先给定，但往往实际中聚类数目是难以估计的；该算法需要根据初始聚类中心开始聚类划分，初始聚类中心的选取不同可能最终的聚类结果不同，若初始中

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

心的选取不当可能会得不到理想的聚类效果。

除了基于用户聚类的协作过滤外，Li 等人提出了基于项目的协作过滤推荐方法[38]。该方法则是对项目进行聚类，然后在每个聚类中找到项目的最近邻居集合。这样，有效的缩小了项目最近邻居的寻找范围，提高了推荐算法性能。

2.3.5 推荐技术比较

以上介绍了四种传统的推荐技术，本节将根据这几种推荐技术的推荐原理，对他们各自存在的优缺点加以对比说明 :①基于人口统计学的推荐单纯利用个人的基本信息产生对用户进行推荐，这样的方法设计简单，但是满足不了当今网络中用户对个性化的需求。用户需要推荐系统能智能的识别自己的兴趣偏好，为自己推荐满意的资源 ;②基于内容的推荐技术会分析资源本身的属性特征，推荐的结果与用户使用过的资源是内容相似的，这样迎合了用户已表现出来的兴趣偏好，但是无法为用户发掘新的兴趣点，不能给用户带来惊喜。而且，针对不同资源可能需要设计不同的资源属性特征表示模型 ;③基于关联规则的推荐技术通过分析用户行为历史，挖掘资源之间的关联关系，得到的推荐资源在内容上可能与用户已使用资源是很不一样的，这样的推荐可以为用户发掘新的兴趣点，但是推荐过程极大依赖于规则的质量。同时，规则的挖掘过程会多次遍历资源库，算法开销很大。

基于协作过滤的推荐技术是推荐领域内应用最广泛的推荐技术，它能协作配合利用用户历史行为信息，寻找用户或者资源的相似关系，基于这样的相似关系为用户生成推荐，User-CF 和 Item-CF 是协作过滤推荐的基础算法，以下通过场合、实时性等方面对比了这两个算法。

- 场合：过多的用户或者资源数量会增加 User-CF 或者 Item-CF 的相似度计算开销，所以在 Item-CF 更适用于物品数目明显小于用户数的场合，User-CF 适用于用户较少的场合；在资源更新频繁的系统中，资源相似度也需要频繁更新，维护难度大，而用户相似度对资源更新不那么敏感，所以 User-CF 更适用于资源实时性要求高的场合。
- 冷启动：User-CF 和 Item-CF 是利用用户-资源评分矩阵中的用户评分向量或项目评分向量来展开推荐过程，新用户或者新物品没有任何观看记录和评分行为，所以它们都无法对新用户做出推荐，也不能把新资源作为推荐结果。
- 推荐理由：Item-CF 容易给用户做出让人信服的推荐解释，比如系统给用户推荐视频甲的推荐解释可以是因为用户之前喜欢视频乙。视频甲和乙是相似视

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

频，Item-CF 才做出了这样的推荐。而 User-CF 很难提供令用户信服的推荐解释。

2.4 推荐技术评测

一种推荐技术的质量和性能需要有相关指标对推荐结果进行评测，对推荐技术评价指标的研究也是推荐系统领域的重要理论部分。推荐技术的评测指标通常分为评分预测评测指标和 Top-N 推荐评测指标两类，常用的评测指标有：准确度、覆盖率和新颖性等。

◆ 评分预测评测指标

传统的协作过滤推荐是基于评分预测的推荐，为用户推荐那些评分预测值较高的项目。因此，对评分预测的准确程度的衡量就是对评分预测推荐质量的评测。平均绝对偏差（Mean Absolute Error，简称 MAE）是评价评分预测准确度的一个经典方法[39-40]。

$$MAE = \frac{1}{c}\sum_{\alpha=1}^{c}\left|v_{i\alpha} - r_{i\alpha}\right| \tag{2-11}$$

其中，c 为测试集中目标用户 i 的评分项目个数，$r_{i\alpha}$ 为测试集中用户的实际评分，$v_{i\alpha}$ 为系统的预测评分。该公式计算的是系统预测评分与用户实际评分之间的偏差程度。系统用户的 MAE 平均值就可以表示系统的评分预测准确度。与平均绝对偏差类似，均方根误差（Root-Mean Square Error，简称 RMSE）也是重要评分预测准确度指标。

$$RMSE = \sqrt{\frac{1}{c}\sum_{(i,\alpha)}\left|v_{i\alpha} - r_{i\alpha}\right|^2} \tag{2-12}$$

该公式在在线视频公司 Netflix 举办的 Netflix Prize 比赛中得到了广泛采用，并使其成为了评分预测推荐系统的主要评测指标。

◆ Top-N 推荐评测指标

Top-N 推荐是不同于评分预测的推荐方式，它的目标是通过分析用户的历史行为数据找到用户未使用过的资源中用户兴趣度预测值最高的前 N 个资源，推荐给该用户。亚马逊科学家 Linden 于 2009 年在 Communications of the ACM 网站上发表的一篇文章[41]中，指出 Top-N 推荐可能比预测评分（用预测的评分表示用户兴趣）更具有现实意义，因为电影推荐的目的是找到用户最有可能感兴趣的电影，而不是预测用户看了电影后会给电影做什么样的评分，也许有一部电影用户看了

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

之后会给出很高的评分，但是用户看的可能性却非常小。

这类推荐系统的推荐精度可以准确率（Precision）和召回率（Recall）进行评测：

$$\text{Precision} = \frac{\sum_u |R_u \text{ I } T_u|}{\sum_u |R_u|} \tag{2-13}$$

$$\text{Recall} = \frac{\sum_u |R_u \text{ I } T_u|}{\sum_u |T_u|} \tag{2-14}$$

其中，R_u 代表对用户 u 的推荐资源集合，T_u 代表出现在测试集中用户 u 喜欢的项目集合。它们类似于信息检索领域的查全率和查准率，准确率指的是有效推荐资源（有效推荐资源指的是系统为某用户推荐的资源中的，用户在测试集上曾喜欢的资源）的数量占实际推荐数量的比例，而召回率指的是有效推荐数量占实际感兴趣项目的比例，它们共同描述了 Top-N 推荐的准确程度。

覆盖率（Coverage）指标描述了 Top-N 推荐发掘长尾资源的能力。高覆盖率的推荐系统能够保证系统中相对冷门的资源也能推荐给合适的用户，一方面，这些用户可能对某些冷门资源很感兴趣，另一方面系统的资源也能得到全面调度。覆盖率指标度量了推荐系统中有多大比例的资源推荐给了系统用户。覆盖率定义如下，其中，R_u 代表对用户 u 的推荐结果集合，U 为用户集合，V 为资源集合。

$$\text{Coverage} = \frac{|\text{U}_{u \in U} R_u|}{|V|} \tag{2-15}$$

推荐系统除了关心推荐的资源是否符合用户兴趣偏好外，也需要向用户推荐一些并不那么符合用户已有偏好特征的新颖资源。当有这样的推荐产生时，用户会惊喜地发现推荐系统为自己找到了新的兴趣点。推荐技术的新颖性正是描述推荐系统这一能力的指标。只是新颖性没有固定的公式化表达形式，其评测方法要视具体的推荐系统可定，同时也可以通过用户调查，通过用户反馈信息来评测推荐结果的新颖性。

2.5 推荐系统实验和常用数据集

2.5.1 推荐系统实验方法

推荐系统中，主要有三种评测推荐效果的实验方法，包括离线实验（offline

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

experiment)、用户调查（user study）和在线实验（online experiment)。

1、离线实验

离线实验的方法一般有如下几个步骤构成：

1）通过日志系统获得用户行为数据，并按照一定格式生成一个标准数据集；

2）按照一定的规则将数据集分成训练集和测试集；

3）在训练集上训练用户兴趣模型，生成推荐；

4）通过事先定义的离线指标评测算法在测试集上的评测推荐结果。

从上面的步骤可以看到，推荐系统的离线实验都是在一个从实际系统日志中提取的数据集上完成的。这种实验方法的好处是不需要真实用户参与，可以直接快速地计算出推荐结果，从而方便、快速地测试不同的推荐算法。同时，它的主要缺点是无法获得商业上关注的指标，如点击率、转化率等，找到和商业指标非常相关的离线指标也是很困难的事情。表 2-2 简单总结了离线实验的优缺点。

表 2-2 离线实验的优缺点

优 点	缺 点
不需要有对实际系统的控制权	无法计算商业上关心的指标
不需要用户参与实验	离线实验的指标和商业指标存在差距
速度快，可以测试大量算法	

2、用户调查

商业上的推荐系统，在上线测试前一般需要做一次称为用户调查的测试。用户调查需要一些真实的用户，让他们在需要测试的推荐系统上完成一些任务，并观察和记录他们完成任务时的行为，然后让他们回答一份设计好的调查问卷。最后，通过分析真实用户的行为和调查问卷来了解推荐系统的性能。

用户调查的主要作用是，弥补离线实验无法评测与用户主观感受相关的性能指标这一缺陷，比如用户满意度和用户信任度。但用户调查只能选择一些有代表性的用户，因此会使其实验结果的统计意义不足。

3、在线实验

商业化的推荐系统在最终上线运营之前，需要把推荐系统上线，做最后的在线实验。AB 测试是一种常用的在线评测实验方法。它将用户随机分成几组，并对不同组的用户采用不同的算法，然后通过统计不同组内用户的各项评测指标来比较不同算法。AB 测试的缺点是周期比较长，必须进行较长时间的实验才能得到可靠的结果。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.5.2 实验数据集

推荐算法实验所采用的数据集一般都是从实际业务系统中积累起来的，使得推荐技术学术研究尽量与行业实际情况接轨。目前已有不少实际的业务系统公开了自己采集的数据集供研究者们使用。下面简单介绍两个本文后续章节的实验中采用的数据集，它们在视频推荐技术研究中也最为常用。

（1） Netflix 电影评分数据集

Netflix 公司在 2005 年底开启 Netflix Prize 比赛的同时，公布了用于比赛和学术研究的 Netflix 电影评分数据集。该数据集包含了 Netflix 系统从 1999 开始的 6 年间记录的近 50 万用户对 17000 多部电影的评分数据。系统使用 5 分评分制，评分值可为：1、2、3、4 和 5。数字越大则代表用户对电影评价越高。Netflix 电影评分数据集包含了一下几个文件：

- MOVIE TITLE：包含数据集中所有影片的一些基本信息，如影片名称，上映时间等。
- TRAINING DATASET：作为提供给参赛者或者研究者进行推荐算法实验的训练集，包含了用户 ID、电影 ID 和用户评分等关键信息。
- QUALIFYING DATASET：未包含评分的测试数据集合，提供给参赛者作为其设计系统的评分预测对象。
- PROBE DATASET：上个数据文件的包含用户真实评分的版本，以让学术研究者对评分预测结果进行准确度评测。

（2） MovieLens 电影评分数据集

电影推荐系统 MovieLens 由 Minnesota 大学开发。MovieLens 电影评分数据集包括了该电影推荐系统中用户对电影的评分等相关数据，评分值同样为 1 到 5 的整数，该数据集按数据规模的不同，存在 3 个版本：

- MovieLens 10M：规模最大的版本，记录了 72000 个用户对 10000 部电影的一千万条评分及相关数据，总大小为 10M 左右。
- MovieLens 1M：中等规模版本，记录了 6000 个用户对 4000 部电影的一百万条评分及相关数据，总大小为 1M 左右。
- MovieLens 100K：规模较小的版本，记录了约 1000 个用户对近 1700 部电影的十万条评分及相关数据，总大小超过 100K。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.6 本章小结

视频推荐是推荐系统技术在视频业务领域的重要应用。本章从推荐系统的定义和作用开始，对推荐系统典型框架和常用技术做了详细阐述和应用对比分析。简单介绍了推荐技术的主要实验方法和评测指标后，对视频推荐系统实验常用的数据集做了简单说明。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第三章 基于视频流行度和用户偏好变化的协作过滤算法

在视频业务系统中，不断积累的用户行为数据并不是随机的，而是蕴含着很多模式和规律。传统的协作过滤技术片面地考虑用户对视频的显式打分行为，以评分数据作为用户偏好特征来对用户做出推荐，忽略了用户行为数据中体现出的视频流行程度，用户活跃程度和用户偏好随时间变化等信息和特征，传统推荐技术主要的缺陷在于以下两点：一是缺少对于用户观看视频行为的挖掘，仅依赖于用户主动的打分行为，没有考虑用户观看视频时长、次数和观看时间等信息；二是忽略了视频节目流行程度的不同和用户偏好随时间的变化等客观事实，其推荐质量存在很大的提升空间。由此，本章将在充分挖掘用户行为的条件下，提出应用于 Top-N 推荐的，基于视频流行度和用户偏好变化的协作过滤推荐算法，并进行实验分析。

3.1 用户偏好的获取与建模

3.1.1 用户行为信息分类和表示

用户行为数据在网站上最简单的存在形式就是日志，推荐系统和视频业务网站会汇总原始日志生成描述用户行为的会话日志，并将他们存储在分布式数据仓库中。这些日志记录了用户的各种行为，如在电子商务网站中这些行为主要包括网页浏览、商品购买、点击、评分和评论等。

用户行为在个性化视频推荐系统中一般分为两种：

- 显式反馈行为（explicit feedback）
- 隐式反馈行为（implicit feedback）

显式反馈行为包括用户主动明确表示对视频节目喜欢的行为。比如亚马逊和豆瓣等系统使用 5 分的评分系统来让用户直接表达对物品的喜好，优酷和土豆等系统则使用简单的“喜欢”或者“不喜欢”按钮获取用户的兴趣。

和显式反馈行为相对应的是隐式反馈行为。隐式反馈行为指的那些并非用户主动评价的不能直接反应用户喜好的行为，在视频业务系统中，主要包括如下三种隐式反馈信息：用户观看视频时长、观看某个视频的次数和观看视频的时间。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

相对于显式反馈，隐式反馈虽然不直接表明用户偏好，不容易区分正回馈（喜欢）还是负反馈（不喜欢），但数据量大，蕴含信息丰富。某些系统中，可能有些用户没有主动去评分，则没有显式反馈数据，系统只能通过其隐式反馈数据来挖掘用户偏好。表 3-1 从几个不同方面比较了显式反馈数据和隐式反馈数据。

表 3-1　显式反馈数据和隐式反馈数据的比较

	显式反馈信息	隐式反馈信息
用户兴趣	明确	不明确
数量	较少	庞大
存储	数据库	分布式文件系统
正负反馈	都有	只有正反馈

在视频推荐系统中，用户行为种类各异，经过对用户行为的归纳分析，本文给出一种用户行为统一表达方式的定义。

定义 3-1：（用户行为）视频推荐系统中，用户行为指的是 6 个方面的内容，包括产生行为的用户、用户行为对象、行为的种类、产生行为的上下文和行为的内容。

该定义可以由表 3-2 表示。

表 3-2 用户行为的统一表示

user id	产生行为的用户的唯一标识
item id	产生行为的对象的唯一标识
behavior type	行为的种类（比如是评分还是观看次数）
context	产生行为的上下文，包括时间或者地点等
behavior content	行为的内容（比如评分多少，观看时长等）

3.1.2 用户偏好变化的度量方式

用户行为反应了用户的偏好，丰富的用户行为信息除了直接体现出用户兴趣外，还包含着用户兴趣变化的信息。图 3-1 是通过 Google Insights 得到的两个关键词（Madonna，Adele）自 2005 年以来的搜索频率曲线，从图中可以看到，Adele 的搜索量从 2011 年开始从接近于零的状态直线飙升，在 2012 年到达顶峰后开始下降，而 Madonna 的搜索量在 2006 年就达到高峰，并也在 2012 年再次达到高峰，但增长趋势低于 Adele。这些变化的产生都源于用户兴趣随时间的变化。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

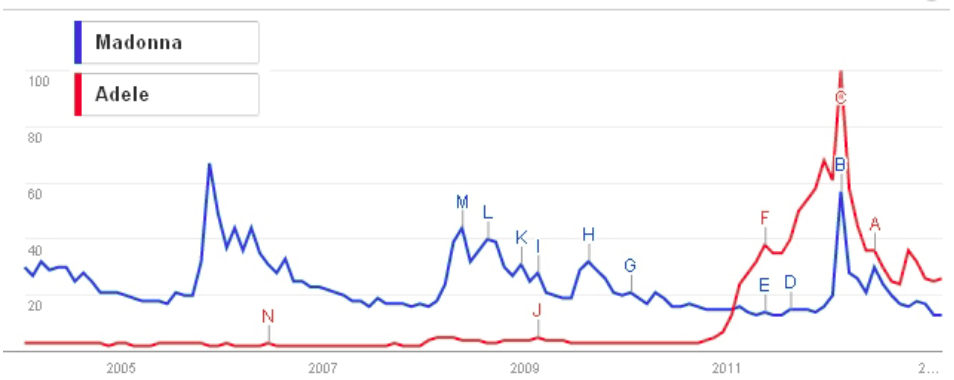

图 3-1 Madonna 和 Adele 的搜索变化曲线（截取自 Google Insights）

时间信息对用户兴趣的影响是视频推荐领域的研究热点，并取得了一些研究成果。比如 Lu 修改了传统的协作过滤算法，将用户和物品的特征向量看作一个随时间变化的动态特征向量，从而将时间作为新的一维建模到推荐模型中去[42]。Ding 在文献[43]中提出，用户在未来的兴趣主要受他近期兴趣的影响，所以设计的推荐算法加重了用户近期行为对最终结果的影响[43]。但是以上对时间信息的处理都相对片面，对它们对兴趣变化产生的客观原因阐述不明。基于此，本文提出两种基于时间信息的用户偏好变化度量方式。

- 基于横向用户时间跨度的用户偏好度量

正如前面提到，就单个用户而言，用户近期观看过的视频节目对推荐给该用户未来可能感兴趣的节目起着比较重要的作用，而早期的行为记录对推荐的影响相对较小，所以在预测用户对为观看节目的兴趣时，要削减用户早期行为对用户未来兴趣值的影响，从而提高用户最近行为的作用。

- 基于纵向用户时间跨度的用户偏好度量

除了考虑到个体用户的兴趣变化外，不同用户行为之间的时间跨度同样体现了用户兴趣的变化特性。两个用户兴趣相似是因为他们喜欢相同的节目，或者对相同的节目产生过行为。但是，用户产生行为的时间是不一样的，比如用户 A 在 2008 对节目 i 感兴趣，在 2010 年对节目 j 感兴趣，用户 B 在 2008 年对节目 j 感兴趣，在 2010 年对节目 i 感兴趣，而用户 C 和用户 A 一样，在 2008 对节目 i 感兴趣，在 2010 年对节目 j 感兴趣。那么，如果忽略时间因素对兴趣的影响，用户 A 和用户 B 的兴趣相似度等于用户 A 和用户 C 的兴趣相似度。但显然在现实条件下，人们会认为用户 A 和用户 C 的兴趣相似度大于用户 A 和用户 B 的兴趣相似度，因为用户 A-B 观看同一节目的时间跨度大于用户 A-C 观看同一节目的时间跨度。由

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

此，在计算用户兴趣相似度时，要削减不同用户间有较大时间跨度的相似行为对用户相似度的影响，提高为目标用户推荐的准确度。

两种基于时间跨度的用户偏好变化的度量方式分别应用于预测用户对未产生行为的视频节目的兴趣值和用户之间的兴趣相似度中，具体实施方式会在本文后续的章节中作进一步说明。

3.1.3 用户偏好模型的建立

用户偏好模型的表示与建立决定了是否能真实客观的反应用户的兴趣偏好，直接影响到系统最终的推荐结果，因此在推荐系统中，建立适合相应推荐算法的、客观准确的用户偏好模型是推荐过程中必不可少的环节。在视频业务系统中，用户的行为信息一般包括用户对视频的评分（5 分制评分）、用户观看视频的时长和用户观看视频的次数等，传统的协作过滤算法只利用了用户显示的评分数据来建立用户偏好空间向量模型，而忽略了用户的隐式反馈行为。基于此，本文提出一种综合利用用户显式和隐式反馈行为信息的用户偏好空间向量模型建立算法，算法最终形成用户-节目二进制关联矩阵模型。算法流程如图 3-2 所示。

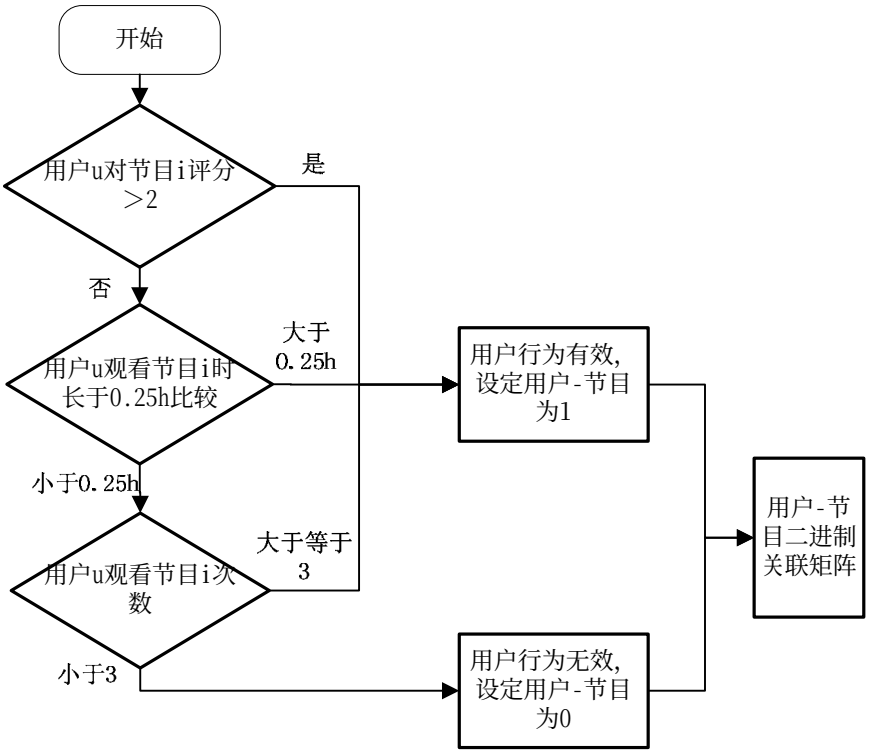

图 3-2 混合反馈信息的用户偏好向量模型建立算法流程图

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

算法描述如下：

输入：结构化用户行为信息（包括用户显示评分、用户观看节时长和观看节目次数）

输出：用户-节目二进制关联矩阵

算法步骤：

第一步 处理视频系统通过 5 分制评分方法搜集到的用户评分信息，用户 u 对节目 i 打分记为 $score_{ui}$，当 $score_{ui} \geq 3$ 时，称用户对该节目的行为有效，并设定用户-节目关联 $b_{ui}=1$，然后停止处理该用户对节目 i 的行为数据处理；否则继续处理其他行为数据。

第二步 处理用户观看时长数据，用户 u 对节目 i 观看时长记为 $length_{ui}$，当 $length_{ui} \geq 0.25h$ 时，称用户对该节目行为有效，并设定 $b_{ui}=1$，然后停止处理该用户其他数据行为；否则继续处理其他行为数据。

第三步 处理用户观看次数数据，用户 u 对节目 i 观看时长记为 $freq_{ui}$，当 $freq_{ui} \geq 3$ 时，称用户对该节目行为有效，并设定 $b_{ui}=1$；否则 $b_{ui}=0$。

第四步 处理完所有用户行为数据后，把 m 个用户偏好向量组成用户-节目二进制关联矩阵 $B(m,n)$，表达形式如下：

$$B(m,n)=\begin{bmatrix} b_{11} & \cdots\cdots & b_{1k} & \cdots\cdots & b_{1n} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ b_{s1} & \cdots\cdots & b_{sk} & \cdots\cdots & b_{sn} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ b_{m1} & \cdots\cdots & b_{mk} & \cdots\cdots & b_{mn} \end{bmatrix} \tag{3-1}$$

该矩阵包括了 m 个用户对 n 个视频节目的二进制关联数据，行对应用户，列对应视频节目。

3.2 视频流行度和用户活跃度

很多关于互联网数据的研究发现，互联网上的很多数据分布都满足幂律分布（Power Laws），这个分布在很多领域也称为长尾分布。如果随机变量或者序数 x 的出现频次

$$f(x)=cx^{-a} \tag{3-2}$$

则该变量服从长尾分布，如图 3-3 所示。如果对幂律分布的公式两遍取对数，

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

可以得到

$$\log f(x) = \log c + (-a)\log x \tag{3-3}$$

这说明x的频次分布在对数坐标上是一条直线, 这可以用来判断一个分布是否服从幂律分布。长尾分布其实很早就被统计学家注意到了。1932 年，哈佛大学的语言学家 Zipf 在研究英文单词出现频率时，发现如果把单词出现的频率按由大到小的顺序排列，则每个单词出现的频率与它的排列名次的常熟次幂存在简单的反比关系，这个分布称为 Zipf 分布。这个现象表明，只有很少数的词被经常使用，如图 3-3 中的“ 长尾 ”部分，而绝大多数词很少被使用，如上图中的“ 主体 ”部分。

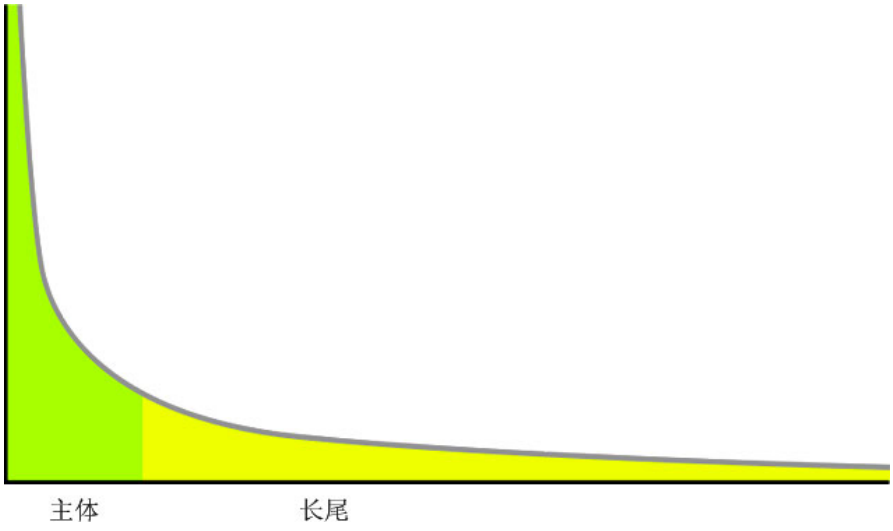

图 3-3 长尾分布模型图

很多研究人员发现，用户行为数据也蕴含着这种规律。首先，这里给出本文中的视频流行度和用户活跃度的定义：

定义 3-2：（视频流行度）对某视频产生过有效行为的用户总数为该视频的流行度。

定义 3-3：（用户活跃度）某用户产生过有效行为的视频总数为该用户的活跃度。

根据定义，两者都可以从用户行为数据中得到：基于本文的用户-节目二进制关联矩阵模型，视频流行度就是矩阵中某视频对应列中“ 1 ”的个数，即对该视频产生过有效行为的用户总数；用户活跃度就是矩阵中某用户对应行中“ 1 ”的个数，即该用户发生过有效行为的视频总数。研究表明，视频业务系统中，视频流行度和用户活跃度同样服从长尾分布。如图 3-4，是针对某视频网站部分视频的流行度

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分析图，该图是将视频播放数的频次分布画在对数坐标上。

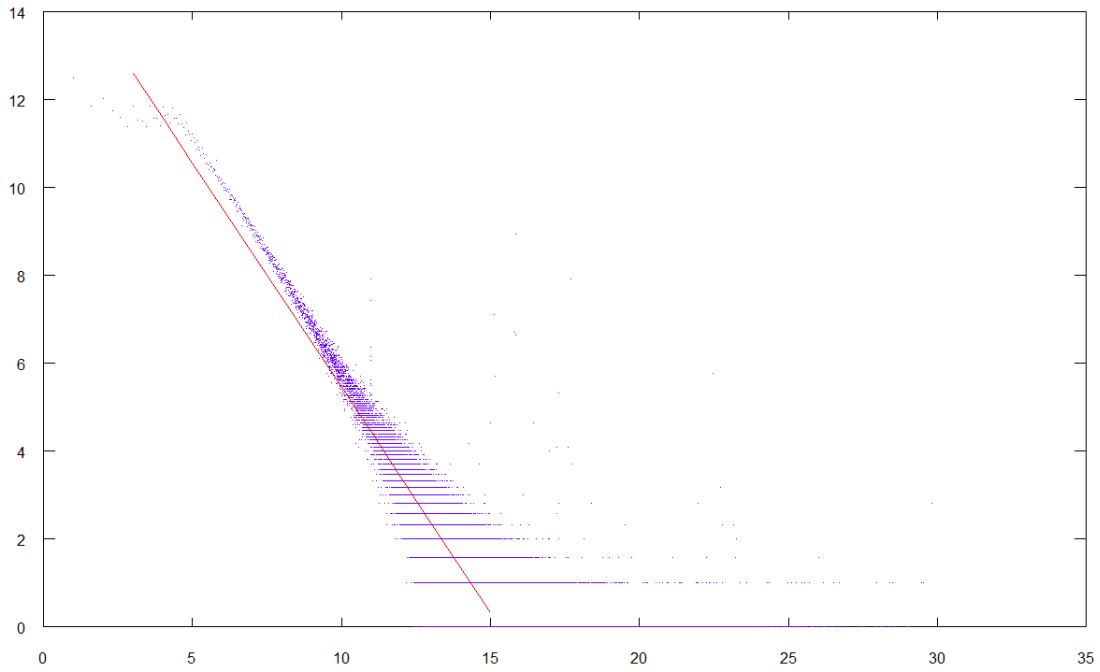

图 3-4 视频流行度的长尾分布

可以看出，除了头部和尾部的少数噪声数据外，中间绝大部分排成形成一条直线排列，这说明视频业务系统中的视频播放次数也就是视频流行度是近似服从长尾分布的：热门的视频非常少，越冷门的视频数量反而越多。另外，对上图中的点分布稍加说明：图上靠下部分很多点分布在了几条横线上，比如纵坐标 1 对应的那条横线，这表明有很多大约在横坐标对应的 2^13 到 2^18 之间的播放数刚好出现了两次，在对数坐标上显示出来就是一条横线了；图中的点越到下面发散，说明只出现几次的播放数，随机性比较大，容易偏离直线比较远。

既然视频的流行度和用户的活跃度各异，并且蕴含着分布规律，那么在视频推荐算法中考虑它们对推荐质量的影响，是客观并有助于提高推荐质量的。

3.3 Top-N 推荐中基于领域的协作过滤算法

基于领域的协作过滤算法是业界应用最广泛的推荐算法，基于领域的算法分为两大类，一类是基于用户的协作过滤算法，另一类是基于物品的协作过滤算法。本章将提出应用于 Top-N 推荐的，基于领域的协作过滤视频推荐算法。算法是将用户行为数据调入内存，通过计算搜寻一组和目标用户或者物品相似的邻居组成

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

最近邻集合，然后根据这组邻居的行为数据预测目标用户对未产生有效行为项目的兴趣度，根据得到的兴趣度大小生成 Top-N 推荐。

3.3.1 基于用户的协作过滤

在个性化推荐系统中，当一个用户 A 需要个性化推荐时，可以先找到和他有相似兴趣的其他用户，然后把那些用户喜欢的、而用户 A 没有听说过的物品推荐给 A。这种方法称为基于用户的协作过滤算法。举个例子，每年新学期开始，刚进实验室的师弟总会问师兄如“我应该买什么书”、“我应该看什么论文”等问题，师兄们一般会给他们做出一些推荐，师弟可能会请教很多师兄，然后做出最终的判断。例子中，师弟之所以请教师兄，一方面因为互相认识且信任对方，但主要的原因是师兄和师弟有共同的研究领域和兴趣。

从上面的例子中可以看出，视频推荐系统中，基于用户的协作过滤算法主要步骤如图 3-5 所示，首先通过计算用户相似度找到和目标用户兴趣相似的用户集合，然后通过预测用户兴趣值，找到相似用户集合中用户喜欢的、且目标用户没有产生有效行为的视频推荐给目标用户。

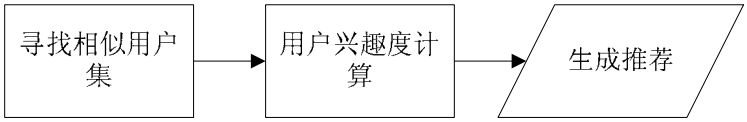

图 3-5 基于用户的协作过滤算法主要步骤

1）基于用户偏好向量的相似度计算方法包括主要有以下几种：欧几里得距离，余弦相似度和 Pearson 相关系数。计算方法如表 3-3 所示。

表 3-3 相似度计算方法

度量方法	计算公式
欧几里得距离	$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
余弦相似度	$c(x,y)=\frac{x\mathrm{g}y}{\|x\|\times\|y\|}=\frac{\sum_{i=1}^{n}x_iy_i}{\sqrt{\sum_{i=1}^{n}x_i^2}\sqrt{\sum_{i=1}^{n}y_i^2}}$

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Pearson 相关系数	$p(x,y)=\frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-(\sum_{i=1}^{n}y_i)^2}}$ （n 为向量 x 和 y 的维度）

本文研究在用户-视频二进制关联矩阵模型下基于用户的协作过滤算法，简称为 BinaryUserCF。用户相似度的计算采用余弦相似度的计算方法，用户 u 和用户 v 的相似度计算公式如下：

$$sim_{uv}=\frac{\left|I(u)\cap I(v)\right|}{\sqrt{\left|I(u)\right|\left|I(v)\right|}} \tag{3-4}$$

其中，$I(u)$ 和 $I(v)$ 表示用户 u 和用户 v 有过有效行为的视频集合。例如由表 3-4 给出的行为数据示例。

表 3-4 用户-视频二进制关联数据示例

	视频 a	视频 b	视频 c	视频 d	视频 e
用户 A	1	1	0	1	0
用户 B	1	0	1	0	0
用户 C	0	1	0	0	1
用户 D	0	0	1	1	1

利用公式 3-3 计算用户 A 和用户 B 的兴趣相似度为：

$$sim_{AB}=\frac{\left|\{a,b,d\}\cap\{a,c\}\right|}{\sqrt{\left|\{a,b,d\}\right|\left|\{a,c\}\right|}}=\frac{1}{\sqrt{6}}$$

$$sim_{AD}=\frac{\left|\{a,b,d\}\cap\{c,d,e\}\right|}{\sqrt{\left|\{a,b,d\}\right|\left|\{c,d,e\}\right|}}=\frac{1}{3}$$

2）得到用户之间的兴趣相似度后，选取 K 个与目标用户相似度最大的用户作为最近邻居集合，然后预测目标用户对没有过有效行为视频的兴趣大小，最后选取前 N 个兴趣值最大的视频形成推荐。预测公式如下：

$$p(u,j)=\sum_{v}sim_{uv}b_{vj},v\in S(u,K)\cap U(j) \tag{3-5}$$

其中，$S(u,K)$ 表示和用户 u 的 K 最近邻居集合，$U(j)$ 表示对视频 j 产生过有

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

效行为的用户集合，用户 v 是在用户 u 的最近邻居中对节目 j 产生过有效行为的用户。sim_{uv} 是用户 u 和用户 v 的兴趣相似度，b_{vj} 表示用户 v 对视频 j 的兴趣，因为本文采用的用户-视频二进制关联，所有的 $b_{vj}=1$。

根据表 3-4 的示例数据，以用户 A 作为目标用户，选取 $K=3$，用户 A 对视频 c、e 没有过有效行为，根据公式 3-5，用户 A 对视频 c、e 的兴趣值为：

$$p(A,e)=sim_{AC}+sim_{AD}=0.7416,\quad p(A,c)=sim_{AB}+sim_{AD}=0.7416$$

3.3.2 基于物品的协作过滤

基于物品的协作过滤推荐算法最初是由著名的电子商务公司亚马逊提出的，该算法给用户推荐那些和他们之前喜欢的物品相似的物品。不过该算法并不利用物品的内容属性计算物品之间的相似度，二是通过分析用户的行为信息来计算物品之间的相似度，比如在视频推荐系统中，视频 A 和视频 B 具有很大的相似度喜欢视频 A 的用户中很多也喜欢视频 B。

Top-N 推荐下基于物品的协作过滤推荐主要步骤如图 3-6 所示，首先通过计算视频相似度找到和对应视频相似的视频集合，然后通过预测用户兴趣值，找到相似视频集合中目标用户未来会产生兴趣，且没有产生过有效行为的视频推荐给目标用户。

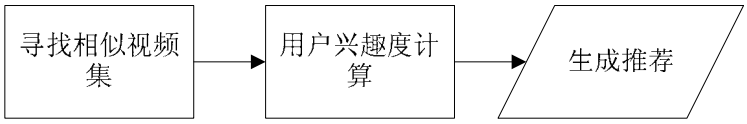

图 3-6 基于物品的协作过滤算法主要步骤

1） 本文中，在用户-视频二进制关联矩阵模型下，基于物品的协作过滤算法简称为 BinaryItemCF。视频相似度计算公式如下：

$$sim_{ij}=\frac{\left|U(i)\cap U(j)\right|}{\sqrt{\left|U(i)\right|\left|U(j)\right|}} \tag{3-6}$$

其中，$U(i)$ 和 $U(j)$ 表示对视频 i 和视频 j 有过有效行为的用户集合。

2）得到视频相似度后，选取 K 个与相应视频相似度最大的视频作为最近邻居集合，然后预测目标用户对没有过有效行为视频的兴趣大小，从而形成推荐。预测公式如下：

$$p(u,j)=\sum_{i} sim_{ij}b_{ui}, i\in S(j,K)\cap I(u) \tag{3-7}$$

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中，$S(j,K)$表示和视频j的K最近邻居集合，$I(u)$表示对用户u有过有效行为的视频集合，视频i是在用户u有过有效行为的视频中且属于视频j的K最近邻居的视频。sim_{ij}是视频i和视频j的相似度，b_{ui}表示用户u对视频i的兴趣，因为本文采用的用户-视频二进制关联，所有的$b_{ui}=1$。

3.3.3 基于视频流行度和用户偏好变化的协作过滤

根据本章前面的分析和讨论，可以发现在协作过滤推荐算法中加入视频流行度和用户兴趣变化的影响是客观必要的，并且可以预见它们提升推荐质量的可能性。本文将针对基于用户的协作过滤，提出基于视频流行度和用户偏好变化的协作过推荐算法，简称为 BUser-HTF，该算法将从用户相似度计算和用户兴趣预测两方面进行改进。

（1） 用户相似度的改进策略

- 视频流行度权重

传统的协作过滤算法中，所有视频都被同等地位的看待，没有考虑视频流行度对用户相似度的影响，不符合客观事实。比如，两用户都通过视频系统观看了电影《泰囧》，表面上他们行为相似，但是由于《泰囧》这部电影非常流行，这就像两个学生都购买了《新华词典》，由于物品的高流行度，用户对它们产生的相似行为几乎不能说明用户兴趣相似，而对相对冷门物品的相似行为更能说明用户兴趣的相似性。由此，在用户相似度计算中需要削弱对流行度偏高视频的用户相似行为的作用，于是设计视频流行度权重公式如下：

$$\omega_h=\frac{1}{\log(1+|U(i)|)},i\in I \tag{3-8}$$

其中I表示系统中所有视频的集合，$U(i)$表示对视频i有过有效行为的用户集合，根据本章对视频流行度的定义可知，$|U(i)|$就是视频流行度值。

- 纵向用户偏好变化权重

视频业务系统中，用户对视频的相似行为一般都不会发生在同一时刻，比如，用户 A ，用户 B 和用户 C 都观看过视频 a，用户 A 和用户 B 观看视频 a 的时间间隔为一周，而用户 A 和用户 C 观看时间间隔为一年。虽然三个用户都产生了相似行为，但是用户 A 和用户 C 行为时间间隔更长，用户 A 与用户 C 的兴趣相似度明显小于用户 A 与用户 B 的兴趣相似度。由此，在用户相似度计算中需要削弱较大时间跨度内的用户相似行为的作用，于是设计纵向用户偏好变化权重公式如下：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\omega_t = \frac{1}{\log\left(1+\alpha\left|t_{us}-t_{vs}+1\right|\right)}, s \in I(u) \cap I(v) \tag{3-9}$$

其中视频 s 表示用户 u 和用户 v 都有过有效行为的视频，t_{us} 和 t_{vs} 分别表示用户 u 和用户 v 对视频 s 产生有效行为的时间（如果对某视频产生过多次有效行为，取最近一次有效行为的时间），α（$0<\alpha<1$）为偏好变化影响系数，α 越大，则削减影响越大。

有了以上两个权重因子，可以得到本文改进算法中用户相似度的计算公式，如下：

$$sim_{uv} = \frac{\sum_{s} \omega_h \omega_t}{\sqrt{\left|I(u)\right|\left|I(v)\right|}}, s \in I(u) \cap I(v) \tag{3-10}$$

这样，在计算用户相似度时，通过以上两个权重的引入，同时削减了对高流行度视频和大时间跨度内的相似行为对用户相似度的贡献，使得用户相似度计算更准确。

（2） 用户兴趣预测的改进策略

- 横向用户偏好变化权重

从本章中的讨论和分析可以知道就单个用户而言，用户最近行为会比较长时间以前的行为对用户未来兴趣的影响更大，因此在根据目标用户最近邻居集内用户行为预测目标用户兴趣时，应削弱用户距离预测时刻较长时间以前行为对目标用户未来兴趣的影响，于是设计横向用户偏好变化权重公式如下：

$$\omega_t^{'} = \frac{1}{\log\left(1+\beta\left|t_0-t_{vj}+1\right|\right)} \tag{3-11}$$

其中，t_0 当前时刻，j 为目标用户未产生有效行为的视频，t_{vj} 为目标用户最近邻居集中对视频 j 产生有效行为的时间，β（$0<\beta<1$）为横向用户偏好变化影响系数，β 越大，则削减影响越大。

有了横向用户偏好变化权重因子，考虑和目标用户兴趣相似用户的最近兴趣，可以得到本文改进算法中用户兴趣预测公式，如下：

$$p(u,j) = \sum_{v} sim_{uv} b_{vj} \omega_t^{'}, v \in S(u,K) \cap U(j) \tag{3-12}$$

该改进的目标用户兴趣预测公式考虑了与目标用户相似用户兴趣的变化，是的兴趣预测更客观准确，有利于提升推荐质量。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

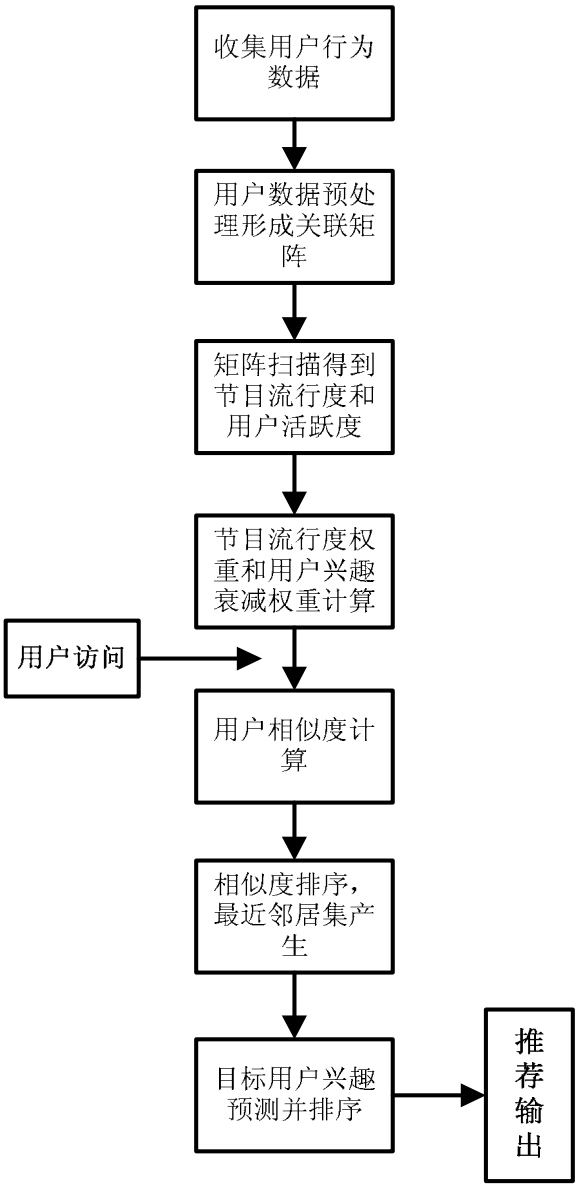

图 3-7 BUser-HTF 算法流程示意图

（3） 算法流程

改进算法的流程图如上图 3-7 所示。

流程描述为：

Step1 收集用户行为数据

Step2 处理用户行为数据，得到用户-视频二进制关联矩阵，实际系统中，可按照本章 3.1 中描述算法处理。

Step3 分别对矩阵进行列扫描和行扫描, 列扫描找到相应视频产生过有效行为

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的用户集合，从而得到视频流行度；行扫描找到相应用户有过有效行为的视频集合，从而得到用户活跃度。

Step4 计算视频流行度权重、纵向用户偏好变化权重和横向用户偏好变化权重。

Step5 计算用户相似度。

Step6 相似度排序，选取与目标用户相似度较大的 K 个用户组成最近邻集合。

Step7 预测目标用户对未产生行为视频的兴趣值，并排序形成推荐列表。

3.4 实验设计及结果分析

3.4.1 实验数据及分析流程

（1） 实验数据和预处理

本章实验中将采用美国 Minnesota 大学 GroupLens 研究小组提供的电影评分数据集 MovieLens 100K，该数据集包含了 943 个用户对 1682 部电影的 10 万条评分数据，每个用户至少对 20 部电影进行了评分、平均对 100 个资源进行了评分。该数据集的评分分值分为 5 个不同的等级：$[1,2,3,4,5]$，按照评分等级整理 10 万条评分数据，可以得到如图 3-8 所示的柱状图。可以看到评分值为 1 的评分项仅占全部评分项的 6.21%，大多数评分项都是 3-5 的评分值，占了超过总数的 80%。

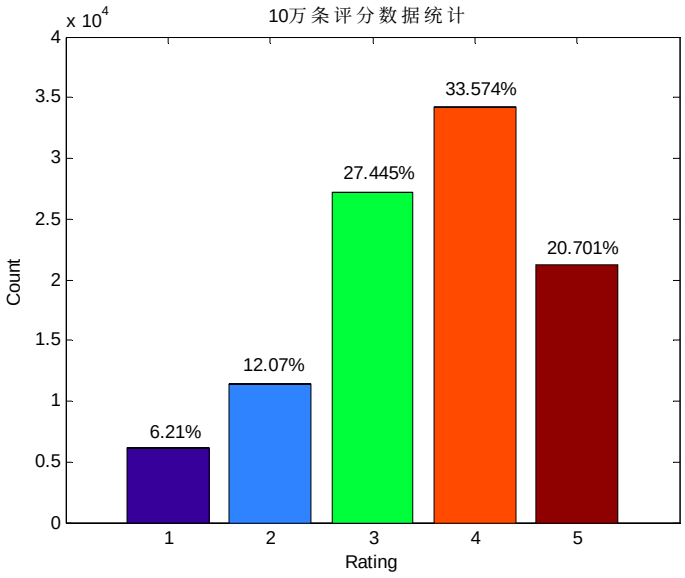

图 3-8 实验数据评分情况统计

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

这些评分数据包含四项：①用户 ID，②电影 ID，③评分值，④评分时间。如图 3-9 所示为整理得到的部分评分数据示例。

user_id	movie_id	rating	timestamp
186	302	3	1998-04-04 19:22:22
22	377	1	1997-11-07 07:18:36
305	451	3	1998-02-01 09:20:17
6	86	3	1997-12-31 21:16:53

图 3-9 MovieLens 数据集评分数据集示例图

要测试本章描述的推荐算法，需要对实验数据进行如下预处理：

- 将 MovieLens 100K 数据集均匀分成 M 份（本章实验取 M=8），挑选一份作为测试数据，形成文件 u.test；将剩下的 M-1 份组成训练数据，形成文件 u.base。
- 本章算法是使用用户-视频二进制关联数据做 Top-N 推荐，所以需要对数据集中的评分数据做一个转换处理：评分值大于等于 2 分的数据项视为用户对视频的行为有效，对应的用户-视频二进制关联规定为“1”；而没有评分值或者评分值为 1 分的数据项视为用户对视频的行为无效，对应的用户-视频二进制关联为“0”。由此，得到用户-视频二进制关联数据。
- 两种用户偏好变化权重计算公式中，用户偏好变化影响系数α和β越大，则权重因子对计算结果的影响越大。本实验中，设定它们的折中值：α=0.5，β=0.5。

（2） 实验分析流程

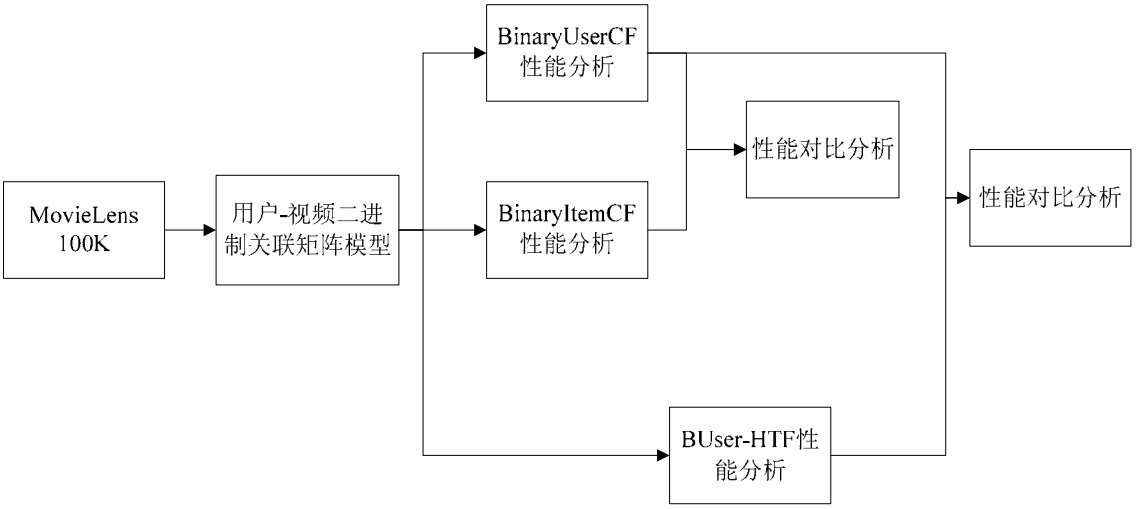

图 3-10 分析流程示意图

在接下来的小节中将根据图 3-10 的实验分析流程，以上述数据预处理为基础

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

进行实验分析，包括推荐的准确率、召回率等。为了避免选择测试数据的偶然性和某次实验结果的过拟合，每次使用不同的测试数据，进行 M 次实验，将实验测出的评测指标的平均值作为最终的评测指标。

3.4.2 实验评测指标

介绍评测指标前，先对两个数据做符号定义：

- I：数据集中所有的视频集合。
- U ：数据集中所有的用户集合。
- $R(u)$：对用户 u 推荐的 N 个视频集合。
- $T(u)$：用户 u 在测试集上有过有效行为的视频集合。

本章实验将从准确率、召回率、覆盖度和流行度四个方面评测本章推荐算法的性能：

◆ 准确率

预测准确率度量一个推荐算法预测用户行为的能力，它无疑是最重要的推荐算法评测指标，任何推荐系统的评测都少不了准确率，本章实验也不例外。推荐结果的准确率定义为：

$$\text{Precision}=\frac{\sum_{u}\left|R(u)\cap T(u)\right|}{\sum_{u}\left|R(u)\right|},\quad u\in U \tag{3-13}$$

准确率描述的是：最终的推荐列表中有多少比例是发生过的用户-视频有效行为记录，考察推荐列表的准确程度。

◆ 召回率

$$\text{Recall}=\frac{\sum_{u}\left|R(u)\cap T(u)\right|}{\sum_{u}\left|T(u)\right|},\quad u\in U \tag{3-14}$$

召回率描述的是：测试数据中有多少比例的用户-视频有效行为记录包含在最终的推荐列表中，描述了推荐列表反映用户实际兴趣的程度，考察推荐列表反映用户兴趣的完整性。

◆ 覆盖率

覆盖率反映了推荐算法挖掘长尾的能力，覆盖率越高，说明推荐算法越能够将长尾中的物品推荐给用户。覆盖率定义为：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\text{Coverage}=\frac{\left|\bigcup_{u\in U} R(u)\right|}{|I|} \tag{3-15}$$

该覆盖率表示最终的推荐列表中包含系统中多大比例的视频。如果所有的物品都被推荐给至少一个用户，那么覆盖率就是 100%。

◆ 流行度

流行度反应了推荐的新颖程度，可以用推荐列表中视频的平均流行度度量推荐结果的新颖程度，如果推荐出的视频都很热门，说明推荐新颖度较低。

3.4.3 实验结果分析

（1） 非个性化基础算法实验

为了说明本章算法在该数据集上的性能特点，首先对两种非个性化基础算法做性能测试。Random 算法每次都随机挑选 10 个用户没有产生过有效行为的视频推荐给当前用户，MostPopular 算法按照视频的流行度给用户推荐他没有产生过有效行为的视频中最热门的 10 个视频。实验结果如表 3-5 所示。

表 3-5 两种基础算法在 MovieLens 数据集下的性能

	准确率	召回率	覆盖率	流行度
Random	0.631%	0.305%	100%	4.3855
MostPopular	12.79%	6.18%	2.60%	7.7244

MostPopular 算法的准确率和召回率远高于 Random 算法，达到了接近 13%的准确率。但它的覆盖率非常低，推荐结果都非常热门；Random 算法的准确率和召回率很低，但覆盖率很高，平均流行率比较低，推荐结果新颖度高。

（2） 算法对比实验结果分析

BinaryUserCF 只有一个重要的参数 K，即每个用户选出 K 个和他兴趣最相似的用户, 然后推荐那 K 个用户感兴趣的视频。实验测试在不同 K 值下 BinaryUserCF 算法的性能指标，结果如表 3-6 所示。

表 3-6 MovieLens 数据集中 BinaryUserCF 在不同 K 值下的性能

K	准确率	召回率	覆盖率	流行度
5	16.99%	8.21%	51.33%	6.813293
10	20.59%	9.95%	41.49%	6.978854

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

20	22.99%	11.11%	33.17%	7.10162
40	24.50%	11.83%	25.87%	7.203149
80	**25.20%**	**12.17%**	**20.29%**	**7.289817**
160	24.90%	12.03%	15.21%	7.369063

可以看到，推荐算法的精度指标（准确率和召回率）并不和参数 K 成线性关系。在 MovieLens 数据集中，选择 $K=80$ 左右会获得比较高的准确率和召回率，达到了 25%的准确率，远高于两种非个性化基础算法的推荐准确率；另一方面，K 越大，BinaryUserCF 推荐结果就越流行、新颖度越底，这是因为 K 决定了算法在给用户做推荐时参考了和你兴趣相似的其他用户的兴趣，那么如果 K 越大，参考的用户越多，结果就越趋近于全局热门的视频；还有，K 越大，BinaryUserCF 推荐结果的覆盖率越低，它的降低是因为流行度的增加，随着流行度增加，算法越来越趋向于推荐热门的视频，从而对长尾物品的推荐越来越少，因此造成了覆盖率越低。

同样，针对 BinaryUserCF 选取不同的 K 值，即最相似的 K 个视频，测试该算法的性能，结果如表 3-7 所示。

表 3-7 MovieLens 数据集中 BinaryItemCF 在不同 K 值下的性能

K	准确率	召回率	覆盖率	流行度
5	21.47%	10.37%	21.74%	7.172411
10	**22.28%**	**10.76%**	**18.84%**	**7.254526**
20	22.24%	10.74%	16.93%	7.338615
40	21.68%	10.47%	15.31%	7.391163
80	20.64%	9.97%	13.64%	7.413358
160	19.37%	9.36%	11.77%	7.385278

BinaryItemCF 的推荐精度也不是和 K 成正相关或者负相关的，因此选择合适的 K 对获得较高的精度对算法来说是非常重要的，由表看出，选择 $K=10$ 会取得最高的准确率和召回率，22%的准确率同样远高于两种非个性化基础算法的推荐准确率；同时，参数 K 对 BinaryItemCF 推荐结果流行度的影响不是完全正相关的，随着 K 的增加，结果流行度会逐渐提高，但当 K 增加到一定程度，流行度就不会再有明显变化；同样，K 增加会降低系统的覆盖率。

下面选取准确率、覆盖率和流行度，以折线图的方式对比 BinaryUserCF 和

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

BinaryItemCF 算法的性能。

- 精度对比

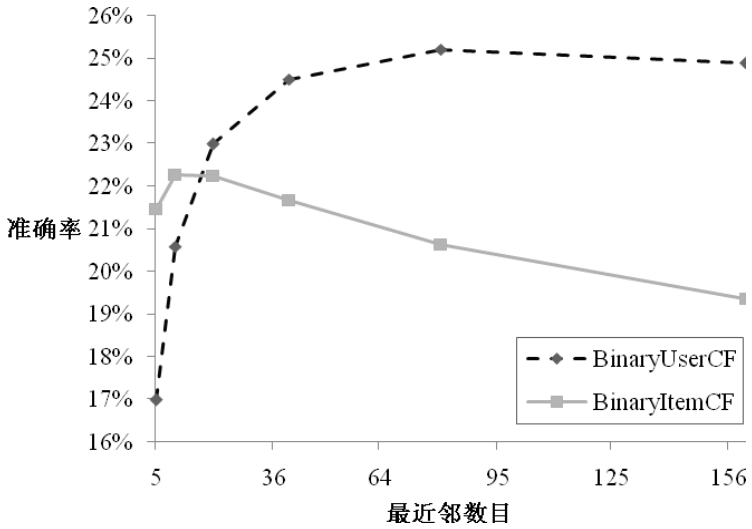

图 3-11 BinaryUserCF 和 BinaryItemCF 算法在不同 K 值的准确率曲线

推荐结果的准确率和召回率共同描述了推荐的精度。这里选取推荐的准确率评测结果绘制 BinaryUserCF 和 BinaryItemCF 算法的准确率曲线，如图 3-11。可以看出不管是基于用户还是基于物品的协作过滤推荐，推荐的精度都不是随着邻居数目的增加而一直提升的，它们的推荐结果准确率曲线类似于正态分布曲线，在某个 K 值点，精度达到最大。如图，在 MovieLens 100K 数据集下，BinaryItemCF 在最近邻居（相似物品）数目为 10 时，BinaryUserCF 在最近邻居（相似用户）数目为 80 时，会取得最大的推荐精度。在最近邻数目小于该值时，精度与最近邻居数目正相关，但在最近邻数据大于该值时，精度与最近邻居数目负相关。本实验中 BinaryItemCF 的推荐最大精度为 22.28%，小于 BinaryUserCF 的推荐最大精度 25.20%。其原因正如第二章推荐技术比较中所述，MovieLens 100K 数据集中，用户数量（943）远小于视频总数（1682），基于用户的协作过滤取得了更好的推荐精度指标。

- 覆盖率和流行度对比

BinaryUserCF 和 BinaryItemCF 算法在不同 K 值的覆盖率曲线和流行度曲线如图 3-12 和图 3-13 所示。先讨论流行度，两种算法的推荐结果平均流行度值基本都随最近邻数目的增加而增大，在最近邻较少时增幅较大，随着最近邻数目超过 50 后，增幅变得较为平缓。原因正如前文所述，随着邻居数目的增加，基于领域的

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

协作过滤算法参照的用户\项目就越多，推荐结果就越趋向于热门的视频，流行度自然就会越大，即为新颖度越小。另外，对于覆盖率，它与流行度指标相反，与最近邻数目成负相关。这可以用流行度的变化来说明，流行度的增大说明推荐结果越趋向于热门的视频，算法发掘长尾资源的能力越小，自然推荐结果的覆盖率越低。可见，在追求高精度推荐结果的同时，总是牺牲了部分推荐结果的覆盖率和新颖度。同样的，可以看到，BinaryUserCF 推荐算法的覆盖率和新颖度指标也优于 BinaryItemCF 推荐算法。

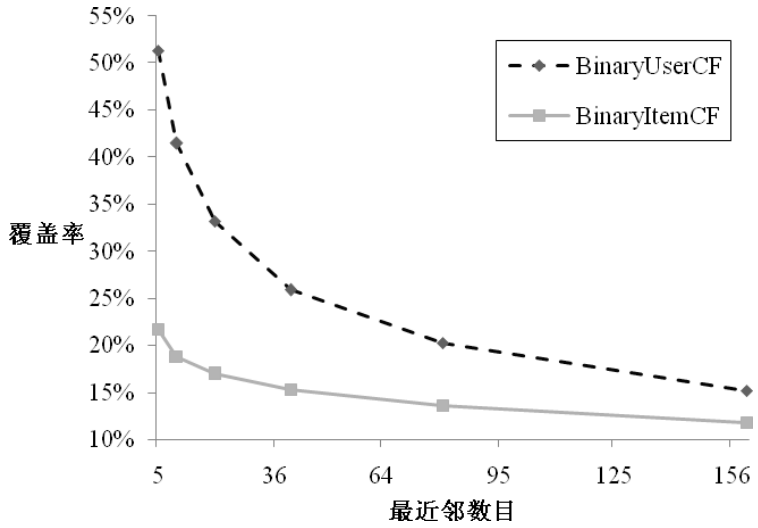

图 3-12 BinaryUserCF 和 BinaryItemCF 算法在不同 K 值的覆盖率曲线

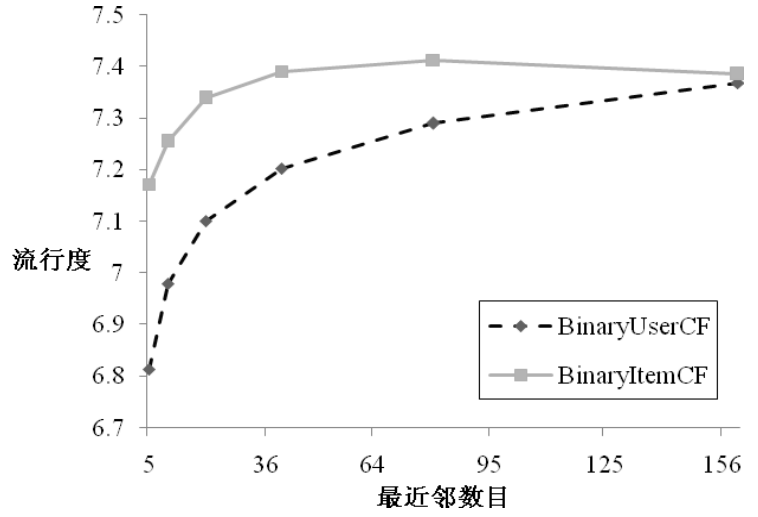

图 3-13 BinaryUserCF 和 BinaryItemCF 算法在不同 K 值的流行度曲线

下面将对基于的视频流行度和用户偏好变化的 BinaryUserCF 改进算法

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

（BUser-HTF）做实验分析。

（3） 算法改进实验结果分析

本节将通过实验评测 BUser-HTF 的推荐性能，并将其和 BinaryUserCF 进行对比以观察改进算法对性能的提升效果。上一节试验中，$K=80$ 时 BinaryUserCF 的性能最好，因此本实验直接选取 $K=80$，评测结果如表 3-8 所示。

表 3-8 MovieLens 数据集中 BinaryUserCF 和 BUser-HTF 算法的改进对比

	准确率	召回率	覆盖率	流行度
BinaryUserCF	25.20%	12.17%	20.29%	7.289817
BUser-HTF	26.74%	13.70%	23.29%	7.151151

根据评测性能结果，以柱状图的形式展示改进算法的主要性能提升效果，如图 3-14 所示。

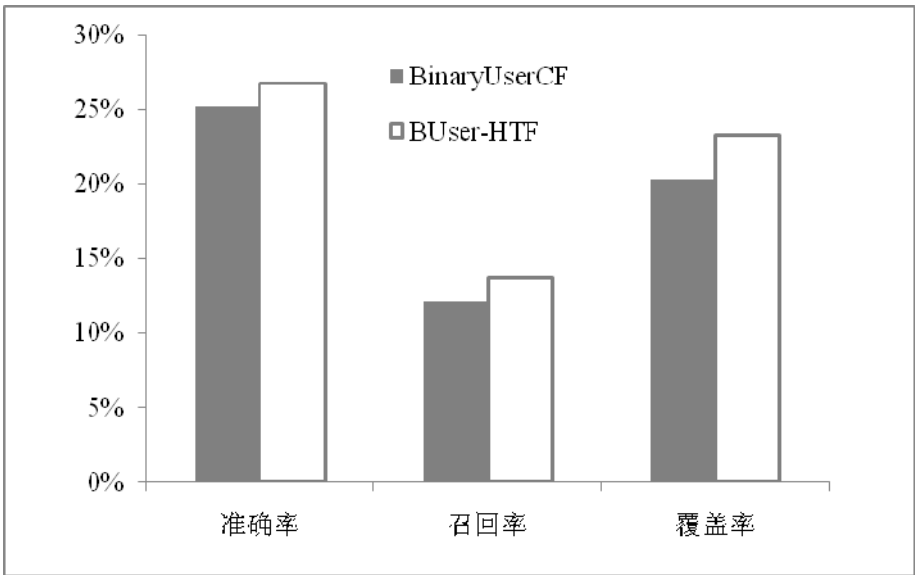

图 3-14 BUser-HTF 算法的性能提升对比图

可以看到考虑视频流行度和用户偏好变化的 BUser-HTF 算法在各项性能上都优于 BinaryUserCF，准确率上提升了 1.54 个百分点，覆盖率上提升了 2 个百分点，平均流行度降低了 0.14，降低了流行度即是提高了推荐结果的新颖度。由此，改进算法 BUser-HTF 对推荐性能取得了较为理想的提高，说明了推荐算法中考虑视频流行度和用户兴趣变化有助于提升推荐结果的质量。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3.5 本章小结

本章以推荐系统领域的 Top-N 推荐为核心，设计了用户-节目二进制关联的用户偏好模型，在提出了视频流行度和用户活跃度这两个概念的基础上，设计了视频流行度和用户偏好变化的衡量方式。接着，提出在用户-视频二进制关联模型下的基于领域的协作过滤推荐算法，并引入视频流行度和用户偏好变化，提出改进算法，然后进行了推荐实验分析和对比。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第四章 基于矩阵分解的协作过滤算法

本文第三章研究的是 Top-N 推荐，即为一个给定的用户预测并推荐其兴趣值最高的前 N 个视频。然而，评分预测问题作为推荐系统研究的另一核心，从 GroupLens 到 Netflix Prize 到 KDD cup，一直受到广泛的重视。视频推荐系统中的评分预测能够给用户计算一个分数表明用户是否喜欢其未看过的视频，而这个分数可以帮助系统决策是否给用户推荐这部电影。如何提高这个分数的预测精度就是评分预测要解决的关键问题。本章将在从评分预测问题中的数据稀疏性入手，提出评分预测问题中，基于矩阵分解的协作过滤推荐的改进算法，并进行实验分析。

4.1 评分预测中的数据稀疏性问题

4.1.1 稀疏性问题

用户评分数据是评分预测实现推荐的主要依据，评分数值反映了用户对不同项目的喜好程度。在基于领域的协作过滤（User-CF 和 Item-CF）评分预测推荐中，如果有两个用户在现实中有着相同的兴趣偏好，但他们在推荐系统中从没有对同一视频有过评分，那么系统也会把他们视为不相似；如果两个实际类似的视频从没有被同一个用户评分过，那么系统也无法判定这两个视频是相似的。

许多大型的推荐系统的用户数量庞大，但每个用户涉及的信息却相当有限，例如在亚马逊的推荐系统中，评分数最多的用户也只评价了系统资源数目 1%-2%。如果用用户和项目已由的选择关系占所有可能存在的选择关系的比例来衡量系统的稀疏性，那么可以得到推荐系统领域研究较多的数据集的稀疏度数[44]，如表 4-1 所示。

造成推荐系统中数据稀疏性问题的原因主要包括两个方面，一个来自用户，一个来自系统本身。

- 来自用户的成因：正如前文所述，单个用户的行为信息有限，其评价过的资源只能是系统资源的很少部分。
- 来自系统的成因：在大型的资源服务推荐系统中，系统所提供的资源项目规模可达数以百万甚至上亿的数量级，并且逐日增加，其增幅可能远超过系统用

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

户数量的增幅。高维的用户-项目评分矩阵同样带来了数据稀疏性的问题。

可以看到，用户-评分矩阵不仅是高维的，而且是稀疏的，这使得很多评分预测算法难以发挥理想的效果，评分预测的推荐质量大打折扣。由此，能够处理稀疏数据的算法将被视为更有前途[44]。

表 4-1 常用数据集的稀疏度

	稀疏度
MovieLens	4.5%
Netflix	1.2%
Bibsonomy	0.35%
Delicious	0.046%

4.1.2 现有的解决方案

稀疏性问题对协作推荐的影响是贯穿于推荐算法的各个环节中,因此对稀疏性问题的解决可以从不同角度进行。分析研究人员们初具规模的研究成果，可以看出主要有两种策略来解决稀疏性问题，一类方法是直接处理用户评分矩阵，通过设定缺省值或者评分矩阵降维等技术，提高评分矩阵的数据密度。而另外一类解决办法则属于间接方法，通过改进协作推荐的关键步骤，如相似度计算、近邻选择和评分预测，以缓解稀疏性问题对协作过滤算法的影响。

1）设定缺省值。

最简单的降低数据稀疏性的方法是对每个用户未评分的项目按统一的方式给定一个缺省值，以填充缺少的评分。一般包括如下几种设定方式：评分制的评分中值, 用户平均评分 (使用用户的平均评分填充该用户对应的未评分项目的评分), 或者项目平均评分（使用项目的平均评分填充该项目对应的缺失的用户评分)。设 $\overline{r_u}$ 为用户平均评分，$\overline{r_i}$ 为项目平均评分，r_{ui} 为用户 u 对未评分项目 i 的评分缺省值，结合上述两种设定方法，也可以得到如下两种方法来计算 r_{ui}。

① 用户平均评分加权

$$r_{ui}=\overline{r_u}+\frac{\sum_{k=1}^{K}\left(r_{ki}-\overline{r_i}\right)}{K} \tag{4-1}$$

其中，K 为系统用户对项目 i 的评分总数。

② 项目平均评分加权

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$r_{ui}=\bar{r}_i+\frac{\sum_{q=1}^{Q}\left(r_{uq}-\bar{r}_u\right)}{Q} \tag{4-2}$$

其中，Q为系统用户u对系统项目的评分总数。

设置缺省值的方法较为粗糙，考虑的因素太少，用户对不同项目的评分不可能完全一样，这样的设定虽然解决了稀疏性问题，但是也可能降低推荐的质量。

2）聚类分析。

聚类分析是先对系统项目按某种方式进行聚类，从而将“最近邻居”的搜索范围限制在目标项目所在的聚类中，这样通过缩小邻居寻找范围提高了协作过滤推荐的速度，但是通过分析项目的内容属性，使得寻找到的最近邻都是在内容上相似的，失去了协作过滤推荐的新颖性优势。另一种聚类分析方法（ICHM：Item-based Clustering Hybrid Method）能够更好地解决稀疏性问题，该方法在计算项目的相似度时结合了用户评分和项目的内容属性，采用基于项目的协作过滤方法进行评分预测。该算法的大致流程为：

1. 构建组-评分矩阵。该矩阵的行为项目，列为聚类，矩阵元素为项目属于某一聚类的概率$P(j,k)$。概率$P(j,k)$是指项目j属于某一聚类k的概率，是由基于项目的内容信息的模糊聚类分析计算而得。

2. 计算项目相似度。项目相似度是由原始项目-评分矩阵和组-评分矩阵中项目的相似度进行加权求和而得到。

3. 预测评分。按项目相似度大小搜索到缺失用户评分的项目对应的邻居项目，将这些邻居项目得分的相似度加权和作为用户的评分预测值。

可以看出，即使在原始的用户-项目评分矩阵中难以得到某两个项目间的相似度，也可能通过对组-评分矩阵分析得到他们的相似关系，从而缓解了数据稀疏性问题对相似度计算的影响，但是基于项目内容的模糊聚类在某种程度上牺牲了推荐质量，评分预测准确度欠佳。

3） 信任度传播。

信任度传播的方法需要基于系统用户间的信任度评价构建一个全局信任网络，然后根据用户全局信任网络计算用户之间的相似度，进而通过某种算法预测活动用户的评分值，这种相似度计算方法可以忽略非信任用户的评分数据干扰，从而一定程度上避免了稀疏性问题，但这种方法要求用户相互了解，并且信任度评价也必须用户主动参与配合，在实际应用中可行性较低。

类似于上述方法，可以在用户相似度计算时结合用户的社会关系网络，而非

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

单纯的利用用户-项目评分矩阵，也能有效地避免数据稀疏性问题。同时，社会关系网络的取得可以是通过系统分析用户长期积累的行为数据，也可以用户按系统设计主动提交给系统。

4） 传递关系。

Bergholz 提出了用于 Xerox 公司的 Knowledge Pump（KP）文档推荐系统两种推荐方法[45]，。一种方法利用多对用户间的串联起来的关系计算两端用户间的关系，即所谓关系传递。当用户 i 与用户 j 之间的关系受数据稀疏性影响而无法直接计算时，该方法可以利用用户 i 与用户 k 、用户 k 与用户 j 之间的关系间接求得用户 i 与用户 j 之间的关系，间接计算的公式为：

$$r_I(i,j)=\frac{1}{n}\sum_{k}\frac{r(i,k)+r(k,j)}{2} \tag{4-3}$$

另外一种方法引入了人工用户代理，这些代理按照预定的规则对缺失评分的项目进行评分。KP 系统创建了两大类人工用户代理，一类与 KP 系统业务有关，如 recent first rating，high number of visits，high number of ratings 等；另一类与 KP 系统文档格式有关，如 PDF，HTML 等。代理的打分主要依据自身的创建类别，如该方法中，recent first rating 按文档的首次评分时间进行评分补缺，文档的首次评分时间在 2001 年后，则该文档的补缺评分为 5 分，首次评分时间在 2000 年与 2001 年之间，补缺为 4 分，依次类推；根据系统文档格式创建的代理对它所对应的文档类型相同的文档评分补缺为 5 分，其他补缺为 2 分。这种方法针对的是特定的系统，其实用性有限。

5）降维技术。降维技术可以完成的用户-项目评分矩阵从高维到低维的映射，能够有效降低矩阵规模及数据稀疏性，直接对评分矩阵的处理可以从源头上解决数据稀疏性问题。用于协作过滤评分预测的降维技术包括了简单降维方法以及奇异值分解。

- 简单降维方法

例如忽略系统中行为信息极少的用户以及极度冷门的项目的评分数据，删除没有做出任何评分的用户和没有被任何用户评分的项目，都属于简单降维方法。这样的方法可以一定程度上避免数据稀疏性问题，但可能会使删除或忽略的项目/用户无法被推荐给相关的用户，或者不能得到推荐服务。

一种相对有效的方法是将用户-项目矩阵转换为用户-类别矩阵的方法，并且用系统用户对某类别中所有项目评分的平均值作为用户-类别矩阵中对应元素的值。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

由于把项目总数缩小为项目类别数，大幅度且有效地降低了矩阵维度，增加了矩阵数据密度，并且还有助于改善冷启动问题。

➢ 奇异值分解

矩阵分解（Matrix Factorization）是利用数学的方法分析用户-评分矩阵的特征，通过数学上的矩阵分解处理发掘用户和项目的内在关联，进而达到对矩阵中遗漏评分值的预测作用。奇异值分解（SVD：Singular Value Decomposition）是应用最广泛的矩阵分解模型，它在引入到推荐系统算法中以前，在信息检索、机器学习、数据压缩等领域都有着广泛的应用。

Daniel Billsus 等人最早提出了基于奇异值分解的协作过滤评分预测算法[46]。给定 m 个用户和 n 个项目，和用户对项目的评分矩阵 $R \in \square^{m \times n}$。首先需要对评分矩阵中的缺失值进行简单的补全，可以用前文提到的用户/项目平均值补全，然后得到补全后的矩阵 $R^{'}$。接着，用 SVD 分解将 $R^{'}$ 分解成如下形式。

$$R^{'}=U \sum V^{T} \tag{4-4}$$

其中，$U \in \square^{m \times k}$，$V \in \square^{n \times k}$ 是两个正交矩阵，称为左右奇异矩阵。$S \in \square^{k \times k}$ 是对角矩阵，对角线上的每一元素都是矩阵的奇异值。奇异值在矩阵 $\sum$ 中从大到小排列，很多情况下，前 10%左右的奇异值的和就占了全部奇异值之和的 99%以上，大部分的奇异值对矩阵特征的影响是很微小的。所以该方法中，可以取最大的 f 个奇异值组成对角矩阵 $\sum_{f}$，并且找到这 f 个奇异值中每个值在 U 中对应的列，在 V 中对应的行，得到 $U_{f} \in \square^{m \times f}$，$V_{f} \in \square^{n \times f}$。从而在保留了绝大多数用户评分矩阵的数据特征情况下，对奇异值分解的三个矩阵起到可观的降维作用，最后可以得到一个新的与原评分矩阵近似的评分矩阵：

$$R_{f}^{'}=U_{f} \sum_{f} V_{f}^{T} \tag{4-5}$$

其中，$R_{f}^{'} \in \square^{m \times n}$，$R_{f}^{'}(u,i)$ 就是用户 u 对项目 i 评分的预测值。

同时，可以为用户评分矩阵进行奇异值分解后的三个矩阵意义，作出如下解释。

U ：左奇异矩阵的每一行表示一个 user 的特征，共有 k 个特征分量。行中的值代表了该用户对应某特征的重要性（或者说相关性）程度。

V ：右奇异矩阵的每一列表示一个 item 的特征，共有 k 个特征分量。列中的值代表了该项目对应某特征的重要性（或者说相关性）程度。

$\sum$ ：每个奇异值表示对应的 user 和 item 的特征相关性。

对奇异值矩阵 $\sum$ 的降维处理，就是忽略用户和项目相关性很小的部分特征分

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

量，从而可以有效而不失准确度的进行评分预测。

SVD 分解是较早期推荐系统评分预测常用的方法，不过该方法存在以下缺点，因此很难在实际中应用。

① 该方法首先要对稀疏评分矩阵进行预补全。一旦补全，一是让评分矩阵的存储开销大增，大量的空间需求在实际系统很难接受；二是简单的补全也会影响推荐的准确度。

②该方法依赖的 SVD 分解方法的计算复杂度很高，特别是在稠密的大规模矩阵上更是非常慢。一般来说，SVD 分解用于 1000 维以上的矩阵就会非常慢，而实际系统往往是上百万的项目和用户，所以这一方法实用性不高。

下一节中将阐述另一种带局部优化求解的矩阵分解方法，及其改进策略。这种方法是由上述的 SVD 分解发展而来，并且在推荐系统评分预测问题中比 SVD 分解更具有可行性。

4.2 基于矩阵分解的协作过滤推荐

4.2.1 矩阵分解模型

矩阵分解模型是在 SVD 分解的基础上发展而来的应用于协作过滤评分预测的方法，传统的 SVD 分解需要在数学上直接分解得到左右奇异矩阵和奇异值矩阵进而进行评分预测，尽管分解出来的精度比较高，但是计算复杂度太高，很难在实际的推荐算法中应用。本节将阐述一种可以转换为优化问题求解的矩阵分解模型，并提出一种加入用户评分偏置的改进算法。

4.2.1.1 矩阵分解模型的描述

该模型由 Simon Funk 在 2006 年的博文中提出, 它是建立在 SVD 分解基础上，用左右奇异矩阵把中间的奇异矩阵吸收进来。过程可描述为：

$$R=(U\Sigma)V^T=U(\Sigma V^T)=(U\sqrt{\Sigma})(\sqrt{\Sigma}V^T) \tag{4-6}$$

然后合并括号内的分项，消去Σ，得到：

$$R=UV^T \tag{4-7}$$

这样合并后，U 表示用户特征矩阵，V 则表示项目特征矩阵。同样，在只取 f 个最大的奇异值，其余用 0 代替的情况下，可以用 $\hat{R}$ 近似表示 R，得到：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\hat{R} = P^T Q \tag{4-8}$$

其中 $P \in \mathbb{R}^{m \times f}$ 和 $Q \in \mathbb{R}^{n \times f}$ 是两个降维后的矩阵，分别代表用户和项目的内在特征。那么，对于用户 u 对物品 i 的评分预测值 $\hat{R}(u,i) = \hat{r}_{ui}$，可以通过如下公式计算：

$$\hat{r}_{ui} = \sum_f p_{uf} q_{if} \tag{4-9}$$

其中，$p_{uf} = P(u,f)$，$q_{if} = Q(i,f)$。

矩阵分解模型为什么可以用于协作过滤的评分预测，本文经过分析总结，提出如下两个方面的模型解释。

- 解释一：隐性因子模型

矩阵分解的过程实际上是从用户评分矩阵中分别提取了用户和项目的隐性特征到用户特征矩阵和项目特征矩阵中，从而用户对项目的喜欢程度用两个矩阵中对应向量的内积表示。只是特征的提取是通过数学方法完成的，并且它们是蕴藏于用户评分矩阵的隐性特征，没有一个统一的方式为这些特征命名，针对不同的项目对象可以有不同的理解。以视频推荐系统为例，这些隐性因子可以描述为电影的搞笑因子、恐怖因子，爱情因子等。用户特征矩阵中的元素值表示用户对对应隐性因子的喜欢程度，项目特征矩阵中的元素值则表示项目与对应隐性因子的相关程度。下面通过一个简单示例说明，例子中的用户评分矩阵包括了两个用户对 3 部电影的评分数据，选取电影的两个隐性特征因子。

表 4-2 视频推荐矩阵分解示例

	电影 1	电影 2	电影 3
用户 A	5	3	?
用户 B	2	4	5

（A）

	搞笑	恐怖
用户 A	1	0.1
用户 B	0.2	1

（B）

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

	搞笑	恐怖
电影 1	5	0
电影 2	3	3
电影 3	0	5

（C）

按照示例评分矩阵（A）看来，用户 A 可能不喜欢恐怖电影，应该对电影 3 给出较低的分数，按照矩阵分解结果（B）和（C）预测用户 A 对电影 3 的评分为 0.5，正好是一个较低的评分值。

- 解释二：降维

这是把矩阵分解引入到推荐系统评分预测算法的初衷，通过降维可以解决评分数据稀疏的问题。可以看到，矩阵分解把维度为 $m \times n$ 的评分矩阵分解为用户特征矩阵 $U \in \mathbb{R}^{m \times k}$ 和项目特征矩阵 $V \in \mathbb{R}^{n \times k}$ 的乘积，其中 $k = \min(m, n)$。通过选取最具代表性的 f 个奇异值，得到降维后的用户特征矩阵 $P \in \mathbb{R}^{m \times f}$ 和项目特征矩阵 $Q \in \mathbb{R}^{n \times f}$，其中 $f < k$。由此，数据集的大小由 $m \times n$ 降低为 $m \times k + n \times k$，这样大大降低了需要处理的数据维度，减小了数据存储空间开销，解决了原始评分数据的高维稀疏性问题。

4.2.1.2 矩阵分解模型的求解

传统的 SVD 分解是通过数学上的直接对填充后的评分矩阵直接进行矩阵分解，这样效率低且复杂度高，难以在实际中应用。一种利用最小化目标函数的方法是把矩阵分解问题转为为最优化问题，使得矩阵分解得到极大简化。

如前文所述，$\hat{r}_{ui}$ 是算法得到预测评分值，从而评分偏差和评分误差平方和可以分别表示为：

$$e_{ui} = r_{ui} - \hat{r}_{ui} \tag{4-10}$$

$$SE = \sum e_{ui}^2 = \sum \left(r_{ui} - \sum_{s=1}^{f} p_{us} q_{is} \right)^2 \tag{4-11}$$

算法的目标找到最优的矩阵分解形式，使得得到的 $\hat{R}$ 最接近于原始的评分矩阵，原始的评分矩阵为训练集（用 D_{train} 表示）中的数据，那么 $\hat{R}$ 中对应于原始评分矩阵中未评分项的评分数据，即预测的评分数据将会最接近于测试集（D_{test}）中的评分数值。由此，矩阵分解问题可以转为为通过最小化误差平方和，找到最优

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

的矩阵分解形式。该误差评分和就是需要优化的目标函数。

由于数据的稀疏性，直接优化上面的目标函数可能会导致学习的过拟合，因此还需要加入防止过拟合项$\lambda\left(\|P\|^2+\|Q\|^2\right)$，其中$\lambda$是正则化参数，从而得到最终的目标函数：

$$C\left(p_u,q_i\right)=\sum_{(u,i)}\left(r_{ui}-\sum_{s=1}^{f}p_{us}q_{is}\right)^2+\lambda\left(\|p_u\|^2+\|q_i\|^2\right),(u,i)\in D_{train} \tag{4-12}$$

最小化上面的目标函数，可以用最优化理论的梯度下降法。它首先通过求参数的偏导数找到函数值下降最快的方向，即梯度方向。然后选择这样的方向不断更新特征矩阵P和Q的值，使得目标函数逐步变小，直到收敛。目标函数的梯度方向为：

$$\frac{\partial C}{\partial p_u}=-2q_i\left(r_{ui}-\sum_{s=1}^{f}p_{us}q_{is}\right)+2\lambda p_u=-2e_{ui}q_i+2\lambda p_u \tag{4-13}$$

$$\frac{\partial C}{\partial q_i}=-2p_u\left(r_{ui}-\sum_{s=1}^{f}p_{us}q_{is}\right)+2\lambda q_i=-2e_{ui}p_u+2\lambda q_i \tag{4-14}$$

然后根据梯度下降法，需要将参数沿着梯度方向推进，可以得到如下迭代公式：

$$p_u=p_u-\gamma\cdot\frac{\partial C}{\partial p_u}=p_u+2\gamma\left(e_{ui}q_i-\lambda p_u\right) \tag{4-15}$$

$$q_i=q_i-\gamma\cdot\frac{\partial C}{\partial q_i}=q_i+2\gamma\left(e_{ui}p_u-\lambda q_i\right) \tag{4-16}$$

其中为γ为迭代步长，也称为学习速率，它需要在算法之初设定一个初值，随着迭代过程进行，迭代步长需要逐步衰减（通常衰减因子取为0.9，即$\gamma=0.9\gamma$），衰减的理由形象的说就是：如果需要在一个区域找到极值，一开始可能需要大范围搜索，但随着搜索的进行，搜索范围要逐渐缩小。这就是由迭代步长控制的。

除了初始化γ之外，在开始梯度下降法优化前，还需要对正则化参数λ和两个特征矩阵进行初始化，初始化特征矩阵一般是将这两个矩阵用随机数填充，并且随机数需要和$1/sprt\left(F\right)$成正比，F为算法选取的隐性特征因子数目。

4.2.2 引入评分偏置的矩阵分解模型

实际的视频系统中，不同用户有着不同的打分倾向，两个用户在同样喜欢某

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

个节目的情况下，他们也可能打出不同的分数。同时，视频的品质也存在差异，有些视频品质较高，总能得到比较高的评分。基于此，本文在上述评分预测模型中加入评分偏置项，提出改进的矩阵分解模型（BiasMF），使得评分预测模型更加接近真实的情况，以提高模型的评分预测准确度。

- o（Overallmean）是训练集中所有已有评分记录的全局平均数。在不同网站的推荐系统中，因为网站的定位和面向人群的不同，网站业务项目的整体评分分布也会显示出一些差异。全局平均数代表了网站系统本身对用户评分的影响。
- b_u 是用户评分偏置（user bias）项。这一项表示和具体项目无关的用户个人的评分倾向习惯。比如有些用户要求比较苛刻，对系统资源的要求很高，那么他的评分就会偏低，而有些用户则比较宽容，对很多项目都觉得不错，那么他的评分就会偏高。
- b_i 是项目评分偏置（item bias）项。这一项表示了项目接受评分时的自身质量影响因素。比如有些项目本身品质比较高，因此获得的评分相对都比较高，而有些视频本身品质较差，哪怕它们因为某些原因比较流行，它们获得的评分都会比较低。

于是，在评分预测模型中引入以上偏置项后，可以得到新的评分预测公式：

$$\hat{r}_{ui} = o + b_u + b_i + \sum_f p_{uf} q_{if} \tag{4-17}$$

基于引入评分偏置的矩阵分解模型的协作过滤评分预测（BiasMFCF）的求解同样可以采用最优化目标函数的方法，新的目标函数构造为：

$$C\left(b_u, b_i, p_u, q_i\right) = \sum_{(u,i)} \left(r_{ui} - o - b_u - b_i - \sum_{s=1}^{f} p_{us} q_{is} \right)^2 + \lambda \left(\left\| b_u \right\|^2 + \left\| b_i \right\|^2 + \left\| p_u \right\|^2 + \left\| q_i \right\|^2 \right), (u,i) \in D_{train} \tag{4-18}$$

其中，o 可以由训练集评分数据直接求得。偏置向量 b_u 和 b_i 则需要和特征向量一样，通过迭代训练求得，迭代公式为：

$$b_u = b_u - \gamma \cdot \frac{\partial C}{\partial b_u} = b_u + 2\gamma \left(e_{ui} - \lambda b_u \right) \tag{4-19}$$

$$b_i = b_i - \gamma \cdot \frac{\partial C}{\partial b_i} = b_i + 2\gamma \left(e_{ui} - \lambda b_i \right) \tag{4-20}$$

具体的算法流程如算法 4.1：

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

算法 4.1 改进的基于矩阵分解的协作过滤评分预测算法（BiasMFCF）

计算训练集中评分的全局平均数 o

初始化特征矩阵 U 、V 和偏置向量

设置迭代步长 γ，正则化参数 λ 和特征向量的维度 f

do

foreach r_{ui} in D_{train} **do**

计算评分误差 e_{ui}

计算目标函数 C 的梯度

更新迭代步长 γ

更新偏置向量 b_u 和 b_i %按公式 4-18 和 4-19

更新特征向量 p_u 和 q_i %按公式 4-14 和 4-15

end

按照更新的特征矩阵计算目标函数

while 还未满足终止条件（目标函数减小的幅度小于某个阈值）

foreach r_{ui} in D_{test} **do**

计算预测评分与实际评分误差 e_{ui}

end

计算均方根误差（RMSE）并输出

4.3 实验设计及结果分析

4.3.1 实验数据

本章将采用 Netflix 公司公开的电影评分数据集。该数据集包括了电影信息、训练集（Training Set）、测试集（Probe Set）和评估集（Qualifying Set）。训练集和测试集是专门给相关领域的研究者做离线试验使用的，评估集专用于 Netflix Prize 比赛，里面的评分数据为空，参赛者需要提交预测结果才能评测算法预测性能。本章实验使用就是其中的训练集和测试集。

训练集 D_{Train} 为包括用户 ID、电影 ID、用户评分日期和评分值得表格，其中评分值取值同 MovieLens 数据集一致，取值区间为 $[1,2,3,4,5]$。

测试集 D_{Test} 的形式与训练集相同，Netflix 生成该测试集的方法为：对于每一

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

个用户，把这个用户最新的 5%的评分记录作为测试数据，这 5%中的最新的 2/3 放入到评估集中，而剩下的 1/3 则放入到训练集中。

4.3.2 实验评测指标

在推荐系统评分预测问题中，推荐质量评测指标就是对评分准确性的评测。本章实验采用均方根误差（RMSE）来评测预测结果的准确性。RMSE 公式如下：

$$RMSE=\sqrt{\frac{\sum_{(u,i)\in D_{Train}}\left(r_{ui}-\hat{r}_{ui}\right)^2}{\left|D_{Train}\right|}} \tag{4-21}$$

其中，$\hat{r}_{ui}$ 是算法对用户 u 对视频 i 的评分的预测值。RMSE 越小，表明预测精度越高。

4.3.3 实验方案及结果分析

本实验在 Netflix 电影评分数据集上对两种评分预测算法进行考查：基于矩阵分解的协作过滤算法（MFCF），基于评分偏置矩阵分解的协作过滤算法（BiasMFCF)。通过三组实验对比两种算法，以分析验证改进算法对评分预测准确性的提高和在效率方面的改进。

影响以上两种算法性能的主要因素有：用户/项目特征向量的维度 f，迭代步长（学习速率）γ。因此，本节为验证改进算法对预测准确度和执行效率的优化，按如下三组实验对两算法性能进行对比分析。

- 实验 1：固定迭代步长 γ 和正则化参数 λ，查看在不同特征向量维度 f 的情况下，算法的准确度性能比较。本组实验中，迭代步长初始值设定为 $\gamma=0.008$，正则化参数 $\lambda=0.03$，特征向量维度 f 分别取 20、30、40、60、90。

实验评测结果如下表 4-3 所示：

表 4-3 不同特征向量维度下算法的性能（RMSE）对比

	$f=20$	$f=30$	$f=40$	$f=60$	$f=90$
MFCF	1.913	1.465	1.193	1.049	1.009
BiasMFCF	1.899	1.431	1.151	1.006	0.949
提升	0.014	0.034	0.042	0.043	0.060

根据表中的预测结果的 RMSE 评测结果，以特征维度 f 为横坐标，RMSE 为

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

纵坐标，两种算法的性能对比曲线如下图 4-1 所示：

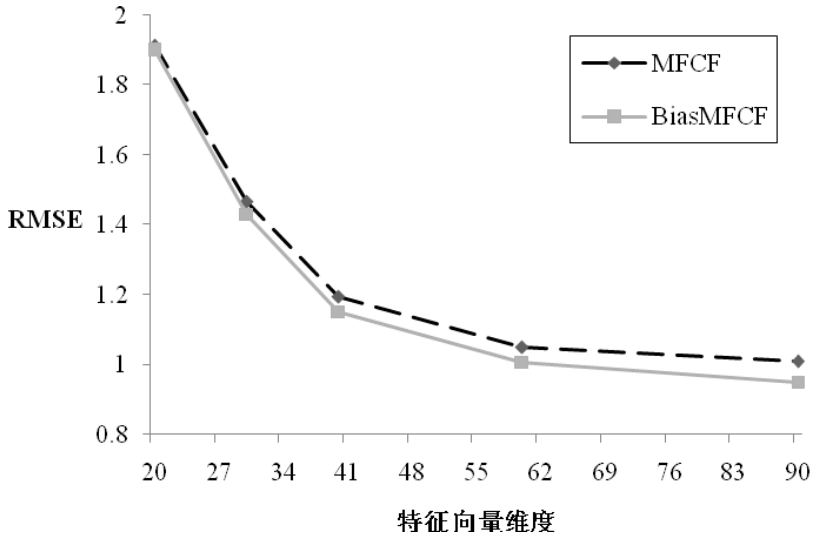

图 4-1 不同维度下的 MFCF 与 BiasMFCF 性能曲线对比图

可以看到两算法的预测准确度都随着维度 f 的增大而提高，这符合算法的基本原理，因为维度 f 代表的选取的特征数，特征数越多，越接近于原始的训练集评分数据，预测准确度就越高。开始准确度提高较快，到了 $f=60$ 左右后，提高幅度变得平缓，说明 $f=60$ 的特征维数就能较好体现该数据集评分矩阵的内在特征。同时，本文提出的改进算法在各维度下都起到了提升预测准确度的作用，在 $f=20$ 是提升了 0.014，随着特征向量维度的增大，该进算法的预测准确度提升效果越好，到 $f=90$ 最大提升了 0.060。各维度下平均提升了 0.0386。

- 实验 2：固定特征维度 f 和正则化参数 λ，查看在不同迭代步长初值 γ 的情况下，算法的准确度性能比较。由于在上组实验中 $f=60$ 时性能表现较好且趋于稳定，因此本组实验中，设定特征维度 $f=60$，正则化参数 $\lambda=0.03$，迭代步长 γ 分别取为 0.015、0.01、0.005、0.001。

实验评测结果如表 4-4 所示。

表 4-4 不同迭代步长下算法的性能（RMSE）对比

	$\gamma=0.015$	$\gamma=0.01$	$\gamma=0.005$	$\gamma=0.001$
MFCF	1.154	1.011	0.987	1.058
BiasMFCF	1.114	0.953	0.945	1.045
提升	0.040	0.058	0.042	0.013

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

根据表中的预测结果的 RMSE 评测结果，以迭代步长γ为横坐标，RMSE 为纵坐标，两种算法的性能对比曲线如下图 4-2 所示：

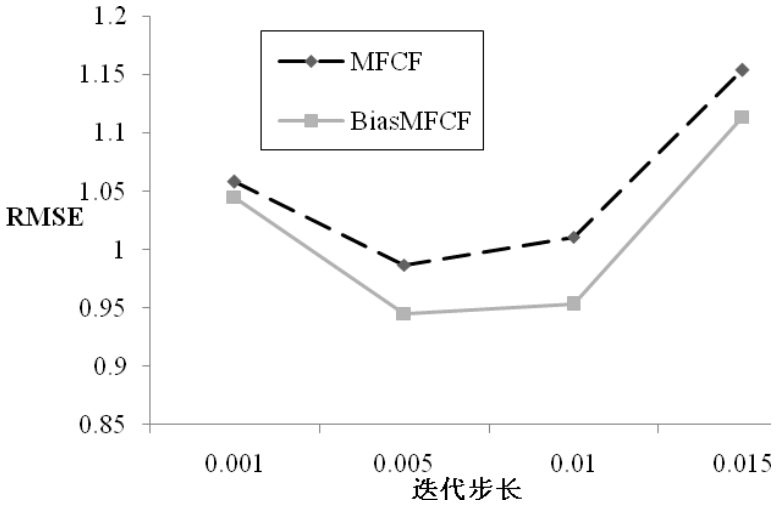

图 4-2 不同迭代步长下的 MFCF 与 BiasMFCF 性能曲线对比图

可以看到在同样的特征维度下，两种算法在$\lambda = 0.015$时的 RMSE 较大，随着迭代步长初值减小，RMSE 也逐渐减小，评分预测准确度提高，直到$\lambda = 0.005$时达到了较高的评分预测准确度。但通过实验可以发现，预测准确度并不是随着迭代步长减小而正比例提高，$\lambda = 0.001$时，预测准确度反而减小了，改进算法对预测准确度的提升也相对微弱。可以发现迭代步长的选择比较有讲究：如前文所述，梯度下降过程中，按衰减系数 0.9 衰减的迭代步长控制着对目标函数极小值搜索范围：先在一个较大范围内搜索，然后逐步缩小这个搜索区间，从而找到目标函数极小值（这个极小值可能是一个接近于真正极小值的局部极小值）。如果初始的搜索范围过小，就很有可能搜索到的局部极小值距离目标函数真正的极小值还比较远，优化程度不够理想，最后的得到 RMSE 评测指标则相对较差。所以，并不是迭代步长越小，预测准确度越高。

同时，本文的改进模型在各迭代步长下，对评分预测准确度都有不同程度的提升，迭代步长为 0.1 的时候，准确度提升最大，RMSE 减小了 0.058。在迭代步长为 0.005 的时候，RMSE 仅减小了 0.013，所以在迭代步长很小的时候，算法的改进效果也会受到影响。各迭代步长下改进算法的预测准确度平均提升了 0.038。

- 实验 3：根据以上两个实验的结果，选取适当的特征维度f，迭代步长γ和正则化参数λ，查看在固定这些参数的条件下，两种算法的运行效率特征。本组实

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

验中，设定特征维度 $f=40$，迭代步长初始值设定为 $\gamma=0.005$，正则化参数 $\lambda=0.03$。

以算法运行的关键迭代步数为横坐标，以对应的推荐结果 RMSE 为纵坐标，得到两种算法的学习效率曲线示意图，如图 4-3 所示。

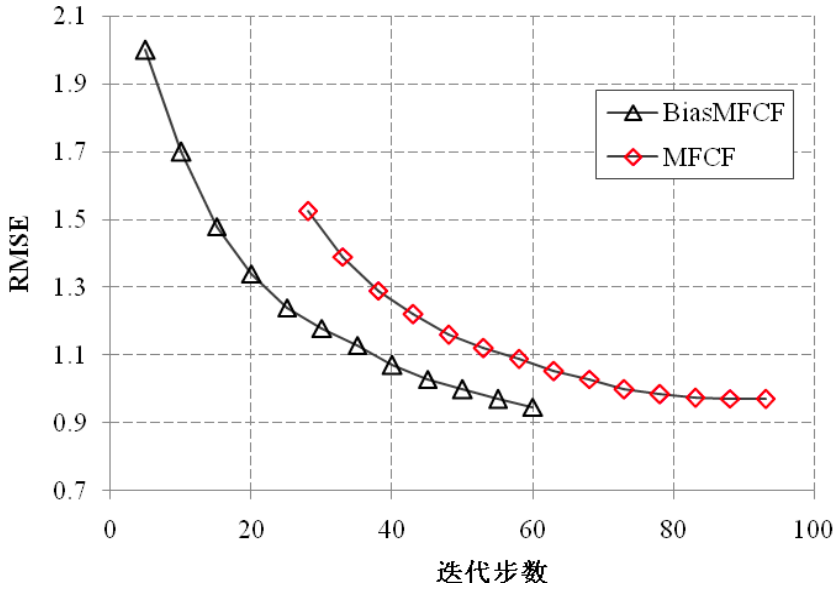

图 4-3 MFCF 与 BiasMFCF 学习效率性能曲线对比图

（$f=40$，$\gamma=0.005$，$\lambda=0.03$）

首先，可以看出基础算法 MFCF 的学习效率曲线比较平缓，改进算法 BiasMFCF 的学习效率曲线更加陡峭，即是说增加同样的迭代步数，改进算法对应的 RMSE 下降得更快，说明改进算法能够在较少的迭代步数内，快速达到 RMSE 的收敛效果。本实验中，MFCF 的 RMSE 收敛为 0.971，BiasMFCF 的 RMSE 收敛为 0.946。其次，BiasMFCF 对应的效率曲线总在 MFCF 的下方，说明当两种算法执行到相同的迭代步骤时，改进算法的推荐性能总是优于原有算法；当两种评分预测算法要达到同样的 RMSE 时，BiasMFCF 会比 MFCF 更有效率。

另一方面，虽然改进算法增加偏置项后，每次迭代都会比 MFCF 消耗更多的存储空间和时间，但是增加的时空开销相对于迭代本身的代价来说是微不足道的。所以，可以说 BiasMFCF 是牺牲了每次迭代一定的时空消耗，带来了算法整体效率和评分预测准确度的提升。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

4.4 本章小结

本章以推荐系统评分预测问题中数据稀疏性问题为切入点，介绍了现有的解决方案。由传统的基于降维思想的 SVD 分解引入到基于矩阵分解的协作过滤评分预测，提出并分析一种引入评分偏置的矩阵分解模型，并通过实验对改进模型的性能作了分析和讨论。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

第五章 总结及展望

5.1 总结

本课题研究旨在针对视频业务领域，针对协作过滤推荐算法中未考虑视频流行度和用户偏好变化的问题，以及数据稀疏性问题，在对现有方法进行分析研究的基础上从用户兴趣度预测的 Top-N 推荐和用户评分预测推荐两个方面设计改进的，适用于视频业务推荐的算法模型。

本文主要研究内容为基于用户偏好的视频推荐技术，从领域背景出发，概括了国内外研究现状，肯定了本课题研究的价值和意义；接着，简要阐述了视频推荐系统的相关技术；然后，从 Top-N 推荐和评分预测推荐两个方面对协作过滤视频推荐进行深入研究和探讨，并进行了详细的实验分析。总结起来，全文的主要价值和创新点在于：

①详细分析了视频推荐系统中用户偏好变化的特征，创新地提出了两种用户偏好变化的度量方式：基于横向用户时间跨度的用户偏好度量和基于纵向用户时间跨度的用户偏好度量，并成功运用于推荐算法的设计中。

②就 Top-N 推荐问题，设计了用户-视频二进制关联模型下基于领域的协作过滤算法模型，并通过引入视频流行度权重和用户偏好变化权重提出了对应的改进算法，使推荐算法的质量得到有效的提升。

③就评分预测问题，针对评分数据稀疏性的典型解决方案—基于矩阵分解的协作过滤，提出了引入评分偏置项的改进模型，通过实验分析，验证了改进算法能够提高评分预测精度。

5.2 进一步工作及未来研究方向

视频推荐技术在当前视频类业务系统中需求旺盛，也是学术界和工业界的研究热点。本文关于协作过滤推荐的改进模型主要集中在学术研究中，还不能成熟的应用到工业中，究其原因如下，它们同时也需要在未来工作中进行进一步研究：

①实际的推荐系统需要处理数量庞大的用户/资源数据，必须要有强大的服务器支撑，对推荐算法的效率有着很高的要求，本文研究内容更多的集中在合理性

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

和硬指标的评测上，对实际系统中最为实际的计算时效问题涉及较少。因此，下一步工作拟将在大规模环境下，借助于成熟的 Map-Reduce 编程模型，对于相关推荐算法进行改进，这应该是一个不错的解决途径。

②实际的视频业务系统中，除了单一的视频业务外，可能还提供了时下非常流行的社交功能。使得用户数据不仅只是用户操作行为数据，还可能包括了用户的社交信息，以及用户的上下文信息（观看视频的时间、地点和心情等）。如何综合利用这些丰富的用户信息来提高实际系统的推荐质量也是下一步需要深入研究的方向。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

致　谢

在硕士论文即将完成之际，回首在电子科大学三年充实而忙碌的研究生学习生活，内心感激与感慨良多。在此，我谨向在科大的岁月里，给予我悉心教导的老师，帮助支持的同窗、好友，还有默默关心疼爱我的家人致以诚挚的谢意。

首先，我要感谢的是我的导师阳小龙教授。阳老师以其扎实的专业知识、勤恳的治学态度和谦逊低调的为人教育和熏陶着我，对我在研究生期间乃至于以后的学习和工作产生着深刻的影响。此外，还要感谢孙健、徐杰和彭云峰三位老师，他们都曾给予了我热切关怀和细心指导，特别是孙老师，常常询问科研与学习情况，给予朋友般的关心。同时还对研究室的隆克平，刘健等几位老师也表示真挚的谢意。

其次，感谢2班的各位同窗，感谢已经毕业的师兄师姐，是他们勤奋的学习态度和优秀的学习方法影响并提高了我。衷心感谢对我在学习上的帮助和生活中的陪伴。也感谢我的好友们，三年的研究生生活中，一起奋斗，相互鼓励，学习之余一起嬉闹，让我的生活充实而多彩。

最后，感谢我的爸爸、妈妈、哥哥，我的女友何雨熹，以及挚友刘潇，是你们在我的研究生生涯中给予我最大的支持，让我不断前进，才能一路顺利走来，谢谢你们。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

参考文献

[1] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval[M]. New York: Addison-Wesley Publishing Co., 1999

[2] Richard MacManus, A guide to Recommender Systems[OL], Read and Write Website. 2010

[3] Mark Levene, An Introduction to Search Engines and Web Navigation, Second Edition, Wiley, 2010

[4] 许海玲,吴潇,李晓东,阎保平.互联网推荐系统比较研究[J]. 软件学报, 2009,2(20): 350-362

[5] Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an informationtapestry[J]. 1992,Commun.ACM35(12), 61–70

[6] Mahmood, T., Ricci, F.: Improving recommender systems with adaptive conversational strategies[J]. In: C. Cattuto, G.Ruffo, F.Menczer (eds.) Hypertext, 2009, 73–82

[7] McSherry,F.,Mironov,I.: Differentially private recommender systems: building privacy into the net[C]. In: KDD ’09: Proceedings of the 15th ACM SIGKDD international conference on Knowledgediscoveryand datamining, 2009, 627–636

[8] Anand, S.S., Mobasher, B.: Intelligent techniques for web personalization[J]. In: Intelligent TechniquesforWebPersonalization,2005, 1–36

[9] Xiwang Yang, Harald Steck, Yong Liu. Circle-based Recommendation in online Social Networks[C]. KDD August 2012

[10] Jie Tang, Sen Wu, Jimeng Sun, Hang Su. Cross-domain Collaboration Recommendation [C]. KDD August 2012

[11] Ziad Al Bawab, George H.Mills, Jean-Francois Crespo. Finding Trending Local Topics i n Search Queries for Personalization of a Recommendation System[C].KDD August 2012

[12] Amit Goyal, Laks V. S. Lakshmanan, RecMax:Exploiting Recommender Systems for Fun and Profit[C]. KDD August 2012

[13] Greg Linden, Brent Smith, Jeremy York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, vol.7(1):76-80

[14] Miller BN, Albert I, Lam SK, Konstan JA, Riedl J. MovieLens unplugged: Experiences with an occasionally connected recommender system[C]. In Proceedings of the International Conference on Intelligent User Interfaces, 2003, 263-266

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[15] James Davidson, Benjamin Liebald, Junning Liu. The Youtube Video Recommendation System[C]. RecSys2010, September 26-30, 2010

[16] Paul Resniek, Hal R. Varian.Recommender systems[J]. Comunications of the ACM, 1997, Vol.40(3):56-58

[17] Gediminas Adomavicius, Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and Possible Extensions[J]. IEEE Transaction on knowledge and Data Engineering, 2005, Vol.17(6):734-749

[18] 王元涛. Netflix 数据集上的协同过滤算法[D]. 北京:清华大学,2009

[19] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl. An algorithmic framework for performing collaborative filtering[M]. 1999

[20] PBo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, PMingjing Li. Online video recommendation based on multimodal fusion and relevance feedback[C]. Proceedings of the 6th ACM international conference on Image and video retrieval, 2007, 73-80

[21] Yohan Jin,Minqing Hu. MySpace Video Recommendation with Map-Reduce on Qizmt[C]. IEEE International Conference on Semantic Computing, 2010, 126-133

[22] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters[J]. Communications of the ACM, 2008, Vol.51(1): 107-113

[23] Kim, J. What is a recommender system?[OL]. In: Proceedings of Recommenders06.com 2006, 1-21

[24] Thompson, C. (2008): If You Liked This, You're Sure to Love That[J]. In: The New York Times Magazine ,November 21, 2008

[25] Agrawal, R. and Srikant, R., Fast Algorithms for Mining Assocaitions Rules[C]. Proc.20th Int.Conf. Very Large Data Bases, VLDB,1994, 487-499

[26] 蔡伟贤.关联分析在入侵检测中的研究与应用[D]. 北京:广东工业大学,2011

[27] 韩家炜.数据挖掘的概念与技术[M]. 北京：机械工业出版社，2012，147-149

[28] Shardanand U, Maes P. Social information filtering: Algorithms for automating of mouth[C]. In Proceeding of the ACM CHI conference, 1995, 210-217

[29] 邹显春，谢中，周彦晖. 电子商务与 web 数据挖掘[J]. 计算机应用，2001，21(5):21-23

[30] 娄兰芳,潘庆先.基于集合运算的频繁集挖掘算法[J].山东大学学报, 2008, 43(11): 54-57

[31] Agrawal, R., Lmielinski, T., Mining Associations Between Sets of Items in Large Databases[C]. Proc.of the ACM SIGMOD Int'l Conference on Management of Data, May, 1993, 207-216

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[32] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. Grouplens: an open architecture for collaborative filtering of netnews[C]. In Proceedings of the 1994 ACM conference on Computer supported cooperative work, NY, USA, 1994, 175–186

[33] Greg Linden, Brent Smith, Jeremy York. Amazon.com Recommendations: Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003, vol.7(1):76-80

[34] Herlocker, J. Understanding and Improving Automated Collaborative Filtering System[D]. Ph. D. Thesis, Computer Science Dept., 2000

[35] Bhatt Chidansh Amitkumar, Kankanhalli Mohan S. Multimedia Data Ming state of art and challenges[J]. Multimedia Tools and Applications, 2011, Vol(51).1: 35–76

[36] Chee, S.H.S., RecTree: A Linear Collaborative Filtering Algorithm[D]. M.Sc.thesis, Computing Science Dapartment, 2000

[37] Chee, S.H.S., Han, J., and Wang, K., RecTree: A Linear Collaborative Filtering Method[C]. Proceedings of Data Warehouse and Knowledge Discovery, Sept.,2001

[38] QiLin Li and Mingtian Zhou, Research and design of an efficient collaborative filtering predication algorithm[J]. Parallel and Distributed Computing, Applications and Technologies, 2003, 171-174

[39] 刘建国,周涛,郭强,汪秉宏.个性化推荐系统评价方法综述[J].复杂系统与复杂性科学, 2009,6(3): 1-9

[40] Linyuan Lu, Matus Medo, Chi Ho Yeung, Yi-Cheng Zhang,Zi-Ke Zhang, Tao Zhou. Recommender systems[J]. Physics Reports 2012, 1-49

[41] Greg Linden, What is a Good Recommendation Algorithm. March[J], Communications of the ACM , 2009

[42] Zhengdong Lu, Deepak Agarwal, and Inderjit S. Dhillon. A spatio-temporal approach to collaborative filtering[C]. In Proceedings of the third ACM conference on Recommender systems, RecSys ’09, New York, NY, USA, 2009, 13–20

[43] Yi Ding and Xue Li. Time weight collaborative filtering[C]. In Proceedings of the 14th ACM international conference on Information and knowledge management, CIKM ’05, New York, NY, USA, 2005, 485–492

[44] 周涛.个性化推荐十大挑战[OL].中国科学网, 2012 年 7 月 4 日

[45] Andre Bergholz, Coping with Sparsity in A Recommender System[C]. Webkdd, 2002, 86-99

[46] Daniel Billsus, Michael J.Pazzani. Learning Collaborative Information Filters[C]. Proceeding of the 15th International Conference on Machine Learning. 1998, 46-54

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

攻硕期间取得的研究成果

获奖情况：

[1] 2010 年，电子科技大学通信与信息工程学院研究生新生入学一等奖学金；

[2] 2011 年，电子科技大学通信与信息工程学院硕士一等奖学金；

[3] 2012 年，电子科技大学通信与信息工程学院硕士三等奖学金；

申请专利：

[1] 孙健，唐明，徐杰，隆克平. 一种引入视频流行度和用户兴趣变化的协作过滤推荐方法. 中国，发明专利（201310111179.3）

[2] 孙健，艾丽丽，谢发川，隆克平，唐明. 一种基于用户关联性的资源个性化推荐方法. 中国，发明专利（201210179907.X）

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

硕士学位论文

MASTER THESIS

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net