江苏大学

硕士学位论文

电子商务个性化推荐算法设计与实现

姓名：刘芳先

申请学位级别：硕士

专业：计算机应用技术

指导教师：宋顺林

20100611

摘 要

电子商务系统在为用户提供越来越多选择的同时，商品信息过载现象越来越严峻，其结构也变得日益复杂，用户想要从商品海洋里迅速便捷地找到自己真正需要的商品越来越困难，于是电子商务个性化推荐系统应运而生。

推荐算法是推荐系统中最核心的部分，在很大程度上决定了推荐系统性能的优劣。协同过滤推荐根据与用户兴趣爱好相同或相似的其他用户的评价产生推荐，个性化程度高，是目前应用最广泛、最成功的推荐技术，但它在实际应用中还存在很多缺陷：如预测评分中用户相似性计算不准确，实时性差、推荐精度不高等。

本文针对推荐系统的实时性要求，提出了基于项目簇偏好的用户聚类算法。该算法首先基于项目属性对项目聚类，得到用户对不同项目簇的偏好，然后利用K-means聚类算法对用户进行聚类，将相同或相似兴趣的用户分到同一类中，这样可以找到离当前用户最近的几个聚类，然后在这几个聚类中搜寻最近邻居，避免了在整个用户群上搜寻，提高了实时响应速度。

K-means聚类算法由于随机选取初始聚类中心，这样得到的聚类结果随机性很大。本文将用户在不同项目簇上的评价差异作为用户距离，采用克鲁斯卡尔（kruskal）算法生成初始聚类中心，使得初始中心靠近类中心，这样得到的聚类更符合实际。

针对传统方法没有考虑项目之间的内容关系而影响推荐精度问题，本文提出了基于项目相关性的协同过滤算法。该算法首先将项目相似性引入到预测评分中的用户相似性计算，避免了不相关项目对用户相似性计算的干扰，其次，在预测评分中增加时间权限，使得越新的用户兴趣在推荐过程中的权值越大。

最后利用MovieLens数据集进行两个实验：最近邻居搜寻效率实验和协同过滤算法实验。前者的度量方法是最小空间内搜索到更多的邻居，实验结果表明基于项目簇偏好的的K-means聚类算法可以在更小的用户空间内搜索到更多的邻居用户，提高了查找用户最近邻的效率和精度；后者以MAE作为评价指标，对本文设计算法和传统算法进行性能比较，实验结果表明本文设计算法得到了更好的推荐效果。

关键词：个性化推荐；协同过滤；项目属性；项目簇偏好；K-means初始聚类中心

ABSTRACT

E-commerce system gives users more and more choices, meanwhile, information overload is grim increasingly and framework of system becomes more complex, then it becomes more and more difficulty for users to find what they like, then e-commerce personalized recommedation system appears.

The recommendation algorithm is the core of the recommedation system, and it determines recommendation results to a great extent. Collaborative filtering system gathers ratings from people of the same interest with the target user and then creates recommendations, and it has a high degree of personalization, so it is the most successful and popular method. However, there are still many deficiencies in practical application, such as inaccurate calculation in user similarity, the real-time response, new-item and accuracy problems.

This paper proposed a clustering users algorithm based on users preference for item sort to meet the needs of real-time. The algorithm firstly clusters items based on attributes, and gets users preference for item sort. Then it uses k-means clustering to cluster users, and lets the users with the same interest in the same class. We can find the user's nearest neighbor from several nearest clusters to avoid the entire users base, and enhance the real-time response speed.

Because the first center of k-means clustering is random, it will result that user clusters are random. This paper uses kruskal algorithm with user difference evaluation on item sort to produce the first centers, and lets the first centers are near to class centers, then gets clusters with high accuracy.

The user-based CF algorithm doesn't consider item relevance,which affects the accuracy, and takes the user's interests in different time into equal consideration, which leads to the lack of effectiveness in the given period of time. In order to revolve these issues, this dissertation advances a CF algorithm based on item relevance. The algorithm adds item relevance to calculate user similarity, then avoids disturbance of unrelevant item, at the same time, it adds time as a weight for computing missing ratings, and makes the interests approaching the gathering time have bigger weight in recommendation process.

In the end, this paper takes two experiments with MovieLens data sets: experiment of searching for nearest neighbor and CF algorithm experiment. The first experiment uses minimum space searching for more neighbors to estimate result, and experiment results show

that k-means clustering based on preference for item sort can find more neighbours from minimal space than k-means clustering, and it improves accuracy of finding neighbour; The second experiment useds MAE to evaluate recommedation quality. Compared the improved recommedation algorithm and traditon recommedation algorithm, experimental results show that the improved algorithm is more precise and gives better prediction in accuracy.

Key words: personalized recommendation; collaborative filtering; item attributes; preference for item sort; the first center of k-means clustering;

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权江苏大学可以将本学位论文的全部内容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和汇编本学位论文。

保　密 □，在　　年解密后适用本授权书。

本学位论文属于

不保密 ☑。

学位论文作者签名：刘芳芳　　指导教师签名：

2010年6月17日　　2010年6月17日

独 创 性 声 明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已注明引用的内容以外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：刘芳光

日 期：2010年6月17日

第一章 绪论

1.1 课题的背景及意义

电子商务是信息时代中产生和发展起来的新事物，也是信息技术和信息化建设的必然产物。随着互联网的普及和企业信息化程度的不断提高，电子商务正以令人难以置信的速度蓬勃发展。电子商务使得个人或企业通过网络，采用数字化电子方式进行数据交换和开展商务活动，目前已拥有在线购物、网上银行、在线支付结算系统、电子票据、网上商情广告等多种类型的电子商务形式。

但是，电子商务系统在为用户提供越来越多选择的同时，商品信息过载的现象越来越严峻，其结构也变得日益复杂，用户在大量的商品信息空间中无法快速便捷地找到自己真正需要的商品。如何对电子商务信息进行有效的组织利用，如何尽可能地了解顾客的兴趣爱好，以优化网站设计，从而方便顾客购物，成为电子商务发展迫切需要解决的问题。于是个性化推荐系统应运而生。

所谓个性化服务，是在顾客浏览 Web 站点时，系统尽可能地迎合每个顾客的浏览兴趣并且通过不断地调整自身布局来适应顾客的兴趣变化，使得每个顾客都有为该 web 站点唯一顾客的感受[1]。其作用主要表现在以下三个方面：将电子商务网站的浏览者转变为购买者；加强电子商务网站的交叉销售能力；提高客户对电子商务网站的忠诚度[2]。

个性化推荐系统使得网站主动适应每个客户的特定需求，为每个客户创建适应客户个性化需求的电子商店，从而为每个客户提供不尽相同的个性化购物环境，为电子商务系统实现“一对一营销”的个性化服务提供了可能。目前，几乎所有大型的电子商务系统，如 eBay、Anlazon、CDNOw、淘宝网、当当网等都不同程度地使用了各种形式的推荐系统。研究表明，电子商务的销售业务使用个性化推荐系统后，销售额能提高 2%至 8%，尤其是书籍、电影、音像、百货等相对价廉且种类繁多的商品[3]。电子商务个性化推荐系统具有良好的发展和应用前景。在日趋激烈的竞争环境下，电子商务推荐系统能够有效保留老客户，发展新客户，提高企业的销售额。成功的电子商务推荐系统将会产生巨大的经济效益和社会效应。

顾客在浏览电子商务网站时都会产生大量的数据信息，不仅有本次的交易信息，还有利用搜索引擎以及在站点内浏览的相关数据，这些数据中包含了对市场分析及预测非常有益的潜在信息。在日益激烈的电子商务竞争中，任何与消费者行为有关的信息对商家来说都是非常宝贵的，但是这些数据资源中所蕴涵的大量有益信息至今却未能得到充分地挖掘和利用。数据挖掘技术为研究用户浏览行为提供了工具，能对电子商务网站上的各种数据进行分析，挖掘出具有实际应用价值的知识模式，使得企业更有效地改善客户关系、更好的运作站点和向客户提供更优质的个性化推荐服务，从而为企业带来更好的效益，有利于提高商业站点的竞争力，同时也方便了用户浏览商品和购物，可谓一举两得。因此，将数据挖掘技术应用于电子商务推荐系统具有非常重要的现实意义。

1.2 国内外研究现状

自 1997 年 Resnick 和 Varian 提出世界上第一个电子商务推荐系统以来，推荐系统在电子商务、网络经济学和人类社会学等领域一直保持很高的研究热度并逐渐成为一门独立的学科。各种推荐算法涵盖包括认知科学、近似性理论、信息检索[4]、管理科学[5]、市场营销建模[6]等在内的众多研究领域[7]。

国外很多研究机构在推荐系统上投入大量精力。ACM 从 1999 年开始每年召开一次电子商务研讨会，其中有很多文章都是研究电子商务推荐系统的。同年，SIGKDD 小组设立 WEBKDD 研讨组，研究内容主题集中在电子商务中的 web 挖掘技术和推荐系统技术上。第 7 届国际人工智能联合会议 IJCAI 把 E-Business&the Intelligent Web 作为一个独立的研讨小组，而 ACM 下面的信息检索特别兴趣组 SIGIR 在召开的第 24 届研究和发展会议上，开始专门把推荐系统作为一个研讨主题。与此同时，第十五届人工智能会议、第一届知识管理应用会议 PAKM 等也纷纷开始将电子商务推荐系统作为研究主题。近几年来，国际学术界出现了大量关于计算机网络信息整合的推荐研究：AcM 设立推荐系统年会；计算机领域的人机交互、数据挖掘和机器学习顶级会议(如 sIGcHI，KDD，sIGIR 等)中，推荐算法的文章逐年增加；国际数据分析领域的高阶期刊(如 IEEE，ACM 等)刊载数篇推荐系统方面的文章。纽约大学(AlexanderTuzhilin)、美国密歇根大学(PaulResnick)、卡内基梅隆大学(Jaimecallan)、微软研究院等都在研究信息领域

的推荐系统，其中，美国密歇根大学在2006年开授了由PaulResnick主讲的推荐系统课程。到目前为止国外已有许多成型系统，如：NEC公司的“V 5-7820”系统，IBM公司P.S.Yu等人研究的SpeedTracer系统等[8]。

个性化服务技术在国内是自2000年以来逐渐成为研究热点的，目前国内学者和研究机构开发了一些个性化服务的原型系统，而一些信息服务商也在其数据库产品中推出了简单的推荐服务功能。清华大学推出的混合推荐系统openBookmark通过集中管理用户群的Bookmark来实现混合推荐；南京大学的潘金贵等人设计并实现了个性化信息检索智能体DOLTRI Agent系统；上海理工大学的陈世平、周福华等研究和开发了面向领域的个性化智能检索系统Myspy，它可实现基于智能代理的信息过滤和个性化服务，其利用同义词词典、蕴涵词词典和辅助词典，对查询词进行概念搜索，返回与查询需求相似的文档；万方数据的iLib系统具有相似资源推荐的功能，可根据用户当前访问的文献资源推荐内容相似的其他资源；国内CNKI的中国期刊全文数据库，除了提供相似资源推荐外，还具有根据文献的引用信息、作者信息进行引用文献、被引文献、同作者文献等推荐的功能。但总体来说，我国电子商务推荐系统相对国外差距较大，起步晚、理论研究落后是影响我国推荐技术发展的直接原因，现有的推荐系统在推荐深度、规模和质量方面都落后于国外。

1.3 个性化推荐系统面临的主要挑战

电子商务个性化推荐是一个新兴的领域，尽管目前已经取得了一定的研究成果，但仍面临很多挑战，主要包括以下几个方面：

(1) 稀疏问题。它是推荐技术中的重要问题之一[9]。在任何大型的推荐系统中，用户和项目的数量非常庞大，并且随着时间的推移会越来越多，那么用户对项目的评价数据也应该越来越多，但是，实际上每个用户不可能对每个项目都进行评价，据统计，一般用户购买商品的总量仅占网站总商品量的1%-2%左右，用户对项目的评价数据也仅如此，造成用户—项目评价矩阵非常稀疏（即稀疏矩阵），这种情况带来的问题是得到用户间的相似性不准确，邻居用户不可靠。

如表1.1和表1.2描述的是不同稀疏程度的用户—项评价矩阵。很明显，表1.1比表1.2的数据更稀疏，那么根据表1.2得到的邻居用户肯定比表1.1准确，

而现实的协同过滤推荐系统中的用户—项评价矩阵的稀疏程度却跟表 1.1 类似，这样得到的邻居用户不准确，从而影响推荐效果。

表 1.1 稀疏的用户—项评价矩阵

项目 / 用户	项目 1	项目 2	项目 3	项目 4	项目 5	项目 6
用户 1	1					
用户 2		4				5
用户 3				3		
用户 4	2				1	
用户 5			2			

表 1.2 稠密的用户—项评价矩阵

项目 / 用户	项目 1	项目 2	项目 3	项目 4	项目 5	项目 6
用户 1	2	5		4	1	2
用户 2		3	1		2	5
用户 3	4		2	3		2
用户 4	2	4		3	1	
用户 5	5		2		4	3

目前解决这问题主要有三种方法：设置初始评分，基于人工智能的方法和基于降维思想的方法。如，采用 Horting 图[10]、聚类[11]、贝叶斯网络[12]及粗糙集[13]等手段，增加用户在项目空间上重叠的数目，以降低数据稀疏性；采用奇异值分解[14]、潜在语义索引[15]、矩阵划分等技术降维，使数据变得更稠密些。

(2) 冷启动问题。它分为新项目和新用户两种问题。在推荐系统中，新项目加入数据库后必须等待一段时间才有用户查看或评价，在评价达到一定数量之前无法对此项目进行推荐，即新项目问题，这在协同过滤推荐系统中尤为突出。目前，一般考虑使用组合推荐的方法来应对。新用户问题是指，系统没有或很少存储新用户的信息，包括查看项目的历史记录和对项目的评价，基于模型的方法无法获得训练数据而基于规则的方法难以进行推理，使得对新用户的推荐无法进行。近期，有用到对象熵、受欢迎程度、用户个性属性等来解决此问题。

(3) 可扩展性问题。由于用户没有对足够多类别的项目进行评价，推荐系统往往无法完全掌握用户每个方面的兴趣和需求，于是就有过拟合现象，即系统推荐给用户的项目与用户刚刚看过的不太相似或不相关。该问题本质上来自于数据的不完备性，这在实际应用中无法完全避免。在信息检索领域，这类问题普遍存

在，解决的主要方法是引入随机性，使算法收敛到全局最优或者逼近全局最优。

(4) 推荐准确度。现有的个性化推荐系统不能很好地根据用户的历史信息和当前的会话进行分析和判断，得出准确的推荐方案，使得经常推荐一些不符合用户兴趣或需求的商务信息或商品，要么干脆一视同仁，给每个人推荐一样的信息，这就导致推荐准确度不高。即使能达到预期效果，那也是在用户额外提供信息以及部分人工分析的情况下得到的，智能化程度有待提高。

(5) 实时性。互联网上存储的信息以指数级增长，使用网络的用户也越来越多，这样要为大量的在线用户提供个性化推荐，实时性很难保证。此外，推荐系统的推荐准确度和实时性是一对矛盾，大部分推荐技术为了保证实时性，是以牺牲推荐系统的推荐质量为代价的。在提供实时推荐服务的同时，如何有效提高推荐系统的推荐质量，有待进一步的研究。

(6) 有效数据挖掘。用户在商务网站上浏览或购物过程中，都会产生大量数据信息，不仅有本次的交易信息，还有利用搜索引擎的信息以及在站点内进行浏览的相关数据，但当前大部分电子商务推荐系统都只利用了这些信息的极小部分进行推荐，从而影响了推荐效果。

(7) 推荐结果解释。推荐系统为了说服用户选择其推荐，需要向用户解释推荐产生的原因，但目前的个性化推荐系统只是通过简单的浏览排行、销售排行以及其他用户对项目的评价信息等方式来达到上述目的。需要进一步研究更加有效的方法向用户解释产生推荐的原因，来增强用户对推荐系统的信任度，从而说服用户选择推荐系统的推荐。

1.4 本文的主要研究内容

本文通过对协同过滤及其在电子商务推荐系统中的应用、面临的问题和挑战以及相应的解决方法进行了详细的分析与研究，提出了一种基于项目簇偏好的用户聚类方法，改进了预测评分过程中的用户相似性计算方法，并考虑了用户兴趣随时间变化的情况，这样便保证了系统的实时性，解决了新项目问题，提升了推荐精度，实验证明，达到了预期效果。

本论文所做的主要工作如下：

(1) 构建项目属性矩阵，基于项目属性相似性进行项目聚类，得到用户对不

同项目簇的偏好，并将项目相似性引入预测评分过程中，避免了不相关项目的干扰，提升了推荐精度，并间接解决了新项目问题。

(2) 基于用户—项目簇矩阵，通过改进的 K-means 聚类算法对用户进行聚类，将有着相同或相似品味的用户分到同一类中，这样就可以在跟目标用户最近的几个聚类中搜寻最近邻，避免了在整个用户群上搜寻，提高了实时响应速度。

(3) 传统推荐算法将用户不同时间的兴趣等同考虑，时效性不足，本论文在预测评分中增加时间权限，使得越新的用户兴趣在推荐过程中越重要，提高了推荐的准确度。

(4) 利用 MovieLens 数据集进行实验，采用绝对偏差 MAE 作为评价指标，对改进算法和原算法进行性能比较。

1.5 本文的组织结构

全文共分为六章，文章结构和各章节主要内容如下：

第一章 绪论

本章主要介绍了本课题的研究背景及意义，国内外研究现状，并简单介绍了本文研究的主要内容和文章组织结构。

第二章 基本概念及相关技术

本章主要介绍了电子商务个性化推荐系统的基本理论，并简单介绍了协同过滤推荐技术以及其中涉及到的数据挖掘、聚类等技术。

第三章 基于项目簇偏好的用户聚类

本章将项目属性相似性引入项目聚类，在此基础上利用用户对不同项目簇的偏好信息，采用改进的 K-means 聚类算法将有相同或相似品味的用户聚为一类，缩小了搜索最近邻范围，减少了搜索最近邻的时间，满足系统实时性要求，并间接解决了新项目问题。

第四章 基于项目相关性的协同过滤推荐

本章针对传统方法中不相关项目对用户相似性计算的干扰问题，提出基于项目相关性的用户相似性计算方法，并在预测评分的过程中增加时间权限，使得接近采集时间的用户兴趣在推荐过程中具有更大权值。

第五章 实验与分析

在前两章给出改进算法的基础上，应用 MovieLens 数据集对改进后的算法进行验证，并对实验结果进行详细分析。

第六章 总结

对本课题的研究工作进行总结，并对下一步的工作提出展望。

第二章 电子商务个性化推荐系统

电子商务正成为贸易发展的新方向，它不再受时空的限制，改变了贸易形态，加速了商品流通，缩短资金周转时间，有效地降低企业生产成本，使得企业从有限的资源中获得更大的利润，提高了竞争力。电子商务已成为世界经济市场中必不可少的组成部分，但在购物过程中，它需要用户逐个浏览商品及商家信息，这与人们日常的购买行为是有差异的，具体表现在以下方面：(l)用户获得的各种商品信息仅仅是事先定义好的静态子目录，有时也会有些图片或文字描述信息，但用户却不能近距离观察或触摸商品，这样缺乏真实感；(2)用户需要花费很多时间来浏览商品信息，并对所有商品进行比较。显然，这种方法是低效率的，而且随着商品信息的增多，信息过载越来越严重，用户往往花费大量时间获得的却不是自己所需要的信息，这势必影响用户的购物兴趣。这些问题集中反应了第一代电子商务系统在智能化和自动化程度上的不足，理想的做法是把客户真正需要的信息直接提供给客户，使得以商品为中心转变为以客户为中心，创建个性化服务的电子商务推荐系统，于是，第二代电子商务系统——电子商务个性化推荐系统应运而生。

2.1 个性化推荐系统概述

电子商务推荐系统正式的定义是Resnick & Varian在1997年给出的：“它是依据电子商务网站向客户提供的商品信息，帮助用户决定应该购买什么商品，模拟销售人员帮助客户完成购买过程”[16]，它根据用户的兴趣爱好推荐符合用户兴趣爱好的商品，因此也称电子商务个性化推荐系统。

2.1.1 个性化推荐系统的作用

电子商务个性化推荐系统和销售系统、决策支持系统既有相同之处又有区别。销售系统是帮助销售人员把商品销售出去；而决策支持系统是帮助生产者决定什么时候生产什么产品，其目的是为产品生产企业服务。而推荐系统是帮助用户决定购买什么商品，是面向用户的系统。推荐系统的服务对象是用户，系统的

目标是为用户提供项目推荐。用户是指推荐系统的使用者，即电子商务网站中的用户。项目是被推荐的对象，即电子商务网站中的商品或服务，也就是最终推荐系统推荐给用户的内容。

电子商务个性化推荐系统不仅能为用户服务，而且能给电子商务网站带来丰厚的商业利益。主要体现在以下几个方面：

(1) 提升电子商务网站的服务质量。推荐系统可以挖掘用户兴趣，帮助网站的设计者调整站点的逻辑映射，达到方便用户的目的，增加了用户的满意度。

(2) 提高用户的忠诚度。在电子商务环境下，用户要去浏览竞争者的网站只需简单的几次点击操作，因此提高用户的忠诚度是商业竞争中的一个重要营销策略。要提高用户的忠诚度，就要增加站点的吸引力，这除了有更好的站点内容外，还需要为用户提供一个方便快捷浏览兴趣商品的途径。试想，如果用户每次购买商品的时候，推荐系统都可以对其进行高效的商品推荐，无疑用户下次会继续在该网站上进行商品选购。而且，一对一的个性化推荐系统还可以延长用户在站点的逗留时间，增加了商品销售的可能性。

(3) 将浏览者转变为购买者。有时站点的访问者只是随便浏览，并无购买意向，如果这个时候推荐系统能够有针对性地向其提供高质量的商品推荐，就有可能引起访问者的购买兴趣，从而从访问者转变成购买者。

(4) 增加交叉销售。推荐系统可以根据用户当前购物车中的商品向他们推荐同这些物品相关的商品。比如，用户购买了笔记本电脑，网站可以向他推荐软件光盘等。这样很有可能提高站点的交叉销售量。

2.1.2 个性化推荐系统的框架及流程

电子商务个性化推荐系统的完整框架主要由用户交互代理、推荐引擎、推荐模型库、数据仓库、数据挖掘引擎、操作数据库等构成，如图 2.1 所示。

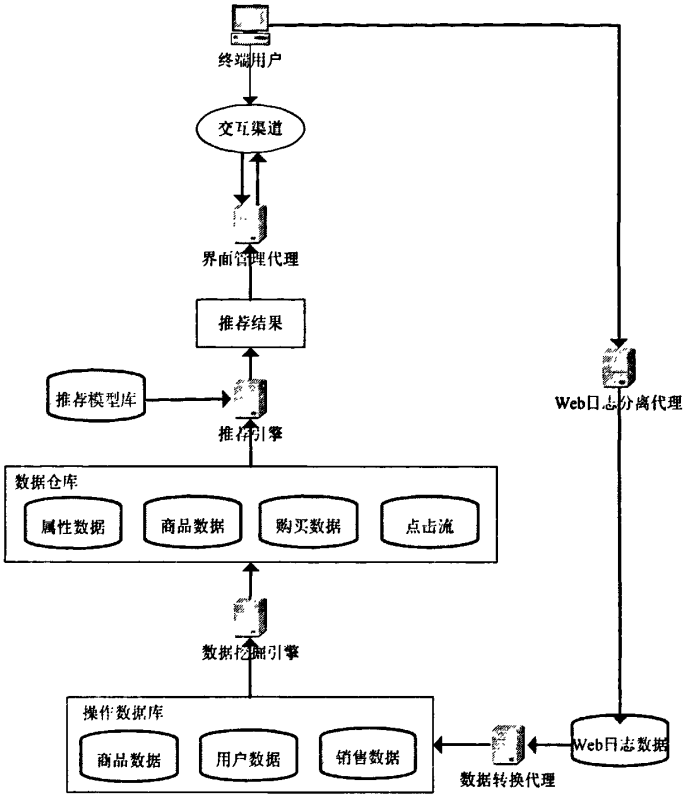

图 2.1 个性化推荐系统整体框架

- 用户交互代理：管理用户界面，接收用户的请求，并提供推荐结果给用户。
- 推荐引擎：主要功能是接收推荐请求，运行推荐策略，产生推荐结果。推荐引擎对外提供了统一的推荐服务接口，对内规范了推荐算法的运行环境，方便了推荐算法的编制。
- 推荐模型库：存储推荐算法，不同推荐技术采用不同推荐算法。
- 数据仓库：存储推荐系统直接操作的数据，即那些经过清洗和初步挖掘后的规整数据，包括属性数据、商品数据、购买数据、点击流等。
- 数据挖掘引擎：初步挖掘操作数据库中的数据，从中抽取出具有一定关联性且能被推荐算法直接采用的有意义的数据。
- 操作数据库：存储用户操作需要使用的数据，包括了商品数据库、用户数据库、销售数据库等。

个性化推荐系统需要完成从数据信息采集到产生推荐的一系列工作，其完整的应用流程具体来说包括以下几个部分：

(1) 清洗、转换和加载数据：将经过数据转换代理清洗、转换的数据，送入

数据挖掘引擎进行初步挖掘，然后再加载到数据仓库中成为规整数据。所选数据形式多种多样，可以是评分数据，也可以是交易数据，应该选择什么样的数据由具体的推荐应用决定。

(2) 生成模型：根据具体的推荐应用，提取对应的规整数据，选择适当的推荐模型产生针对此具体推荐应用的模型，并将其存储在推荐模型库中，作为一个可用模型。怎样选择适当的推荐模型产生新的模型要视具体的推荐应用而定。

(3) 配置推荐策略：是指推荐过程的配置，其中包括推荐算法和推荐模型。具体的推荐功能是由推荐引擎运行相应的推荐策略来实现的，所以推荐引擎要实现推荐服务，就必须有已经配置好的推荐策略。配置工作主要是根据具体推荐应用修改推荐策略，这包括选择不同的推荐算法和推荐模型，并请求推荐引擎启动或重载此策略。

(4) 访问推荐服务：电子商务系统直接向推荐引擎提供当前用户的信息，并请求用指定的推荐策略产生商品的推荐列表。推荐引擎则根据电子商务系统的请求运行对应的推荐策略，产生合适的推荐结果。

(5) 更新操作数据：电子商务系统在开展网络商业活动和提供推荐服务的同时，新用户、新商品在不断的增加，而且用户的活动也是不断变化的，那么操作数据库也是在发生变化的，为了能进行有效地推荐，则需要及时更新操作数据库。

整个个性化推荐系统应用流程是一个不断循环的过程，当操作数据库变化到一定程度的时候，就要更行数据仓库、推荐模型，以便能及时的反映出当前用户的兴趣变化。模型的更新由具体的应用要求决定，一般采用周期性更新，也有采用推荐效果反馈闭值进行控制的。

2.1.3 个性化推荐系统的评价指标

目前，绝大多数个性化推荐系统都是利用准确度评价推荐系统的好坏。由于不同推荐系统的目标不同，而且评价指标缺乏标准化，因此很难对不同系统的推荐效果进行比较。针对不同的系统，已有的准确度指标有预测准确度、分类准确度、排序准确度、预测打分关联、距离标准化指标和半衰期效用指标等等，下面简单介绍常用的预测准确度、分类准确度。

(1) 预测准确度：衡量推荐系统的推荐评分和用户对对应项目的实际评分之

间的差别。它包括三个常用的标准：平均绝对偏差 MAE[17]、根平均方差 RMSE 和标准平均绝对误差 NMAE[18]。

- ✓ 平均绝对偏差 MAE：它通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性。MAE 越小，推荐准确度越高。假设预测的用户评分集合表示为 $\{p_1,p_2,\ldots,p_n\}$，对应的实际用户评分集合为 $\{q_1,q_2,\ldots,q_n\}$，则平均绝对偏差 MAE 的计算公式如（2.1）所示。

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{2.1}$$

- ✓ 根平均方差 RMSE：它使得偏差大的分量在最终的偏差中所占的比重较大。和平均绝对偏差一样，根平均方差越低，推荐准确度越高。它的计算公式如（2.2）所示。

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(p_i-q_i)^2}{n}} \tag{2.2}$$

- ✓ 标准平均绝对误差 NMAE：它在评分值区间内作标准化，使得可以在不同的数据集上对推荐效果进行比较。NMAE 越低，推荐质量越好。其计算公式如（2.3）所示。

$$NMAE=\frac{MAE}{(q_{\max}-q_{\min})} \tag{2.3}$$

其中，$q_{\max}$ 和 $q_{\min}$ 分别为用户评分区间里的最大值和最小值。

(2) 分类准确度：是指用户是否喜欢某个产品的判定正确的比例。广泛使用的这类指标有准确率、召回率。准确率定义为系统的推荐列表中用户喜欢的产品和所有被推荐产品的比率；召回率指推荐列表中用户喜欢的产品与系统中用户喜欢的所有产品的比率。准确率和召回率在一定程度上是一对相对矛盾的指标，为了平衡两者，通常采用综合评价指标 F-measure[19]：

$$F-measure=\frac{2}{1/precision+1/recall} \tag{2.4}$$

其中，*precision* 为准确率，*recall* 为召回率。

在实际应用中，发现准确率高的推荐系统并不能保证用户对其推荐结果的满意度。推荐系统不仅需要高的准确率，还需要得到用户的认可，而后者才是更本质的，因此除了准确率之外，度量推荐系统的评价指标还包括推荐的流行性和多样性、覆盖率、新鲜性和意外性以及用户的满意度等指标。

2.1.4 个性化推荐系统的分类

个性化推荐系统有不同的划分。根据推荐对象的特点，目前主要有两类个性化推荐系统：一种是以网页为对象的个性化推荐系统，主要采用 Web 数据挖掘技术，为用户推荐符合其兴趣爱好的网页；另一种是以商品为推荐对象的个性化推荐系统，为用户推荐符合其兴趣爱好的各类商品，这种推荐系统是一般意义上的电子商务个性化推荐系统。

依据个性化推荐系统采用的技术不同，可将其分为：基于聚类分析的个性化推荐系统、基于规则的个性化推荐系统、基于知识的个性化推荐系统和基于Agent的个性化推荐系统等。

由于推荐系统有两个分类标准，即自动化程度和持久性程度，如果个性化推荐系统也依此为标准的话，则可分为：基于商品特征的个性化推荐系统、基于相关商品的个性化推荐系统和基于相关客户的个性化推荐系统。自动化程度是指用户为了得到推荐系统的推荐需要显性输入信息的程度；持久性程度是指推荐系统产生的推荐是基于用户当前的单个会话还是多个会话，是暂时的还是持久的。基于商品特征的个性化推荐系统是根据用户输入的其偏好的商品特征进行推荐；基于相关商品的个性化推荐系统主要根据商品的聚类，推荐用户偏好商品的相似商品；基于相关客户的个性化推荐系统即基于用户的协同过滤推荐系统，它根据用户对商品的评价找到有相似爱好的用户，再使用相似用户的观点对目标用户产生推荐。

2.2 个性化推荐方法

个性化推荐方法是个性化推荐系统中最核心的技术，很大程度上决定了推荐系统性能的优劣。电子商务个性化推荐方法大致可以分为主动式推荐和被动式推荐。主动式推荐是指系统根据对用户信息和行为的分析，给出符合用户需要的商

品或信息；而被动式推荐是指用户通过自己的努力在系统的帮助下获得所需要的商品或信息，如网络信息的浏览、关键字查询等。被动式推荐方法主要有分类浏览和关键词搜索，智能化程度低，不能发现用户的潜在兴趣与需求，因此目前研究比较多的是主动式推荐，主要有协同过滤推荐、基于内容的推荐、基于关联规则的推荐、基于用户统计信息的推荐、基于效用的推荐、基于知识的推荐等，下面进行简单的介绍。个性化推荐技术分类如图 2.2 所示。

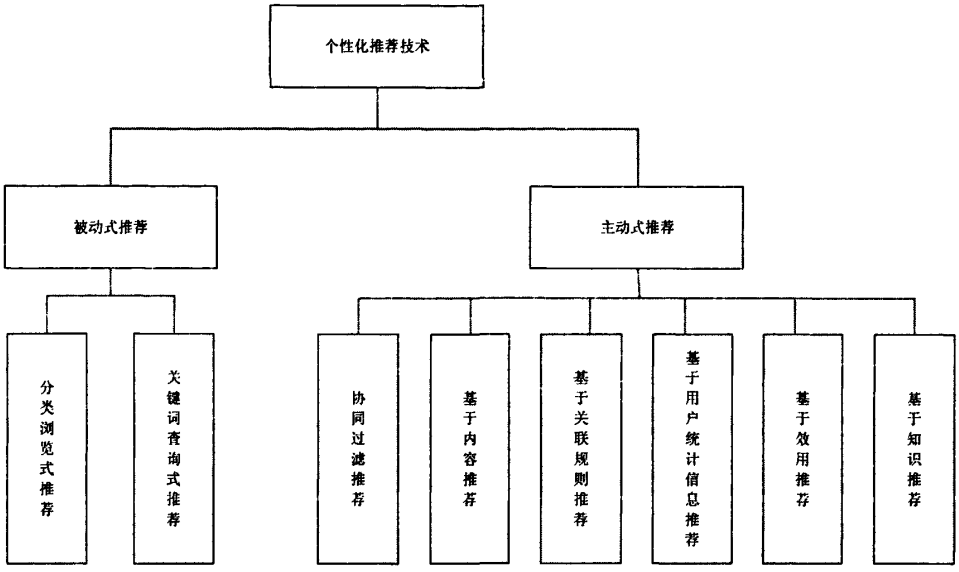

图 2.2 个性化推荐技术分类图

2.2.1 协同过滤推荐

协同过滤推荐是目前研究最多、应用最广的推荐技术，它的基本思想是根据与当前用户具有相似观点的用户的行为对该用户进行推荐或者预测，个性化程度高。协同过滤推荐就是根据一个用户对其它项目的评分以及相似用户群的评分记录来预测这个用户对某一未评分项目的评分。对协同过滤最早的研究有 Grundy system[20]，后来的研究成果包括 Tapestry system[21]，GroupLens[22]，Ringo[23]，PHOAKS system[24]，Jester system[25]等。Tapestry 是最早提出的个性化协同过滤推荐系统。用户需要明确指出与自己兴趣爱好相似的其他用户，推荐系统根据指定的其他用户对商品的评价产生推荐结果。GroupLens 是最早提出的的自动个性化协同过滤推荐系统，用于从大量的新闻中搜索用户感兴趣的新闻列表。

由于协同过滤推荐只需知道用户对项目的评价，而无需关心项目的具体内

容，所以其最大优点是对推荐对象即项目没有特殊要求，能处理非结构化的复杂对象，如音乐、电影等，它可以实现跨领域的推荐，能发现内容上完全无关的项目，用户对推荐的内容是预料不到的。但也存在许多问题，如用户对商品的评价矩阵非常稀疏即稀疏问题、冷开始问题，随着系统用户和商品的增多，系统的性能越来越低即可扩展性问题等等。

图 2.3　协同过滤推荐系统构成

协同过滤系统出输入，预测引擎，输出结果三部分构成，如图 2.3 所示。协同过滤推荐系统的输入可以是用户当前的行为，也可以是用户的访问历史。在大型的电子商务系统中，为了产生高质量的推荐，推荐系统可能需要多种类型的输入信息。协同过滤推荐系统的输出形式主要包括相关商品信息、个体对商品的评分等等。

假设一个推荐系统有 m 个用户和 n 个项目，那么这个系统可表述为一个 $m\times n$ 的用户—项评价矩阵 $R=(r_{ij})$，其中 r_{ij} 表示第 i 个用户对第 j 个项的评价值，即用户 i 对项目 j 的兴趣度，具体指用户是否浏览了该项或者对该项的喜好程度，如果用户 i 没有对项目 j 进行过评分，一般我们令 r_{ij}=0。那么协同过滤推荐可以看成是预测评价矩阵 R 中缺失元素，并选出预测评价值最高的几个项目进行推荐。典型的协同过滤推荐流程如图 2.4 所示：

(1) 用户输入他对项目的评价信息。这是协同过滤系统的关键之一，因为后面的输出结果是以此为依据的。评价的方式有两种：显式和隐式。显示方式就是需要用户的额外工作来表明他的兴趣，比如要求用户直接给出对某一项目的评价值或填写问答式表格主动告诉系统他的兴趣。隐式方式是指在用户寻找信息过程中，系统同步跟踪用户的行为，然后根据行为科学和心理学的研究结论来分析用户兴趣，这个过程是用户不知道的。比如，根据用户对项目的收藏或浏览时间长短，再或者是用户的购买记录来断定用户对哪些项目是感兴趣的。

(2) 预测引擎收集所有评价值，从现有信息中归纳出用户兴趣模板，然后根据兴趣模板利用协同过滤技术从海量信息中过滤掉用户不感兴趣的。预测引擎为

目标用户返回一个 top-N 的有序序列，或者是当一个目标用户提供给预测引擎一个项目列表后，预测引擎返回这个列表中各个项目的预测评价值。一般情况下，用户不知道推荐结果是怎么产生的，也就是说预测引擎对用户而言犹如“黑盒”。

(3) 输出预测结果。预测结果主要有两种形式，一种是推荐项目列表，另一种是预测项目评价值。推荐项目列表是一个提供给目标用户的具有 N 项他最喜欢的项目列表，典型的如 top-N 推荐集。预测项目评价值是系统对给定项目的一个评价值。

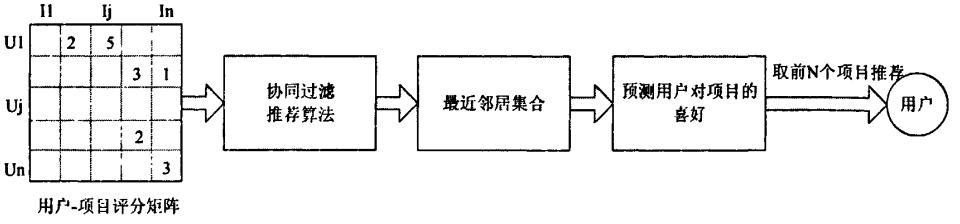

图 2.4 典型协同过滤推荐流程图

协同过滤算法大致可分为两类：基于内存的协同过滤算法和基于模型的协同过滤算法。

(1) 基于内存的协同过滤算法[26]：基于用户—项评价矩阵，用统计方法计算得出跟目标用户具有相似兴趣爱好的邻居用户，再根据邻居用户的评价计算得出项目预测值，选出满足一定条件的项目进行推荐。依据算法过程中所使用的不同事物的关联性，它又分为基于用户的协同过滤算法和基于项目的协同过滤算法。

基于用户的协同过滤算法的核心是假设人和人之间的行为具有某种程度的相似性，即购买行为类似的顾客，那么他们很可能会购买类似的产品。随着电子商务系统规模的扩大，用户和项目数目指数级增长，评价矩阵极端稀疏，使得计算得到的邻居用户不准确，推荐质量急剧下降。于是 Sarwr 教授提出基于项目的协同过滤算法，该算法基于这样的假设：如果大部分用户对一些项的评分比较相似，则目标用户对这些项的评分也比较相似。它的思想是根据用户对相似项的评价预测该用户对目标项的评价。

(2) 基于模型的协同过滤算法[27]：先用诸如用户对商品的评价信息、用户购买信息等历史数据得到一个模型，再用此模型进行预测。其广泛采用的技术包括聚类技术、贝叶斯网络、机器学习方法、关联规则、神经网络、潜在语义索引等。

基于内存的协同过滤算法是对各个单个用户进行预测，而聚类技术则是将具有相近相似度的用户分为一类，然后以此进行预测。该技术通过观察与分析将用户集或项目集划分为多个类或簇，使得同一簇中的对象具有较高的相似度，不同簇中的对象相似度很低。基于聚类的协同推荐首先将用户或项目进行聚类，依据每类用户具有相近的兴趣或每类项目有相似特性，于是就依据所属类的共同喜好或特性向用户进行推荐。

贝叶斯网络建立的是一个概率模型，在训练得到的网络结构中，每个节点都有一组对其有影响的父结点，一旦知道父节点的值，就可以预测该节点的值。贝叶斯网络建立的模型小巧快捷，推荐精度不亚于基于内存的协同过滤算法，但模型建立的时间复杂度较高，适合于变化较少的环境。

2.2.2 其他推荐技术

1. 基于内容的推荐

基于内容的推荐技术[28]是信息过滤技术的延续与发展，它利用信息检索技术(IR)分析项目的内容，通过属性特征来定义项目，应用邻居函数和分类技术分析和聚类项目，系统基于用户对项目的评价来学习用户的兴趣，最后依据用户兴趣与待预测项目的匹配程度进行推荐。基于内容的推荐主要集中在文本信息推荐领域，比如，新闻组过滤系统 NewsWeede[29]、Personal Web Watcher[30]、InfoFinder[30]等。

基于内容的推荐技术对每个用户独立操作，不需要考虑其他用户的兴趣，也不存在评价级别问题，即使是没有任何用户购买或评价的商品也能推荐给用户，这就不存在稀疏问题、新项目问题和级别问题。另一方面，分析商品的属性和相关性可以脱机进行，因而推荐响应时间快。但是它也存在一定的局限性，首先只能获得项目特征的部分信息，通常是文本信息，其它的信息如图像、音频、视频等内容被忽略了。其次，由于是针对用户的浏览或购买历史寻找有相似特征的项目进行推荐，这导致推荐的资源过于狭窄，无法向用户产生跨种类的推荐，缺乏新颖性，局限了用户的视野，再加上用户的兴趣爱好并不是一成不变的，因此此方法适应用户兴趣变化的能力较低。要得到一个准确的用户模型，需要使用大量的用户数据进行长期的训练，才能产生较为准确的推荐。另外，对一些难以用几

个属性描述的商品，这种方法也不适用。

2. 基于关联规则的推荐

基于关联规则的推荐以关联规则为基础，把已购商品作为规则头，推荐对象作为规则体。所谓关联规则[31]，即在一个交易数据库中统计购买了商品集X的交易中有多大比例的交易同时购买了商品集Y，得到的关联规则表示为$X \Rightarrow Y[s\%, c\%]$，$s$表示关联规则的支持度，$c$表示关联规则的置信度。关联规则挖掘技术可以发现不同商品在销售过程中的相关性，在零售业中已经取得了成功，如最有名的购物篮分析，沃尔玛利用此方法分析出“啤酒与尿布”的关联性，将“啤酒”和“尿布”这两个看上去并没有关系的商品摆放在一起进行销售，取得了很好的销售收益，成为营销届的神话。

基于关联规则的系统分为在线和离线两部分，离线部分主要功能是从 Web 日志中提取用户事务，产生频繁项集，将过滤的规则入库；在线部分首先提取当前用户事务，将其送入推荐引擎，推荐引擎再从规则库中提取匹配规则并结合网络结构将最佳推荐页面推荐给用户。其推荐流程如图 2.5 所示。

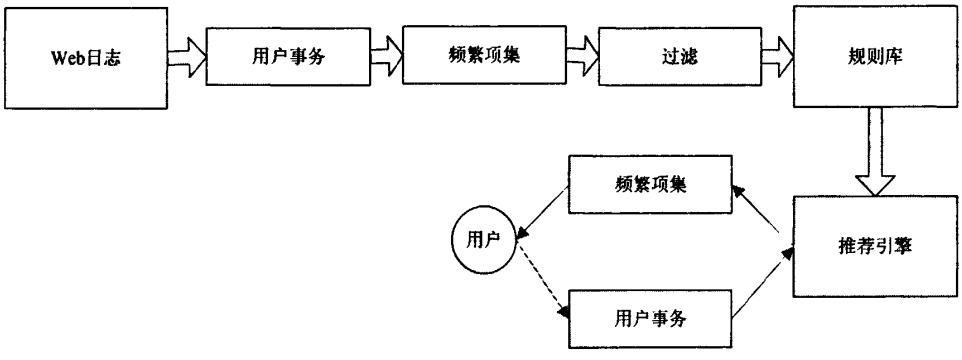

图 2.5 关联规则推荐流程图

使用关联规则的推荐系统有 Surflen 和 ASARM[26]。关联规则的优点在于能发现表面看似无关的项目之间的相关性，由此产生推荐，增加销售。但关联规则的发现最为关键且最耗时，是算法的瓶颈，虽说可以离线进行，但规则质量很难保证且不能动态更新，随着规则数量的增多系统将变得越来越难以管理。另一方面，由于它是根据被购商品之间的关系来建立商品项目之间的关系的，因此个性化程度不高，且在数据集高维稀疏的情况下会导致弱规则。此外，商品名称的同

义性问题也是关联规则的一个难点。

3. 基于用户统计信息的推荐

基于用户统计信息的推荐是基于用户个人属性对用户进行分类，再按分类对各类中的用户进行推荐。这种推荐系统最早的有 Grundy，最近的有 Krulwich，Grundy 是从人机对话获取用户基本信息，在此基础上推荐书籍给用户，而 Krulwich 使用简单问卷调查的方式收集用户信息用于用户分类，再向用户推荐产品或服务。该方法最大的优点是不需要反映用户兴趣爱好的历史数据，此外，由于它不需要其它用户的评价数据，因此也没有冷开始问题。所以，在一些系统中，有在基于用户统计信息方法的基础上，再用机器学习分类器的方法进行推荐的例子。但该方法也需要统计用户的信息，如果获得的用户信息不够多，那它也存在着稀疏性问题，另外，分类效果也不会很好。在电子商务环境中，基于用户信息的推荐系统往往很难实施，因为考虑到个人隐私问题，很多用户不愿意公开自己的个人信息，这为推荐的开展造成了很大困难。

4. 基于效用的推荐

基于效用的推荐需要项目的属性特征数据，得到合适的用户对项目的效用描述函数，然后用该效用函数对所有项目进行排序，取前 N 个项目作为对目标用户的推荐。其核心问题是如何为每个用户创建效用函数，因此，用户资料模型很大程度上是由系统所采用的效用函数决定的。采用这种技术的有 Me-a-Tete 和电子商务网站 PersonaLogic2。基于效用推荐的最大优点在于它能把项目的非自身属性，如提供商的可靠性、产品的可获得性等，考虑到效用计算中，这样能在决策的时候考虑诸如到货时间之类的非项目自身因素问题，体现了推荐的周全性，增加了个性化的程度。基于效用的推荐不是基于历史评价数据的，所以不存在冷开始、数据稀疏和系统可扩展性等问题。但该项技术的关键和难点是如何设计出考虑周全且性能良好的效用函数，且该函数的使用不具有通用性，只能局限在具体的某个推荐系统内，因此缺乏灵活性。基于效用的推荐对只有少数几个特征的项目是适合的。

5. 基于知识的推荐

基于知识的推荐[32]不是建立在用户需求和偏好基础上的推荐，而是利用针对特定领域制定的规则来进行基于规则和实例的推理，所以从某种程度上来说它是一种推理技术，提出了使用功能知识进行推理的概念。所谓功能知识，是指关于某个项目如何满足某个特定用户需求的知识，它能解释需要和推荐之间的关系。在基于知识的推荐看来，用户资料可以是任何支持推理的知识结构，并不一定是用户的兴趣爱好（用户—项目评价矩阵），当然，功能知识也可以是详细的用户需要表示，如 Entree 系统，利用基于案例的推理技术实现基于知识的推荐，文献[33]中的系统根据饭店间的不同菜式，对顾客进行饭店推荐。基于知识的推荐所使用的知识可以有多种形式，如 Google 使用了网页间的超链接，Entree 使用了餐馆之间的相似度。该技术的优点是对用户信息的需求相对较少，也不需要用户的历史偏好数据，因此不受数据稀疏问题的困扰，同时，能在已有的用户知识基础上推荐出尽可能广而全的项目。但最大的缺点是有效知识的获取很难，而且不能发现用户的新兴趣，做出具有“奇异发现”的推荐。当然，如果在建造知识库前能发现产品的相关联知识，那该技术也可以推荐不同类型的项目给用户，但要做到这点比较难。

6. 组合推荐

由于各种推荐方法都有优缺点，所以在实际中常采用组合推荐，从而弥补或避免各种推荐技术的弱点，其中研究和应用最多的是内容推荐和协同过滤推荐的组合[34]。组合推荐一个最重要的原则，就是通过组合后要能避免或弥补各自推荐技术的弱点，以达到更好的推荐效果。由于目前并没有一种非常完美的推荐方法，所以要实现一个现实的推荐系统，组合推荐的思路非常有必要，于是有研究人员在组合方式上提出了七种组合思路[35]：

(1) 加权(weight)：采用多种推荐技术得到对某一项目的预测评分，根据权重相加得到总评分，以此得出推荐结果。

(2) 变换(switch)：随着问题背景和实际情况的变化，采用不同的推荐技术。

(3) 混合(mixed)：同时采用多种推荐技术给出多种推荐结果，为用户提供参考。

(4) 特征组合(feature combination)：组合来自不同推荐数据源的特征用于另一种推荐算法进行推荐。

(5) 层叠(cascade)：先用一种推荐技术产生粗糙的推荐结果，再采用第二种推荐技术在此基础上进一步作更精确的推荐。

(6) 特征扩充(feature augmentation)：一种推荐技术的输出结果作为另一种推荐技术的特征输入。

(7) 模型放大（Meta-level)：用一种推荐方法产生的模型作为另一种推荐方法的输入。

目前，采用的比较多的组合推荐技术有以下三种：Content-based/collaborative feature augmentation hybrid，这是应用研究最早的组合推荐技术，目前的应用有基于内容的 filterbot 系统；Collaborative/content-based meta-level hybrid，用协同的信息去生成用户的全面评价数据集，然后基于该数据集进行不同项目的比较；Collaborative/demographic augmentation hybrid，先利用协同技术识别出邻居用户，再用邻居用户的信息作为用户统计信息。由于用户的评价信息相对容易获得，所以内容推荐和协同推荐的组合研究得比较多。跟基于用户统计信息的推荐的组合研究较少，其中的一个重要原因是存在在线数据保密问题。此外，在采用组合推荐时，必须注意组合策略。如对于基于内容和协同过滤的技术，不论如何组合，总是存在初始化问题，因为它们都需要历史的评价数据；但若把基于协同过滤与基于知识或效用的推荐技术相结合，则能有效解决初始化问题，因为后两种技术不需要许多用户历史数据。

2.3 相关技术

电子商务个性化推荐系统是一个多模块、多功能的大型智能系统，汇集了信息检索、数据挖掘和数据仓库等技术。这些技术相互融合使得电子商务个性化推荐系统高效运作。可以说，这些技术是构成整个电子商务个性化推荐系统的基石。

2.3.1 数据挖掘

数据挖掘又被称为数据库知识发现[36]（Knowledge Discovery in Databases, KDD)，是指从数据源（如数据库、文本、图片、万维网等）中探寻有用的、有

潜在价值的，并且可以被理解的模式或者知识的过程。数据挖掘是一门多学科交叉的学科，它常采用的技术包括机器学习、统计、数据库、数据仓库和 OLAP、人工智能、信息检索、可视化和神经元网络等不同领域的技术，并且其在诸如零售、通信、银行、保险、基因分析、股票市场分析、Web 挖掘等不同领域得到了应用。目前数据挖掘能够从关系数据、对象关系数据、文本数据、多媒体数据、时间序列、空间数据、异质数据等多种数据源中挖掘知识。

一般而言，数据挖掘的过程大致分为如下三个步骤进行：

- ✓ 数据预处理：原始数据通常都不适合直接用来挖掘，其中有多个原因，首先，它可能包含噪音和异常情况，必须经过过滤。其次，数据量可能非常大，并且包含有不相关的属性，需要通过采样和选择特定属性来降低数据量。
- ✓ 数据挖掘：将经过预处理的数据送到数据挖掘算法中，生成模式或知识。
- ✓ 后续处理：在许多应用中，并不是所有被发现的模式都是有用的，这一步就是要识别出对具体应用而言有用的部分。有很多评估和可视化技术可用来做此项决策。

整个数据挖掘过程是可迭代的，一般都要通过多轮迭代才能获得最终结果，来促进现实世界的各项工作。

数据挖掘技术已经成功应用于电子商务个性化推荐系统，通过数据挖掘技术对用户行为和属性进行学习分析，从中获取有价值的知识——用户的兴趣，根据得到的知识产生推荐。电子商务个性化推荐系统中的数据挖掘主要有关联规则挖掘和分类挖掘两类：

- ✓ 关联规则挖掘：用于发现大量数据中项目之间的有用关联性或相互联系。基于关联规则的推荐系统根据生成的关联规则推荐模型和用户的购买行为向用户产生推荐。关联规则推荐模型的建立可以离线进行，因此可以有效保证推荐的实时性要求。
- ✓ 分类挖掘：分类挖掘模型根据对象的相关信息通过多种机器学习方法将对象划分为不同类别，这类方法有：聚类、贝叶斯网络等。聚类技术可将众多用户划分为不同的用户组，并将用户相似兴趣作为不同用户群的特征。基于聚类的推荐系统就是将具有相似兴趣的用户分配到一组中，根据相似用户的兴趣对目标用户进行推荐。聚类一旦产生后，推荐准确

度较高，系统性能好，但是聚类过程要花费很长时间，所以一般离线进行，此外，对处于聚类边缘的用户的推荐效果不理想。贝叶斯网络技术利用训练集创建相应的模型，用决策树表示模型，用节点和边分别表示用户和用户间的关系。这个模型很小，推荐效率高，其推荐效果几乎和最近邻技术一样精确，但建立模型需要花费大量时间，而且不适用顾客兴趣变化大的情况，因为它不能快速反映数据的变化。本文利用分类挖掘中的聚类技术对用户进行分类，为后面的推荐做准备，下面详细介绍聚类技术。

2.3.2 聚类

聚类是数据挖掘中用于发现数据分布和模式的一项重要技术。聚类[36,38]是一个将数据集中在某些方面相似的数据成员进行分类组织的过程，处于相同聚类中的对象具有较高相似度，而处于不同聚类中的对象彼此差别很大。一个公司希望通过市场营销战略来推销自己的产品，最有效地方法是根据每个客户的兴趣和收入情况为每个人制定一套个性化营销策略，但当客户数量很大时，这样做的成本十分昂贵，如果公司为了节约成本而采用对所有客户只制定一套营销策略，这种一视同仁的做法往往效果很差。解决这个问题性价比最高的方案是根据客户的相似度把全部客户划分成兴趣爱好不同的组，然后分别为每组客户制定一套营销策略。因此，在电子商务个性化推荐系统中聚类是很有必要的，其中客户的分类通常采用的是聚类算法。

聚类的形式化定义[39]：对给定的n个数据对象，按照它们之间的相似性划分为k个集合，即满足如下条件：令$S=\{X_1,X_2,...,X_n\}$代表n个数据对象构成的集合，$C_1,C_2,...,C_k$代表k个不同的划分集合（簇或类），则满足：

$$C_i \neq \Phi, \quad i=1,2,\cdots,k;$$

$$C_i \cap C_j = \Phi, \quad i=1,2,\cdots,k; \quad j=1,2,\cdots,k, \text{且} i \neq j;$$

$$\bigcup_{i=1}^{k} C_i = S.$$

聚类与分类都是对一组对象进行归类，但二者的归类方法不同。在聚类开始

之前用户并不知道要把数据分成多少组，也不知道分组的具体标准，聚类是根据一定的聚类规则，将具有某种相同特征的数据聚在一起，是一种无指导学习；而分类，用户开始就知道数据将分为几类，据此总结出判别规则，再将要处理的数据分入不同的类别，它是一种有指导学习。因此，与分类相比，聚类不依赖于预先定义好的类，也不需要训练集。

一个聚类过程的效果取决于度量标准的选择，为了度量对象之间的相似程度，需要定义一些相似度量标准，如用$sim(x,y)$表样本x和样本y的相似度，当x和y相似时，$sim(x,y)$的取值大，当x和y不相似时，$sim(x,y)$的取值小，而且常常将相似度标准化为$0 \le sim(x,y) \le 1$。但在通常情况下，聚类算法是用特征空间中的距离即两个样本间的相异度作为度量标准的。对于某些样本空间来说，距离的度量标准可以是度量的或半度量的，以便用来量化样本的相异度。一般情况下，用样本间的距离$d(x,y)$表示样本x和样本y间的相异度。当x和y相似时，距离$d(x,y)$的取值就小；当x和y不相似时，$d(x,y)$的取值就大。

1.相似度计算

常用的相似度计算方法有余弦相似度、连接相似度及 Pearson 相关系数，下面对这些方法进行简单介绍。

✓ 余弦相似度：

$$sim(x,y)=\frac{x\cdot y}{\|x\|\cdot\|y\|}=\frac{\sum_{i=1}^{n}x_i\cdot y_i}{\sqrt{\sum_{i=1}^{n}x_i^{\,2}}\cdot\sqrt{\sum_{i=1}^{n}y_i^{\,2}}} \tag{2.5}$$

其中x，y为两个样本的n维特征向量，$sim(x,y)$的取值范围为[0，1]。

✓ 连接相似度：

$$sim(x,y)=\frac{|x\cap y|}{|x\cup y|} \tag{2.6}$$

连接相似度实际上是 Jaccard 系数，在基于类别、布尔类型数据的聚类中常用到，它的取值范围也为[0,1]。例如：若x，y分别表示两个顾客的购买商品序列，则连接相似度表示两个顾客共同购买的商品数在他们购买的商

品总数中的比例。

✓ Pearson 相关系数[40]：

$$sim(x,y)=\frac{\sum_{i\in I_{xy}}\left(r_{xi}-\overline{r_x}\right)\left(r_{yi}-\overline{r_y}\right)}{\sqrt{\sum_{i\in I_{xy}}\left(r_{xi}-\overline{r_x}\right)^2}\cdot\sqrt{\sum_{i\in I_{xy}}\left(r_{yi}-\overline{r_y}\right)^2}} \tag{2.7}$$

其中，$sim(x,y)$为用户 x 与用户 y 间的皮尔森相关系数，表示两用户的相似度，集合 I_{xy} 是用户 x 和 y 共同评分的项目集，r_{xi}，r_{yi} 分别表示用户 x、y 对项目 i 的评分，$\overline{r_x},\overline{r_y}$ 分别表示用户 x 和 y 对所有项目的平均评分。

2.距离相异度计算

设 $x=(x_1,x_2,\cdots,x_n)$，$y=(y_1,y_2,\cdots,y_n)$是两个具有 n 维特征的对象，则两对象间的距离常用明可夫斯基距离函数[36]表示：

$$d(x,y)=\sqrt[r]{\sum_{i=1}^{n}|x_i-y_i|^r} \tag{2.8}$$

当 r 取不同的值时，上述公式演化为一些特殊的距离度量。

当 $r=1$ 时，其演变为绝对值距离：

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i| \tag{2.9}$$

当 $r=2$ 时，其演变为欧式距离：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}|x_i-y_i|^2} \tag{2.10}$$

根据数据在聚类中的积聚规则以及应用这些规则的方法不同，聚类算法大致可分为以下几种：划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法和模糊聚类方法等。下面仅简单介绍本文所用到的两种方法。

1. 划分方法

对于具有 n 个数据对象的数据集，采用目标函数最小化的策略把数据集分成 k 个组，每组称为一个簇，这就是划分方法。这种聚类方法必须具备以下两个条件：（1）每组至少包含一个数据对象；（2）每个数据对象必须属于且仅属于一个

组。最著名与最常用的划分聚类算法是 k-menas 算法。

k-menas 算法以 k(最终分类的个数)为参数，把 n 个对象分成 k 个簇，每个簇有个聚类中心，它是这个族中所有对象的均值，使簇内的相似度较高，簇间的相似度尽量低。相似度是指对象跟聚类中心的距离。

k-menas 算法[36]的过程如下：首先，随机选取 k 个对象作为初始的聚类中心；然后计算其余对象与各个聚类中心的距离，将这些对象分配到距离它最近的簇；最后重新计算所有的聚类中心。这个过程将不断重复直到满足某个终止条件，这个条件可以是以下任何一个：（1）没有对象被重新分配给不同的族；（2）没有聚类中心再发生变化；（3）误差平方和（SSE）局部最小，$SSE=\sum_{j=1}^{k}\sum_{x\in C_j} dist(x,m_j)^2$，其中，$C_j$ 表示第 j 个聚类，m_j 表示这个聚类的中心，$dist(x,m_j)$ 表示 x 和 m_j 之间的距离。

2. 模糊聚类方法

在现实世界的问题中，网络对象的聚类不仅没有纯粹的界限，而且有相当一部分的重叠，而且由于各方面的原因，数据集中存在不完整的数据，这将导致聚类效果不理想。基于模糊集理论提供了解决不完整信息和重叠聚类问题的方法。

在 Zadeh[37]于 1965 年提出模糊集之后，模糊集理论在理论发展和应用两个方面都得到了广泛的关注。模糊聚类是利用模糊等价关系将给定的对象分成一些等价类，这种聚类方法不需要事先确定聚类的数目，而是通过一定的阈值来确定对象的相似类别。模糊等价关系满足自反性、对称性，由此得到与关系相应的模糊相似矩阵。然而聚类分析是利用模糊等价关系所对应的矩阵进行分类的，即该模糊矩阵还要满足传递性，因此必须根据相似矩阵求其传递关系的闭包，然后在传递关系的闭包上实现分类。具体步骤如下：

- ✓ 计算对象之间的相似度，按下面的方法建立模糊相似矩阵 $M_{m\times n}$：

  当 $i=j$ 时，$M_{ij}=1$；

  当 $i\neq j$ 时，M_{ij} 为对象 i 和 j 之间的相似度。

- ✓ 计算相似矩阵 $M_{m\times n}$ 的传递闭包，得到等价矩阵。

- ✓ 设置不同的阈值 λ 确定相应的截集，即对矩阵 $M_{m\times n}$ 的模糊相似关系图进

行模糊聚类。

2.4 小结

本章简单介绍了个性化推荐系统的框架、推荐技术和评价标准等，详细介绍了基于协同过滤的个性化推荐系统，为下文奠定了理论基础。数据挖掘技术是个性化推荐系统常采用的技术，也对其有个简单的介绍。最后描述两种常用聚类算法，为下文的聚类提供了理论依据。

第三章　基于项目簇偏好的用户聚类

个性化推荐系统中应用最成功、最广泛的是协同过滤推荐，它通过用户兴趣的相似性决定是否把其他用户的爱好提供给目标用户。该方法被广泛应用，在于它能对快速更新的用户信息做出反应，因为它每次都是根据最新资料重新寻找跟目标用户有着相似兴趣爱好的用户来进行推荐的，也就是说每次都要寻找“同类”用户即近邻用户才能给出推荐结果。随着电子商务系统的进一步扩大，在一个用户和商品均数以万计的系统中，同时为数以万计的用户提供实时推荐服务越来越困难，提高推荐速度迫在眉睫。而协同过滤推荐算法的主要时间在于寻找近邻，那么提高推荐速度的一个有效的方法就是压缩寻找空间，这样就可以大大减少寻找近邻的时间，从而提高整个算法的效率。于是便提出了基于聚类的协同过滤推荐算法。

稀疏矩阵是协同过滤推荐面临的挑战之一，它也是导致推荐精度不高的主要原因之一，因此有必要对用户－项评价矩阵的表示方式进行相应处理，项目之间潜在的内容相似关系是不可忽视的，因为用户往往对内容相似的项目有同样的兴趣，那么应该在一类项目集合上寻找用户的相似兴趣，而不是在每个项目上寻找。再者，相对用户—项评价矩阵而言，项目之间的潜在关系更加稳定可靠。本章首先应用模糊聚类技术从项目的属性特征上对项目进行聚类，在此基础上，依据用户对不同项目集的兴趣偏好对所有用户聚类，使得同一聚类内用户对项目集的评价尽可能相似，而不同聚类间用户对项目集合的评价尽可能不同，然后在和目标用户最相似的几个聚类中搜寻邻居用户，从而在尽量少的用户空间上搜索尽量多的邻居用户，压缩了寻找空间，大大降低了寻找时间，提高了推荐速度，并且最终的推荐是基于用户对项目所属类别的偏好形成的，这便涉及了用户各个方面的爱好，推荐便更全面、更准确。

3.1 基于项目属性特征的项目聚类

从传统的基于用户的协同过滤推荐技术的基本原理可以看出，对用户的推荐仅仅是依赖于用户－项评价矩阵，事实上，用户—项评价矩阵相当稀疏，那么可能存在由于没有相同项目的评价导致两个兴趣爱好相似的用户无法相互推荐的

情况，还有新项目由于没有任何用户的评价而不能推荐给用户，即新项目问题；另一方面，每个项目都有各自的属性特征，所有的属性特征值构成了项目这个实体本身，不同的属性值可以区别不同的项目实体，而且项目之间的属性特征差异很大（即项目的多内容问题[41]），但这并不在传统推荐方法的考虑之中，这些缺陷往往使得邻居用户不够准确，而影响推荐精度。本文将项目的属性特征引入项目聚类，为后文的处理用户一项评价矩阵的表示方式作准备，以改善数据稀疏问题，新项目也得以解决，而且基于属性特征的项目关系更稳定可靠，这有利于准确推荐。

3.1.1 项目的属性相似性

前面已经提到项目有各自的属性特征，并且项目之间的属性特征差异很大，那么也就可以用项目之间的属性值相似度来度量项目相似度。下面以电影为例说明项目之间存在的属性关系，有四部电影《叶问》、《阿凡达》、《风声》及《功夫之王》，分别用 A、B、C、D 来代替，它们有不同的属性特征，如图 3.1 所示。

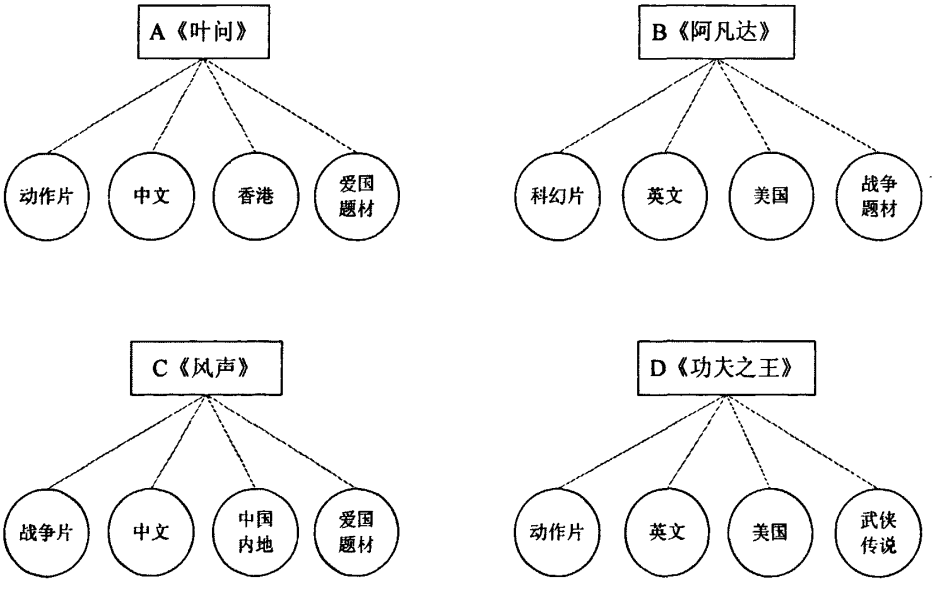

图 3.1 四部电影的属性图

由上图可知，A、D 有一个共同特征：动作片，那么它们有一定的相似性，这种由于项目特征属性带来的相似性叫做项目属性相似性。A、C 具有两个共同属性：中文、爱国题材，一个相似属性—出产地，若不考虑用户评价的影响，则有 A、C 的相似性大于 A、D 的相似性，即 $sim(A,C) \succ sim(A,D)$。由于 B、D 具

有两个相同属性，则有 $sim(B,D) \succ sim(A,D)$，由于 A、D 除了有两个相同属性外还有一个相似属性，则有 $sim(A,C) \succ sim(B,D) \succ sim(A,D)$。

这种原本客观存在的影响因素在传统的推荐算法中没有体现出来，这是因为传统的方法都只是单一的考虑用户评价对项目相似性的影响，即项目之间的相似性仅仅取决于用户对项目的评价，这显然和实际情况有所出入。因此，我们可以利用项目的特征值来计算项目之间的相似性，而且这也解决了新项目问题，因为在传统方法中，新项目因为没有用户的评价而得不到推荐，引入项目特征相似性后，可以根据目标用户评价的其他项目跟新项目的属性相似性对新项目进行推荐。

为了计算项目之间的相似性，我们要构造项目属性矩阵。假设每个项目是一个 t 维向量，每个维度表示项目的一个属性特征，并且每个属性特征有固定的属性值，这里我们采用布尔值(0，1)，那么项目属性矩阵 $A=(a_{ij})$ 如表 3.1 所示，其中，$a_{ij} \in \{0, 1\}$，表示是否具有某种属性，如具有某种属性则值为 1，否则为 0。鉴于不同的应用领域对属性特征有不同的偏重，本文将每个属性特征在最终推荐中所起的贡献不同赋予它不同的权值，然后加权得到两个项目之间的综合相似度。

表 3.1 项目属性矩阵

属性 / 项目	属性 1	属性 2	…	属性 t
项目 1	0	1	…	0
项目 2	1	0	…	1
· · ·	· · ·	· · ·		· · ·
项目 n	1	1	…	0

设项目 i 和 j 在 t 维空间上的属性值分别看作向量 $i=(i_1,i_2,\cdots,i_t)$，$j=(j_1,j_2,\cdots,j_t)$，那么项目 i 和 j 之间的相似性计算公式可表示成：

$$sim(i,j)=\sum_{k=1}^{t}\frac{w_k}{1+d(i_k,j_k)} \tag{3.1}$$

其中，w_k 是指第 k 个属性特征在最终推荐中的贡献值，一般情况下，它由专家或推荐领域的特点来决定。$d(i_k,j_k)$ 为项目 i 和 j 在第 k 个属性特征上的绝对

值距离，见 2.3.2 节的公式（2.9），则$\frac{1}{1+d(i_k,j_k)}$表示项目i和j在第k个属性特征上的相似性。

3.1.2 引入模糊聚类技术

模糊聚类算法建立起的样本对于类型的不确定性描述，能够比较客观地反映现实世界，因此，人们开始用模糊的方法来处理聚类问题，并称之为模糊聚类。实际上，模糊聚类比传统的聚类方法更加有效，这是因为传统聚类方法把目标对象强行地划分到某个簇中，并且它只能属于这个簇，也就是说同一个对象不能属于两个或两个以上的簇，再者，如果目标对象处于簇的边缘，那它也有可能处在另一个簇的边缘，而在具体处理时只将它看成其中一个簇的成员，这显然是不符合客观实际，存在弊端。但模糊聚类不同，是一种软聚类，对象相对于每个簇都有一个隶属度，可以说它属于所有簇，只是跟每个簇的接近程度不同，这更符合现实世界、更准确。同时模糊聚类将项目之间在相似关系上可能含有传递关系的特性考虑在内，所以为了尽可能发掘出项目之间的内在关系，本文采用模糊聚类技术对项目进行聚类，步骤如下：

(1) 利用项目属性矩阵计算项目之间的相似性，建立模糊相似矩阵$M_{n\times n}$。其中，n表示系统中的项目总数，M_{ij}为项目i和j之间的相似度$sim(i,j)$，如公式（3.1）所示。

(2) 计算相似矩阵的$M_{n\times n}$传递闭包，得到等价矩阵。

计算相似矩阵的传递闭包一般采用平方法：$M\to M^2\to(M^2)^2\to\cdots\to M^{2k}=\hat{M}$，但它的时间复杂度为$O(n^3\log_2 n)$，如果$n$值特别大，势必会影响总的计算时间，所以本文采用基于图连通分支计算的模糊聚类最佳算法[42]完成项目的聚类。

建立对应于相似矩阵$M_{n\times n}$的模糊关系的无向图$G(V,E)$，其中V代表项目集合，E是满足$M_{ij}\geq\lambda$的边集合，λ为所选取的聚类阈值。采用计算无向图的连通分支算法求$G(V,E)$的连通分支。方法如下：计算$[x]=\{y\mid M(x,y)\leq\lambda\}$，集合$[x]$即为模糊聚类的等价类，也就是对应相似关系图$G(V,E)$上的连通分量，该分量中的项目集合构成了所求模糊聚类集合。通过设置不同的阈值λ，求模糊相似关

系无向图在不同阈值λ上的连通分量，每个连通分量就是对应这个聚类阈值的一个模糊等价类。

矩阵合成运算的递推如下：

$M_{ij}^{(0)} = M_{ij} \quad 0 \le i, j \le n;$

$M_{ij}^{(k)} = \max\left\{M_{ij}^{(k-1)}, \min\left[M_{ik}^{(k-1)}, M_{kj}^{(k-1)}\right]\right\} \quad 0 \le i, j \le n, 0 \le k \le n.$

这种方法比传统的传递闭包方法的时间复杂度至少降低了$O(n\log_2 n)$，它的$T(n)$满足$O(n) \le T(n) \le O(n^2)$。这种算法在进行大规模数据的模糊聚类时，它的计算时间是可以被实际应用所接受，比较适用于推荐系统领域中的项目数目很大的情形。

模糊聚类完成后，不仅得到了项目的模糊分类集合，即项目在属性特征上的相似性分类；而且还得到了基于传递关系的模糊等价关系矩阵，即项目之间在属性特征上的相似关系值。这一步推荐系统可采用离线的方式构造，这样将不会影响到推荐系统的实时性。

模糊聚类后会形成相应的项目模糊簇，用下列公式计算每个项目相对于某个模糊簇的隶属度。

$$\mu_j(x_i) = \frac{\|x_i - m_j\|}{\sum_{k=1}^{c}(\|x_i - m_k\|)} \quad i = 1,2,\cdots,n;\ j = 1,2,\cdots,c. \tag{3.2}$$

其中，x_i是项目的属性特征向量，m_j是模糊簇中心的属性特征向量，$\|x_i - m_j\|$是指项目i与模糊簇j的相似程度，c是模糊簇的数目。

通过稀疏的用户－项评价矩阵中用户的评价值和项目属于各个模糊簇的隶属度，构造如表 3.2 所示的密集的用户－项目簇的评价矩阵，矩阵中的值代表用户u对项目簇j的评价。

表 3.2 用户一项目簇评价矩阵

项目簇 / 用户	C_1	C_2	…	C_c
u_1	pc_{11}	pc_{12}	…	pc_{1c}
u_2	pc_{21}	pc_{22}		pc_{2c}
⋮	⋮	⋮		⋮
u_k	pc_{k1}	pc_{k2}	…	pc_{kc}

其中，用户 u_i 对项目簇 C_j 的评价为：

$$Pc_{ij}=\frac{\sum_{k\in I_i}P_{ik}\times\mu_j(x_k)}{\sum_{k\in I_i}\mu_j(x_k)}\quad j=1,2,\cdots,c. \tag{3.3}$$

其中，Pc_{ij} 是用户 u_i 对项目簇 C_j 的评价值，P_{ik} 是用户 u_i 对项目 k 的评价值，I_i 是用户 u_i 已评价过的项目集。

在同一应用领域中，同一用户对具有相似属性特征的项目的评价应该是相近或相似的，因此，在属性特征上相似的项目可以形成一个大类，用户对这个大类的评价就代表用户对该类中包含的所有项目的评价。构造用户一项目簇评价矩阵有以下几个优势：一是对用户一项评价矩阵的降维，因为簇的数目是远远小于项目的数目，低维矩阵的计算必定比高维矩阵的计算耗时要少，从而提高推荐方法的可扩展性；二是在实际应用中，一般来说用户对同类项目的喜好程度基本上是一致的，那么由用户对类别的评价代替用户对单个项目的评价并不会影响到用户喜好的表达；三是用户一项评价矩阵是稀疏的，而用户一项目簇评价矩阵是密集的，这可解决由评价数据稀疏性造成的相似群度量不准确的问题；四是新项目问题也得到解决，因为系统中有新项目（即没有任何人评价的项目）时，将新项目基于属性特征聚类到项目簇中，这样便可实现对用户的推荐。

3.2 基于项目簇偏好的用户聚类算法改进

随着电子商务的快速发展，用户数量也是急速攀升的，在整个用户空间内寻找最近邻居，花费的时间比较大，系统实时性则很难保证。本节将利用上节的用

户－项目簇评价矩阵用 K-means 聚类算法对用户聚类，使得具有相似项目类别爱好的用户在同一个簇中，并且只在用户所属的簇中搜索最近邻，从而降低搜寻时间，提高了推荐速度。就目前的技术来看，要想达到最佳推荐效果，还是需要用户注册的，所以本文也是针对注册用户的推荐，那么用户聚类就可以离线完成，这也减少了推荐系统的时间开销。

3.2.1K-means 聚类算法

K-means 聚类算法是目前应用于大规模数据集聚类的最广泛聚类算法，主要是因为算法时间复杂度为 $O(nkl)$，算法空间复杂度为 $O(n+k)$，其中 n 为样本数量，k 为聚类的数目，l 为算法收敛时已迭代的次数。通常 k 和 l 是预先给定的，因此算法的时间复杂度和空间复杂度与数据集的大小呈线性关系，算法的效率非常高。所以在此采用该算法对用户聚类，下面对算法进行简单描述[43]：

(1) 从样本集 $\{x_1,x_2,\cdots,x_n\}$ 中随即选取 k 个点 $c_1,c_2,\cdots,c_k$ 作为 k 个聚类集合的初始中心；

(2) 对每一个样本向量 x_i 计算样本同聚类中心的距离 $\|x_i-c_j\|$，将 x_i 分配给 $\|x_i-c_j\|$ 值最小的那个聚类，其中 $i\in\{1,2,\cdots,n\}$，$j\in\{1,2,\cdots,k\}$；

(3) 根据各聚类集合中的点计算新的聚类中心 $c_1',c_2',\cdots,c_k'$，令 $c_i=c_i'$，其中 $c_i'=\frac{1}{n}\sum_{x_j}x_j$，$i=1,2,\cdots,k$，$x_j$ 为第 i 个聚类中的样本点；

(4) 重复(2)(3)步，直到聚类中心不再发生变化为止。

3.2.2 基于项目簇偏好的 K-means 算法改进

K-means 聚类算法虽然在空间及时间复杂度上有一定的优势，但由于 k 个初始聚类中心是随机选取的，倘若选择的 k 个对象比较接近，各个聚类之间就不会分得很开，这样往往得到的是局部最优解，势必会影响聚类效果。而且从不同的初始聚类中心出发，得到的聚类结果是不同的，准确率也不一样，随机性很大。由此可见，如何选取初始聚类中心对聚类结果至关重要。

本文将针对初始聚类中心的选取提出一种改进的 K-means 聚类算法，以期望得到较好的聚类结果。要想得到全局最优解，就要从整个数据空间分布着手选取初始聚类中心，使得它们分别代表这样的数据集合，就是同一聚类中对象相似度最大，不同聚类中对象相似度最小的数据集合。数据结构中最小生成树算法—克鲁斯卡尔（kruskal）算法的思想是每次选取整个数据集中权值最小的边加入到生成树中，这正是我们所需要的。我们将用户看成是数据空间中的顶点，而用户之间的距离看成是带权值的边，根据 kruskal 算法构造最小生成树的过程来求解初始聚类中心。要指出的是，这里的用户距离是指用户在不同项目簇上评价的差异。

K-means 聚类算法的初始聚类中心形成过程：

(1) 将每个用户表示成空间中的一个顶点，两两用户之间的距离表示成两点之间的连线即边，将用户—项目簇评价矩阵的行向量看成是用户向量，采用欧式距离计算用户之间的距离，如公式（2.10）所示。这样全体用户集合就可以表示成一个连通网 $G=(V,E)$，其中，V 是所有用户顶点的集合，E 是所有边的集合。

(2) 令最开始的状态为只有 n 个顶点而无边的非连通图 $T=(V,\ )$，每个顶点自成一个连通分量。

(3) 在 E 中选择距离最小的边，若该边依附的顶点落在 T 中不同的连通分量上，则将此边加入到 T 中，否则舍去此边而选择下一条距离最小的边。依次类推，直到部分顶点的连线构成了一个环，将构成环的所有顶点添加到同一个集合 M 中，并在集合 T 中删除这些顶点。

(4) 重复第(3)步，直到所有的顶点都分配到 k 个集合中。

(5) 计算各个集合的中心作为 k 个初始聚类中心。

得到 k 个初始聚类中心后，将其应用于 K-means 聚类算法，完成对用户的聚类。

下面举例说明求解初始聚类中心的过程，如图 3.2 所示，系统中有五个用户 v_1,v_2,v_3,v_4,v_5。

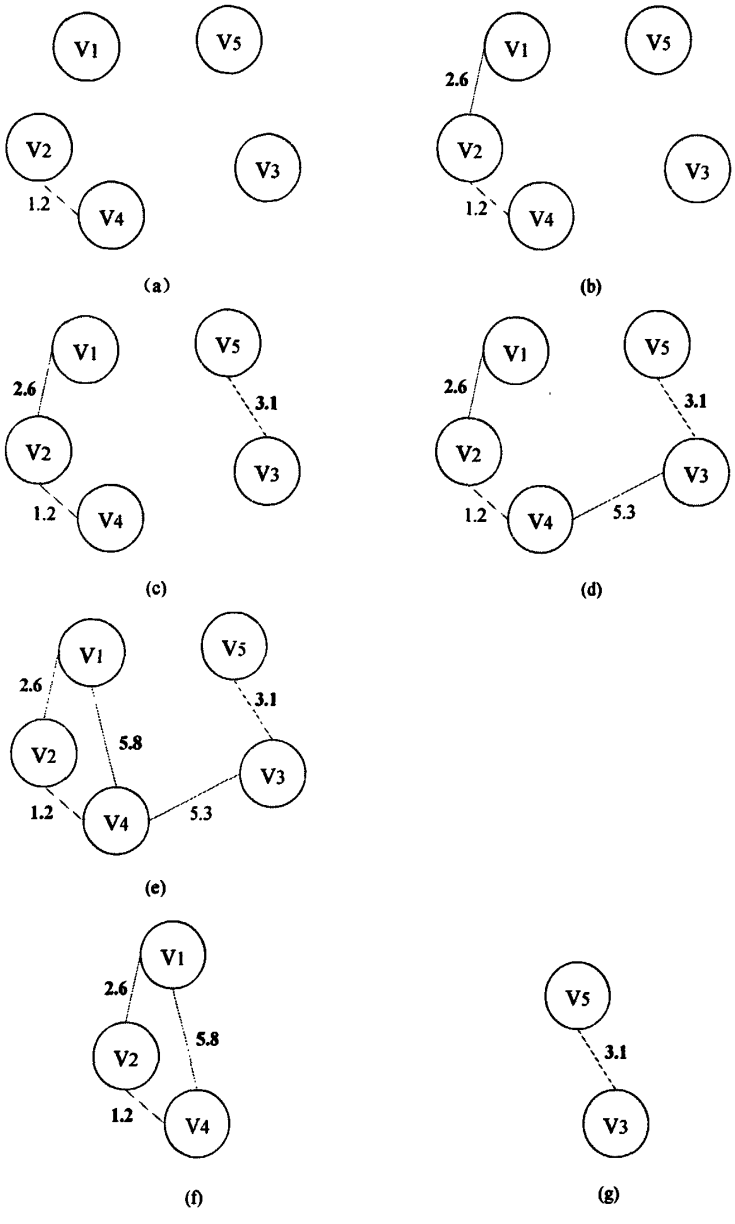

图 3.2 求解初始聚类中心过程举例

首先在用户群中选择相似度最大的两个用户 v_2 和 v_4，将其连线，如(a)所示，接着在用户群中继续选择相似度最大的两个用户，将其连线，如此继续，直到有环生成，如状态(e)，输出组成环的所有节点，最后得到两个集合，如(f)和(g)。每个集合中的对象相似度最大，分别计算这两个集合的平均值，作为的初始聚类中心，这两个聚类中心在空间上的分布与数据的实际分布是一致的，再将它应用

于K-means聚类算法进行用户聚类。

3.3 算法设计

基于用户—项目簇偏好通过K-means聚类算法对用户进行聚类的具体算法表述如下：

输入：聚类数目k。

输出：k个用户聚类。

方法：

(1) 假设系统中有n个用户，令V为所有用户的集合，令E为网络中边的集合，令T为所选边的集合，并初始化T为空，即$T=[\ ]$；

(2) $while(E\neq[\ ])\&\&(|T|\neq n-1)$

{令(u,v)为E中距离最小的边，则将(u,v)从E中删除，即$E=E-(u,v)$；

将u,v归并到集合M中，并从V中删除u,v，即$V=V-u-v$；

$if((u,v)$加入T中会产生环路$)$

输出集合M；

}

(3) 重复执行(2)，知道形成k个集合；

(4) 计算各个集合的中心作为k个初始聚类中心；

(5) 将得到的k个初始聚类中心分别赋值给$c_1,c_2,\cdots,c_k$，将用户—项目簇评价矩阵的n个行向量看成是用户向量，计算n个用户向量同聚类中心的距离$\|x_i-c_j\|$，将用户向量x_i分配给$\|x_i-c_j\|$值最小的那个聚类，其中$i\in\{1,2,\cdots,n\}$，$j\in\{1,2,\cdots,k\}$；

(6) 根据各聚类集合中的点计算新的聚类中心$c_1',c_2',\cdots,c_k'$，令$c_i=c_i'$，其中$c_i'=\frac{1}{n}\sum_{x_j}x_j$，$i=1,2,\cdots,k$，$x_j$为第$i$个聚类中的用户向量；

(7) 重复(5)(6)步，直到聚类中心不再发生变化为止。

这里要对步骤(5)中用户向量同聚类中心的距离$\|x_i - c_j\|$进行说明。3.1.2 节计算得到的用户－项目簇评价矩阵反映的是用户对不同项目簇的评价，那么由此矩阵得到的用户的相似关系群就是与用户品味相似的用户集合。协同过滤推荐算法认为如果两个用户评价趋势相近，就意着这两个用户在选择项目时的兴趣也是相近的，因此，如果两个用户对项目簇的评价一致，意味着用户之间兴趣的一致。在这里我们用用户－项目簇评价矩阵计算用户在项目簇上的相似性，这样将具有相似品味的用户放在同一类中，而不同类的用户的品味差距尽量大。由于用户相似度越大，用户之间的距离越小，所以在步骤(5)中，我们将用户向量分配到其离聚类中心最小的那个聚类，也就是将他分配给相似度最大的那个聚类。用户之间的基于项目簇的相似度见公式（3.4）所示。

$$sim(u_i,u_k)=\frac{\sum_{j=1}^{c}\left(Pc_{ij}-\overline{Pc_i}\right)\left(Pc_{kj}-\overline{Pc_k}\right)}{\sqrt{\sum_{j=1}^{c}\left(Pc_{ij}-\overline{Pc_i}\right)^2}\cdot\sqrt{\sum_{j=1}^{c}\left(Pc_{kj}-\overline{Pc_k}\right)^2}} \tag{3.4}$$

其中，$sim(u_i,u_k)$是用户u_i和用户u_k基于所有项目簇的相似度，Pc_{ij}、Pc_{kj}分别表示用户u_i、u_k对项目簇C_j的评价值，见公式（3.3）。$\overline{Pc_i}$、$\overline{Pc_k}$分别表示用户u_i、u_k对所有项目簇的平均评价值。

第四章　基于项目相关性的协同过滤推荐

分析传统的协同过滤算法，会发现在推荐过程中计算用户相似性只考虑了用户之间是否有共同评价项目，而对项目之间是否相关，以及这种相关对推荐结果是否有影响等问题并未考虑，实际上，项目的相关性会影响推荐结果；另一方面，推荐过程没有体现用户的兴趣是随着时间而不断变化的，没有突出用户某一时段的高兴趣度，从而影响了推荐质量。针对以上问题，本文提出了一种改进的协同过滤推荐算法，针对用户近邻计算和项目评分预测两个关键步骤，提出基于项目相关性的用户相似性计算方法，以便用于推荐的邻居用户更准确，同时在预测评分的过程中增加时间权限，使得接近采集时间的用户兴趣在推荐过程中具有更大权值。本章首先介绍传统基于用户的协同过滤推荐算法，并举例说明该方法存在的不足，在此基础上给出改进的协同过滤算法。

4.1 基于用户的协同过滤推荐算法

基于用户的协同过滤技术的理论基础是人们的从众行为。从社会心理学的角度看，每个人每天都有许多机会成为他人的社会影响对象，而这些社会影响是人们社会互动中的重要组成部分。基于用户的协同过滤算法根据相似用户的观点对目标用户产生推荐，它基于这样的假设：如果用户对一些项目的评分比较相似，则他们对其他项目的评分也将会比较相似。

4.1.1 基于用户的协同过滤推荐过程

基于用户的协同过滤算法通过分析历史数据，生成与当前用户行为兴趣最相近的用户集，将他们感兴趣的项作为当前用户的推荐结果，即 top-N 推荐。基于用户的协同过滤推荐过程可分为 3 个阶段：数据表述；发现最近邻居；产生推荐数据集[44]。

第一步：数据表述

数据表述主要是完成用户评价数据的描述，通常可表述为一个 $m \times n$ 的用户-项评价矩阵 $R=(r_{ij})$，其中 m 表示用户数，n 表示项目数，r_{ij} 表示第 i 个用户对第

j个项的评价值，一般$r_{ij} \in [0,5]$且r_{ij}是整数，该值表示用户对该项的兴趣度，也就是用户对该项的喜好程度，如表4.1所示。

表 4.1 用户—项评价矩阵

项目 / 用户	项目 1	项目 2	...	项目 n
用户 1	1	3	…	2
用户 2	0	4	…	5
.	.	.		.
.	.	.		.
.	.	.		.
用户 m	2	5	…	1

第二步：发现最近邻居

发现最近邻居是指识别目标用户的最近邻居或最相似用户。协同过滤是通过计算用户之间的相似性，识别出当前目标用户的最近邻居集，根据“邻居”的信息进行推荐，因此实现协同过滤推荐的关键就是如何准确地为一个需要推荐服务的目标用户找到最相似的邻居集，即：对一个目标用户u_T，要寻找出一个根据相似度大小排列的邻居集合$Neighbor_{u_T} = \{u_1, u_2, ..., u_n\}$，且$u_T \notin Neighbor_{u_T}$，设用户$u_T$和$u_i$的相似度为$Sim(u_T, u_i)$，并且$Sim(u_T, u_1) \succ Sim(u_T, u_2) \succ ... \succ Sim(u_T, u_n)$。其中，用户之间的相似度$Sim(u_a, u_b)$常用 Pearson 相关系数计算得出，计算公式见公式（2.7）。

第三步：产生推荐数据集

目标用户u_T的最近邻居集$Neighbor_{u_T} = \{u_1, u_2, ..., u_n\}$产生后，基于最近邻居集计算$u_T$对未评分项目$i_T$的预测评分值，同时产生 top-N 推荐集，预测$u_T$对$i_T$的评价值的计算公式[45,46]如下：

$$P_{u_T, i_T} = \overline{r_{u_T}} + \frac{\sum_{u \in Neighbor_{u_T}} Sim(u_T, u) \times (r_{u, i_T} - \overline{r_u})}{\sum_{u \in Neighbor_{u_T}} Sim(u_T, u)} \tag{4.1}$$

其中，$Sim(u_T, u)$表示目标用户u_T与最近邻居用户u的相似性，r_{u,i_T}表示用户u对项目i_T的评分，$\overline{r_{u_T}}$、$\overline{r_u}$分别表示用户u_T和用户u对项目的平均评分值。

4.1.2 基于用户的协同过滤算法分析

分析基于用户的协同过滤算法，我们发现它主要有以下几个不足之处：

(1) 由 4.2 节可知，现有算法在计算用户之间的相似度时，是基于两用户共同评分的所有项目，而项目之间是否相关并没有考虑进去，也就说在用户共同评分的项目中可能存在着跟被预测项目根本无关的项目，但是在计算基于该预测项目的用户相似度时却将其考虑在内，这就影响了寻找合适的邻居用户，从而影响推荐精度。

如表 4.2 所示，表中列出了用户的爱好，其中，“1”表示用户喜欢该项目，“0”表示用户讨厌该项目，已知前四个用户对所有项目的评价，要预测用户 5 对巧克力和科幻电影的可能评价。很明显，用户 5 对巧克力的评价可能跟他对蛋糕的评价有关，而跟他对冒险电影的评价很可能无关，同样，他对科幻电影的评价跟他对冒险电影的评价有关，而同他对蛋糕的评价关系不大。但是在现有的协同过滤算法中，是基于用户共同评分的所有项目计算用户之间的相似度的，例如，在预测用户 5 对巧克力的评价时，计算用户 5 同其他四个用户的相似性过程中，将五个用户共同评分的蛋糕和冒险电影都考虑了进去，但是实际上用户对巧克力的评价同他对冒险电影的评价相关度很低，这样计算出来的用户基于巧克力的相似度受到了他对冒险电影的评价的干扰，这种情况下找到的近邻用户就不准确，从而产生的推荐也就不准确。

表 4.2 用户爱好

项目 / 用户	奶酪蛋糕	冒险电影	巧克力	科幻电影
用户 1	1	1	1	0
用户 2	0	0	1	1
用户 3	1	0	0	1
用户 4	0	1	0	0
用户 5	1	1	?	?

(2) 用户的兴趣是经常变化的，但传统的协同过滤方法不能将这点表现出来，往往推荐的是用户过去感兴趣而现在可能不再感兴趣的东西。例如，一个用户在想买数码相机时，则对数码相机的品牌、性能、价格等信息感兴趣，等该用户买了数码相机之后，则对这些方面不怎么感兴趣了，转而会关心维修点、配件购买点等信息。如果此时还是针对用户以前的兴趣进行推荐的话，就显得没有任

何意义了。这就需要减少用户过去喜好数据的影响，提高对新兴趣的敏感性，从而实现更准确的实时推荐。

(3) 现有的协同过滤算法主要是利用用户-项评价矩阵来计算项目间的相似性的，而没有充分利用项目本身的属性，其结果是相似性度量不够准确，影响了推荐精度。

4.2 改进的协同过滤推荐算法的设计思想

基于上述分析，针对传统协同过滤算法存在的不足，现提出以下解决方案：一、提出基于项目相关性的用户相似性计算方法，即在计算用户相似性时把项目之间的关系考虑进去，使得用于推荐的邻居用户更准确，从而提高推荐精度；二、在预测过程中为了体现用户兴趣随时间的变化，将时间加权函数引入预测评分过程中，使得越早发生的用户兴趣的重要性越小，提高对用户新兴趣的敏感性，实现更准确的实时推荐。

4.2.1 基于项目相关性的用户相似性计算

我们将要计算项目之间的相关性，然后在用户相似性计算中引入项目相关性的权重，使得由此找到的邻居用户更准确。该方法是将基于用户的和基于项目的协同过滤算法有机结合起来，但它不同于以往的将两者线性组合的方式，而是以非线性的方式将两者组合起来。文献[45]提出的一种将基于用户和基于项目的协同过滤算法以线性方式结合起来，即各占一定的比例，而本文提出的是在计算用户相似性过程中引入项目相关性的权重，也就是将基于项目融入基于用户中。同时，项目的相关性不是用 Pearson 相关系数来计算，而是基于项目特征属性来计算的，这样更具客观性。改进后的用户相似性计算公式[47]如下：

$$Sim^{i_T}(u_a,u_b)=\frac{\sum_{i\in I_{ab}}\left(r_{a,i}-\overline{r_a}\right)\left(r_{b,i}-\overline{r_b}\right)\bullet rel(i,i_T)}{\sqrt{\sum_{i\in I_{ab}}\left(r_{a,i}-\overline{r_a}\right)^2\bullet rel(i,i_T)}\bullet\sqrt{\sum_{i\in I_{ab}}\left(r_{b,i}-\overline{r_b}\right)^2\bullet rel(i,i_T)}} \quad (4.2)$$

其中，$Sim^{i_T}(u_a,u_b)$是用户u_a和u_b基于i_T的相似性，i_T是待预测评分项目，$rel(i,i_T)$指项目i和i_T的相关性，其他符号意义见 4.1.1 节公式（4.1）。注意，此

处的$rel(i,i_T)$不同于 3.1.1 节的$sim(i,j)$，后面将具体说明。在计算用户相似性的过程中，加入项目相关性$rel(i,i_T)$，这减少了不相关项目对用户相似性计算的干扰，加强了相关项目在用户相似性计算中的作用，由此产生的邻居用户更准确，那么推荐也就更确切。

传统方法会用 Pearson 相关系数法计算$rel(i,i_T)$，但是该方法是基于用户-项评价矩阵的，而此矩阵是极具主观性的，并不能反映项目之间的真正关系，而前面对项目聚类过程中我们已用到了项目属性矩阵（见 3.1.1 节表 3.1），它更具客观的反映了项目之间的关系，而且鉴于属性矩阵本身计算$rel(i,i_T)$要简单的多，所以这里也利用属性矩阵计算$rel(i,i_T)$。

设项目i_1和i_2在k维空间上的属性值分别看作向量$i_1=\{i_{11},i_{12},...,i_{1k}\}$、$i_2=\{i_{21},i_{22},...,i_{2k}\}$，由于向量各维取值为 0 或 1，则项目$i_1$和$i_2$之间的相似性计算公式[48]可表示成：

$$rel(i_1,i_2)=Sim(i_1,i_2)=\begin{cases}\dfrac{\sum_{j=1}^{k} i_{1j}\wedge i_{2j}}{|attr(i_1)\cup attr(i_2)|}, & i_{1j}与i_{2j}属于不同类\\ 1, & i_{1j}与i_{2j}属于同一类\end{cases} \tag{4.3}$$

其中，$i_{1j}\wedge i_{2j}$表示析取运算，只有项目i_1和i_2同时具有第j项属性即属性值同时为 1 时，运算结果才为 1。$|attr(i_1)\cup attr(i_2)|$表示项目$i_1$和$i_2$具有的属性集并集中属性的个数。在项目聚类中搜索，当两项目属于同一类时，我们就认为这两项目相似度为 1。

4.2.2 基于时间加权的预测评分

为了顺应用户兴趣的变化，以及提高对新项目预测的敏感性，我们要用目标用户最近的兴趣来反映他们将来的喜好，同时，由于旧的兴趣是不太可靠的，因此我们要减少用户过去喜好数据的影响，从而实现更准确的实时推荐。充分考虑到“时间效应”的影响，于是引入时间加权函数$f(t)$（t为时间变量）到预测评

分过程中，将目标用户u_T对未评分项目i_T的加权预测评分改进为P'_{u_T,i_T}：

$$P'_{u_T,i_T}=\overline{r_{u_T}}+\frac{\sum_{u\in Neighbor_{u_T}} Sim^{i_T}(u_T,u)\times\left(r_{u,i_T}-\overline{r_u}\right)\times f\left(t_{ui_T}\right)}{\sum_{u\in Neighbor_{u_T}} Sim^{i_T}(u_T,u)\times f\left(t_{ui_T}\right)} \tag{4.4}$$

其中，t_{ui_T}表示用户u对项目i_T产生兴趣的时间。在预测评分的过程中，考虑到用户的兴趣是随时间发生变化的，即越迟发生的用户兴趣，其重要性越大，应提高它对推荐的影响，那么时间加权函数$f(t)$是单调递增函数，其随时间t增加而增加，且保持在（0，1）范围内，也就说，所有数据都有利于项目推荐，但最新数据贡献更大，旧数据反映的是用户以前的兴趣，则在推荐预测上占较小的权值[4]。指数时间被广泛应用于实际中，它更有希望得到渐进的过去行为变化趋势[49]，本文采用指数时间，时间函数为：

$$f\left(t_{ui_T}\right)=1-0.5\times e^{-t_{ui_T}} \tag{4.5}$$

由（4.5）式知，时间函数$f\left(t_{ui_T}\right)$的取值范围是（0，1），且它随着时间的增加而增加，数据越新，时间函数值越大，符合我们的需求。

4.3 算法设计

基于以上分析，提出的改进的协同过滤算法表述如下：

输入：目标用户u_T，推荐的项目数k，用户集合数m，邻居用户数n，待预测项目集I_{i_T}。

输出：目标用户u_T的k个推荐项目。

方法：

(1) 计算目标用户u_T同用户聚类中心的距离，按从小到大排列，选取前m个作为邻居用户的搜寻空间；

(2) 对搜寻空间内的每个用户u_i，找到同目标用户u_T共同评分过的项目，将其记录在数组$u_i[\ ]$中；

(3) 从待预测项目集I_{i_T}中任选出一个待预测项目i_T，根据项目属性矩阵和

项目聚类，利用公式（4.3）计算待预测项目i_T和所有$u_i[\ ]$中的项目之间的相关性即$rel(i,i_T)$，且$I_{i_T}=I_{i_T}-\{i_T\}$；

(4) 根据公式（4.2）计算目标用户u_T同其他所有用户之间的相似度，并将用户按相似度从大到小排列，取前n个作为u_T的邻居用户；

(5) 基于用户—项评分矩阵得到邻居用户对待预测项目i_T的评分，根据公式（4.4）预测目标用户u_T对项目i_T的评分P'_{u_T,i_T}，返回（2）直到I_{i_T}为空集；

(6) 依据P'_{u_T,i_T}从大到小将项目进行排序，取前k个项目作为目标用户u_T最感兴趣的项目。

第五章 实验与分析

根据前两章对推荐算法中寻找邻居用户阶段和产生推荐阶段的改进，本章进行了两组仿真实验，最近邻居搜寻效率实验和协同过滤推荐算法对比实验。在最近邻居搜寻效率实验中，将基于项目簇偏好的 K-means 算法和传统 K-means 算法进行了最近邻搜寻效率的比较。在协同过滤推荐算法对比实验中，从准确率上验证了改进算法的可行性和有效性，并与基于用户的协同过滤算法进行了对比。

5.1 实验数据

5.1.1 数据集介绍

实验使用的数据集是来自美国 Minnesota 大学 GroupLens 项目组提供的 MovieLens 数据集（http://www.grouplens.org）。MovieLens 是 GroupLens 项目组开发的一个基于 Web 的研究型推荐系统，用于接收用户对电影的评分并提供相应的电影推荐列表。目前，该 Web 站点的用户已经超过 43000 人，可供用户评分的电影超过 3500 部。MovieLens 数据集中，每个用户至少对 20 部电影进行了评分，评分值是从 1 到 5 的整数，数值越高，表明了用户对该电影的喜好程度越高。GroupLens 项目组提供全部 MovieLens 数据集的同时，将其分成 5 个互不相交的子集，然后每次选择一个子集作为 test 数据集，其他四个合为一个 base 数据集，每次选择一对 base 数据集和 test 数据集，使用 base 数据集中的记录作为基本用户，对 test 数据集中的目标用户进行推荐测试。数据库中主要包括 6 张数据表，分别是：USERS、MOVIES、RATINGS、AGE、GENRES、OCCUPATION。

本文实验用到的两张数据表 MOVIES、RATINGS 中的数据内容如表 5.1 所示：

表 5.1 MOVIES、RATINGS 数据表内容

电影数据表（MOVIES）	
MovieID	电影编号
Title	电影名称
Genres	电影种类

评分数据表（RATINGS）	
UserID	用户编号
MovieID	电影编号
Rating	用户对电影的评分

5.1.2 数据集选取

本文从 MovieLens 站点下载一个包括 943 个用户对 1682 部电影的 100000 条评分数据的数据集，从中随机截取了两组实验数据，其中第一组包含 500 名用户对 1637 部电影的 62770 条评分数据，用于验证基于项目簇偏好的 K-means 算法的有效性与合理性；第二组包含 600 名用户对 1653 部电影的 60351 条评分数据，用于验证改进的协同过滤算法是否提高了推荐质量。

实验需要，根据数据表 MOVIES，整理出 1637 部电影中部分电影的 19 个类别属性的描述矩阵 A，属性项依次为：unknown,Action,Adventure,Animation, War, Comedy,Fantasy, Children's,Crime,Sci-Fi,Documentary,Film-Noir,Horror,Musical, Drama,Mystery, Romance, Thriller, Western。

5.2 实验设计

5.2.1 评价标准

评价推荐系统推荐质量的度量主要包括统计精度度量方法和决策支持精度度量方法两类[48]。统计精度度量方法中的平均绝对偏差 MAE(Mean Absolute Error)易于理解，可以直观地对推荐质量进行度量，是最常用的一种推荐质量度量方法。在本文的研究中，我们希望最后的算法能够准确的预测未评价的项目的评

价值,从而为用户做出比较准确的推荐，本文采用平均绝对偏差 MAE 作为度量标准,平均绝对偏差 MAE 通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性,MAE 越小,推荐精度越高。MAE 的定义见 2.1.3 节，计算公式见公式（2.1）。

5.2.2 实验方案

整个实验分两个部分：

实验一、比较基于项目簇偏好的 K-means 算法和传统的 K-means 算法。使用聚类的目的有两个：一是为了提高算法在线处理数据的速度，由于聚类是提前做了推荐的部分工作，而且其大部分工作都可以离线进行，那么和传统的协同过滤推荐算法相比，其在线处理数据的速度明显会有很大的提高，也是说实时性要强很多，在此就不再给出具体的实验数据；二是便于系统在线查询目标用户的最近邻居，对其进行准确的推荐，因此重要的是看聚类算法执行的效果，即搜寻邻居的效率和向目标客户推荐结果的准确性。所以，我们将做最近邻居搜寻效率实验，这里采用的度量方法是最小空间内搜索到更多的邻居[50]。改进聚类算法对推荐结果准确性的影响实验将放到实验二，便于进行性能比较。

实验二、对比改进的协同过滤算法、基于用户的协同过滤算法（即基于用户的 CF 算法）和加入本文用户聚类后的改进协同过滤算法（即改进的基于用户聚类的 CF 算法），利用评价推荐质量的 MAE 标准，验证了改进的协同过滤算法提高了推荐精度，改进的基于用户聚类的 CF 算法更胜一筹。

5.3 实验结果及分析

5.3.1 最近邻居搜寻效率实验

设目标用户为u_T，整个用户空间为U,首先在整个用户空间上作最近邻查找，选择最近邻居数目为 10，查询结果为U_a；然后在与目标用户u_T最相似的前k个聚类（记为$c_1,c_2,\cdots,c_k$）中作最近邻查询，最近邻居数目也选择为 10，查询结果记为U_k，则本算法的有效性可以表示为只需要扫描原始数据集的

$(|c_1|+|c_2|+\cdots+|c_k|)/|U|$（其中，$|c_i|$代表第$i$个聚类中用户的数量，且$i\in\{1,2,\cdots,k\}$；$|U|$代表整个用户空间中的用户数量）就能找到目标用户$u_T$在全部用户空间上$|U_a|/|U_k|$（百分比）的邻居。

在用户聚类过程中，聚类数目的选取非常重要，数目过大，计算目标用户与聚类中心的相似度耗时多，不能有效地提高推荐系统的实时响应速度；数目过小，那么每个聚类中包含的用户数目比较多，即使在与目标用户最相似的若干聚类中进行最近邻居搜寻也需要扫描大量的侯选用户集，也不能有效提高实时响应速度。因此我们分别以30、40作为聚类数目对500个用户进行聚类。

当聚类数目为30时，实验结果如表5.2所示：

表5.2 聚类数为30的最近邻搜寻效率比较

$\frac{\lvert U_a\rvert}{\lvert U_k\rvert}$ 搜索用户比	传统K-means算法	改进的K-means算法
0	0.253	0.214
0.1	0.436	0.493
0.2	0.621	0.706
0.3	0.733	0.885
0.4	0.867	0.920
0.5	0.915	0.925
0.6	0.942	0.946
0.7	0.967	0.970
0.8	0.974	0.981
0.9	0.982	0.99
1	1	1

改进算法同传统算法的最近邻搜寻效率如图5.1所示：

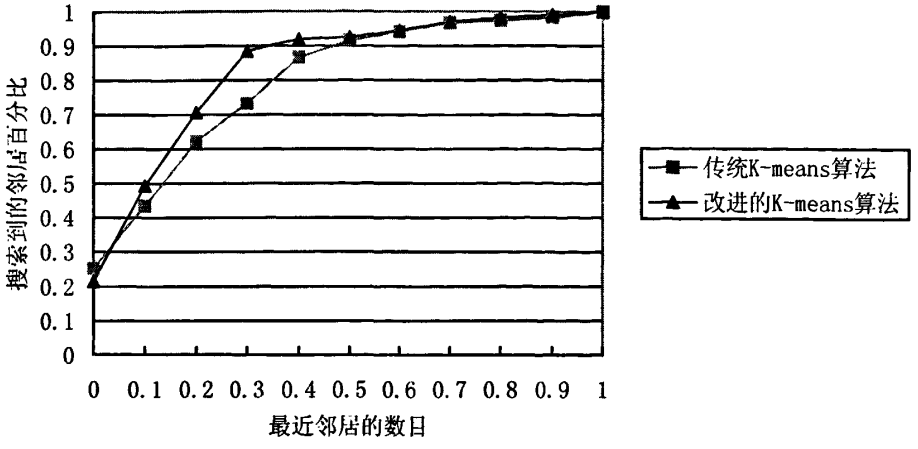

图 5.1 聚类数为 30 的最近邻搜寻效率比较

当聚类数目为 40 时，实验结果如表 5.3 所示，对应的图如图 5.2 所示：

表 5.3 聚类数为 40 的最近邻搜寻效率比较

$\frac{\lvert U_a \rvert}{\lvert U_k \rvert}$ 搜索用户比	传统 K-means 算法	改进的 K-means 算法
0	0.223	0.195
0.1	0.504	0.673
0.2	0.698	0.875
0.3	0.839	0.913
0.4	0.896	0.921
0.5	0.926	0.946
0.6	0.949	0.963
0.7	0.971	0.975
0.8	0.982	0.987
0.9	0.992	0.993
1	1	1

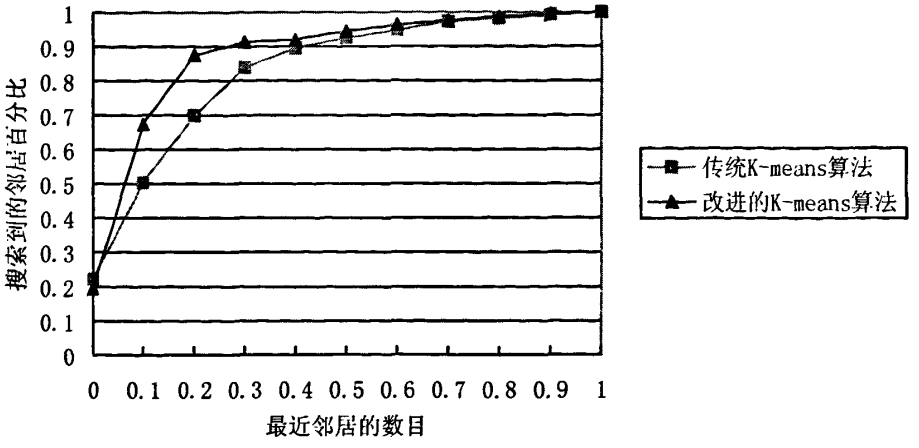

图 5.2 聚类数为 40 的最近邻搜寻效率比较

以上实验结果显示，当聚类数目为 30 时，传统聚类算法在 25.6%的用户空间上搜索到的邻居才 63.8%。，而本文改进的聚类算法在 32.1%的用户空间上却能搜索到 90%的邻居；当聚类数目为 40 时，改进的聚类算法效果更明显，它在 30.3%的用户空间上搜索到了 92%的邻居用户，而传统聚类算法只搜索到 83%的邻居用户。实验结果表明，改进的聚类算法可以在更小的用户空间内搜索到更多的邻居用户，提高了查找用户最近邻的效率和精度，满足了推荐系统的实时性要求。分析实验结果，还会发现聚类数目越大，可以在更小的空间里搜索到更多的邻居，也就是说查找目标用户的最近邻居的速度越快，不过这个结论是在聚类数目远小于用户数目，计算目标用户与聚类中心相似性的时间代价相对于最近邻居查询可以忽略不计的条件下才成立，当聚类数目很大的时侯，计算目标用户与聚类中心相似性的代价并不能忽略不计。

5.3.2 协同过滤算法对比实验

在预测用户对项目的兴趣评分时，参与计算的最近邻居的个数影响着算法的 MAE。实验中，我们采取目标用户的最近邻居个数从 5 增加到 40，间隔为 5，查看不同的最近邻居集大小对预测准确度的影响，实验结果如表 5.4 所示。

表 5.4 改进算法同其它算法的 MAE 对比

MAE / 邻居数目	基于用户的 CF 算法	改进的 CF 算法	改进的基于用户聚类的 CF 算法
5	0.830	0.759	0.746
10	0.819	0.712	0.723
15	0.802	0.720	0.711
20	0.809	0.728	0.722
25	0.813	0.710	0.694
30	0.796	0.703	0.682
35	0.756	0.676	0.631
40	0.741	0.642	0.607

改进算法同其它算法的 MAE 比较如图 5.3 所示：

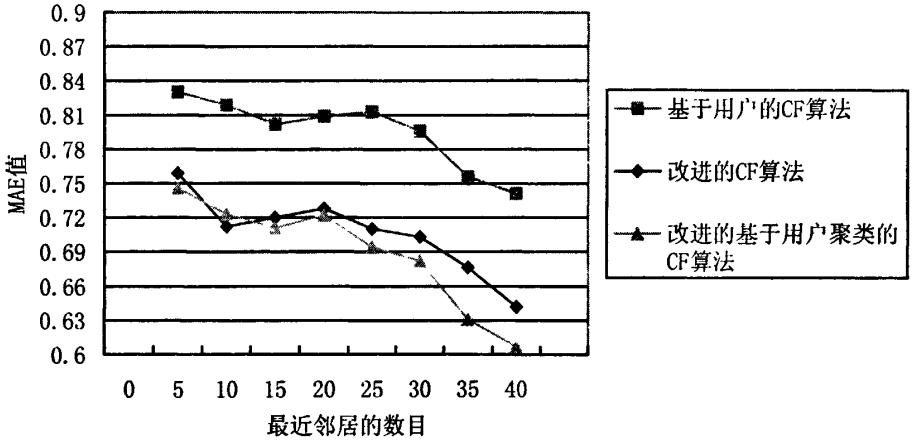

图 5.3 改进算法同其它算法的 MAE 对比

由于 MAE 值越小，推荐精度越高，由图 5.3 可知，改进后的协同过滤算法（即改进后的 CF 算法）的推荐精度比基于用户的协同过滤算法的要高，这是因为本算法借助项目属性矩阵改进用户相似性计算方法，找到更准确的邻居用户进行推荐，同时增加时间权值以体现用户兴趣变化进行实时推荐，改善了现有算法由于用户相似性计算不够准确、时效性不足等问题影响推荐精度的状况，提高了推荐质量，和本文的初衷是一致的。改进的基于用户聚类的 CF 算法比改进的 CF 算法的推荐质量更好，是因为它除了有改进的 CF 算法的优势外，还在用户聚类时就将用户对项目所属类别的偏好考虑进去，这便涉及了用户各个方面的爱好，使得最终的推荐更全面，也就更准确。实验结果还表明，邻居数目越多的情况下，能反映用户兴趣变化的用户越多，本文算法的推荐精度也越高。

第六章 总结

互联网的飞速发展使得电子商务以其独特的优势流行于全世界，它提供给人们越来越多的商品和选择机会，改变着人们的生活方式，与此同时，也带来了新的问题。随着商品数据急剧增长，在浩瀚的商品海洋里挑选出用户真正需要的东西好比大海捞针；另一方面，由于系统无法有效地满足用户需求，那提高用户对网站的忠诚度也仿佛成了空谈。在这样的背景下，电子商务个性化推荐系统应运而生。但就目前的个性化推荐技术而言，推荐系统还存在很多问题，如数据的稀疏问题、冷启动问题、推荐精度以及实时性问题。本文针对个性化推荐系统中出现的问题，改进了传统的协同过滤算法，提出了基于项目簇偏好的用户聚类，缩小邻居用户搜寻空间，在此基础上，将项目之间的属性关系引入推荐过程，使得推荐更准确。

在本课题的研究过程中，主要做了以下几个方面的工作：

(1) 建立项目属性矩阵，基于属性相似性对项目进行模糊聚类，为基于项目簇偏好的用户聚类做准备。这改善了传统的协同过滤算法因未将项目属性关系考虑其中而造成推荐不准确的情况。当系统中出现新项目时，系统将新项目分配到最相似的类中，这便解决了新项目问题。再者，项目之间的属性关系更加稳定可靠，这有利于推荐。

(2) 本文采用 K-means 聚类算法对用户聚类，缩小邻居搜寻空间，这便提高了实时响应速度。K-means 聚类算法由于随机选取初始中心而使得聚类结果不稳定，本文首先采用 kruskal 算法生成初始聚类中心，再用 K-means 聚类算法以用户对不同项目簇的偏好对用户进行聚类，最后从中挑选出离当前用户最近的几个聚类，以此作为搜寻邻居用户的空间，提高实时响应速度的同时，改善了推荐精度。

(3) 针对传统协同过滤推荐算法未考虑干扰项目问题和用户兴趣变化问题，本文提出了基于项目相关性的协同过滤算法，将项目相关性引入到用户相似性计算中，以减少不相关项目对用户相似性计算的影响，同时在预测评分的过程中增加时间权限，使得接近采集时间的用户兴趣在推荐过程中具有更大权值，这样得到的邻居用户更准确，反应了用户的兴趣变化，从而提高了推荐质量。

(4) 本文中的实验以 MovieLens 数据集为测试集，最近邻居搜寻效率实验分别以 30、40 作为聚类数目对 500 个用户进行聚类，将基于项目簇偏好的 K-means

聚类算法和 K-means 聚类算法进行比较，实验结果显示本文算法能在更小的用户空间内搜索到更多的邻居；协同过滤算法实验以 MAE 作为评价标准，将基于用户的 CF 算法、改进的 CF 算法和改进的基于用户聚类的 CF 算法进行性能比较，实验结果显示基于用户聚类的 CF 算法的推荐效果更好 。

尽管在课题研究以及论文写作过程中，本人对个性化推荐系统、协同过滤推荐技术以及聚类技术进行了多方面的学习，但由于能力有限，仍在某些方面存在或多或少的不足，就论文而言，在以下几个方面还需要做进一步的研究完善：

(1) 针对推荐算法过程中的项目相关性，如何抓住对具体用户而言的关键项目属性，采用更适合的方法来计算项目相关性；新用户由于没有偏好信息而不能对其进行合适的推荐，则要解决新用户问题；通过用户对商品的浏览，搜集隐式的用户信息，对用户的浏览信息进行挖掘，从中分析出用户可能的兴趣点。同时，还可以实现对推荐质量的跟踪。

(2) 本文采用了 MovieLens 网站中提供的小规模的数据集，在算法的准确性上有一定的局限性。下一步的工作需要对更大规模的数据集、更多类型的数据进行实验，从而确保算法的扩展性和适应性。

(3) 本文提出了新推荐算法，并通过实验证明了其有效性，但没有构建一个系统来实现算法的在线运行。下一步的工作是要构建一个实验系统，通过在线运行本算法来验证实验结果。

总之，本文在现有技术的基础上进行了一些新的探索和尝试，但本文的完成并不意味着研究工作的结束，这只是个起点，在以后的学习中我们还将对这个课题做更多更深入的研究。

参考文献

[1] 曾子明，余小鹏.电子商务推荐系统与智能化谈判技术.武汉:武汉大学出版社，2008年.

[2] KunChangUe, Soonjae Kwon. Online shopping recommendation mechanism and its influence on consumer decisions and behaviors:A causal map approach. Expert Systems with Applications. 2008,35(4):1567-1574.

[3] 余力，刘鲁．电子商务个性化推荐研究[J]. 计算机集成制造系统-CIMS,2004(10):P1306-1313.

[4] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval. NewYork: Addison-Wesley Publishing Co.,1999.

[5] Murthi BPS, Sarkar S. The role of the management sciences in research on personalization. Management Science，2003，49(10): 1344-1362.

[6] Smith SM，Swinyard WR. Introduction to markating models. 1999. http://marketing.byu.edu/courses/693r/modelsbook/ preface.html

[7] Adomavicius G，Tuzhilin A .Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans.on kowledge and Data Engineering，2005，17(6):734-749.

[8] K.L.Wu,P.S.Yu.A Web usage mining and analysis tool.IBM Systems Journal. 1998,89-105

[9] Aggawal CA, Wolf J L, Wu K L,et al. Horting hatches an egg:a new graph-theoretic approach to collaborative filtering[A]. Proceedings of ACM SIGKDD International Con- ference on Knowledge Discovery and Data Mining[C]. 1999.

[10] 邓爱林，朱扬勇，施伯乐.基于项目预测评分的协同过滤推荐算法[J].软件学报，2002，13(4):1-8.

[11] Amd K，Bemard M.Clustering for collaborative filtering applications[C].Proeeedings of Computational Intelligenee for Modeling, Control and Automation.Vienna，Austria:IOS Press，1999:17-19.

[12] Ansari A Essegaier S,Kohli R. Internet recommendation systems[J].Journal of Marketing Research,2000,37:363-375.

[13] 张巍，刘鲁，葛健. 一种基于粗集的协同过滤算法[J].小型微型计算机系统，2005，26(11):1972-1974.

[14] Sarwar BM, Karypis G, Konstan JA, et al.Application of dimensionality reduction in recommender system a case study [C]. Proc ACM WebKDD 2000 Web Mining for E-Commerce Workshop.New York:ACM Press,2000:82-90.

[15] Rosenstein M,Lochbaum C.Recommending from content: Preliminary results from an E-commerce experiment [C]. Proceedings of Conference on Human Factors in Computing.Heidelberg Berlin:ACM Press,2000:291-292.

[16] Resnick,Varian. Recommender systems[J]. In Communications of the ACM.1997,40 (3):56-58.

[17] Sarwar B,Karypis G,Konstan J,Riedl J. Item-Based collaborative filtering recommendation algorithms[A]. In: Proceedings of the 10th International World Wide Web Conference[C]. [S.l.]:[s.n.],2001, 285-295.

[18] Balabanovic M, Shoham Y. Fab: content-based, collaborative recommendation[J]. Comm ACM, 1997, 40(3): 66-72.

[19] Pazzani M, Billsus D. Learning and revising user profiles: the identification of interesting web sites[J]. Machine Learning,1997, 27:313-331.

[20] RichE.User modeling via stereotypes.Cognitive Science, 1979,3(4):329-354.

[21] Goldberg D，Nichols D，Oki BM，Terry D. Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992,35(12):61-70.

[22] Konstan JA，Miller BN，Maltz D，Herlocker JL，Gordon LR，Riedl J. GrouPLens: APPlying collaborative filtering to usenet news. Cornmunieations of the ACM.1997，40(3):77-87.

[23] Shardanand U，Maes P.Social information filtering: Algorithms for automating "Word of Mouth". In: Proc.of the Conf. on Human Factors in Computing Systems.New York: ACM Press，1995.210-217.

[24] Terveen L，Hill W，Amento B，McDonald D，Creter J. PHOAKS:A system for sharing recommendations. Communicationsof the ACM,1997,40(3):77-87.

[25] Goldberg K，Roeder T，Gupta D，Perkins C. Eigentaste: A constant time collaborative filtering algorithm.Infortnation Retrieval，2001,4(2):133-151.

[26] Schafer JB，Konstan J，Riedl J.Reommender systems in e-commerce.In Proc.of the lst ACM Conf.on Electronie Commeree. NewYork:ACMPress，1999.158-166.

[27] Ben J,Konstan JA, John R. E-commenrce recommendation applications[R]. University of Minnesota, 2001.

[28] 曾艳，麦永浩.基于内容预测和项目评分的协同过滤推荐[J].计算机应用，2004(01):111-112.

[29] Lang K. Newsweeder:learning to filter news'[A]. Proceedings of the 12th International Conference on Machine Learning[C].1995.331-339.

[30] 简士尧.以内容为基础之网络学习导览推荐之研究「D].台湾铭传大学资讯工程学系(硕士学位论文)，2004.

[31] 邢春晓，高凤荣，战思南，周立柱. 适应用户兴趣变化的协同过滤推荐算法[J]. 计算机研究与发展，2007，44(2)：296-301.

[32] Towle.B., Quinn.C.. Knowledge based recommendation systems using explicit user models. In Knowledge- based Electronic Markets, AAAI Workshop, AAAI Technical Report WS-00-04,Menlo Park, CA: AAAI Press, 2000

[33] Burke R.Knowledg-Based recommender systems. Encyclopedia of Library and Information Systems,2000,69(32):180-200."

[34] Marko B,Yoav S. FAB:content-based collaborative recommendation[J]. Communications of the ACM, 1997,40(3):66-72.

[35] Robin B. Hybrid recommender systems: survey and experiments[R]. Department of Information Systems and DecisionSciences, California State University, Fullerton.

[36] Bing Liu 著，俞勇，薛贵荣，韩定一译.Web 数据挖掘[M].第一版. 北京：清华大学出版社，2009.

[37] L.A.Zadeh.Fuzzy sets.Information and Control,1965(8):338–353.

[38] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques[M].Morgan Kaufmann Publishers,2001.

[39] 金微.基于遗传算法的 Kmeans 聚类方法的研究.河海大学硕士论文.2007.

[40] 马宏伟，张光卫，李鹏. 协同过滤推荐算法综述[J]. 小型微型计算机系统，2009，30(7):1282-1288.

[41] Yu Li, Liu Lu, Li Xue feng. A Hybrid Collaborative Filtering Method for

Multiple-Interests and Multiple-Content Recommendation in E-Commerce[J]. Expert Systems with Applications,2005,281:67-77.

[42] 马军, 邵陆. 模糊聚类计算的最佳算法[J].软件学报, 2001:12(04): 578 -581.

[43] 陆林花, 王渡. 一种改进的遗传聚类算法[J]. 计算机工程与应用, 2007.43(21): 170-172.

[44] 赵亮, 胡乃静, 张守志. 个性化推荐算法设计[J]. 计算机研究与发展, 2002,39(8): 986-990.

[45] Linden G, Smith B, York J. Amazon.com Recommendation Item-to-Item Collaborative Filtering[J]. IEEE Internet Computing, 2003,7(1): 76-80.

[46] 肖敏, 熊前进. 基于项目语义相似度的协同过滤推荐算法[J]. 武汉理工大学学报, 2009,31(3): 21-23.

[47] Yun Zhang, Peter Andreae. Iterative Neighbourhood Similarity Computation for Collaborative Filtering[C]. 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[48] 吴发青, 贺樑, 夏薇薇, 任磊. 一种基于用户兴趣局部相似性的推荐算法[J]. 计算机应用,2008,28(8): 1981-1985.

[49] 彭德巍, 胡斌. 一种基于用户特征和时间的协同过滤算法[J]. 武汉理工大学学报, 2009,31(3): 24-28.

[50] 潘宇, 林鸿飞, 杨志豪. 基于用户聚类的电子商务推荐系统[J]. 计算机应用与软件, 2008, 25(4): 25-26.

致 谢

衷心地感谢我的导师宋顺林教授。在选题过程中曾多次得到宋老师的指导，宋老师从很多方面就该课题的科学性以及可行性给出了指导性意见，在论文初稿完成后，宋老师又多次给出修改意见并且亲自对论文部分章节内容及结构进行了修改，可以说从选题到完成，从基础理论到具体语言的表述，无处不渗透着宋老师的心血。三年来，宋老师严谨敏锐的治学作风、谦虚和善的学者风范、诚恳正直的处世原则对我树立正确的治学观念和价值观念都产生了直接而深远的影响。我所有的进步和成绩都是与宋老师的教导分不开的。

感谢与我朝夕相处的实验室同学以及共同生活学习过的同学们，感谢我的家人，感谢所有曾经给我鼓励、帮助的朋友们！

攻读硕士期间发表的论文及参加的科研项目

攻读硕士学位期间，发表和录用的学术论文有：

1. 《改进的协同过滤推荐算法》，计算机工程与应用，中文核心，第一作者

参与的科研项目有：

1. 2009.7-2009.12 镇江市无负压供水设备对管网运行的安全对策及评价系统

电子商务个性化推荐算法设计与实现

作者：刘芳先
学位授予单位：江苏大学