分类号:_____密级:_____

UDC:_____

学 位 论 文

基于 Android 平台的电力交易推荐系统的研究与实现

作 者 姓 名 ：徐振康

指 导 教 师 ：焦明海　副教授

东北大学计算机科学与工程学院

申请学位级别：硕士　　学 科 类 别 ：专业学位

学科专业名称：计算机技术

论文提交日期：2017 年 12 月　　论文答辩日期：　2017 年 12 月

学位授予日期：　　答辩委员会主席：黄卫祖

评 阅 人 ：邓庆绪、库涛

东北大学

2017 年 12 月

A Thesis in Computer Technology

Research and Implementation of Electric Trading Recommendation System Based on Android Platform

By Xu Zhenkang

Supervisor: Associate Professor Jiao Minghai

Northeastern University

Dec 2017

独创性声明

本人声明，所呈交的学位论文是在导师的指导下完成的。论文中取得的研究成果除加以标注和致谢的地方外，不包含其他人已经发表或撰写过的研究成果，也不包括本人为获得其他学位而使用过的材料。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。

学位论文作者签名：

日　期：

学位论文版权使用授权书

本学位论文作者和指导教师完全了解东北大学有关保留、使用学位论文的规定：即学校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。本人同意东北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、交流。

作者和导师同意网上交流的时间为作者获得学位后：

半年□一年□一年半□两年□

学位论文作者签名：　　导师签名：

签字日期：　　签字日期:

摘 要

协同过滤是个性化推荐技术中比较成熟，应用最为广泛的推荐方法。在电力改革的背景下，电力交易模式发生转变，电力交易推荐系统呼之欲出。在此背景下，本文对协同过滤算法做出改进，提出了两种不同的协同过滤方法以应对电力交易推荐系统的实际需求，并设计和实现了电力交易推荐系统的移动端应用程序原型。

本文提出基于时序社交关系的协同过滤算法。在该算法中，主要工作如下：首先利用用户给出评分或发生交易行为的时序关系，挖掘出用户之间的影响关系和从众关系，将它们融入概率矩阵分解算法以提高推荐准确性。给出了推荐框架，并分析了计算复杂度。其次，在此过程中，提出了时序关系下用户社交关系的定量分析方法。最后，基于电力交易过程中数据的特点，提出了评分和时间数据获取与标准化方案，使得算法可以有效处理电力交易数据。通过在真实数据集上对算法做出评估检验可知，该算法具有较高的准确度和较好的性能。

还提出了一个应对稀疏数据的基于用户偏好估计的协同过滤算法，在该算法中，主要工作如下：由于传统方法对待用户间的相似性往往给出的估计值较为粗略，利用核密度估计方法估计用户的偏好分布，通过分析两个用户偏好的 KL 散度的方式来计算它们偏好的相似度，从而提高矩阵填充的准确度。在计算用户相似度的过程中，本文提出了一种基于商品标签的相似度，可以从商品标签的角度描述商品的相似度。最后在真实数据集上验证了该算法应对稀疏数据有较好的性能。

最后本文根据电力改革背景，设计并实现基于 Android 平台下的电力交易推荐系统原型。根据软件工程化思路，从需求分析到到详细设计，分别做了移动端、服务端和网络架构设计以及数据库设计，最终给出电力交易推荐系统的移动端展示平台。

关键词：协同过滤；概率矩阵分解；核密度估计；电力改革；Android 平台

Abstract

Collaborative filtering is widely applied in personalized recommendation techniques. Under the background of electricity reform, the electricity trading model changes, and power trading recommendation system is about to emerge. Under this background, this thesis improves the collaborative filtering algorithm, which proposes two different collaborative filtering methods to deal with the actual demand of power trading recommendation system. The thesis also designs and realizes the mobile terminal application prototype of power trading recommendation system.

This thesis proposes a collaborative filtering algorithm based on time sequence social relations. In this algorithm, the main work of this thesis demonstrates as follows: Firstly, the relationship between users and their relationship is excavated by using the time-series relationship between users' rating or transaction behavior, and they are integrated into the probability matrix factorization algorithm to improve the accuracy of recommendation. The recommended framework is given and the computational complexity is analyzed. Secondly, in the process, a quantitative analysis method of user's social relations under the time sequence relationship is proposed. Finally, based on the characteristics of the data in the electricity trading process, a scheme of obtaining and standardizing rating and time data is proposed so that the algorithm can effectively process the data of electricity trading. By evaluating the algorithm on the real data set, the algorithm is of higher accuracy and better performs.

This thesis also proposes a collaborative filtering algorithm based on user preference estimation to deal with sparse data. In this algorithm, the main work of this thesis shows as follows: Since the traditional methods always gives a rough estimate for users. The methods use kernel density estimation method to estimates the user's preference distribution, and calculates the similarity of their preferences by analyzing the KL divergence of two user preferences, so as to improve the accuracy of the matrix completion. In the process of calculating user similarity, this thesis proposes a similarity based on product labeling, which can describe the similarity of products from the perspective of product’s label. Finally, the real data set shows that the algorithm could better perform on sparse data.

Finally, according to the background of electricity reform, this thesis designs and realizes

the prototype of electricity trading recommendation system based on Android platform. In terms of the idea of software engineering, from the demand analysis to the detailed design, the mobile terminal, server and network architecture design and database design are respectively done. The mobile terminal display platform of the electricity trading recommendation system is given.

Keywords: Collaborative Filtering; Probability Matrix Factorization; Kernel Density Estimation; Electricity Reform; Android Platform

目 录

第1章 绪 论

1.1 研究背景

2015 年 3 月，中共中央、国务院下发了《关于进一步深化电力体制改革的若干意见》（中发〔2015〕9 号）（后文简称《意见》），备受社会各界瞩目的新一轮电力体制改革正式拉开帷幕[1]。《意见》秉承五项基本原则：一是坚持安全可靠；二是坚持市场化改革；三是坚持保障民生；四是坚持节能减排；五是坚持科学监管。其中，市场化改革是核心。长久以来，电力市场交易模式一直处于“垄断行业”状态，传统电力市场中，发电企业与购电企业之间不能直接进行交易，也无法直接电力传送，而需要一层国家电网的调度。在新一轮的电力改革方案启动之后，发电企业凭借自身的发电优势以及相关许可即可直接与购电企业达成交易，中间的竞价和市场中的竞争等环节不再受到国家电网等电力资源管理部门的严格约束，而是在合理的竞争规则内自由进行，这样形成了售电主体和购电用户之间的点对点交易。市场主体的自由性还体现在相互自主确定交易用户、交易电量和价格，交易过程中按照国家规定的关于电价输配的方案来决定过网费和相关手续即可。这样放开竞争可以为工商业用户和企业用户等提供更加优质和经济的电力保障和服务。政府敏锐的意识到市场的竞争应该交给市场自己调控，要减少对市场的控制，只需管住中间而放开竞争。《意见》中强调了电力改革的重要性和紧迫性。从根本上改变传统电力市场中发电厂与电网一体，政企不分的状态，争取形成电力市场自由竞争的多元化格局。竞争性环节电价的有序放开可以推进交易机构相对独立，规范市场运行。在市场竞争主体的范围不断扩大过程中，大用户与交易主体数量会呈几何级数增长，又加之交易行为具有实时性及地域性，竞争符合条件的市场主体可以通过移动终端与互联网技术发生电力交易行为。

随着大用户直购电交易业务的深入开展和市场交易主体模式的多样化，参与清洁能源的直购电交易主体成员数量将会快速增长。随着移动互联网新技术的普及应用，它正在改变社会成员的沟通方式，改变人们的日常生活习惯，并且开始渗透到工业的各个领域，即将形成的“互联网+”的交易双赢模式，必将促进电力工业的社会效益和经济效益。移动互联网技术应用到大用户直购电点对点交易业务场景中，是“互联网+”电力市场交易模式的直接体现，也是贯彻落实“十九大”关于深化供给侧结构性改革的实行。随着电力市场的深入改革，电力交易成员数量的急剧增加，市场各类成员渴望提供更加

弹性和多样化的电力市场交易方式，需要研究移动互联的电力交易用户行为模型和算法。因此，电力交易的移动端用户交互交易方式将成为电力市场用户交互方式的有效补充。

在数据量日益增大的今天，用户在数据的海洋里显得手足无措，数据量已经足够大，但是这些数据的利用率却降低，随之而来的“信息过载”问题亟待解决。目前，针对该问题以用户为主动的解决方案是当今互联网广泛采用的搜索引擎,而推荐引擎则是用户作为被动接受推荐对象的主体,广泛地应用于电子商务等互联网应用上。推荐技术在今天的互联网应用和产品中被广泛采用，比如电子商务的商品推荐、社交网络上的好友推荐等，它们是目前互联网上最常见的智能产品形式。推荐系统是为了解决“信息过载”问题而出现的新技术。从上个世纪 90 年代开始，推荐系统开始被众多学者及领域专家所熟知和研究，内容涉及近似理论、认知科学和信息检索等相关学科。长期以来，推荐系统领域的研究工作重点围绕在用户信息获取和建模、推荐算法研究、推荐系统评价指标、以及推荐系统的应用和社会影响的研究[2]。根据用户的兴趣爱好推荐符合用户兴趣的对象是推荐系统的核心功能。由于推荐系统可以帮助市场主体达到个性化营销而提升销售量，为企业增大利润。推荐系统相关技术获得了众多企业的重视，很多学者相继对推荐系统进行深入的研究，推荐系统领域得到了长足的发展。

基于电力市场急需的供给侧结构性改革，结合推荐系统在“互联网+”上的巨大成果，依托于电力市场的第二次改革的背景，将推荐系统与电力市场有机结合是在即将形成的“互联网+”的交易双赢模式中移动互联网技术融合于实体经济的创新思想。在上述改革的大环境和要求下，本文首先调研发达国家在电力市场服务及移动互联网技术建设方面的先进经验和理念，从建立移动端电力市场交易管理的常态机制，提供完善的移动端电力市场服务产品角度出发，融合推荐系统技术体现的巨大商业价值，以满足统一电力市场交易平台运营管理的不同需求，既保证了供电侧发电供给与购电侧用电行为的平衡，又提高了服务模式的体验。

1.2 研究意义

在以互联网技术为驱动的互联网时代,电力领域的交易模式正在发生着天翻地覆的变化,以往的线下交易发展至如今的线上交易。国家电网作为电力交易的枢纽和服务配备，移动互联网技术作为发电企业和用电企业之间沟通的桥梁，推荐技术则是移动互联网应用于电力交易领域的重要纽带，电力交易模式发生改变，为了提升服务质量，设计一个能用在电力交易领域的服务平台是大势所趋的。由于电力能源作为电力市场特殊商

品，它是一种无法大规模储藏的能源，电力能源的生产、输送和消费都是通过电力网络同时完成的，在电力生产的过程中，即不存在半成品，也不存在库存品[3]。为了使电力生产、流通和消费等环节能很好的相互衔接，电力工业需要采用大量的自动化控制技术和设备，以实现发、输、售、用各个环节的相互紧密配合，协调统一的进行[4]。电力交易服务平台的实施可以从互联网的角度来解决电力输配的供需平衡问题，并且能很好的完成售电和用电，发电和输配等过程的紧密结合，而且做到了信息化，共享化，更易于管理。在这样的平台上建立起电力推荐服务能提高服务的质量，增加可观的盈利，提高市场的利用率，该平台对发电企业以及大用户双方都有想当可观的利益，促进双方的合作和经济的发展。

电力交易行业迈进到“互联网+”的时代浪潮中，不仅体现在技术上的迈进，而且体现在解决传统电力市场问题的角度发生改变，观念开始有了新的突破。该平台的实施既能打破商品交易的中间环节，而且去中介化，打造了创新平台。电力商品传输与销售完全依赖信息垄断的行为来获取超额利润的行业模式完全被打破，电力产品的生产者即发电企业可以更加直接的与购电方发生交易行为，不仅降低了成本，而且提高了效益。另一方面，建立“互联网+”的信息交互平台，即该电力交易领域的服务平台，是在信息平等的基础上，提供满足售电和购电双方信息共享的开放性的供需互动的商业系统，不仅可以满足供需双方基本业务需求，还能提供可靠的服务进而是双方都能从中获取盈利，相信经过市场不断对其迭代会促成一个功能完善，服务体验优质的电力互联网商业带，其发展潜力巨大。上述两层意义恰恰体现了该平台的实施是对《意见》提出了“管住中间、放开两头”思路的落地。推荐策略的引入也可以鼓励多买多卖，激发电力市场的活力，才能真正意义上打破供需用户单一的僵局。有电力改革的政策作为驱动，加之移动互联网技术的落地都会促进电力改革的成果和增加改革的红利。

1.3 研究现状

1.3.1 国内外电力市场交易模式

目前西方各国竞相进行电力市场化改革。美国、英国、北欧、日本等国家和地区通过改革来扩大市场范围，激励竞争，提高资源配置效率。随着可再生能源的发展，清洁能源消纳的需求也在一定程度上刺激市场。电力交易模式主要体现在发电和售电环节引入竞争机制。

美国电力改革集中体现在引入竞争机制，减轻对市场的干涉，降低电力成本，提高输配效率。由于美国国情，不同的州有不同的电力改革方案，共同点在于：在原有的电

力交易模式下阶段性地为大用户提供灵活的用电选择性服务。在商业竞争的背景下，使用用电量和电压使用级别开放用户选择权。售电主体准入由国家层面的监管机构进行审批。在美国1/3的州开放用户选择权。在财务方面，美国德克萨斯州要求售电主体或其担保公司有形资产净值不低于1亿美元[5]。美国售电侧放开的18个州中13个州大工商用户更换供电商的比例在80%以上，但居民用户行使购电选择权的比例普遍不高[6-7]。英国电力市场交易主要是电力交易所进行，电力交易所是独立于电网公司和发电企业的第三方交易市场，在政府的政策和条例约束下由电网公司组建，其主要业务是提供电力的短期交易，多数为当日电力现货交易市场，为电网公司监控供需不平衡的信息，并负责供需平衡的控制。英国对于用户选择权已经全部放开，历时9年[6]。在北欧，世界上唯一一个横跨多国的电力交易市场，其电力市场服务体系拥有自身独特的特点，重点突出在电力市场的开放性。北欧电力交易所提供双边交易的电量信息和电价信息，成交量和成交价格的历时数据，接入的实时数据，统计数据每日现货市场价格数据，甚至还有提供用于研究的科研教学数据，绿色电力数据等等。日本的会员制在国际电力交易市场上别具一格。在市场上交易必须具有交易会员的资格。交易会员的申请需要提供相应的资金。日本电力交易所的市场分为日前现货市场，远期合约市场，自由合约市场三种[8]。对于开放选择权用户上，日本开放范围扩大到全部用户的 60%[6]，日本的垂直一体化电力公司，大用户不太喜欢更换电力供应商，电力公司的市场份额受市场竞争的影响较小[6]，十大供电商所占市场份额的综合高达93.9%[9]。

与国外的电力市场运营模式相比，当前，我国电力交易改革的重点在于发展大用户直接交易模式。主要内容包括市场主体准入机制、售电侧放开电力交易平台、发电商与售电商的交易模式、电价机制、余缺电量平衡机制、监督惩罚机制等方面。为了逐步推动交易趋向市场化发展，全面放开售电侧市场成员交易，需要制定切实有效的交易主体准入机制，利用市场经济的杠杆来优化配置电力资源。售电侧放开电力交易平台实施后，发电商获得自主电力交易能力，使得电力交易形式变得更加多样，国家已经逐步成立了33个电力交易中心。在发电商与售电商的交易模式这块，我国电力交易以签订中长期合约为主，包括双边交易（单一发电商与单一售电商交易为双边交易，体现为“一对一”的模式）和多边交易（“多对一”及“一对多”模式下购电商与售电商的交易为多边交易）。改革的另一重点在于余缺电量平衡机制，当突发事件和外力突发时，用户实际用电量和直接交易的合约电量会有差距，该机制即是应对该类事件的保障性服务机制。通过

上述改革，我国的电力市场交易模式、服务模式等逐渐趋于完善。各部分改革试点中发电企业均能以竞价上网的模式入网，符合交易规则的大用户能直接向发电企业购电，实现点对点交易。这将是目前我国电力市场建设改革的重点。

1.3.2 国内外推荐系统研究现状

在互联网数据爆炸的时代，用户逐渐被淹没在信息的世界中,如何从海量数据中快速而高效地获取用户所需要的信息这一问题变得越来越严峻。推荐系统应运而生。由于推荐系统可以有效地解决信息过载问题[10],因而受到来自学术界和工业界的广泛关注.随着互联网的迅速发展，网络上的信息量呈现井喷式的暴涨，用户逐渐陷入信息的汪洋大海之中，如何快速且准确地在“过载”的信息中为用户找到自己真正需要的信息是互联网时代的主要任务以及当务之急。推荐系统通过分析用户的行为信息，得到用户的偏好模型，为用户提供个性化推荐服务[11]。1992 年 9 月，Xerox Palo Alto 研究中心开发了一套利用相关用户的显式反馈解决信息过载问题的实验系统 Tapestry[12]，用于邮件过滤。1994 年自动推荐的系统 GroupLens 诞生[13]。GroupLens 可以跨网站计算也可以自动完成推荐。GroupLens 是为 Usenet 新闻过滤而产生的。1995 年~1996 年间，由于数据量增大，信息过载问题日渐严重，这一阶段的研究集中在计算性能的提升，降维方法和基于物品的关联规则算法都是这一时期产生的。在 1997 年，“推荐系统”（Recommender System, RS）的概念首次被提出[14]，协同过滤算法是至今为止发展最成熟，应用最广泛的推荐算法。该算法的诞生标志着推荐系统的诞生[15]。其核心思想是：使用先验可用的用户对项目评分集来了解用户和项目之间的相互依赖关系，通过相邻项目的评分（基于邻居的[16-17]）或推测低维嵌入（Low-Dimensional Embedding）（基于潜在因子的[18-19]）来预测用户对项目的评分[20]。在 1997 年以后，推荐系统逐渐被应用到电子商务网站中，比如著名的亚马逊（Amazon.com）。在亚马逊的推荐系统中，率先使用基于物品的协同过滤算法，可以处理超大规模的评分数据，推荐系统的应用为亚马逊带来了空前的效益，推荐系统为其做出的贡献率在 20%~30%[21]。推荐系统还被广泛应用于广告推送，例如社交网站 Facebook，采用了广告推荐，好友推荐等。2000 年到 2005 年互联网泡沫到来，当时的新兴的以推荐系统业务为核心的公司纷纷倒闭，但是推荐技术的研究依然继续。2006 年以后，推荐系统的研究来了一波新的高潮。随着推荐系统技术的不断迭代，其为各行各业带来的效益不断提升。随着互联网的爆炸式扩张，当今推荐系统还有很多问题亟待解决。如数据稀疏性问题等。数据稀疏性是推荐系统最闫军的问题之一[22]。数据稀疏会导致根据目标用户查询出的邻居用户不正确。引发结果准确度偏低。

在国内，互联网发展迅猛，崛起飞快。电子商务是目前国内引入推荐系统最为广泛的商业领域，比如：淘宝、天猫、京东商城等等。这些互联网公司会针对不同用户的不同需求分析用户的偏好，为用户“量身定做”感兴趣的商品列表作为推荐。推荐系统技术可以为企业带来更多的商业价值。社交网络引入推荐系统也比较广泛，如微博、微信、朋友圈等。推荐用户感兴趣的用户。即朋友推荐。推荐系统的应用范围不断扩张，在新的形势下不断提出新需求，这将推进推荐技术不断进步，从而促进社会发展。

1.4 本文组织结构

本文共分为六章，每章主要研究内容如下：

第 1 章绪论。介绍了本文的研究背景和意义，基于电力改革背景，提出电力交易推荐系统设计方案和推荐算法设计方案。介绍了电力交易和推荐系统的国内外研究情况。

第 2 章相关技术。从推荐算法、统计方法和推荐系统三个角度简要介绍本文研究的推荐系统相关的核心技术。

第 3 章提出基于时序社交关系的协同过滤算法。给出了算法的推荐框架。在三组真实数据集上验证了算法的准确性，且效率与传统的推荐算法相比有一定程度的提高。

第 4 章提出基于用户偏好估计的协同过滤算法，并详细阐述算法思想。最后通过在两个真实数据集上对算法的精确性和数据稀疏性问题上做出了评价。

第 5 章设计并实现基于 Android 平台下的电力交易推荐系统原型，使用软件工程的思路阐述电力交易推荐系统移动端原型的设计和实现过程。

第 6 章总结与展望。对本文研究的内容做出总结，并说明本文亟待解决的问题，对下一步工作进行展望。

第2章 相关技术

个性化推荐系统是信息时代解决信息过载问题[10]的关键技术之一。本章将从推荐算法、统计方法和推荐系统三个角度简要介绍本文研究的推荐系统相关的核心技术。

2.1 推荐算法相关技术

本节要讨论现在主流的推荐系统中所广泛使用的几个核心协同过滤算法。这几种算法具有的典型思想将为本文提出的算法提供丰富的参考价值。

2.1.1 协同过滤算法

基于用户的协同过滤算法是推荐系统中最成熟、应用最为广泛的推荐算法[15]。其首先被用于电子邮件的过滤。基于用户的协同过滤算法其实是一种考虑用户的相关性而把相关用户偏好的物品推荐给目标用户的方法，属于基于邻域的推荐算法。基于用户的协同过滤算法的两个步骤如下：

（1） 搜索与目标用户具有相似偏好的用户；

（2） 在搜索用户喜欢的物品中找出目标用户不了解的目标物品推荐给目标用户。

用户与其喜欢的商品构成了用户-物品列表，利用用户-商品列表数据建立物品到用户的列表，如图 2.1 所示。然后建立用户的相似矩阵 W，对物品到用户的列表遍历，认为喜欢同一物品的用户都是类似的，更新他们之间的关系值。例如对于物品v_1，有u_1和u_2两个用户喜欢它，则W_{u_1,u_2}和W_{u_2,u_1}都自增 1。将物品-用户列表扫描结束后，就可以得出一个庞大的用户相似关系的矩阵W了。然后利用常用的相似度计算方式，计算出用户间偏好的相似度。

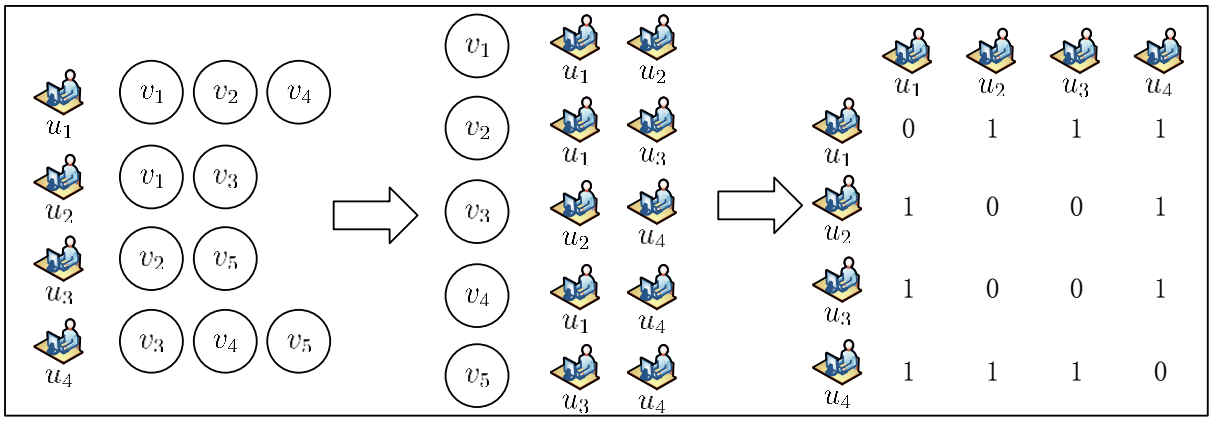


图 2.1 基于用户的协同过滤

Fig. 2.1 User-based collaborative filtering

基于用户的协同过滤算法的思想是：将与目标用户最相似的 Top-*k* 个用户喜欢的物品推荐给该目标用户。用户 *u* 对物品 *i* 的喜爱程度可以定义为：

$$p(u,i)=\sum_{u'\in S(u,k)\cap U_i} w_{u,u'}r_{u',i} \tag{2.1}$$

其中，$S(u,k)$是用户 *u* 偏好最相似的 Top-*k* 个用户集合，U_i表示对物品i表示喜欢的用户集合，$w_{u,u'}$表示用户 *u* 和用户u'的偏好相似度，$r_{u',i}$是用户u'对物品 *i* 的偏好。据此即可给出推荐列表。

基于用户的协同过滤算法思路较为简单，但是推荐性能略差。可以针对用户相似度计算进行改进，是计算出来的用户相似度更加逼真，文献[23]就是利用这种思路提升了推荐质量。

协同过滤技术的另一种算法是基于物品的协同过滤算法，它是目前工业上应用最广泛的推荐算法，首先由亚马逊提出[21]。由于基于用户的协同过滤算法存在一些天生的劣势，如：用户数量增加，产生用户偏好的相似度矩阵的计算时间会非线性提升，另外，也无法对产生的推荐结果做出合理解释。基于物品的协同过滤算法通过分析用户的行为来计算物品之间的相似度，用户对两种物品都喜欢则两种物品的相似度就越高。该算法也主要分为两个步骤：

（1） 计算物品的相似度；

（2） 利用计算的相似度与用户发生的行为对目标用户做出推荐。

计算物品相似度的方法如下：

$$w_{i,j}=\frac{|U_i\cap U_j|}{\sqrt{|U_i|\cdot|U_j|}} \tag{2.2}$$

计算好物品间的相似度后，与式(2.1)类似，基于物品的协同过滤算法通过式计算用户u对物品j的偏好：

$$p(u,i')=\sum_{i'\in S(i',k)\cap I_u} w_{i,i'}r_{u,i'} \tag{2.3}$$

这里，I_u是用户 *u* 喜欢的物品集合，$S(i',k)$是与物品i'最相似的 Top-*k* 个物品集合，$w_{i,i'}$是物品i与物品i'的相似度。基于物品的协同过滤算法也可以从相似度的角度来优化，文献[23]是通过分析用户活跃的程度来对物品相似度度量从而提高推荐质量。另外，还可以对物品相似度进行正规化来提高覆盖率与多样性，文献[24]给出相关的研究结果。

2.1.2 概率矩阵分解

由于传统的协同过滤算法对大数据量的处理性能较差，文献[18]提出了概率矩阵分解

理论。概率矩阵分解的核心思想是线性因子分析模型，用户的偏好建模成几个向量的线性组合，然后使用梯度下降法迭代求出特征向量。

设有 *n* 个用户，*m* 个商品。还有评分矩阵 *R*，其中每个元素$r_{i,j}$表示用户 *i* 对商品 *j* 的评分，用来衡量用户的偏好。引入矩阵分解的思想将评分矩阵 *R* 分解成两个 *k* 维的特征矩阵，即用户特征矩阵$U \in \mathbf{R}^{k\times m}$和商品特征矩阵$V \in \mathbf{R}^{k\times n}$，这$k$个维度，称为隐式特征因子（Latent Trait Factor）[20]，特征矩阵在这 *k* 个维度的嵌入称为隐式特征向量（Lantent Trait Vector）[20]。分解后的用户特征矩阵用 *U* 来表示，U_i代表用户i的隐式特征向量；分解后的商品特征矩阵用 *V* 来表示。V_j表示商品 *j* 的隐式特征向量。现对用户和商品的隐式特征向量的分布情况作出假设，假设用户和商品的隐式特征向量都服从高斯先验分布：

$$p(U|\sigma_U^2) = \prod_{u=1}^{n} \mathcal{N}(U_u|0, \sigma_U^2 \boldsymbol{E}) \tag{2.4}$$

$$p(V|\sigma_V^2) = \prod_{i=1}^{m} \mathcal{N}(V_i|0, \sigma_V^2 \boldsymbol{E}) \tag{2.5}$$

再假设已观测的的评分数据条件概率也服从高斯先验分布：

$$p(R|U, V, \sigma_R^2) = \prod_{u=1}^{m}\prod_{i=1}^{n} \left[\mathcal{N}(R_{u,i}|U_u^T, \sigma_R^2)\right]^{\mathcal{I}_{u,i}^R} \tag{2.6}$$

其中$\mathcal{I}_{u,i}^R$为指示函数，表示：若用户 *u* 对商品 *i* 给出过打分，则该函数值为 1，否则为 0。

由以上几个假设可以得出用户与商品的后验概率如下：

$$\begin{aligned} p(U, V|R, \sigma_R^2, \sigma_U^2, \sigma_V^2) =& p(R|U, V, \sigma_R^2, \sigma_U^2, \sigma_V^2) p(U, V) / p(R, \sigma_R^2, \sigma_U^2, \sigma_V^2) \\ \sim& p(R|U, V, \sigma_R^2, \sigma_U^2, \sigma_V^2) p(U, V) \\ =& p(R|U, V, \sigma_R^2) p(U|\sigma_U^2) p(V|\sigma_V^2) \\ =& \prod_{u=1}^{n}\prod_{i=1}^{m} \left[\mathcal{N}(R_{u,i}|g(U_u^T V_i), \sigma_R^2)\right]^{\mathcal{I}_{u,i}^R} \\ & \times \prod_{u=1}^{n} \mathcal{N}(U_u|0, \sigma_U^2 \boldsymbol{E}) \prod_{i=1}^{m} \mathcal{N}(V_i|0, \sigma_V^2 \boldsymbol{E}) \end{aligned} \tag{2.7}$$

然后，对式(2.7)两端取对数，并且将高斯函数展开，可得到最终的结果：

$$\begin{aligned}\ln p(U,V|R,\sigma_R^2,\sigma_U^2,\sigma_V^2) = &-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{v=1}^{m}\mathcal{I}_{u,v}^{R}\left(R_{u,v}-U_u^TV_v\right)^2\\ &-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^TU_u-\frac{1}{2\sigma_V^2}\sum_{v=1}^{m}V_v^TV_v\\ &-\frac{1}{2}\left(\left(\sum_{u=1}^{n}\sum_{v=1}^{m}\mathcal{I}_{u,v}^{R}\right)\ln\sigma_R^2+nk\ln\sigma_U^2+mk\ln\sigma_V^2\right)\\ &+C\end{aligned} \tag{2.8}$$

最后，对式(2.8)求偏导，可得到梯度。概率矩阵分解过程可以由图 2.1 描述。

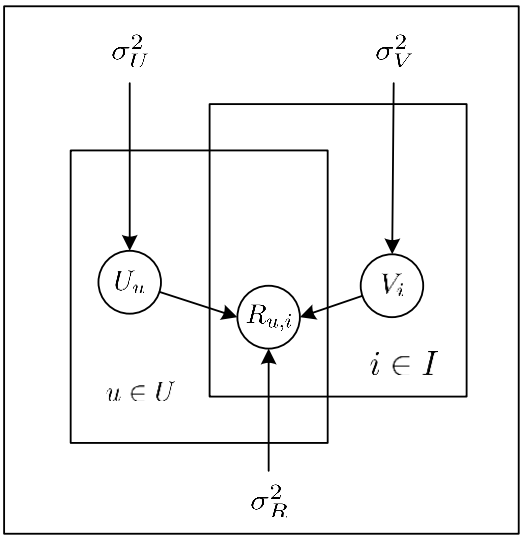


图 2.2 概率矩阵分解示意图
Fig. 2.2 Probability matrix decomposition diagram

经典的概率矩阵分解算法给出了基于概率的用户和商品的特征向量的求解方法，是一种准确性较高的推荐算法。然而，该方法缺少对用户自身以及商品自身相关关系的分析，以至于求解的准确度略有局限。

2.2 统计方法相关技术

人工智能技术的一个关键基石就是统计方法。统计学作为一种研究不确定性问题的理论学科，是数学的一个重要分支。机器学习是统计学与计算机的交叉学科，推荐系统中会广泛的使用机器学习和统计学中的相关技术。本节就与本文讨论的推荐系统相关的核心统计学知识加以梳理。

2.2.1 核密度估计

核密度估计（Kernel Density Estimation）的方法是统计学中一种使用有限的样本来估计概率密度的方法，属于非参数估计的一种。在实际生产生活中，总体的概率密度往往未知。需要通过抽样方法来对总体的概率密度进行估计。常用的估计方法分为两种，

即参数估计（Parameter Estimation）与非参数估计（Non-parametric Estimation）。常用非参数估计有直方图法和核密度估计方法。核密度估计的一般形式如式(2.9)所示：

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{2.9}$$

这里$K(x)$是核函数，核函数要符合概率密度函数的性质，满足非负，其积分为 1，均值为 0，$h > 0$是带宽。核函数有多种，如 Uniform,Triangular,Triweight, Epanechnikov,Normal 等。各种核函数的图形如图 2.3。

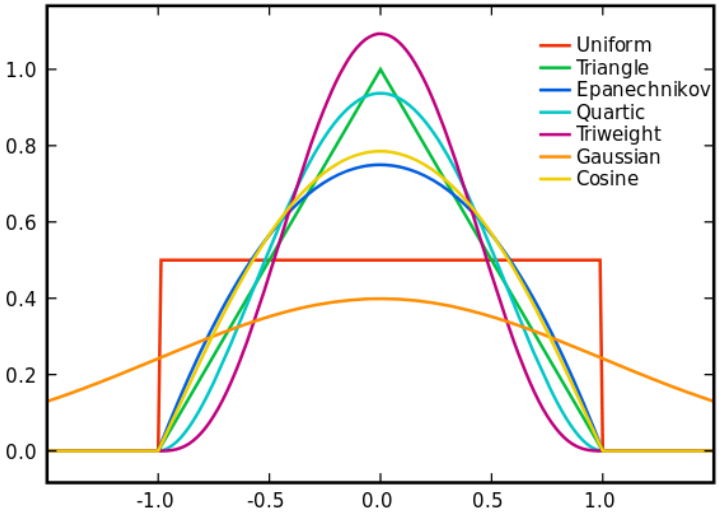


图 2.3 各种核函数图像

Fig. 2.3 Various kernel function images

以高斯核函数和三角和函数为例，这里给出高斯核函数其形式如式(2.10)、三角核函数形式如式(2.11)。

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \tag{2.10}$$

$$K_t(u) = \begin{cases} 0, & |u| > \sqrt{2}h \\ \dfrac{\sqrt{2}h - |u|}{2h}, & otherwise \end{cases} \tag{2.11}$$

如果核函数使用高斯核函数，则带宽h的最优选择为

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \tag{2.12}$$

这就是 Silverman 经验法则[25]。若使用其他的核函数可以最小化 L2 风险函数来决定。由文献[26]知，影响估计准确性的关键性因素不是核函数的形状，而是带宽的选择。带宽越小，越适合于揭示概率密度分布的局部特征，而较大的带宽可以在较为全局的尺度下使热点区域体现得更加明显。对于稀疏型的数据点分布应该采用较大的带宽，而对于密

集型的数据点分布则应考虑使用小一些的带宽。

核密度估计作为一种非参数估计的统计方法被广泛的应用在社会科学、物理科学、生命科学以及各种工程技术领域[25]。可以运用在非参数判别、聚类分析、随机数模拟、多峰性检验等[27]。

2.2.2 常用统计距离度量方法

在统计学中，常常要知道个体之间的差异，如使用相似度计算的方式评价两个个体之间的相似性。常见的两个个体之间的距离可以用曼哈顿距离，欧氏距离，余弦相似度，Jaccard 距离等等。而更加常见的场景是对两个分布进行相似性度量，即距离的测定。本部分将简要描述两种基于分布距离的测量方法：Kullback-Leibler 散度和 F 散度。

（1） Kullback-Leibler Divergence

KL 散度(Kullback-Leibler Divergence,KLD)，也称作相对熵[28]。用来衡量两个分布之间的有向分歧。

$$D_{KL}(P\|Q)=\sum_{x\in X}P(x)\log\frac{P(x)}{Q(x)} \tag{2.13}$$

KL 距离可以解释为在相同的事件空间中，概率$P(x)$和$Q(x)$分布的差异情况。其物理意义是在相同事件空间里，概率分布$P(x)$的事件空间，若用概率分布$Q(x)$编码时，平均每个基本事件（符号）编码长度增加了多少比特。虽然 KL 散度描述了两个分布之间的相异性，但是 KL 散度并不是真正的距离，因为距离需要满足三个条件：非负性、对称性和三角不等式性质。KL 散度不满足对称性和三角不等式性质。

（2） F-Divergence

F 散度，也是一种衡量两个概率分布之间差异的方法。

$P(x)$和$Q(x)$是两种概率密度函数。它们在同一空间中。则它们之间的 F 散度可以表示为如式(2.14)的形式，

$$D_F(P\|Q)=\int Q(x)f\left(\frac{P(x)}{Q(x)}\right)dx \tag{2.14}$$

其中$f(x)$为凸函数，且$f(1)=0$。

由式(2.14)可知 KL 散度定义式(2.13)其实是 F 散度的特殊情况，即$f(x)$取对数 $\log(x)$即可。常用的 F 散度还有 Hellinger 距离、Total Variation 距离[29]等等。它们都是不同于 KL 散度是有界且对称的。

此外，还有很多用于描述分布之间的距离，比如 Wasserstein 距离[30]等。在实际运用中要针对不同的业务场景，选择合适的距离。

2.3 推荐系统相关技术

本节将简述将推荐算法应用到实际的场景中的实施化方案的相关技术。在本文的电力交易推荐系统中，要结合业务场景搭建符合要求的移动端推荐平台，该平台基于 Android 开发。因此，本节首先简要介绍 Android 开发平台，然后对常见的推荐系统服务端的架构给出参考。

2.3.1 移动客户端 Android 平台概述

Android 操作系统是一个软件组件的栈，下面分析 Android 操作系统的架构特点。

Linux 内核是 Android 操作系统的基础。Linux 系统内核能为 Android 系统提供安全性保障，并授权硬件的制造商开发驱动程序。

Android Runtime(ART)是 Android 运行时实例，每个 Android 应用在运行时都会创建该实例。在低内存的设备上运行多个虚拟机，基于这种虚拟机上可以运行 Android 环境下的字节码文件 DEX，这点与 Java 的运行机制类似。ART 上运行着垃圾回收机制，实时的对内存分析，对垃圾回收。

Java API 框架为开发者提供 Android 系统的所有功能，由这些 API 组成构建 Android 应用的所有模块，开发者可以使用这些 API 完成应用开发工作。

2.3.2 推荐系统服务端概述

一个好的推荐系统应该满足以下三点非功能性需求：

（1） 对海量数据能做到快速准确地处理新增数据，并能实时交互；

（2） 可以灵活的加入或更换各种推荐算法；

（3） 延迟较低、响应较高、精准推荐服务。

广泛采用的推荐系统框架如图 2.4，推荐系统分为在线阶段和离线阶段。在线阶段主要负责特征抽取和评分预测；离线阶段负责对样本抽取，划分数据集，训练模型，生成预测评分。

下面就推荐系统服务端应用的主流技术框架做出简要介绍。

目前主流的推荐系统的推荐过程依赖于大数据分布式生态系统，如 hadoop。推荐系统服务端要依赖于三种服务：存储、计算、数据流。存储包含对文件的存储，需要 HDFS 提供分布式文件系统存储服务，还有分布式环境下的数据库，如 HBase，它是一种 Key-Value 形式的数据库，以及基于 HDFS 下的数据仓库，如提供类 SQL 查询语言的 Hive。推荐系统所依赖的分布式计算服务有 MapReduce 框架提供离线分析，还有基于内存的计算模型 Spark，它可以提供实时计算服务。数据流包含分布式消息队列 Kafka，实时流

式计算框架 Storm，以及提供高可用、高可靠的分布式海量日志采集聚合和传输服务的 Flume。

在推荐系统的在线阶段，要求服务引擎具有高并发性、低延迟、高稳定性；负载均衡、扩展性较强。在线的计算可以使用数据缓存（如 Memcache、Redis 等），任务是对数据加载和更新、运算、预测等。

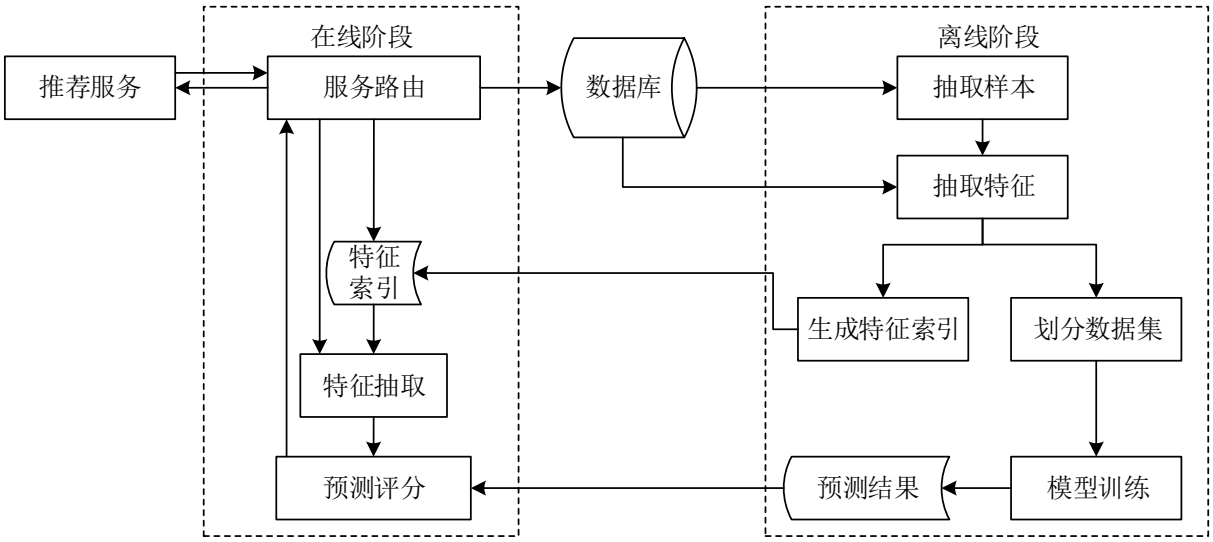


图 2.4 广泛采用的推荐系统框架
Fig. 2.4 Widely used recommended system framework

在离线阶段，先要对用户的兴趣偏好进行建模，对商品建模、以及数据的预处理（如用户、商品聚类，内容去重，数据清洗等），然后使用高精确度的推荐算法对数据处理，给出推荐结果，这部分结果就是在线阶段需要推送给用户的了。

还有一个介于在线阶段和离线阶段中间的阶段，要负责从日志服务器中采集用户的行为数据。该过程可以利用 Flume 的日志采集功能，并将获取到的用户行为数据分发出去。将事件发送给 Kafka。离线阶段将数据存储至 HDFS。

2.4 本章小结

本章介绍了本文研究工作中涉及到的核心技术。从推荐算法、统计方法和推荐系统三个角度展开描述。推荐算法部分简要描述了协同过滤、概率矩阵分解的步骤以及技术特点。统计方法部分介绍了核密度估计和统计距离度量方法。核密度估计是一种非参数估计，统计距离度量方法主要介绍了两种分布之间的距离度量方法。推荐系统部分从推荐系统服务端的架构和技术方面做出简要分析，在此部分还对本文使用的 Android 平台做出了简要介绍。

第3章 基于时序社交关系的协同过滤算法

针对改革后的电力市场交易模式，根据电力推荐系统的需求，本章提出了基于时序社交关系的协同过滤算法。首先综合分析电力交易推荐系统拟解决问题，然后提出利用用户的交易时序信息挖掘用户的影响关系和从众关系，然后结合概率矩阵分解算法，使用随机梯度下降法求解用户和商品的特征因子向量。给出了效率相对较高的推荐框架，并分析计算复杂度。最后在三个真实数据集上验证了本章提出的算法的准确性和效率等与传统的推荐算法相比有一定程度的提高。

3.1 问题定义

本节介绍基于时序社交关系的协同过滤推荐算法所要解决的问题，先引入问题，说明问题的来源，然后介绍电力交易推荐系统需要解决的问题，最后对问题给出形式化的描述。

3.1.1 问题引入

伴随着互联网的发展，协同过滤算法是至今为止发展最为成熟、应用最为广泛的推荐算法。该算法的诞生标志着推荐系统的诞生[15]，也是本文的理论依据[31]。本文的推荐算法主要思想是：挖掘用户与商品的评分数据中的用户行为信息，然后通过评分预测的方式对目标用户的偏好行为进行预测。本文所指“用户”，即大用户。为可接入较高的电压等级。具备一定购电规模的电力用户[32]。“商品”即电力能源，发生交易的前提是大用户直购电工作的开展。在电力体制中，大用户直购电是电厂和终端购电大用户之间通过直接交易的形式协定电量和购电价格[33]，然后委托电网企业将协议电量由发电企业输配终端购电大用户，并另支付电网企业所承担的输配服务[33]。

随着电力改革的推行，大用户可以直接与发电企业达成交易，电网企业不再是众多大用户的唯一售电方。发电企业面对的销售对象逐渐增加以及这样所产生的竞争效应，势必会大大提高电力生产技术和价格定位的灵活性。减少对电力市场的约束，让市场自动调节，使生产和消费双方获得双赢。

3.1.2 电力交易推荐系统拟解决问题

大用户通过自身需求与一个或者多个发电企业进行直接的自主选择交易，这种直购电模式在试点地区运营过一段时间后带来了一定的改革红利。电力交易推荐系统的最终目的是针对电力改革后对电力市场中要出现的新大用户推荐令其满意的发电企业。为电

力市场中新出现的大用户推荐其满意的发电企业作为其要进行交易的候选对象。电力交易的推荐系统主要解决以下几个问题。

（1） 为用户生成满意度较高的推荐列表

为用户生成推荐列表首先要获取用户偏好，要了解用户究竟想要什么。最好的办法是用户注册系统时就主动描述其偏好告知系统，但是这样存在三个不可行之处：首先，用户的偏好无法用当前的自然语言处理技术完全理解；其次，用户的偏好是不断产生变化的，总会产生新的变化，然而用户却不会经常在系统中更新他们的偏好；另外，有些偏好无法用语言来描述，以至于某些用户无法明确自己真实的偏好[31,34]。基于以上问题，用户的历史行为和偏好是推测用户未来行为和偏好的宝贵资料，利用这些数据可以巧妙的避开上述不可行之处，因此，大量的用户数据是推荐系统的重要元素，要从现有的数据中挖掘出用户不断变化的、难以表达和理解的偏好。如何准确把握用户的偏好是影响准确率（推荐系统的重要指标）的关键问题。

（2） 评分矩阵数据稀疏性问题

数据稀疏指的是评分矩阵中只有少数单位（在大量的记录中）被有效地用来表示典型的数据向量[35]，而实际上有大量单位的值为零，只有少数单位的值是非零的。数据稀疏问题作为推荐系统的经典挑战已经成为众多研究者急于攻克的问题之一。在协同过滤算法中，通过分析用户交易行为以及用户交易后对所购商品的评分来预测用户对新商品的偏好评分进而给用户推荐商品，如果用户评分积极性较高则用户-商品评分矩阵会有较多有价值数据；然而，即使用户积极性再高，也无法针对所有商品给出有价值的评分，而评分矩阵中每个单位是某个用户对某一个商品的评分，因而评分矩阵会有很多空位；加之多数用户并不会有充分的积极性来对商品评分，数据的稀疏性可想而知。在电力交易系统中，由于电力改革前的交易模式单一，大用户与发电企业之间的交易一直被电网公司驱动，大用户的购电需求以及满意度评价往往只针对电网公司并非发电企业，历史遗留的原因造成数据稀疏性的问题更加严重。电力改革后，大用户可以对发电企业自主进行点对点交易，其选择要综合多方面因素，若大用户对发电企业给出的电能越满意则相应的交易次数就会增加，反之，交易次数会减少，因此将大用户与发电企业的交易次数设定为初始评分矩阵中的元素是完全可行的。其实，数据稀疏的问题在本质上是无法完全克服的，为了解决这个问题，很多文献中已经提出并使用了许多稀疏措施[36]例如：扩散算法[37]、迭代寻优算法[38]、转移相似性算法[39]。本文结合概率矩阵分解算法可以一定程度上解决数据稀疏性问题。

（3） 新大用户进入系统的冷启动问题

正如前面提及的,用户的历史行为和偏好是推测用户未来行为和偏好的宝贵资料,因此大量的用户历史行为数据就成为推荐系统的极其重要的组成要素以及先决条件[15]。这些问题在多数互联网公司里面或许算不上问题，因为经过长时间的运营这些互联网公司的互联网应用积累了大量的用户行为数据。但是对于电力推荐系统来说，大用户历史行为数据的缺乏是关键性问题，“巧妇难为无米之炊”，这类由于用户历史行为数据缺失而导致推荐不准确的问题，称之为“冷启动问题”（Cold Start）。一般来说，冷启动问题主要分为三类：用户冷启动、物品冷启动、系统冷启动[15]。用户冷启动就是前面提到的缺乏用户历史行为数据而造成无法借此对其个习惯化推荐；物品冷启动意味着将新的商品推荐给可能对它感兴趣的用户；系统冷启动代表在一个全新开发的网站上，没有用户以及用户交易行为，也要能给新用户带来个性化推荐的服务体验。针对这三类问题，有不同的解决方案，大体上有几类：提供非个性化推荐服务、利用新用户注册时提供的基本信息做粗糙的个性化和利用社交网络导入社交网站上好友信息等[15]。

在电力交易推荐系统中，由于新大用户没有明确的历史交易行为作为依据，无法提供准确的推荐给这类用户。因此，本文给出这样一种解决方案：当新用户注册时，即需要添加一些需求偏好，利用这些偏好数据作为它的历史交易行为数据，而匿名登录用户（即未注册用户）仅有浏览一定范围内的信息的权限，仅提供交易量较多的发电企业作为推荐结果。

3.1.3 问题描述

形式上，设有 n 个用户。其构成的集合为 U。用$u_i \in U$表示第i个用户，$\|U\| = n$。有 m 个商品。其构成的集合为 I。用i_j表示第 j 个商品，$\|I\| = m$。评分矩阵为$R = [r_{u,i}]_{n \times m}$。其中，$r_{u,i}$代表用户 u 对商品i的评分。利用概率矩阵分解模型。学习用户和商品的特征向量。然后利用特征向量预测评分。

假设$U \in \mathbf{R}^{k \times m}$和$V \in \mathbf{R}^{k \times n}$代表用户和商品的特征矩阵，其中，$U_u$和$V_i$代表某个特定用户 u 和商品 i 的 k 维特征向量。利用概率矩阵分解模型。学习用户和商品的特征向量。这是其核心思想。根据以上的定义，已有评分数据的条件概率定义如下：

$$p(R|U,V,\sigma_R^2) = \prod_{u=1}^{n}\prod_{i=1}^{m}[\mathcal{N}(R_{u,i}|g(U_u^T V_i),\sigma_R^2)]^{\mathcal{I}_{u,i}^R} \tag{3.1}$$

其中,

$\mathcal{N}(x|\mu,\sigma^2)$表示平均值为$\mu$,方差为$\sigma^2$的高斯分布。

$\mathcal{I}_{u,i}^R$是一个指示函数。若用户 u 对商品 i 给出评分。该函数值为 1，否则为 0。

$g(x)$将$U_u^T V_i$的值映射到$[0,1]$内，本文中$g(x)=(1+e^{-x})^{-1}$。

假设用户与商品的特征向量都服从$\mu=0$的高斯先验，

$$p(U|\sigma_U^2)=\prod_{u=1}^{n}\mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E}) \tag{3.2}$$

$$p(V|\sigma_V^2)=\prod_{i=1}^{m}\mathcal{N}(V_i|0,\sigma_V^2\boldsymbol{E}) \tag{3.3}$$

再假设已观测的评分数据条件概率也服从高斯先验分布，即

$$p(R|U,V,\sigma_R^2)=\prod_{u=1}^{m}\prod_{i=1}^{n}\left[\mathcal{N}(R_{u,i}|U_u^T,\sigma_R^2)\right]^{\mathcal{I}_{u,i}^R} \tag{3.4}$$

根据上述两个假设可得

$$\begin{aligned}&p(U,V|R,\sigma_R^2,\sigma_U^2,\sigma_V^2)\propto p(R|U,V,\sigma_R^2)p(U|\sigma_U^2)p(V|\sigma_V^2)\\&=\prod_{u=1}^{n}\prod_{i=1}^{m}\left[\mathcal{N}(R_{u,i}|g(U_u^T V_i),\sigma_R^2)\right]^{\mathcal{I}_{u,i}^R}\prod_{u=1}^{n}\mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E})\prod_{i=1}^{m}\mathcal{N}(V_i|0,\sigma_V^2\boldsymbol{E})\end{aligned} \tag{3.5}$$

对其取对数，求极大，可得在已知超参数σ_R^2, σ_U^2, σ_V^2和现有的评分矩阵的前提下可能性最大的 U和 V的隐式特征矩阵。

3.2 基于时序分析的用户社交关系选择

协同过滤算法的本质思想是使用先验可用的用户对项目评分集来了解用户和项目之间的相互依赖关系，通过相邻项目（Neighbor-based）的评分[16-17]或推测低维嵌入（Low-dimensional Embedding）来预测用户对项目的评分[20]，后者是基于潜在因子（Latent Factor-based）的矩阵分解[18-19]。概率矩阵分解算法（Probabilistic Matrix Factorization, PMF）是协同过滤算法的一类重要的分支，通过学习低维嵌入分解出潜在因子矩阵（Latent Factors Matrix, LFM）进行推荐，可以高效的处理海量数据。然而，传统的概率矩阵分解算法存在一些缺陷：概率矩阵分解算法的预测准确性较高，但其推荐精度会受初始评分矩阵稀疏特性的影响[40-41]；在概率矩阵分解算法中。使用了潜在因子。不能给出各个因子的解释。也无法给出推荐的解释。此外，概率矩阵分解方法还必须做出属性因子之间满足独立同分布条件的假设，没有覆盖用户与产品间关联关系对矩阵分解的误差影响[42]；概率矩阵分解模型学习数据之后。评分量很少的用户。其特征向量近似于先验分布的平均值。最终导致矩阵评分预测更趋向于商品的均分[43]。其中最重要一点是用户之间或产品之间的关联关系是影响推荐效果的关键性因素[11]，需要通过评定用户或产品之间的关联关系而找到近邻用户或近邻商品，这样能更准确地识别用户的个人偏好，从而有效提

高算法的精确度。

3.2.1 用户影响关系和从众关系的定量分析

当用户考虑是否购买某一件商品时，最直接的影响因素是该用户关注的用户是否在上一时刻购买了该商品，其次就是看看大家是否足够喜欢该商品，因此，在分析用户的购买行为时要综合分析对其影响最大的用户和从众心理因素。

（1） 用户间影响关系的定量分析

为了评定用户或产品之间存在的隐式特征关系，在经典的概率矩阵分解算法中没有考虑用户或产品的交易时间信息，通过分析用户的时序行为会挖掘出隐藏着的用户交易规律，从而提取出用户或产品之间的关联关系。

定义 3.1：用户影响因子 设用户u_i和用户u_j购买的商品集合分别为I_i和I_j，在一定的时间阈值τ内，用户u_i的购买行为发生在用户u_j的购买行为之前，用户u_i对用户u_j具有影响作用，用户u_i对用户u_j的影响因子l_{u_i,u_j}为

$$l_{u_i,u_j} \triangleq \frac{w_{i,j}}{\|I_i \cup I_j\|} \tag{3.6}$$

其中，$w_{i,j}$是影响权重，即用户u_i和用户u_j在一定的时间阈值τ内先后购买了同一个商品则影响权重$w_{i,j}$增加一个单位，那么用户u_i对用户u_j的影响因子就是影响权重与两个用户购买商品的并集的商。由于购买行为存在先后顺序，而这种影响关系也具有方向性。

（2） 用户的影响与从众关系的定量分析

定义 3.2：用户影响力 设F_u为用户 *u* 的 Followers 集合，代表用户 *u* 可以影响的用户集合为F_u，则用户 u 的影响力

$$l_u \triangleq \frac{\|F_u\|}{\|U\|} \tag{3.7}$$

其中，U为全体用户。

定义 3.3：用户从众因子 设L_u为用户 *u* 的 Leaders 集合，即用户u受到L_u集合中的所有用户的影响，则用户 u 的从众因子为

$$f_u \triangleq \frac{\|L_u\|}{\|U\|} \tag{3.8}$$

其中，U为全体用户

若f_u越大，则用户 *u* 受到其他人影响的可能性就越大，而l_u只是描述用户u所能影响到的用户范围大小。因而结合公式(3.6)和公式(3.8)与概率矩阵分解可以提高协同过滤推荐算法的准确度。

3.2.2 用户影响关系集合的获取

定义如下符号表：

表 3.1 部分符号表
Table 3.1 Part symbol table

符号	描述
D_r	评分数据集
S	训练集
T	测试集

将D_r划分为训练集S和测试集T，从D_r中提取每一个用户的 *Followers* 集合和 *Leaders* 集合，分别构成矩阵 *F* 和 *L*（矩阵的第*i*行代表用户 *i* 的 *Followers* 集合或 *Leaders* 集合，即$F[i]=F_i$,$L[i]=L_i$）。

Algorithm 1 Find Followers&Leaders

Input:
S : the train set;
τ : the threshold of time;
γ : the threshold of rating.
Output:
F : the followers set;
L : the leaders set.

```
1: F ← φ, L ← φ // initialize F and L with φ.
2: for each (u_0, i_0, r_0, t_0) ∈ S do
3:    T ← T ∪ {(u_0, i_0, r_0, t_0)} /* T which is initialized with φ,
      is the finished set of current iteration. */
4:    for each (u, i, r, t) ∈ S\T do
5:       if i == i_0 then
6:          Δt ← t − t_0, Δr ← |r − r_0|
7:          /* If Δt is positive, u follows u_0, otherwise, u leads u_0.
            f_u0 and l_u0 which are initialized with φ in this iteration,
            are the set of follow and lead tuples. */
8:          if 0 < Δt < τ ∧ Δr < γ then
9:             f_u0 ← f_u0 ∪ {(u, i, r, t)}
10:         else if −τ < Δt < 0 ∧ Δr < γ then
11:            l_u0 ← l_u0 ∪ {(u, i, r, t)}
12:         end if
13:      end if
14:   end for
15:   F ← F ∪ {(u_0, f_u0)} // update F, add the follow tuple to F.
16:   L ← L ∪ {(u_0, l_u0)} // update L, add the lead tuple to L.
17: end for
18: return F, L
```

在该算法中，第 2 行到第 17 行是对训练集每一个元组都进行处理；在第 3 行和第 4 行的 for 循环是为了对训练集中没有处理过的元组进行处理，处理过程为第 5 行到第 13 行。每次针对一个元组要判断时间是否达到时间阈值τ，评分是否在指定范围内。如果

符合条件就将元组存入临时变量，有两个临时变量f_{u0}和l_{u0}分别来存储 *Follows* 集合元组和 *Leaders* 集合元组。迭代寻找结束后，就将对应的结果加入到 *Follows* 集合和 *Leaders* 集合，如第 15~16 行所示。

3.3 SeqSoPMF 推荐算法描述

本节提出基于时序社交关系的协同过滤算法（Collaborative Filtering Algorithm Based on Time Social Relationship，SeqSoPMF）。算法核心分为五个步骤：导入数据、构建用户社交关系网络图、分析用户从众因子、构建概率矩阵分解模型求解特征向量和重构评分矩阵实现个性化推荐。本节给出概率矩阵分解过程推导，以及 SeqSoPMF 算法框架，最后还对算法的复杂度进行分析。

3.3.1 基于社交关系的概率矩阵分解

通过使用用户发生交易的时序信息挖掘出用户间的影响关系和用户的从众关系应用于概率矩阵分解模型，用户的特征向量会受到其近似的用户的影响以及自身具有的从众属性，那么近似的用户会具有近似的特征向量，若从众关系越大，则用户受到其他人影响就越大，那么近似的特征向量作为估计量，可以得出用户的特征向量为，

$$\hat{U}_u = \sum_{u' \in S_u} l_{u',u} f_u U_{u'} \tag{3.9}$$

其中，$\hat{U}_u$表示用户近似的特征向量，S_u表示用户u的近邻集合。公式(3.9)不仅考虑了用户自身特征和相近用户特征的影响，还考虑了用户自身具备的从众因子的影响。将时间顺序分析融入了用户的特征向量，用户间的影响关系比较明确，为了防止过拟合，每个用户的特征向量服从$\mu = 0$的高斯分布，而且与关联用户的特征向量相近邻，由式(3.5)，可得，

$$\begin{aligned} & p(U|l,\sigma_U^2,\sigma_l^2) \propto p(U|\sigma_U^2)p(U|l,\sigma_l^2) \\ =& \prod_{u=1}^{n} \mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E}) \prod_{u=1}^{n} \mathcal{N}\left(U_u \middle| \sum_{u' \in S_u} l_{u',u} U'_u, \sigma_l^2 \boldsymbol{E}\right) \end{aligned} \tag{3.10}$$

相似的，由公式(3.5)的推导方法可以得出，

$$\begin{aligned} & p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2) \propto p(R|U,V,\sigma_R^2)p(U|l,\sigma_U^2,\sigma_l^2)p(V|\sigma_V^2) \\ =& \prod_{u=1}^{n}\prod_{i=1}^{m} \left[\mathcal{N}(R_{u,i}|g(U_u^T V_i),\sigma_R^2)\right]^{\mathcal{I}_{u,i}^R} \prod_{u=1}^{n} \mathcal{N}\left(U_u \middle| \sum_{u' \in S_u} l_{u',u} U_{u'}, \sigma_l^2 \boldsymbol{E}\right) \\ & \times \prod_{u=1}^{n} \mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E}) \prod_{i=1}^{m} \mathcal{N}(V_i|0,\sigma_V^2\boldsymbol{E}) \end{aligned} \tag{3.11}$$

最后，对$p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)$两侧取对数，并将上述结果代入式(3.11)，整理得

$$\begin{aligned}\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2) = &-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2\\ &-\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\ln\sqrt{2\pi}\sigma_R\\ &-\frac{1}{2\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2-\sum_{u=1}^{n}\ln\sqrt{2\pi}\sigma_l\\ &-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^2-\sum_{u=1}^{n}\ln\sqrt{2\pi}\sigma_U\\ &-\frac{1}{2\sigma_V^2}\sum_{i=1}^{m}V_i^2-\sum_{i=1}^{m}\ln\sqrt{2\pi}\sigma_V\end{aligned}\tag{3.12}$$

为了便于计算，将上式化简为式(3.13):

$$\begin{aligned}\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2) = &-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2\\ &-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^2-\frac{1}{2\sigma_V^2}\sum_{i=1}^{m}V_i^2\\ &-\frac{1}{2\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2\\ &-\frac{1}{2}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln\sigma_R^2\\ &-\frac{1}{2}\left(nk\ln\sigma_l^2+nk\ln\sigma_U^2+mk\ln\sigma_V^2\right)+C\end{aligned}\tag{3.13}$$

其中，C为常数，$C=-\frac{1}{2}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln(2\pi)-(n\cdot k)\ln(2\pi)-\frac{m\cdot k}{2}\ln(2\pi)$.

对上式左右同时乘以$(-\sigma_R^2)$可得

$$\begin{aligned}(-\sigma_R^2)\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2) = &\frac{1}{2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_U^2}\cdot\sum_{u=1}^{n}U_u^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_V^2}\cdot\sum_{i=1}^{m}V_i^2\\ &+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_l^2}\cdot\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2+\frac{1}{2}\sigma_R^2\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln\sigma_R^2\\ &+\frac{1}{2}\sigma_R^2\left(nk\ln\sigma_l^2+nk\ln\sigma_U^2+mk\ln\sigma_V^2\right)+C_1\end{aligned}$$

令风险函数为$L(R,l,U,V)=(-\sigma_R^2)\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)$，最大化后验概率就相当于最小化风险函数$L$，即

$$L(R,l,U,V)=\frac{1}{2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_U^2}\cdot\sum_{u=1}^{n}U_u^2 + \frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_V^2}\cdot\sum_{i=1}^{m}V_i^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_l^2}\cdot\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2 \tag{3.14}$$

梯度的计算方法如下，

$$\frac{\partial L}{\partial U_u}=\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}V_ig'(U_u^TV_i)(g(U_u^TV_i)-R_{u,i})+\frac{\sigma_R^2}{\sigma_U^2}U_u + \frac{\sigma_R^2}{\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)\left(1-\sum_{u'\in S_u}l_{u',u}\right) \tag{3.15}$$

$$\frac{\partial L}{\partial V_i}=\sum_{u=1}^{n}\mathcal{I}_{u,i}^{R}U_ug'(U_u^TV_i)(g(U_u^TV_i)-R_{u,i})+\frac{\sigma_R^2}{\sigma_V^2}V_i \tag{3.16}$$

其中，$g'(x)=\dfrac{e^{-x}}{(1+e^{-x})^2}$，为$g(x)$的一阶导数。然后利用随机梯度下降可以得到用户或商品的特征向量。

更详尽的推导过程参见附 A。

3.3.2 SeqSoPMF 推荐算法框架

本文的 SeqSoPMF 推荐算法是一种协同过滤算法，引入了基于时序分析的用户社交关系挖掘模型，利用概率矩阵分解算法给出推荐项集，可以更精确地预测用户行为。综合来看，本文提出基于时序社交关系的概率矩阵分解算法的推荐框架分为 5 个阶段（如图 3.1）：

（1） 导入数据

导入包含用户基础信息、评分信息以及评分时间信息的基础数据。

（2） 构建用户社交关系网络图

根据导入的数据构建用户社交关系网络图（如图 3.1（2））。其中每个节点代表一个用户，节点旁边的数字代表该用户已经发生交易的商品数量，即式(3.6)中的I_i；边代表用户之间的 *Follow* 关系，边上的权值代表两个用户发生 *Follow* 关系的条件下，他们交易的相同商品的数量，即式(3.6)中的$\omega_{i,j}$，从而计算出影响力最大的近邻集合。

（3） 分析用户从众因子

根据算法 1（Find Followers&Leaders）分析用户间的 *Follow* 和 *Lead* 关系，确定 *Followers* 集合和 *Leaders* 集合，根据式(3.8)求出用户的从众因子。

（4） 构建概率矩阵分解模型求解特征向量

将第 2 步得出的近邻集合与第 3 步得出的从众因子结合进概率矩阵分解模型，利用

随机梯度下降方法分解出用户特征向量与商品特征向量。

（5） 重构评分矩阵实现个性化推荐

根据第（4）步得出的用户特征向量和商品特征向量预测评分矩阵，利用预测的评分矩阵生成推荐列表给用户。

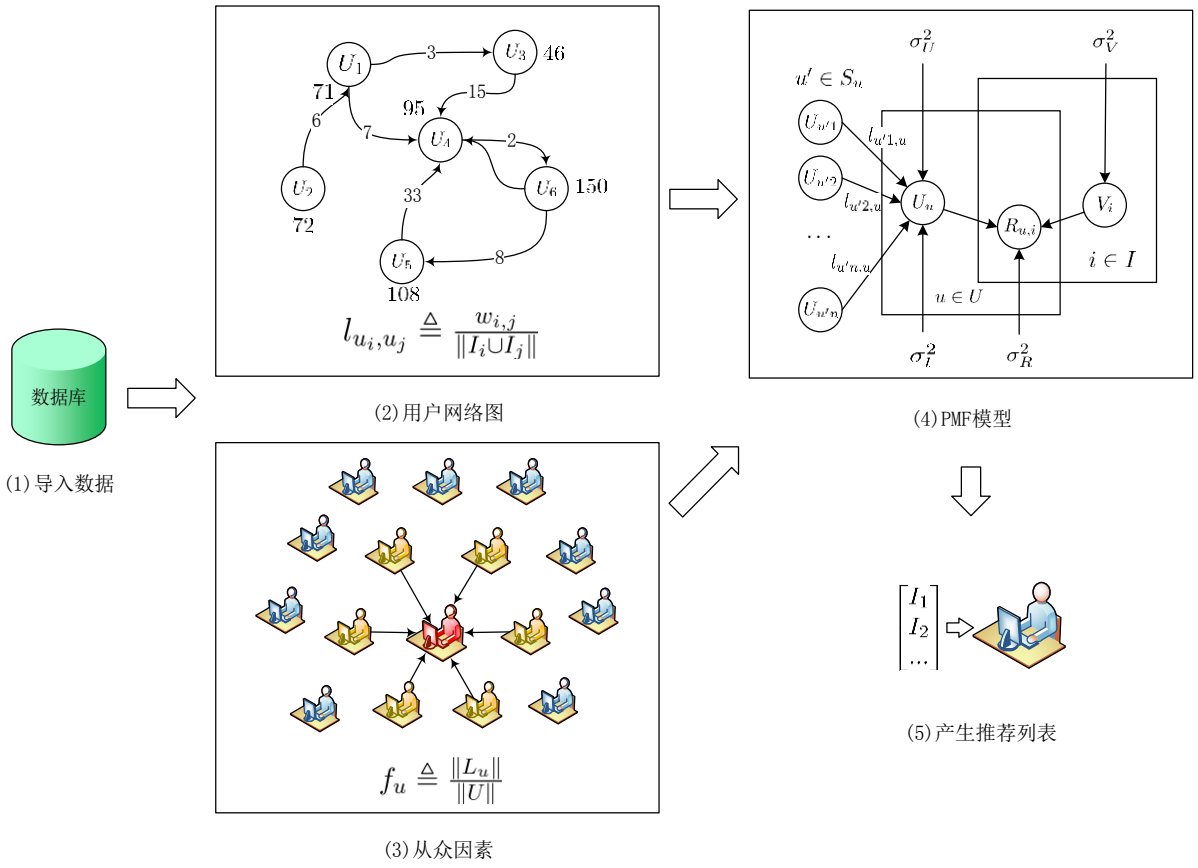


图 3.1 SeqSoPMF 推荐算法框架

Fig. 3.1 SeqSoPMF Recommend algorithm frame

3.3.3 复杂度分析

从 SeqSoPMF 算法框架可以看出，算法复杂度集中体现在第（2）步、第（3）步和第（4）步，即“构建用户社交关系网络图”、“分析用户从众因子”和“构建概率矩阵分解模型求解特征向量”。

在“构建用户社交关系网络图”中，若平均每个商品被α个用户购买，即平均用户数量为α，然后遍历所有边，判断其权值是否提升。那么，对每一个商品构建的用户社交关系网络图的时间复杂度为$O(\alpha^2)$，因而构建用户社交关系网络图的时间复杂度为$O(n\cdot\alpha^2)$。在找其最大近邻集合的过程需要先挖掘用户影响关系。假设每一个用户平均消费β个商品，计算出每个节点对其他节点的影响力，对影响力排序，排序的时间复杂度为$O(n\cdot\beta^2)$，故在“构建用户社交关系网络图”阶段中，需要时间复杂度为

$O\left(n\cdot(\alpha^2+\beta^2)\right)$.

在“分析用户从众因子”阶段，要迭代求出 Followers 集合和 Leaders 集合，基本操作要执行$n(n-1)/2$次，其复杂度刚好为$O(n^2)$.

在“构建概率矩阵分解模型求解特征向量”阶段中，假设平均每个用户的评分数量为γ，平均每个用户直接影响的邻居数量为δ，估计式(3.14)的$L(R,l,U,V)$的计算复杂度为$O(nk\gamma+nk\delta)$，由于评分矩阵R和影响力矩阵$\boldsymbol{I}$都非常稀疏，γ和δ相对较小，因此求解估计式(3.14)的$L(R,l,U,V)$计算复杂度与对应的社交关系网络中的用户数量线性相关。获得式(3.15)与式(3.16)梯度的计算复杂度为$O(nk\gamma+nk\delta^2)$.

综合三个阶段，网络的构建只需遍历评分数据，而影响力矩阵和评分矩阵相对稀疏，γ和δ相对较小，在“分析用户从众因子”阶段，复杂度为 2 阶，在电力数据以及中等数据水平上基本可以容忍，因此 SeqSoPMF 算法框架的时间复杂度不高。

3.4 实验结果与分析

本节将在三个真实数据集上对本算法的运行时间和不同参数下的均方误差做出实验分析,验证本文提出的算法具有较高的准确度和较好的运行效率。

3.4.1 数据集描述与评价指标

采用 Movielens-1m 数据集、Movielens-10m 数据集和 2015 年蒙东地区大用户与发电企业交易的数据。

三个数据集属性见表 3.2。

表 3.2 实验数据集属性
Table 3.2 Experiment Dataset features

Dataset Name	Numbers of Users	Numbers of Items	Ratings
Movielens-1m	6040	3900	1000209
Movielens-10m	71567	10681	10000054
蒙东地区交易数据集	2530	272	68840

Movielens-1m 数据集文件包含 6040 个 MovieLens 用户对 3900 部电影做出的 1,000,209 个评分数据、用户信息数据和电影信息数据。

Movielens-10m 数据集文件包含 71567 个 MovieLens 用户对 10681 部电影做出的 10000054 个评分数据、用户信息数据和电影信息数据。

蒙东地区交易数据集中描述了 2530 个大用户与 272 个发电企业产生的交易。根据交易行为数据转化为对应的评分数据。（转化方法见 0。）抽取 1/3 作为测试集，2/3 作为

训练集。部分大用户与发电企业的直接交易数据见表 3.3。

表 3.3 蒙东地区电力交易部分数据集[1]
Table 3.3 Partial dataset of electric trade of MengDong

购方名称	售方名称	成交电量（MWh）	购方电价（元/MWh）	直接交易电价（元/MWh）	发电类型
A 集团	机组 1#2#	72200	355.6	206.5	火电
B 集团	机组 1#2#	20000	360.55	211.45	火电
C 化工	发电厂 1	20000	344.1	195	火电
D 集团	发电厂 2	14700	334.66	211.56	火电
E 集团	发电厂 3	10000	289	165.9	火电
F 集团	热电公司	77800	288.6	165.5	火电
G 企业	机组 1#2#	7500	324.05	174.95	火电
E 科技	兴安热电公司	77800	288.6	165.5	火电
F 焦化	鄂温克 1#2#	7500	324.05	174.95	火电

本实验主要采用均方根误差（Root Mean Square Error, RMSE）来评价实验指标。

$$RMSE = \frac{\sqrt{\sum_{i,j \in T} (r_{ij} - \hat{r}_{ij})^2}}{|T|} \tag{3.17}$$

其中，T是测试集，r_{ij}是真实的评分，$\hat{r}_{ij}$是估计的评分。

实验运行环境见表 3.4。

表 3.4 实验运行环境
Table 3.4 Experimental running environment

项目	参数
CPU	Inter® Core™ i5-4200U CPU @ 1.60GHz 2.30GHz
内存	2.44GB
操作系统	Windows7 旗舰版 32 位
编译软件	Eclipse neon3 PyDev plugin
编程语言	Python2.7

[1] 蒙东地区电力交易数据集中包含涉密信息，对其中部分数据做保密处理。

3.4.2 实验评分数据获取与标准化

对于标准数据集中的数据，已经有了合适的用户信息、商品信息、评分信息和时间戳信息，因而不需要再对其进行预处理。而针对电力市场上的交易数据来说，大用户与电厂间的交易并无评分信息和评分时间信息，因而本文提出一种简单有效的评分转化方案——由交易数据转化为评分数据和评分时间信息。

电力改革后，大用户可以对发电企业自主进行点对点交易，其选择要综合多方面因素，若大用户对发电企业给出的电能给出的满意度与相应的交易次数成正比，即大用户对发电企业越满意，他们的交易次数就越多，反之越少。因而这里讲大用户对发电企业的评分规定为：大用户与发电企业的交易次数设定为初始评分，可用发生交易的指示向量的 1 范数来度量，如式(3.18)。由于评分的过程需要在多次交易发生后完成，因而规定评分的时间信息由多次交易中最后一次交易的时间戳决定，可用发生交易的时间向量的无穷范数来度量，如式(3.19)。

$$r_{ij} = \left\| \boldsymbol{\mathcal{I}}^{T_r} \right\|_1 = \sum_{(i,j,\tau)\in T_r} \mathcal{I}_{i,j}^{T_r} \tag{3.18}$$

$$t_{ij} = \left\| \boldsymbol{T}_{Sr} \right\|_\infty = \max_{(i,j,\tau_{ij})\in S_r} \tau_{ij} \tag{3.19}$$

其中，r_{ij}代表大用户 i 对电厂 j 的评分；T_r代表交易数据集，交易数据集的每个元素为一个元组(i,j,τ)；τ代表交易时间戳；$\mathcal{I}_{i,j}^{T_r}$为指示函数，当大用户 i 与电厂 j 发生交易则其值为 1，否则为 0；$\|\cdot\|_p$代表p范数，$\boldsymbol{\mathcal{I}}^{T_r}$表示由$\mathcal{I}_{i,j}^{T_r}$构成的集合；$t_{ij}$代表大用户 i 对电厂 j 产生评分的时间戳；S_r表示在T_r中，有交易信息的数据集合，即$S_r = \{(i,j,\tau)\big|\mathcal{I}_{i,j}^{T_r} = 1\}$；$\boldsymbol{T}_{Sr}$表示在$S_r$中由每个元组中的$\tau$项构成的向量，$\tau_{ij}$表示大用户$i$与电厂$j$发生交易的时间戳。

由于计算后的评分参差不齐，没有统一的尺度，因而要对其标准化，本文对其标准化到 0 至 100 之间，如式(3.20)。

$$r_{ij}^* = \frac{r_{ij}}{\|S_r\|} \times 100 \tag{3.20}$$

3.4.3 参数设定与对比算法

本文选取了 3 种方法作为对比算法：

PMF 算法：文献[18]提出的一种概率矩阵分解算法，没有考虑到用户间的影响关系以及商品间的影响关系。

SequentialMF 算法：文献[11]提出的一种基于时序行为的协同过滤算法，其考虑了用

户和商品的双重影响关系，但是没有引入用户的从众因素。

SocialMF 算法：文献[44]提出的一种基于社交网络关系的协同过滤算法，其只考虑了用户的影响关系而没有引入用户的从众因素。

在 SeqSoPMF 模型中。为了降低模型计算复杂度。本文设定$\sigma_R^2/\sigma_U^2 = \sigma_R^2/\sigma_V^2 = 0.001$。另外，本模型设定选取 Top-20 最相近邻的用户作为邻居。特征向量U和V初始化为$\mu = 0$的高斯分布抽样值，然后U和V迭代更新，直至收敛。在 SequentialMF 算法中，本文设定 Top-20 最近邻用户作为邻居，以做对比。

表 3.5 各数据集上部分参数设定
Table 3.5 Partial parameters Setting of Dataset

模型	ml-1m		ml-10m		蒙东地区交易数据集	
	σ_R^2/σ_U^2	σ_R^2/σ_V^2	σ_R^2/σ_U^2	σ_R^2/σ_V^2	σ_R^2/σ_U^2	σ_R^2/σ_V^2
PMF	0.01	10000	10	100	0.1	0.1
SocailMF	0.001	0.001	0.001	0.001	0.001	0.001
SequentialMF	0.001	0.001	0.001	0.001	0.001	0.001
SeqSoPMF	0.001	0.001	0.001	0.001	0.001	0.001

3.4.4 实验结果

本实验有三组：参数σ_R^2/σ_l^2对算法的影响比对试验、特征向量维度k对算法的影响试验和算法运行时间试验。

（1） 参数σ_R^2/σ_l^2对算法的影响比对试验

在 SeqSoPMF 算法中，参数σ_R^2/σ_l^2用来描述用户受到用户影响关系所影响的程度，其值反映了用户的影响关系对推荐效果的作用。在实验中，设定σ_R^2/σ_l^2值分别为 0.01、0.1、0.5、1、5、10、20；设用户特征向量和商品特征向量维度$k = 5$。如图 3.2，说明了参数σ_R^2/σ_l^2对算法的影响有至关重要的作用，当σ_R^2/σ_l^2值增加时，算法的精确度不断加大，说明引入用户影响关系对算法精确度的影响较大。而由实验结果可知，当σ_R^2/σ_l^2过大会导致过拟合而造成精度下降。三种不同数据集存在着不同的精确度则是由不同的数据集的数据量所致。

（2） 特征向量维度 k 对算法的影响试验

本实验对比不同算法在不同特征向量维度 k 取值下造成的 RMSE 值。在实验中，分别设定$k = 5, 10, 20$，如图 3.3 为在 ml-1m 数据集上的对比结果，图 3.4 为在 ml-10m 数据集上的对比结果。

由实验结果可知：

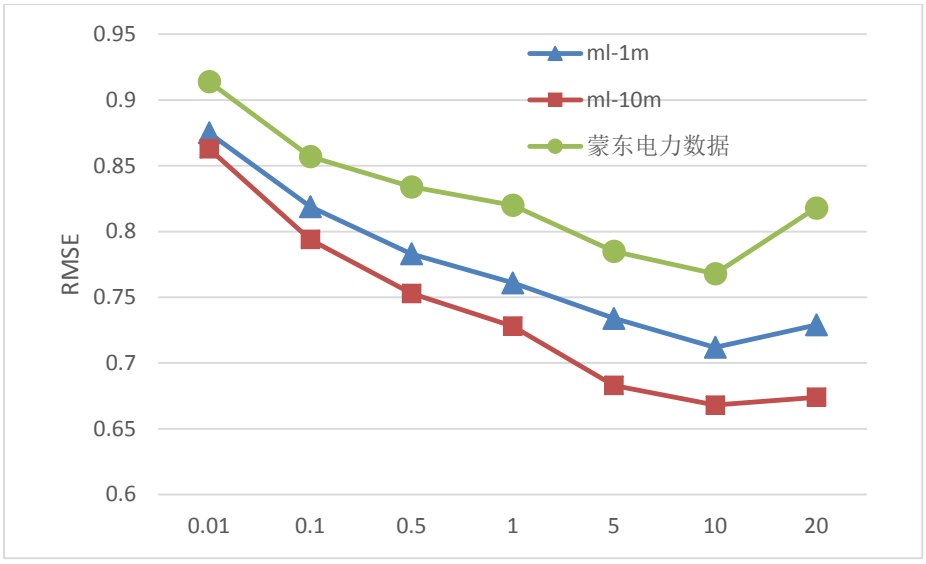


图 3.2 参数σ_R^2/σ_I^2对算法的影响比对试验
Fig. 3.2 The Compartment Influence of Parameter σ_R^2/σ_I^2

1) 随着k值的不断加大各个算法精确度都有所提升；

2) 添加了时序社交关系后精度有明显提高，PMF 的精确度要明显低于其他三个算法的精确度，可见时序社交关系的引入能提升算法的精确度；

3) 由于 SocialMF 考虑了用户间影响关系而没有引入商品间影响关系和从众因素，因而导致其精度略低于 SequentialMF 和 SeqSoPMF 算法，SeqSoPMF 算法考虑了用户间

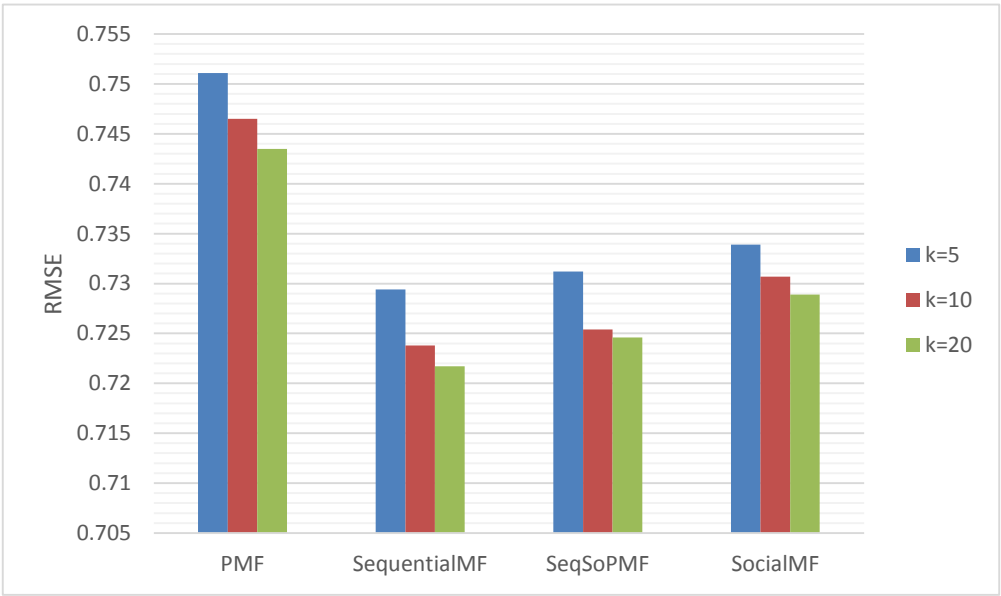


图 3.3 在 ml-1m 数据集上不同算法的不同维度 *k* 取值造成的 RMSE 比较试验
Fig. 3.3 The RMSE Comparision of different dimension of
k-value different algorithms in Training on ml-1m dataset

影响关系以及用户的从众关系，但是忽略了商品间的影响关系因而其精度略低于 SequentialMF 算法，而其相差水平并不太大，而 SequentialMF 引入的商品间的影响关系

很大程度上增加了计算复杂度；

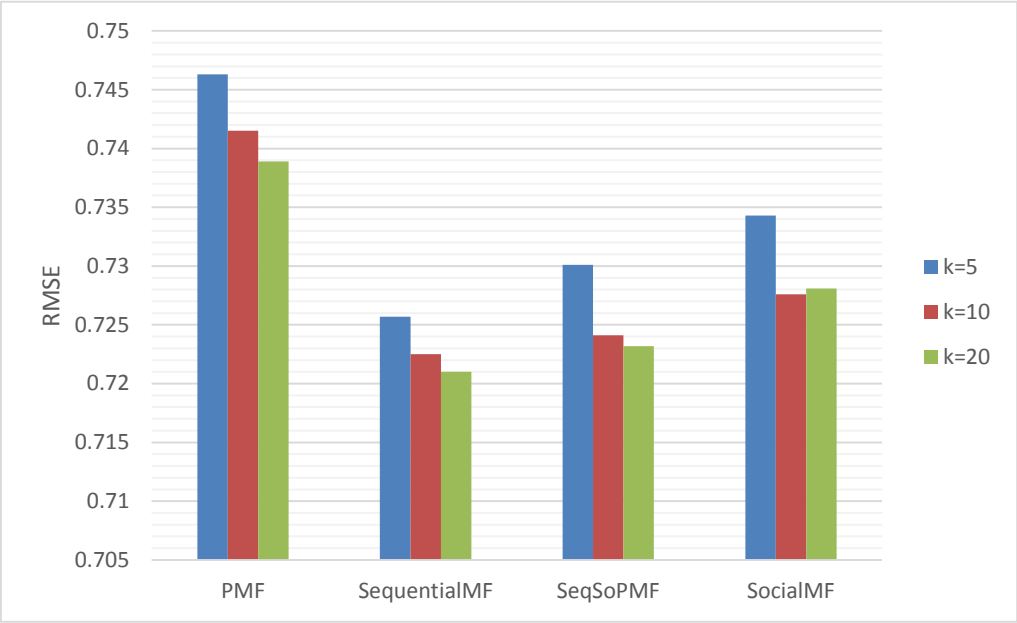


图 3.4 在 ml-10m 数据集上不同算法的不同维度 k 取值造成的 RMSE 比较试验
Fig. 3.4 The RMSE Comparision of different dimension of
k-value different algorithms in Training on ml-10m dataset

4) 对比两组数据集，由于评分数据量的不同导致精度有所差异。

（3） 算法运行时间试验

在 3.3.3 节，分析了 SeqSoPMF 算法的复杂度，在该实验中，通过比较不同算法在迭代一轮所耗费的运行时间来分析本算法的运行时间。设定特征向量维数为$k = 5$。

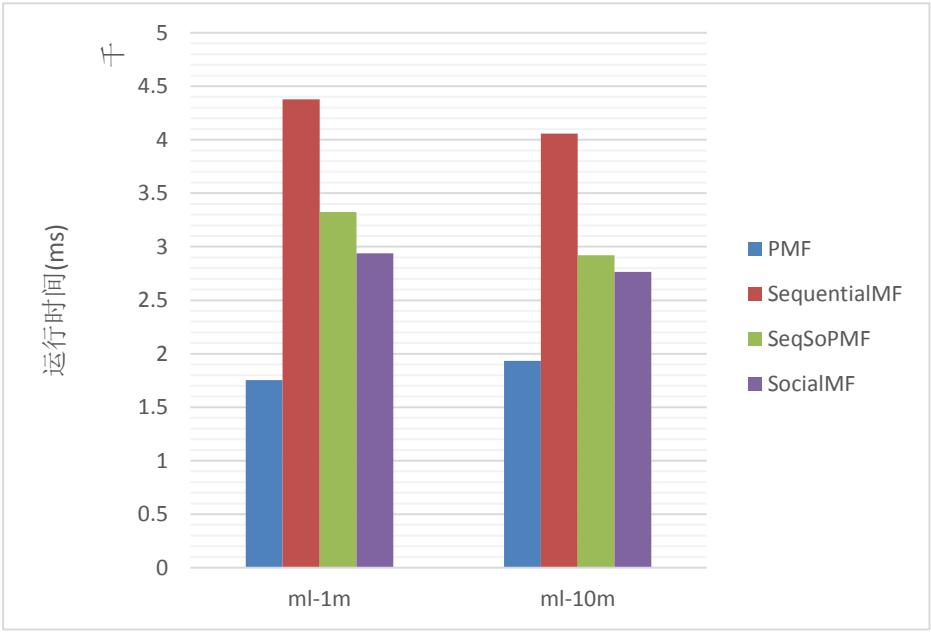


图 3.5 不同算法的运行时间比对试验
Fig. 3.5 The Runtime Comparision of different algorithms in Training

实验结果如图 3.5，从实验结果可以看出，四个算法的运行时间关系满足：PMF < SocialMF < SeqSoPMF < SequentialMF.由文献[11]，算法 SequentialMF 计算了用户的影响关系之外，还计算了商品的影响关系，但是它没有引入用户间的从众关系，因而，其处理时间长于 SeqSoPMF 算法；而 SocialMF 没有考虑商品的影响关系和用户从众关系，因而，其运行速度较快。另外，PMF 算法没有考虑用户或商品之间的影响关系，其计算量小于 SocialMF 算法，因而造成计算时间最少。

3.5 本章小结

本章提出了基于时序社交关系的协同过滤算法。首先分析了电力交易推荐系统拟解决问题，将该推荐系统要解决的问题引入并形式化描述；然后利用用户的交易时序信息挖掘和定量分析用户间的影响关系和从众关系，并寻找最近邻的集合，然后融合概率矩阵分解算法中，分析用户和商品的特征因子向量，再使用随机梯度下降法求解。提高了评分预测精确度，同时通过分析计算复杂度给出了效率相对较高的推荐框架。最后在三个真实数据集上验证了本章提出的算法的准确性和效率等与传统的推荐算法相比有一定程度的提高。该方法可以推广应用到另外一些矩阵分解算法中去，该方法的使用也是为矩阵分解的推荐算法提供新的思路。本章提出的推荐算法不仅可以应用在电力交易领域也可以应用于其他推荐系统当中。在后续的章节里，将提出一种基于核密度估计的方法预测用户偏好，而给出推荐意见的推荐算法。

第4章 基于用户偏好估计的协同过滤算法

本章针对电力领域的数据稀疏性问题，提出了基于用户偏好估计的协同过滤算法。首先给出基于用户偏好的推荐模型的建立过程，然后提出基于商品标签的相似度，基于用户偏好估计的协同过滤算法的过程总体包含三个阶段：用户偏好密度估计、用户相似性计算和评分数据预测及填充。最后通过在两个真实数据集上对算法的精确性和数据稀疏性问题上做出了评价，证明该算法在数据稀疏的情况下能给出精确度较高的推荐结果，这也是本章提出的算法与上一章提出的算法的不同之处。该算法可以应用在推荐系统数据非常稀疏的情况，一般是系统中用户较少或商品较少的情况。

4.1 基于用户偏好推荐模型的建立

电力改革之前，大用户与发电企业之间没有交易行为数据，他们之间不会达成交流，大用户的购电需求以及满意度评价只是针对电网公司，由于此种原因导致大用户对发电企业的评价数据几乎为零。这对于推荐系统来说，无疑是灾难性的。本节将针对电力交易过程中稀疏数据问题给出基于矩阵填充的处理方法，首先给出经典的用户偏好相似性度量方法，然后利用经典的评分预测方法给出推荐模型。该方法步骤简单，处理方式单纯，但是存在一些导致评分预测不准确的问题。后续内容里面将针对这些问题给出处理办法。

4.1.1 数据描述与问题定义

设有 *n* 个用户，构成集合 *U*；*m* 个商品，构成商品集合 *I*。用户 *u* 对商品 i 的评分为 $r_{u,i}$，则用户-商品评分矩阵为

$$\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,m} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,m} \end{bmatrix}$$

由于数据稀疏性原因，该矩阵不是所有评分数据都存在，而要解决数据稀疏性问题是利用已经存在的 $r_{u,i}$ 而尽可能地预测和填补缺失的评分数据。本系统要解决的问题正是要对缺失的评分数据做出填充以给出推荐列表。

4.1.2 相似性度量与评分预测

在推荐系统中，相似性的度量是众多推荐算法中应用广泛而且非常关键的步骤。较常用的有余弦相似度和皮尔逊相关系数。相似度度量是进行评分预测的关键步骤。

（1） 余弦相似度

设I_u为用户 *u* 评分过的商品集合，即$I_u = \{i|i \in I, \mathcal{I}_{u,i}^R = 1\}$，$\mathcal{I}_{u,i}^R$为指示函数，当用户 *u* 对商品$i$给出打分则其值为 1，否则为 0；余弦相似度[45]是将用户的评分作为 *m* 维商品空间上的向量，利用两个用户评分向量夹角的余弦值来衡量两个用户的相似性，如式(4.1)所示。

$$corr_{u,v} = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} r_{v,i})}{\sqrt{\sum_{i \in I_u} (r_{u,i})^2} \sqrt{\sum_{i \in I_v} (r_{v,i})^2}} \tag{4.1}$$

由于使用余弦相似度来衡量两个用户的相似性没有考虑用户评分尺度的差异，因此，通过将所有评分与用户的商品的平均评分作差来消除这种差异，如式(4.2)所示。

$$corr_{u,v} = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_v} (r_{v,i} - \bar{r}_v)^2}} \tag{4.2}$$

（2） 皮尔逊相关系数

皮尔逊相关系数[46]是余弦相似度在维度值缺失情况下的一种改进,由皮尔逊相关系数计算用户相似性的方法如式(4.3)

$$corr_{u,v} = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \bar{r}_v)^2}} \tag{4.3}$$

其中，$\bar{r}_u$为用户u给出评分的平均值。

正如前面两种相关性的描述，这些相关性度量方法都只考虑那些有共同的商品，对于那些缺失的数据中，缺少评分的商品会隐藏着用户潜在的偏好，只是没有做出评分而已。因此，若数据的稀疏性问题足够大则基于此计算出的相似性会有很大偏差。针对以上两种相关性的缺点,本文提出一种基于商品标签的相似度度量方法(详见第 4.2 小节)。

利用上述相似度度量方法，可以对评分进行预测。下面是目前广泛采用的评分预测规则。

利用用户间相似度能在数据集中搜索到目标用户的邻居用户集合，然后对该邻居集合中所用用户对目标商品的评分的加权平均值，权重为用户间相似度。那么目标用户对

目标商品的评分的预测值就可以定义为该加权平均值。又由于用户间的评价尺度的差异，所以在评分的预测值方法中引入偏置量$\bar{\mu}$，如式

$$\hat{r}_{u,i} = \bar{\mu} + \frac{\sum_{v \in N_u} corr_{u,v} \cdot (r_{v,i} - \bar{\mu})}{\sum_{v \in N_u} |corr_{u,v}|} \tag{4.4}$$

其中，$\hat{r}_{u,i}$为用户 *u* 对商品 *i* 的评分预测值，$\bar{\mu}$为用户 *u* 的平均评分，N_u为用户 *u* 的邻居集合。此步骤中的评分预测将是该推荐框架的最后一步。

4.1.3 电力交易稀疏数据矩阵填充

矩阵中的数据表示用户对事物的偏好程度，往往这些表示偏好的数据会缺失很多，低稚矩阵（Low Rank Matrix）是数据稀疏性问题的集中体现[47]。

如图 4.1，在初始矩阵 *R* 中，用户u_1对商品v_2，v_3的评分数据缺失，然而，用户u_2，u_3等等，对商品v_2，v_3的评分数据是存在的。利用已存在的评分数据对未知的评分数据进行估计，就是矩阵填充。计算用户u_1，u_2等用户的相似度，然后根据用户间的相似度和其对商品的评分最终对缺失的用户对商品的评分做出估计。其实，著名的协同过滤算法也是矩阵填充的一种特殊形式。

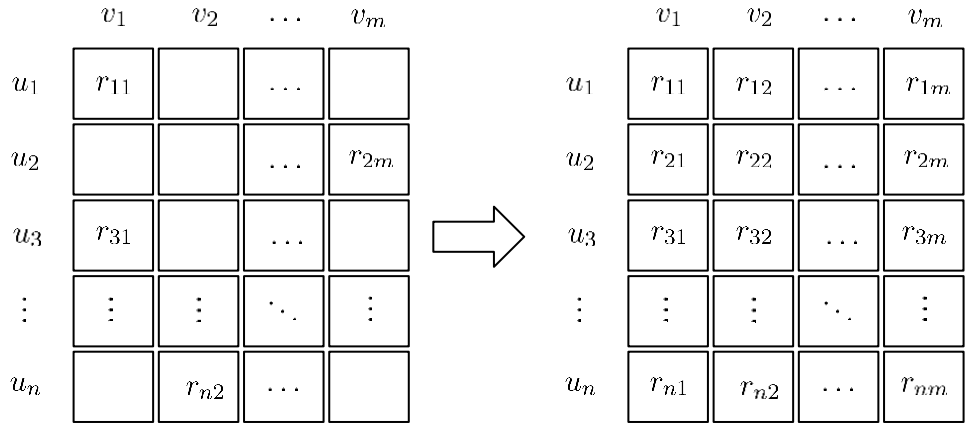


图 4.1 矩阵填充示意图

Fig. 4.1 Diagrammatic drawing of matrix completion

矩阵填充是解决推荐系统中数据稀疏与冷启动问题的重要方法之一。若对冷启动的用户或商品利用预设的初始评分来进行矩阵填充，则为默认值填充；另一种办法是使用已知的评分数据的均值来对缺失数据进行填充，这是均值填充。由于已存在的基础方法都是使用“粗暴的”估值方法来对未知数据进行填充，带来的误差会很大，因而其准确性不言而喻。为了减少这种误差，本文将充分挖掘用户之间偏好相似度，使用非参数密度估计的方法来对用户偏好分布做出估计，再利用估计出的偏好分布情况求出用户间的

偏好相似度，按照这种估计出的偏好相似度来进行有效的评分填充，其准确度会有很大幅度的提升。

4.2 基于商品标签的相似度

在上一小节本文列举了常用的几种计算相似度的方法，如余弦相似度、皮尔逊相关系数等。本节将提出的一种基于商品标签的相似度计算方法。该方法更有针对性的计算两种商品之间基于标签的相似度。并给出了标签提炼的方法。

4.2.1 基于商品标签的相似度定义

在推荐系统中的商品往往会有标签或类别信息，如电影评价网站将电影分成若干类别，动作片、喜剧片等。在电力交易领域，商品的标签或类别往往是发电企业的某些属性，如：机组信息和发电类型（火电、风电等）等。一个商品往往会同时隶属于多个标签之下，如问鼎 2017 年票房冠军的《战狼 2》既属于动作片又属于军事片和战争片。

定义 4.1：基于商品标签的相似度 设商品的标签集合为 L，$l_k \in L$为其中的某一个标签，商品 i 和 j 的所属标签集合分别为L_i和L_j，$L_i \subseteq L$，$L_j \subseteq L$。则基于商品标签的相似度$sim_{i,j}^{L}$为：

$$sim_{i,j}^{L} \triangleq \frac{|L_i \cap L_j|^3}{|L_i| \cdot |L_j| \cdot |L_i \cup L_j|} \tag{4.5}$$

该相似度的度量综合考虑了三个因素：两个商品所属标签中重合部分的比例，重合部分在第一个商品所属标签集合中的比例，重合部分在第二个商品所属标签集合中的比例。

由于两个商品的标签相似度越小则其在商品空间上的距离就越大，极端地，当两个商品的基于标签相似度为 1 时（即两个商品的标签集合完全一致），则其在商品空间上的距离就为 0；而当两个商品的相似度为 0 时，则其在商品空间上的距离就为 1。因此，定义商品间的距离为，

$$dis_{i,j} = 1 - sim_{i,j}^{L} \tag{4.6}$$

4.2.2 电力标签提炼方法

与传统的商品标签不同，改革后的电力市场电力能源作为一种特殊的商品，其标签的概念不是其他商品的类别，因为电力能源这种商品只是一种能源的存在，所有大用户使用的电力能源都是相同的，电力是一种能源载体。那么所有电力能源就不存在类别标签的不同了，因而无法对其分类。然而，对电力价格和售电量有影响以至于对电力市场有影响的因素也可以看做电能的不同属性。为挖掘出电力能源的不同属性，本文针对电

能的生产商的某些属性做出挖掘，用来代表其生产的商品——电能的类别标签。对蒙东大用户交易数据分析可知，发电机组的属性除基础的名称信息以外，还有：机组类型（风电、火电、水电等），机组子类型（燃煤、抽水蓄能、径流等），供热类型（背压式等），地理区域（赤峰市、白山市等）。这些信息都可作为发电企业的类别标签，其生产的电能也可以看做是属于这些类别标签下的商品。为此，本文提出一种针对电力市场下的商品标签提取方法。

假设 M 为发电机组集合，其中每个元素即某种机组为 k 维向量，上述各个影响因素为这些元素的某一个维度，那么显然可得如下标签收集算法 Labels Collecter：

Algorithm 2 Labels Collecter

Input:

M : the genset information set;

$indexes[\dots]$: the key affect dimensionality indexes.

Output:

L : the Labels set;

$\hat{M}$: the genset information set

which included labels set that the machine group belongs to.

1: $\hat{M} \leftarrow \phi$, $L \leftarrow \phi$ // initialize $\hat{M}$ and L with ϕ.
2: **for** each $m \in M$ **do**
3: $Labels \leftarrow \phi$
4: **for** each $index \in indexes$ **do**
5: $Labels \leftarrow Labels \cup \{e[index]\}$ /*this method adds the labels into set $Labels$.*/
6: L.put(e[index]) /*this method filtrates the different labels.*/
7: **end for**
8: $\hat{M} \leftarrow \hat{M} \cup \{(m, Labels)\}$
9: **end for**
10: **return** L, $\hat{M}$

利用该算法对机组信息集合遍历一次，可以得到全体标签集合和新的机组信息集合 $\hat{M}$,该集合包含了每一个发电机组的基本信息，还包含该机组所属的标签集合。

算法的输入部分为发电机组的信息集合，每个元素为一个发电机组，包含机组的各种属性，如机组名称、机组所属地区、机组类型等等。这些属性的一部分会是影响电能的关键属性，通过对以往的交易数据的离线分析可以获得哪些属性会是影响电能交易的关键性较强的因素，这些因素的索引的列表作为第二部分输入的参数。算法的输出为标签集合和发电机组的基本信息集合，输出的发电机组基本信息集合不仅仅包含了发电机组最初的基本信息，还容纳了一个重要的集合——该发电机组的标签集合，即式(4.5)中的L_i集合。算法第 2 行到第 9 行为对机组信息集合的遍历部分，第 4 行到第 7 行的每次迭代过程会把机组的关键属性值添加的全局标签集中，添加过程自动排除重复元素，还要把当前机组的关键属性值添加到当前机组的标签集合中。

4.3 UserPreferedCF 算法描述

基于用户偏好估计的协同过滤算法（Collaborative filtering algorithm based on user's preference estimation，UserPreferedCF）的过程总体包含三个阶段：用户偏好密度估计、用户相似性计算和评分数据预测及填充。在用户偏好密度估计阶段，首先要对商品在商品空间上的距离做出度量，然后根据核密度估计方法对用户偏好进行估计。在用户相似性计算阶段，要使用 KL 散度来计算用户的偏好分布之间的距离，以度量用户偏好的相似度。最后预测评分数据阶段，利用对用户偏好的测量值对评分进行预测，可以得出推荐结果。

4.3.1 用户偏好密度估计

由于传统的基于用户相似性的推荐算法中，都只利用公共评分过的商品，而忽略了一些未评分的商品。但众所周知，用户尚未作出评分的那些商品也会体现用户的偏好信息。更加拟合实际情况的方案是：估计用户在商品空间上的兴趣概率密度分布情况，然后计算用户在其兴趣的概率密度分布下的相似性。出于数据稀疏性的考虑，评分矩阵中绝大多数的数据都是缺失的，上述方案的考虑是符合数据稀疏性的要求的。估计用户兴趣密度分布时，本文采用核密度估计方法(Kernel Density Estimation)。在非参数估计方法中，核密度估计方法又称作 Parzen 窗法，是一种未知分布的密度估计方法。用户兴趣密度由多方面因素造成，各种因素服从的分布情况均未知，由此易知其密度分布具有多个局部最优值，利用核密度估计方法拟合其分布效果较好。

如 2.2.1 节所述，核密度估计属于统计学中非参数估计的一类方法，是一种使用有限的样本来估计其概率密度函数的方法。若$X_1,X_2,\cdots,X_n$是总体X的独立同分布的若干样本，设$K(x)$为$\mathbb{R}$上的一个给定的概率密度函数，$h_n>0$是一个和n有关的常数，则X的概率密度函数$f_h(x)$的核密度估计$\hat{f}_h(x)$的定义如下[27]：

$$\hat{f}_h(x)=\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h_n}\right) \tag{4.7}$$

其中，$K(x)$称核函数，h_n为带宽(Bandwidth).核函数有多种形式，较为常用的有高斯核函数、三角核函数、直方图等等。文献[26]指出核函数的选取与带宽的选取对估计的结果影响前者远小于后者。众所周知，独立同分布且期望和方差有限的随机变量序列的标准化和的极限为标准正态分布，这就是中央极限定理。用户兴趣分布正是一种由多种因素构成的独立同分布且数学期望和方差有限的随机变量序列，因此先采取高斯核函数，即

$$K(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}u^2} \tag{4.8}$$

据此，采用高斯核函数估计用户u的兴趣密度，即综合式(4.7)和式(4.8)有：

$$\hat{f}_h^u(j) = \left(|I_u|\sqrt{2\pi}h\right)^{-1} \sum_{i \in I_u} r_{u,i} e^{\frac{dis_{i,j}^2}{2h^2}} \tag{4.9}$$

由 Silverman 经验法则[25]，当使用高斯核函数时，带宽h的最佳选择为：

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \tag{4.10}$$

除上述高斯核函数外，本文给出基于三角核函数的用户兴趣分布如式(4.11)，此处定义带宽与高斯核函数的带宽相同。

$$K_t(u) = \begin{cases} 0, & |u| > \sqrt{2}h \\ \dfrac{\sqrt{2}h - |u|}{2h}, & otherwise \end{cases} \tag{4.11}$$

对应的偏好密度分布为

$$\hat{f}_h^u(j) = \sum_{i \in I_u} \frac{r_{u,i} \cdot \left(\sqrt{2}h - dis_{i,j}\right) \cdot c\left(\sqrt{2}h - dis_{i,j}\right)}{|I_u| \cdot 2h^2} \tag{4.12}$$

其中，$c(x)$为阶跃函数：

$$c(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{4.13}$$

4.3.2 用户相似性计算

上述步骤通过核密度估计方法对用户兴趣分布作出估计，利用相对熵来对两个已知的分布进行差异性比较，即使用信息相对熵计算用户偏好分布的相似度。

本节利用信息散度计算用户相似度，设P_u为核密度估计方法获取的用户u的兴趣密度函数，则P_u对P_v的 KL 散度定义为：

$$D_{KL}(P_u \| P_v) = \sum_{i=1}^{K} P_u(i) \ln \frac{P_u(i)}{P_v(i)} \tag{4.14}$$

由于 KL 散度具有非对称性，度量的标准需要满足对称性，这显然是矛盾的，为了解决这一问题，本文采用式(4.15)计算用户间相似度：

$$corr_{u,v} = \frac{1}{2}\left(D_{KL}(P_u \| P_v) + D_{KL}(P_v \| P_u)\right) \tag{4.15}$$

由 KL 散度定义式(4.14)可以分析出，如果想让$D_{KL}(P_u \| P_v)$值比较小，那么P_u大的地方P_v也一定要大，否则$P_u(i)/P_v(i)$值也会很大；然而，P_u小的地方$D_{KL}(P_u \| P_v)$值对P_v的影响不那么敏感。

下面讨论利用 F-divergence 来衡量用户相似度，利用 F-Divergence 的 Total Variation

Distance 形式，P_u对P_v的 F-Divergence 定义为：

$$corr_{u,v} \triangleq \sup |P_u(i) - P_v(i)| \tag{4.16}$$

其中 sup 为上确界函数。这里仅讨论使用信息相对熵来描述两个兴趣密度的差异性，类似的，还有 Wasserstein Distance 等。

4.3.3 算法框架

UserPreferedCF 算法遵循着一条较为传统的推荐方法：相似度计算，用户偏好估计以及生成推荐列表。在这个过程中，该算法又引入的新的方法来估计用户偏好分布，评价用户相似度。

根据上面对标签相似性和用户偏好密度估计的讨论，可以给出算法框架，如图 4.2 和算法 3 所示。框架可以概括成如下几个步骤：

（1） 利用式 (4.6)计算商品与其他商品的在商品空间上的距离；

（2） 由式 (4.9) 或者式 (4.12) 计算用户的偏好在商品空间上的分布情况；

（3） 重复步骤（1）和步骤（2），计算所有用户的偏好分布情况；

（4） 由式 (4.15) 计算两个用户间的相似度；

（5） 利用式 (4.4) 作出最终预测.

Algorithm 3 UserPreferedCF

Input:

R : the ratings matrix.

Output:

$\hat{R}$: the predicting ratings matrix.

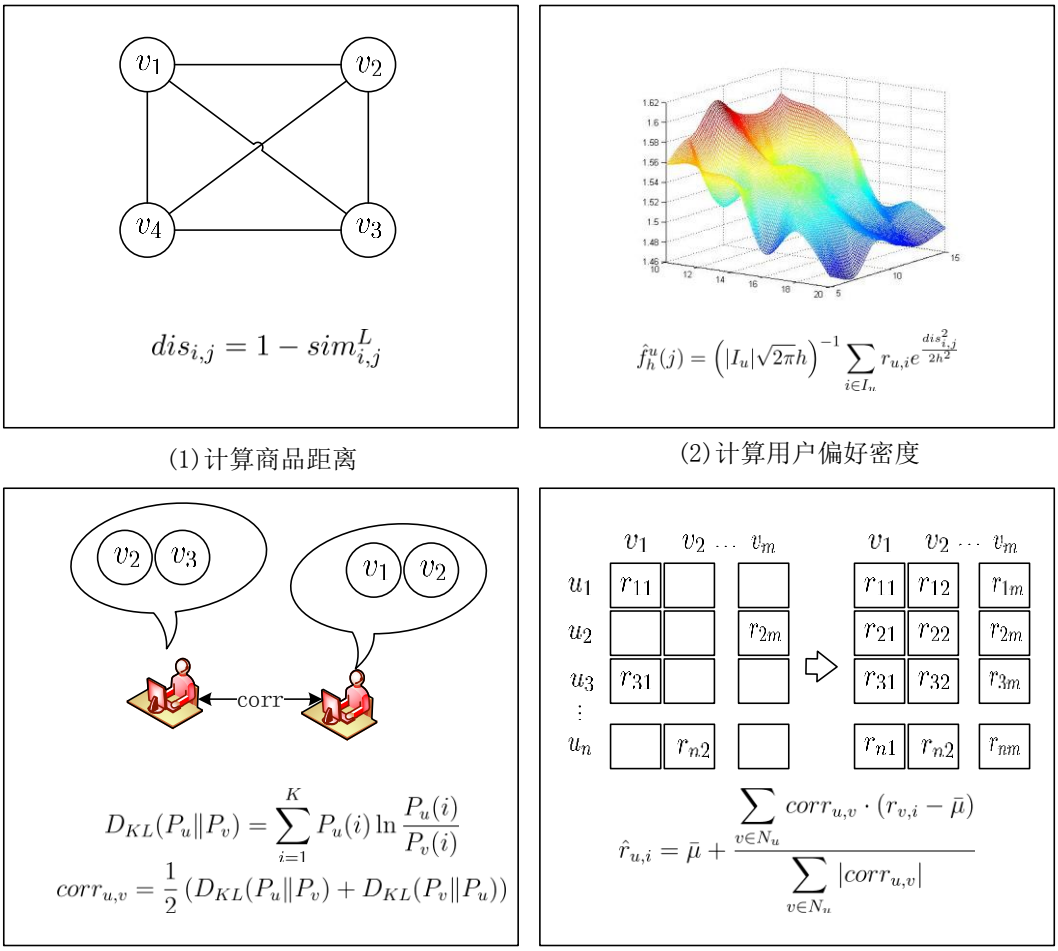
1: $F \leftarrow \phi$, $L \leftarrow \phi$ // initialize F and L with ϕ.
2: **for** each $u \in U$ **do**
3: $D \leftarrow \phi$ /* D is a matrix consist of $dis_{i,j}$, calculate by eq(4.6). Now initialized by ϕ.*/
4: D calculate by eq(4.6)
5: calculate user u's interests distribution on items space by eq(4.9) or eq(4.12).
6: **end for**
7: **for** each $u, u' \in U$ **do**
8: calculate similarity of user u and u' by eq(4.15).
9: $\hat{R} \leftarrow$ generate predicting ratings by eq(4.4)
10: **end for**
11: **return** $\hat{R}$

跟随着传统的协同过滤推荐步骤，本文对其做了如下几点创新：

（1） 从矩阵填充的角度来看，传统的方法只是采用基础填充技术，如均值填充和默认值填充；由于这种方式的误差较大，本文采取一种挖掘用户偏好的方式来对缺失数据进行填充；

（2） 从协同过滤的角度来看，传统方法对待用户间的相似性往往给出的估计值较为粗略，本文结合统计学中的非参数估计方法——核密度估计方法，来综合分析用户的偏好分布，通过分析两个用户的 KL 散度的方式来计算他们的相似度。

（3） 在计算用户相似度的过程中，本文提出了一种基于商品标签的相似度，可以从商品标签的角度描述商品的相似度，也是分析商品在商品空间上的距离的一个方向。



(1)计算商品距离　(2)计算用户偏好密度

(3)用户相似度计算　(4)评分预测与矩阵填充

图 4.2 UserPreferedCF 算法框架

Fig. 4.2 The frame of UserPreferedCF algorithm

4.3.4 复杂度分析

在上一节中给出了 UserPreferedCF 算法的整体框架，本节针对两种数据对该算法的复杂度进行简要分析。

该算法分为两个部分,第一部分是对给定的数据进行用户偏好估计.设每个用户做出评分的平均商品数量为α,用户数量为n,利用式(4.6)计算商品在商品空间上的距离,再利用式(4.9)或式(4.12)计算用户偏好的分布的相似度.该过程的复杂度为$O(n \cdot \alpha)$。

第二部分是利用用户间相似度对指定用户对商品评分的估计.该过程只需要对用户

进行一次遍历求出与指定用户的相似度,由此可以生成该用户的评分向量。该过程的复杂度显然为$O(n)$。

针对电力市场电力标签数据的提炼过程，一般发电机组即电力商品数量，设为m。影响因素平均数量为β，则该过程的复杂度为$O(m \cdot \beta)$。

总体上，该算法的复杂度为$O(n \cdot \alpha) + O(n) = O(n \cdot \alpha)$。可见影响整体复杂度的关键过程是第一个迭代过程。然而事实上，在第一阶段，由于数据稀疏的原因，往往用户平均评分数据量不会很大，因而α的值不会很大。针对标准的推荐系统应用数据，不用将标签集获取作为步骤，因此，标签集获取的复杂度不作考虑。而针对电力交易数据，电力评分数据获取的方法参见 3.4.2 节，而电力标签数据集的提炼是有一定复杂度的，因而需要将上述分析考虑进去，即总体时间复杂度为$O(n \cdot \alpha) + O(n) + O(x \cdot \beta) = O(n \cdot \alpha + m \cdot \beta)$。

4.4 实验结果及分析

本节针对前几节提出的 UserPreferedCF 算法的性能做出实验分析，在两个真实数据集上，通过交叉验证的方式分别分析了不同距离度量方法对推荐结果的影响、不同数据稀疏程度下对推荐结果的影响和不同核函数在不同的带宽下对推荐结果的影响。

4.4.1 数据集描述

本章使用 GroupLens 提供的 ml-100k 数据集[2]和 2015 年蒙东大用户交易数据集作为数据集，对 UserPreredCF 算法进行实验。在 ml-100k 数据集中包含 943 个用户对 1682 部电影的 100000 个评分数据以及包含电影类别标签的电影信息数据，评分范围在 1~5 分。蒙东大用户交易数据与前一章所使用的数据集相同，通过 3.4.2 节提出的方法生成评分数据集。不仅如此，这里还要用到发电机组信息数据，该数据集存在于蒙东大用户交易数据中。部分维度的部分发电机组信息数据集见表 4.2。

为了充分验证本文提出的算法在数据稀疏情况下的性能。在 ml-100k 数据集上，对每个用户随机抽出 10 个评分记录作为测试集，其余作为训练集。在生成的蒙东大用户交易评分数据中，将其 1/3 作为测试集，2/3 作为训练集。然后在这两个数据集的训练集中随机筛选较稀疏的训练子集，其评分数量大约为原始数据集的 90%以下。为了进行交叉验证,将处理后的数据集划分成 5 组,每组按训练集和测试集的比例为 4:1 进行划分。

实验运行环境见表 4.1。

[2] 数据集下载地址：https://grouplens.org/datasets/movielens/100k/

表 4.1 实验运行环境
Table 4.1 Experimental running environment

项目	参数
CPU	Inter® Core™ i5-4200U CPU @ 1.60GHz 2.30GHz
内存	2.44GB
操作系统	Windows7 旗舰版 32 位
编译软件	Eclipse neon3 PyDev plugin
编程语言	Python2.7

表 4.2 部分维度的部分机组信息的数据集[3]
Table 4.2 Partially dimensioned partially Genset information dataset

机组名称	所属市场成员	机组类型	机组子类型	地理区域	机组额定容量	首次并网日期
AA 华电#1	AA 售电有限公司	火电	燃煤	锡林郭勒盟	600	2010-08-08
AA 华电#2	AA 售电有限公司	火电	燃煤	锡林郭勒盟	600	2010-09-01
BB 热 B#1	BB 售电有限公司	火电	燃煤	赤峰市	135	2006-12-24
BB 热 B#2	BB 售电有限公司	火电	燃煤	赤峰市	135	2007-08-19
BB#4	BB 售电公司	火电	燃煤	赤峰市	12	1988-07-25
BB#5	BB 售电公司	火电	燃煤	赤峰市	12	1988-11-24
BB#6	BB 售电公司	火电	燃煤	赤峰市	25	1994-08-25
BB#7	BB 售电公司	火电	燃煤	赤峰市	24	1999-11-25
KKW 机组	KKW 抽水蓄能电站	水电	抽水蓄能	白山市	150	2013-01-01
KKB 机组	KKW 抽水蓄能电站	水电	抽水蓄能	白山市	150	2013-01-01
FF#1	FF 煤电销售有限责任公司	火电	燃煤	呼伦贝尔市	500	1998-10-01
FF#2	FF 煤电销售有限责任公司	火电	燃煤	呼伦贝尔市	500	1999-08-01

4.4.2 评估指标

本章使用平均绝对误差(Mean Absolute Error, MAE)来评估实验的误差结果。MAE 值越大表示生成的推荐列表质量越差。如式(4.17)。

$$MAE = \frac{\sum_{i,j \in T} |r_{ij} - \hat{r}_{ij}|}{|T|} \tag{4.17}$$

其中 T 是测试集，r_{ij}是真实的评分，$\hat{r}_{ij}$是估计的评分。

4.4.3 实验结果

本文设计三组实验，分别是，商品空间上不同的距离函数对推荐结果的影响，不同

[3] 蒙东大用户交易数据包含涉密信息，部分字段不予显示，部分记录保密处理。

数据稀疏程度下对推荐结果的影响和不同核函数在不同的带宽下对推荐结果的影响。

（1） 商品空间上不同的距离函数对推荐结果的影响

对用户偏好进行估计时，商品空间上商品间的距离是一个重要影响因子。使用式(4.6)完成基于商品标签的距离计算。并且与皮尔逊相似度度量方法做出对比。实验结果如图 4.3。

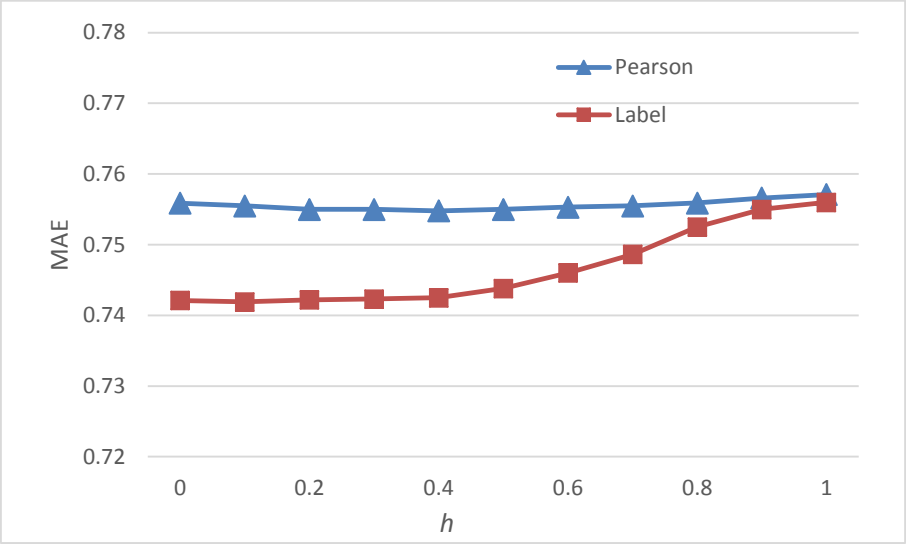


图 4.3 商品空间上不同的距离函数对推荐结果的影响
Fig. 4.3 The Influence of Different Distance Functions on the Product Space to the Recommended Results

通过实验结果可以分析得出：基于商品标签的相似度的计算方法在推荐质量上要优于皮尔逊相似度得出的推荐结果。因为皮尔逊相关系数的计算中，没有考虑商品间的相关性，往往商品空间中的商品特征向量并非正交关系。

（2） 不同数据稀疏程度下对推荐结果的影响

本实验考察在一定的稀疏情况下，UserPreferedCF 算法的性能。使用 2015 年蒙东大用户交易数据与 ml-100k 数据分别对测试 UserPreferedCF 算法的面对稀疏数据的处理能力。图 4.4 为 ml-100k 数据集实验结果，图 4.5 为蒙东大用户交易数据集实验结果。通过与基于用户余弦相似度的协同过滤算法的对比，可以看出，传统的算法基于用户的余弦相似度的矩阵填充方法(Cosine 曲线)，和基于用户评分的均值做填充的算法(user ratings 曲线)总体上的推荐质量都比 UserPreferedCF 算法要差，基于 ml-100k 数据集上，在数据稀疏率达到 0.2 之前，反而传统的填充效果都要比 UserPreferedCF 算法要好些，但是波动性较差，当稀疏率达到 0.2 之后，算法的推荐质量有明显提高。而针对蒙东大用户交易数据集，这种情况却不明显。因此，由于数据集自身的特性，会导致传统矩阵填充的推荐结果不稳定，而 UserPreferedCF 算法的推荐结果仍然较为稳定。

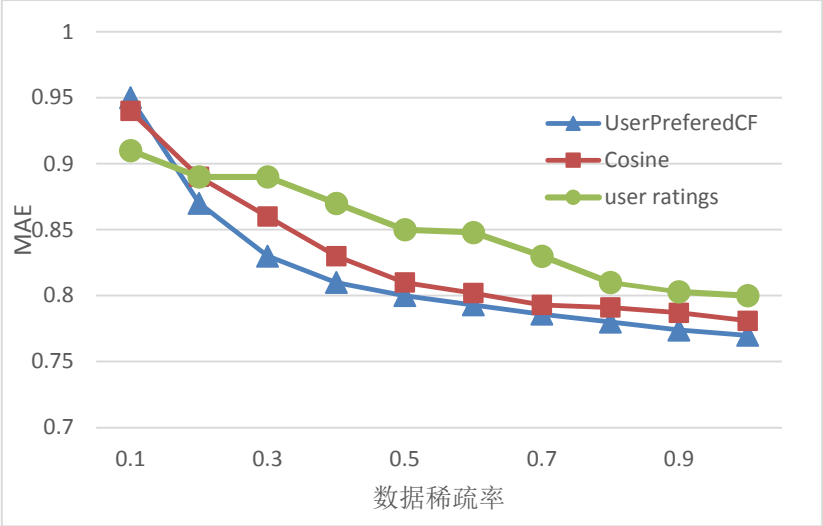


图 4.4 ml-100k 数据集不同数据稀疏程度下对推荐结果的影响
Fig. 4.4 The Effect of Different Data Scatter Degree
on the Recommendation Results on ml-100k dataset

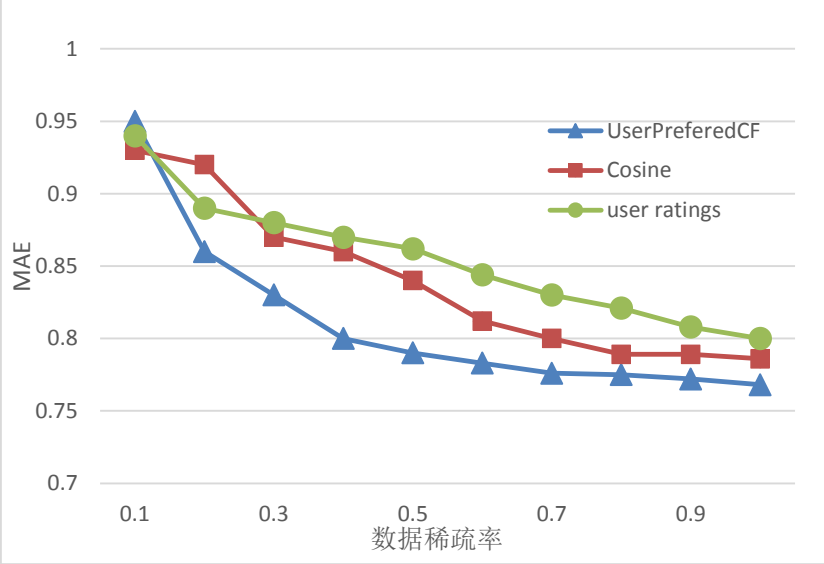


图 4.5 蒙东交易数据集不同数据稀疏程度下对推荐结果的影响
Fig. 4.5 The Effect of Different Data Scatter Degree
on the Recommendation Results on MengDong trading dataset

（3） 不同核函数在不同的带宽下对推荐结果的影响

本实验室考察使用核密度估计时，核函数的选取和不同带宽对结果的影响。图 4.6 是当带宽 h 选取从 0.1 到 1.0 之间的 10 个取值下高斯核函数和三角和函数的 MAE 值对比。从该图上可以看出，带宽不断增加的过程中，三角核函数的结果比高斯核函数的效果略好，但当带宽趋近于 1 时，两者的结果又趋于一致。

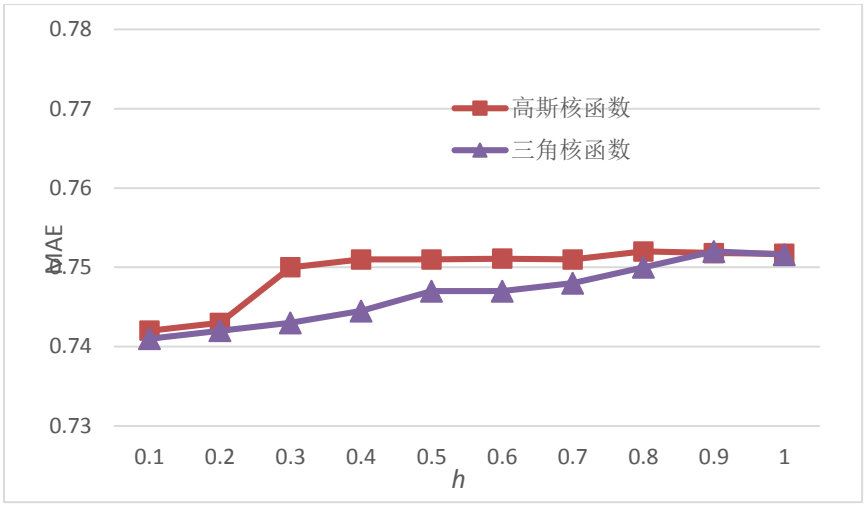


图 4.6 不同核函数在不同的带宽下对推荐结果的影响
Fig. 4.6 The effect of different kernels on the recommended results under different bandwidths

另外，随着带宽不断增大，两种核函数下的推荐质量逐渐降低。因为带宽越大，则对偏好估计的平滑程度越弱，表现这种偏好的尺度就越大，对偏好估计的就越粗糙，在估计商品的时候会把用户兴趣估计到与该商品相关性不大的商品上。可见，核函数的选取所带来的影响与带宽所带来的影响相比，后者更明显。

4.5 本章小结

本章提出了基于用户偏好估计的协同过滤算法。首先给出基于用户偏好的推荐模型的建立过程，然后提出基于商品标签的相似度。基于用户偏好估计的协同过滤算法的过程总体包含三个阶段：用户偏好密度估计、用户相似性计算和评分数据预测及填充。最后通过在两个真实数据集上对算法的准确性和数据稀疏性以及算法使用的核函数问题上做出了评价，该算法在数据稀疏的情况下能给出精确度较高的推荐结果。

第5章 电力交易推荐系统移动端设计与实现

如今移动互联网技术更迭日新月异，移动智能设备远远超过个人微型计算机的数量。大多数互联网产品都扩展了面向移动互联网的产品线。随着智能设备的普及，主流移动操作系统呈现竞争趋势，目前市场占有量最大的是 Android 操作系统。本章根据相关电力改革背景，设计并实现基于 Android 平台下的电力交易推荐系统原型，使用软件工程的思路阐述电力交易推荐系统移动端原型的设计和实现过程。

5.1 需求分析

软件需求分析是软件设计至关重要的阶段。该部分是软件开发的方向性的指导方针。主要任务是对软件的需要和要求进行归纳和整理。本节将分析电力交易推荐系统的功能需求。

使用本系统的角色包含三类：大用户、发电企业、电网企业。其中，电力交易推荐系统移动端的使用角色仅包含大用户。大用户可以在移动端浏览查询交易信息，接收推荐信息，管理合同信息，以及进行个人用户信息管理等。因此移动端需求可通过如图 5.1 的例图来描述。

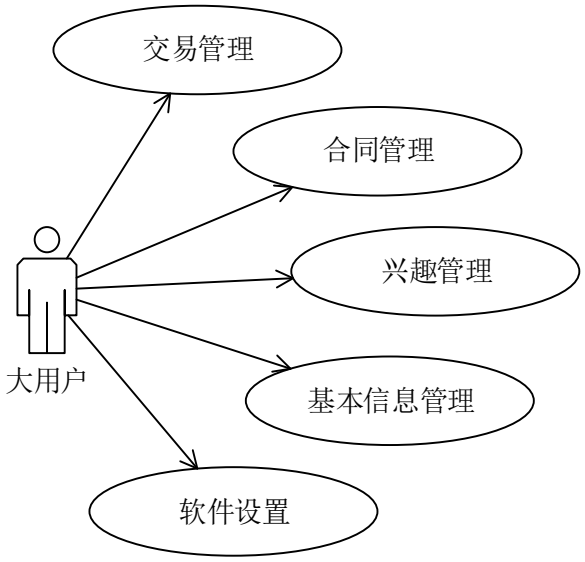


图 5.1 系统移动端用例图
Fig. 5.1 System mobile use case diagram

（1） 交易管理：含有交易公告、交易申报、交易结果等主要功能需求。

主要实现当前用户参与的包括交易公告详细内容、申报状态、交易发布时间、申报截止时间、交易附件；申报数据结果，根据不同交易类型展示数据申报结果，主要内容包括电价、电量、时间段等信息以及交易类型、交易名称、交易形式、成交电量、成交

均价等交易结果。

（2） 兴趣管理：包含当前用户感兴趣的售方展示的功能需求。

主要实现根据用户的行为进行推荐，为用户推送感兴趣售方供用户选择。

（3） 合同管理：含有当前合同、历史合同、合同分析等功能需求。

主要实现对当前用户的合同信息查询，并且提供当前合同的执行追踪情况及统计分析。

（4） 基本信息管理：对市场成员、机组、联系人等信息的管理需求。

含有市场成员信息、机组信息、用电单元信息、联系人信息、准入用户信息、市场成员历史信息、机组历史信息、用电单元历史信息以及统计信息等功能，实现了对当前用户的信息全生命周期统计查询。

（5） 软件设置：含有密码修改、客户端缓存清除等需求。

主要实现密码修改功能，清除客户端缓存功能。

5.2 总体设计

总体设计是软件开发过程中的一个设计性的工作，主要包含对系统架构的设计。本节介绍电力交易推荐系统的系统架构设计情况。

5.2.1 服务端架构设计

服务端的架构总体包含三个层次：数据层、业务层和视图层。如图 5.2 所示。

在服务端主体采用 MVC 架构模式。按照功能职责可以分为数据层、业务层和视图层。数据层中包含对用户行为数据的存储单元，为上层提供存储服务。作为服务端系统的底层，其核心需求是确保数据的安全性和数据读写的性能。安全性体现在底层的高容错性，比如数据可以自动保存为多个副本，副本丢失后可以自动恢复。另外由于用户的行为日志数据会源源不断的产生，因此数据层还要有面向大规模数据的处理能力。GB、TB 级数据可以很安全的存储在数据层。业务层包含日志引擎和推荐引擎。日志引擎是用来采集用户行为数据，并将用户行为数据持久化到数据层。其核心功能在于两个方面：一是读取移动端用户数据，对用户行为过滤，并将用于核心推荐业务的用户行为数据统一格式化；二是对格式化的用户行为数据持久化，核心业务在于将行为数据写入数据层，该过程涉及 I/O 操作，吞吐率和时间性能会是该功能的最大瓶颈。最上层就是用户直接接触的移动端，即视图层。该层是用户直观上满意度的集中体现，用户的满意度往往在于用户界面的友好度。移动端程序运行的流畅度、界面美观性、操作便捷性等，都是视图层质量的体现。

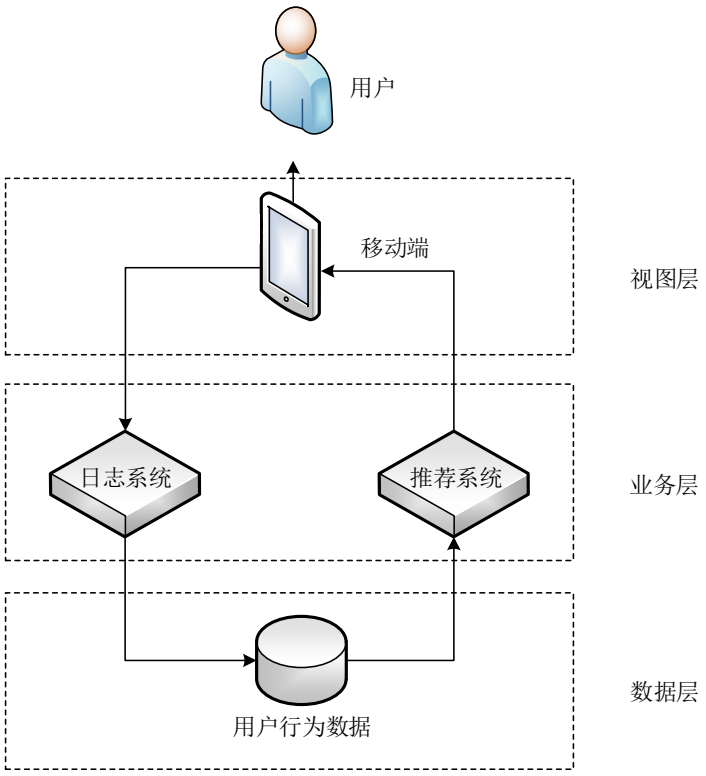


图 5.2 服务端架构设计
Fig. 5.2 Server-side architecture design

视图层是与客户之间直接发生交互的移动端程序，数据层用来存储用户行为的数据，业务层是负责视图层和数据层发生交互的中间程序，提供业务支持，包括对用户行为数据的采集系统，即日志系统，还包括本系统的核心服务，即推荐系统。三者之间发生交互，相互合作，共同为用户提供推荐服务。

推荐服务主体由推荐系统产生，由用户行为驱动。当用户产生行为数据，首先由日志系统对这部分数据进行收集，本部分采用流式数据处理引擎做日志处理（如 Flume）。收集后的用户行为日志数据会存储到数据层，使用 HDFS 提供的分布式文件存储服务完成。由于本系统的用户实时性不那么明显，没有实时推荐需求，因而对获得的日志文件通过推荐系统的离线分析来生成推荐结果。生成的推荐列表由服务层发送到视图层，即用户的 Android 智能设备上。

5.2.2 移动端架构设计

在移动端整体开发遵循 MVC 架构模式[48-49]。移动端架构设计如图 5.3。

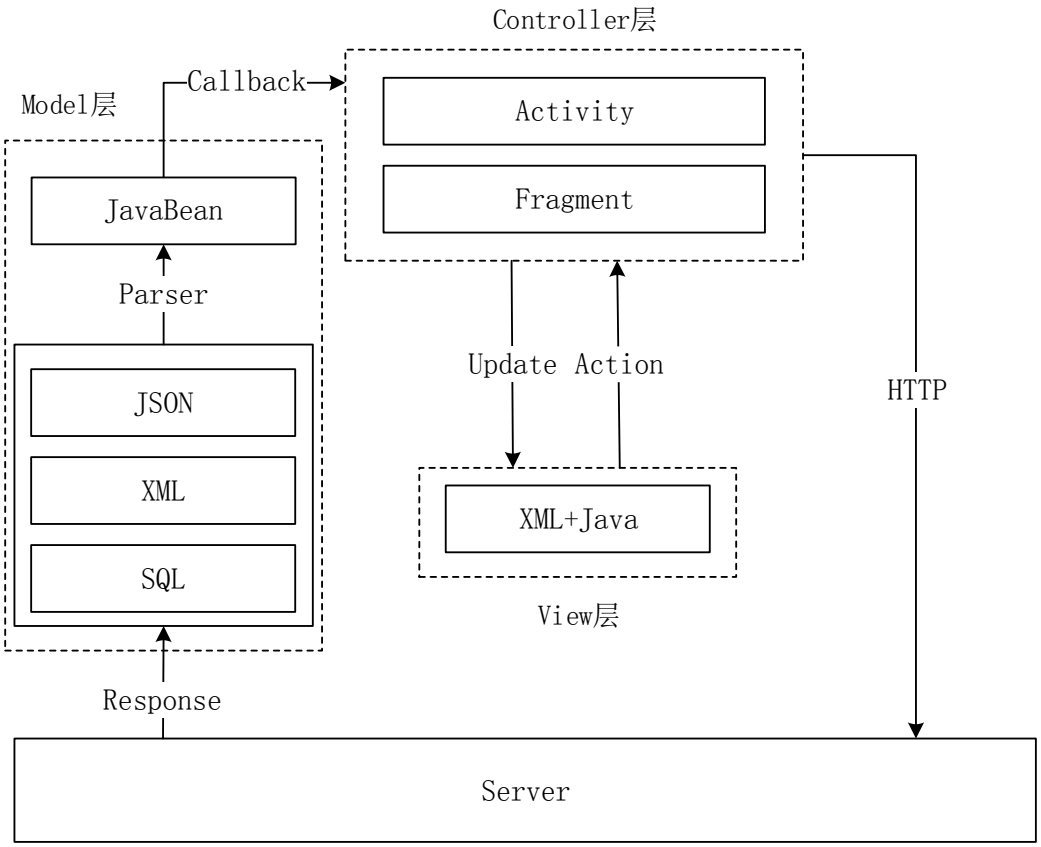


图 5.3 客户端 MVC 架构设计
Fig. 5.3 Client MVC architecture design

视图层(View)是对用户显示数据的一个展示层，主要负责显示界面，包括显示用户的注册，登录等操作功能界面。针对本系统的移动端视图，在 res/layout 目录下完成界面的布局，在 android.view 包下完成界面的组合封装。视图层的开发使用 Xml 语言和 Java 语言共同完成。

控制层(Controller)是调用模型层层的相关代码来实现客户端的业务逻辑控制。用于捕获用户请求并控制请求转发。在移动端的架构中，控制层发送 HTTP 协议请求到服务器端获取相应数据，并根据模型层的回调把数据响应到视图层，用户每次对视图层的触发会产生一个 Action 给控制层，控制层调用模型层代码实现业务逻辑控制。本系统客户端程序使用 Android 的 Activity 和 Fragment 共同完成控制层的编写。

模型层(Model)主要是负责业务逻辑以及数据库的交互。在客户端会有一个本地数据库，在 Android 系统中，采用的是 SQLite，一个基于 MySQL 的关系型数据库。使用本地数据库可以提高系统的响应速度，存储本地系统的数据模型。服务端每次的响应会发送出 JSON 数据，模型层捕获 JSON 数据加以封装，存储于本地数据库 SQLite 中，为控制层提供数据模型。

5.2.3 网络架构设计

电力交易移动应用同时提供 WIFI 和移动网络两种接入方式，参与市场交易的市场成员用户采用 VPN 方式接入，可以保证数据传输安全，公众大用户不需要使用 VPN 通道，直接接入外网代理服务器。外网部署日志服务器和代理服务器，对外统一提供代理服务，内网部署移动应用服务器提供电力交易应用服务，内外网隔离方式有效保证内网数据的安全。内网移动应用服务器与电力交易系统相连接，实时交互业务数据。如图 5.4 所示为网络架构示意图。

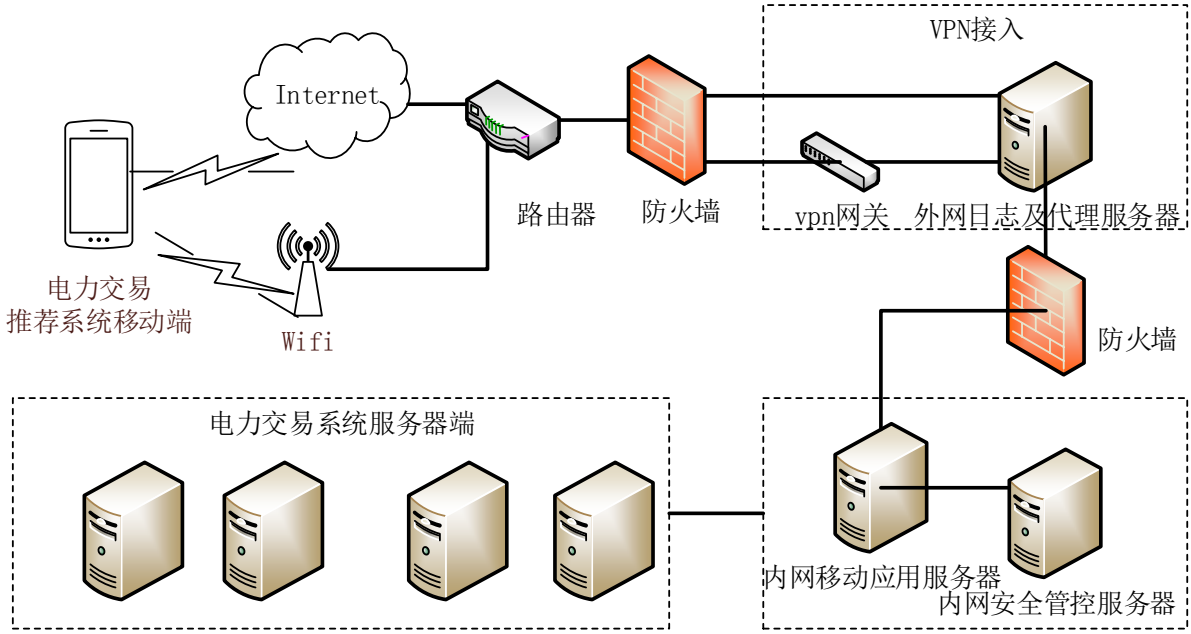


图 5.4 电力交易移动应用网络交互架构示意图

Fig. 5.4 Electric trading mobile application network interaction architecture diagram

5.3 详细设计

本节详细介绍移动端原型的开发流程，包括功能模块设计、数据库的逻辑模型和物理模型设计等。

5.3.1 移动端功能模块设计

由系统的需求分析可以获取到五大根本的模块：交易管理、兴趣管理、合同管理、基本信息管理和软件设置。在这些模块中，可以划分出若干个子功能。移动端的功能模块设计如图 5.5。其中，最核心的推荐功能部署在兴趣管理模块，该模块主要负责对推荐结果的展示功能。移动端原型设计将实现其中较为核心的若干模块。

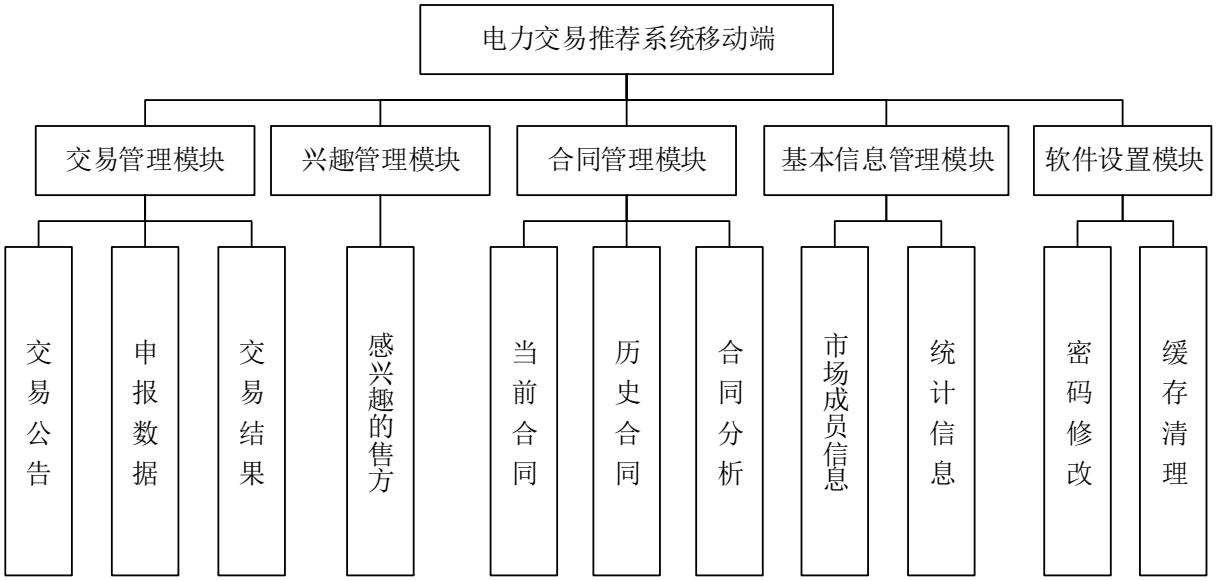


图 5.5 电力交易推荐系统移动端模块设计

Fig. 5.5 Mobile Terminal Module Design for Electric Trading Recommendation System

现对图 5.5 中的各个子模块作出分析：

交易公告：展示当前用户参与的包括交易公告详细内容、申报状态、交易发布时间、申报截止时间、交易附件。

申报数据：展示申报数据结果，更具不同交易类型展示数据申报结果，主要内容包括电价、电量、时间段等信息。

交易结果：展示交易结果，内容包括交易类型、交易名称、交易形式、成交电量、成交均价等信息。

当前合同：展示合同有效日期包含当前时间的合同，包括合同详细信息和合同文本附件，合同详细信息中可查看合同名称、合同类型、开始时间、截至时间等内容，同时还可以在合同信息中查看合同按月分解的电量信息、计量点信息、合同分段电量电价信息、合同输电信息、合同机组信息。

历史合同：展示合同有效日期不包含当前时间的合同，包括合同详细信息和合同文本附件，合同详细信息中可查看的内容与“当前合同”相同。

合同分析：大用户根据电厂发电类型对自己合同电量、电价进行分析，对省内其他电厂电量、电价进行分析。

感兴趣的售方：根据用户行为分析，推荐用户感兴趣的售方。

市场成员信息：展示市场成员信息，内容包括企业全称、市场成员类型、入市时间、地理区域、企业法人名称、开户银行、开户名称、开户账号、税务登记证号、通信地址、邮政编号、联系人姓名、办公电话等。

统计信息：图表展示装机情况统计信息，包括柱状图展示装机分类情况、饼图展示直调装机占比情况、表格展示新增机组情况、折线图展示总装机曲线。

密码修改：修改用户密码。

缓存清理：清除软件垃圾缓存。

5.3.2 数据库逻辑模型设计

根据功能模块的设计，以及需求分析的要求，可以设计出数据库的逻辑模型。如图 5.6 所示为部分主要的实体-联系图。

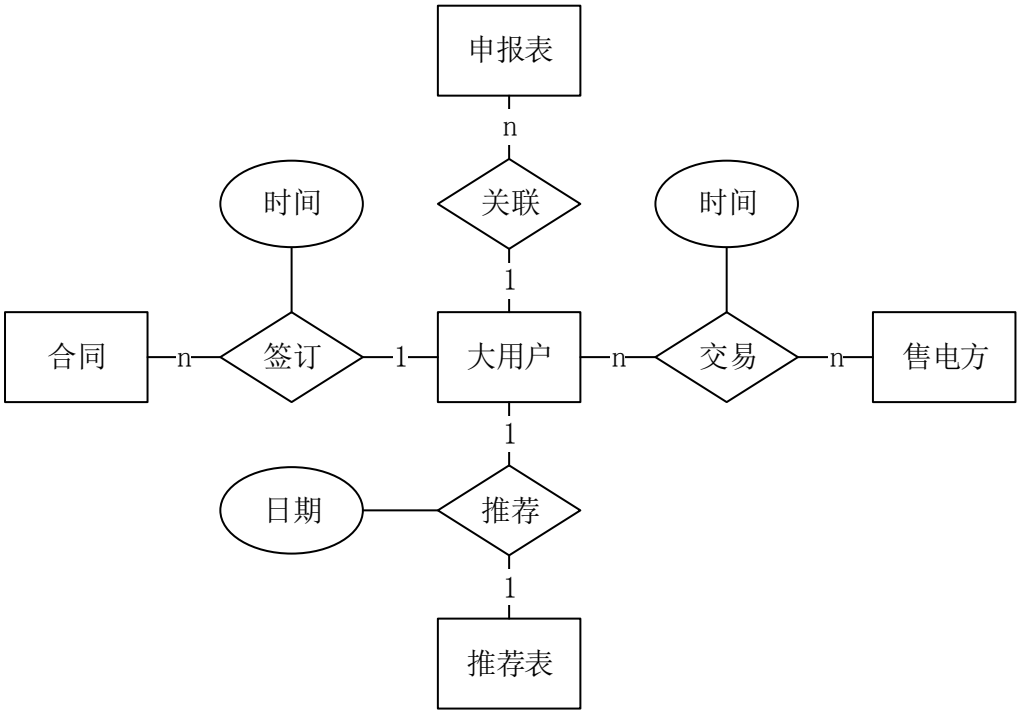


图 5.6 部分核心实体-联系图

Fig. 5.6 Some core entities - contact diagram

大用户作为最核心的实体，与多个实体发生联系。分析交易模块，大用户与售电方发生交易的同时会产生时间属性，一个大用户可以喝多个售电方发生交易，同样，一个售电方也可以与多个大用户发生交易，因此是多对多关系。同理，分析合同管理模块，大用户签订合同，可以签订多个合同，可见是一对多关系。分析其他的各个模块，可以得出该 E-R 图。

5.3.3 数据库物理模型设计

针对上述的 E-R 图逻辑模型设计，可以分析得出该系统移动客户端的物理表设计。每个实体可以抽象出一张实体表，实体物理表将实体的各个属性加以封装；每个联系也可以抽象出关联表，关联生成时可能会产生新的属性，该属性也要容纳入关联表中。作

为示例，本文列举了大用户表、交易表、推荐表与合同表的部分主要字段，见表 5.1、表 5.2、表 5.3 和表 5.4。

表 5.1 大用户表的部分字段
Table 5.1 Some of the large user table fields

字段	主键	数据类型	长度	非空	说明
UID	是	varchar	8	是	表主键
UName	否	varchar	25	是	企业全称
Legal	否	varchar	12	是	企业法人名称
Region	否	varchar	25	是	企业地理位置
EnterTime	否	date	-	是	入市时间
Address	否	varchar	50	否	通信地址

表 5.2 交易表部分字段
Table 5.2 Some of the Transaction table field

字段	主键	数据类型	长度	非空	说明
UID	是	varchar	8	是	大用户 ID
PID	是	varchar	8	是	售电方 ID
DealTime	否	date	-	是	交易时间

表 5.3 推荐表部分字段
Table 5.3 Some of the Recommended table field

字段	主键	数据类型	长度	非空	说明
UID	是	varchar	8	是	大用户 ID
PID	是	varchar	8	是	售电方 ID
RecTime	否	date	-	是	推荐生成时间

表 5.4 合同表部分字段
Table 5.4 Some of the Contract table field

字段	主键	数据类型	长度	非空	说明
CID	是	varchar	8	是	合同 ID
UID	外	varchar	8	是	大用户 ID
CType	是	varchar	10	是	合同类型
StartTime	否	date	-	是	开始时间
CutTime	否	date	-	是	截止时间

5.4 系统展现

本系统移动端的设计遵循 MVC 设计模式原则[49]，界面布局简约，图 5.7 和图 5.8 分别为系统的数据申报、感兴趣的售方、合同信息界面详情，图中数据仅做测试使用。

图 5.7 移动端部分界面展示
Fig. 5.7 Mobile part of the interface display

图 5.8 移动端部分界面展示
Fig. 5.8 Mobile part of the interface display

5.5 本章小结

本章相关电力改革背景，设计并实现了基于 Android 平台下的电力交易推荐系统原型。对该系统移动端的核心做出设计与实现。首先对移动端的功能需求做出分析，明确了系统的功能需求，移动端的设计划分为五个基本需求，然后对系统的总体设计做出了规划，包括服务端架构的设计，移动端架构的设计以及网络架构设计。移动端整体遵循 MVC 设计模式，重点在于界面布局和数据展现。在移动端的详细设计阶段，根据需求分析的结果做出移动端功能模块设计与规划，并给出了数据库的逻辑模型和物理模型的设计，完成编码后，对系统原型的基本页面做出展示。最终实现电力交易推荐系统的移动端原型。

第6章 总结与展望

6.1 本文总结

本文以电力改革为研究背景，对比国内外电力改革情况，并围绕电力交易推荐系统展开了推荐算法的技术讨论。以电力交易推荐系统为应用背景，提出了两种适合于不同情境下的推荐算法以应对电力推荐系统的实际需求。

本文提出基于时序社交关系的协同过滤算法，在该算法中，本文的主要创新与贡献有三点：首先，利用用户发生评分或交易的时序关系，挖掘出用户之间的影响关系和从众关系，将获得的两种关系融入概率矩阵分解算法以提高其准确率，并给出计算复杂度分析。其次，在此过程中，提出了时序关系下，用户社交关系的定量分析方法。最后，基于电力交易过程中数据的特点，提出了评分和时间数据获取与标准化方案，使得算法可以有效处理电力交易数据。通过在真实数据集上对算法做出评估检验可知，该算法具有较高的准确度和较好的性能。

本文还提出了一个应对稀疏数据的基于用户偏好估计的协同过滤算法，在该算法中，本文的主要创新与贡献有三点：本文采取一种挖掘用户偏好的方式来对缺失数据进行填充，来提高矩阵填充的准确度。传统方法对待用户间的相似性往往给出的估计值较为粗略，结合统计学中的非参数估计方法——核密度估计方法，来综合分析用户的偏好分布，通过分析两个用户的 KL 散度的方式来计算它们的相似度。在计算用户相似度的过程中，本文提出了一种基于商品标签的相似度，可以从商品标签的角度描述商品的相似度。最后在真实数据集上验证了该算法应对稀疏数据有较好的性能。

最后本文根据电力改革背景，设计并实现基于 Android 平台下的电力交易推荐系统原型。根据软件工程化思路，从需求分析到总体设计，再到详细设计，分别做了移动端、服务端和网络架构设计以及数据库设计，最终给出电力推荐的移动端展示平台。

6.2 未来工作

本文以电力交易推荐系统为应用背景，提出了对协同过滤算法的不同的改进算法，并完成了电力交易推荐系统移动端原型的设计与实现。在所改进的算法中还存在一些问题有待进一步研究。首先，在基于时序社交关系的协同过滤算法中，本文只是联系了用户之间的社交关系，而没有融合商品之间的内在联系，可见算法的准确性方面还可以提升，未来工作可以朝向融合商品内在联系的方向对算法进行改进。另外，该算法针对稀疏数据的处理精度还不够，未来可以着重对稀疏数据的处理性能方面做深入研究。其次，

在基于用户偏好估计的协同过滤算法中，本文使用核密度估计做用户偏好估计方法，用户的偏好可能存在参照尺度不同的问题，因而可以考虑使用自适应带宽的核密度估计方法对用户偏好进行估计。这部分工作还有待后续研究。另外，本文的系统实现过程只是针对性的对移动端做出设计与实现，后续工作要考虑服务端对大数据量的应对，以及对实时性问题的解决。

附A SeqSoPMF 推导过程

SeqSoPMF 概率矩阵分解过程推导如下。

对$\ln\prod_{u=1}^{n}\prod_{i=1}^{n}\left[\mathcal{N}(R_{u,i}|g(U_u^TV_i),\sigma_R^2)\right]^{\mathcal{I}_{u,i}^R}$的推导过程：

$$\mathcal{N}(R_{u,i}|g(U_u^TV_i),\sigma_R^2)=\frac{1}{\sqrt{2\pi}\sigma_R}\cdot e^{-\frac{(R_{u,i}-g(U_u^TV_i))^2}{2\sigma_R^2}} \tag{D.18}$$

对加上指数$\mathcal{I}_{u,i}^R$，可得：

$$\mathcal{N}(R_{u,i}|g(U_u^TV_i),\sigma_R^2)^{\mathcal{I}_{u,i}^R}=\left(\frac{1}{\sqrt{2\pi}\sigma_R}\right)^{\mathcal{I}_{u,i}^R}\cdot e^{-\frac{(R_{u,i}-g(U_u^TV_i))^2\cdot\mathcal{I}_{u,i}^R}{2\sigma_R^2}} \tag{D.19}$$

对等号两侧取对数，有：

$$\ln\mathcal{N}(R_{u,i}|g(U_u^TV_i),\sigma_R^2)^{\mathcal{I}_{u,i}^R}=-\frac{(R_{u,i}-g(U_u^TV_i))^2\mathcal{I}_{u,i}^R}{2\sigma_R^2}-\mathcal{I}_{u,i}^R\ln\sqrt{2\pi}\sigma_R \tag{D.20}$$

因此，

$$\begin{aligned}&\ln\prod_{u=1}^{n}\prod_{i=1}^{n}\left[\mathcal{N}(R_{u,i}|g(U_u^TV_i),\sigma_R^2)\right]^{\mathcal{I}_{u,i}^R}\\=&-\sum_{u=1}^{n}\sum_{i=1}^{m}\left(-\frac{(R_{u,i}-g(U_u^TV_i))^2\mathcal{I}_{u,i}^R}{2\sigma_R^2}+\mathcal{I}_{u,i}^R\ln\sqrt{2\pi}\sigma_R\right)\\=&-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^R(R_{u,i}-g(U_u^TV_i))^2-\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^R\ln\sqrt{2\pi}\sigma_R\end{aligned} \tag{D.21}$$

对$\ln\prod_{u=1}^{n}\mathcal{N}\left(U_u\middle|\sum_{u'_u}l_{u',u}U'_u,\sigma_l^2\boldsymbol{E}\right)$的推导过程：

$$\mathcal{N}\left(U_u\middle|\sum_{u'_u}l_{u',u}U'_u,\sigma_l^2\boldsymbol{E}\right)=\frac{1}{\sqrt{2\pi}\sigma_l}e^{-\frac{1}{2\sigma_l^2}\left(U_u-\sum\limits_{u'\in S_u}l_{u',u}U_{u'}\right)^2} \tag{D.22}$$

对等号两侧取对数，有：

$$\ln\mathcal{N}\left(U_u\middle|\sum_{u'_u}l_{u',u}U'_u,\sigma_l^2\boldsymbol{E}\right)=-\frac{1}{2\sigma_l^2}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2-\ln\sqrt{2\pi}\sigma_l \tag{D.23}$$

因此，

$$\begin{aligned}\ln\prod_{u=1}^{n}\mathcal{N}\left(U_u\middle|\sum_{u'_u}l_{u',u}U'_u,\sigma_l^2\boldsymbol{E}\right)&=\sum_{u=1}^{n}\ln\mathcal{N}\left(U_u\middle|\sum_{u'_u}l_{u',u}U'_u,\sigma_l^2\boldsymbol{E}\right)\\&=-\frac{1}{2\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2\\&\quad-\sum_{u=1}^{n}\ln\sqrt{2\pi}\sigma_l\end{aligned}\tag{D.24}$$

同理，对$\ln\prod_{u=1}^{n}\mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E})$的推导过程和$\ln\prod_{i=1}^{n}\mathcal{N}(V_i|0,\sigma_V^2\boldsymbol{E})$的推导后，得，

$$\begin{aligned}\ln\prod_{u=1}^{n}\mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E})&=\sum_{u=1}^{n}\ln\mathcal{N}(U_u|0,\sigma_U^2\boldsymbol{E})\\&=-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^2-\sum_{u=1}^{n}\ln\sqrt{2\pi}\sigma_U\end{aligned}\tag{D.25}$$

$$\begin{aligned}\ln\prod_{i=1}^{n}\mathcal{N}(V_i|0,\sigma_V^2\boldsymbol{E})&=\sum_{i=1}^{m}\ln\mathcal{N}(V_i|0,\sigma_V^2\boldsymbol{E})\\&=-\frac{1}{2\sigma_V^2}\sum_{i=1}^{m}V_i^2-\sum_{i=1}^{m}\ln\sqrt{2\pi}\sigma_V\end{aligned}\tag{D.26}$$

最后，对$p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)$两侧取对数，并将上述结果带入，整理得：

$$\begin{aligned}\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)=&-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2\\&-\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\ln\sqrt{2\pi}\sigma_R\\&-\frac{1}{2\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2-\sum_{u=1}^{n}\ln\sqrt{2\pi}\sigma_l\\&-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^2-\sum_{u=1}^{n}\ln\sqrt{2\pi}\sigma_U\\&-\frac{1}{2\sigma_V^2}\sum_{i=1}^{m}V_i^2-\sum_{i=1}^{m}\ln\sqrt{2\pi}\sigma_V\end{aligned}\tag{D.27}$$

为了便于计算，将式(D.27)可整理为如下后验概率（由于$\ln\sqrt{2\pi}\sigma_R=\frac{1}{2}\ln(2\pi\sigma_R^2)$，其他类似），

$$
\begin{aligned}
&\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)\\
=&-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^2-\frac{1}{2\sigma_V^2}\sum_{i=1}^{m}V_i^2\\
&-\frac{1}{2\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2-\frac{1}{2}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln\sigma_R^2\\
&-\frac{n\cdot k}{2}\ln\sigma_l^2-\frac{n\cdot k}{2}\ln\sigma_U^2-\frac{m\cdot k}{2}\ln\sigma_V^2\\
&-\frac{1}{2}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln(2\pi)-(n\cdot k)\ln(2\pi)-\frac{m\cdot k}{2}\ln(2\pi)
\end{aligned}
\tag{D.28}
$$

令$C=-\frac{1}{2}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln(2\pi)-(n\cdot k)\ln(2\pi)-\frac{m\cdot k}{2}\ln(2\pi)$，$C$为常数，则式(D.28)可以化简为，

$$
\begin{aligned}
\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)=&-\frac{1}{2\sigma_R^2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2\\
&-\frac{1}{2\sigma_U^2}\sum_{u=1}^{n}U_u^2-\frac{1}{2\sigma_V^2}\sum_{i=1}^{m}V_i^2\\
&-\frac{1}{2\sigma_l^2}\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2\\
&-\frac{1}{2}\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln\sigma_R^2\\
&-\frac{1}{2}\left(nk\ln\sigma_l^2+nk\ln\sigma_U^2+mk\ln\sigma_V^2\right)+C
\end{aligned}
\tag{D.29}
$$

对上式左右同时乘以$(-\sigma_R^2)$可得

$$
\begin{aligned}
&(-\sigma_R^2)\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)\\
=&\frac{1}{2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_U^2}\cdot\sum_{u=1}^{n}U_u^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_V^2}\cdot\sum_{i=1}^{m}V_i^2\\
&+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_l^2}\cdot\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2+\frac{1}{2}\sigma_R^2\left(\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}\right)\ln\sigma_R^2\\
&+\frac{1}{2}\sigma_R^2\left(nk\ln\sigma_l^2+nk\ln\sigma_U^2+mk\ln\sigma_V^2\right)+C_1
\end{aligned}
\tag{D.30}
$$

令损失函数为$L(R,l,U,V)=(-\sigma_R^2)\ln p(U,V|R,l,\sigma_R^2,\sigma_U^2,\sigma_V^2)$，最大化后验概率就相当于最小化该损失函数，即

$$
\begin{aligned}
L(R,l,U,V)=&\frac{1}{2}\sum_{u=1}^{n}\sum_{i=1}^{m}\mathcal{I}_{u,i}^{R}(R_{u,i}-g(U_u^TV_i))^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_U^2}\cdot\sum_{u=1}^{n}U_u^2\\
&+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_V^2}\cdot\sum_{i=1}^{m}V_i^2+\frac{1}{2}\cdot\frac{\sigma_R^2}{\sigma_l^2}\cdot\sum_{u=1}^{n}\left(U_u-\sum_{u'\in S_u}l_{u',u}U_{u'}\right)^2
\end{aligned}
\tag{D.31}
$$

L分别对U和V求偏导，梯度的计算如下：

$$\begin{aligned}\frac{\partial L}{\partial U_u} =& \sum_{i=1}^{m} \mathcal{I}_{u,i}^{R} V_i g'(U_u^T V_i)(g(U_u^T V_i) - R_{u,i}) + \frac{\sigma_R^2}{\sigma_U^2} U_u \\ &+ \frac{\sigma_R^2}{\sigma_l^2} \sum_{u=1}^{n} \left(U_u - \sum_{u' \in S_u} l_{u',u} U_{u'} \right) \left(1 - \sum_{u' \in S_u} l_{u',u} \right)\end{aligned} \tag{D.32}$$

$$\frac{\partial L}{\partial V_i} = \sum_{u=1}^{n} \mathcal{I}_{u,i}^{R} U_u g'(U_u^T V_i)(g(U_u^T V_i) - R_{u,i}) + \frac{\sigma_R^2}{\sigma_V^2} V_i \tag{D.33}$$

其中，$g'(x) = \dfrac{e^{-x}}{(1+e^{-x})^2}$，为$g(x)$的一阶导数。然后利用随机梯度下降可以得到用户或商品的特征向量。

参考文献

[1] 中发〔2015〕9 号, 关于进一步深化电力体制改革的若干意见[S].

[2] 刘鲁, 任晓丽. 推荐系统研究进展及展望[J]. 信息系统学报, 2008(1):82-90.

[3] 刘杰. 电力市场力的量度及抑制理论方法研究[D]. 华北电力大学(北京), 2004.

[4] 程其麟. 贵州电网企业发展研究[D]. 贵州大学, 2006.

[5] Public Utility Commission of Texas.25.107 Certification of retail electric providers(REPs)[EB/OL].[2015-11-20].

[6] 张晓萱, 薛松, 杨素,等. 售电侧市场放开国际经验及其启示[J]. 电力系统自动化, 2016, 40(9):1-8.

[7] American Public Power Association.2014 retail electric rates in deregulated and regulated States[R].2015.

[8] 郭立夫, 崔新宇. 日本电力交易所的设计与运营[J]. 商场现代化, 2007(14):80-81.

[9] ASANO H.Regulatory reform of the electricity industry in Japan:what is the next step of deregulation?[J].Energy Policy,2006,34(16):2491-2497.

[10] Maes P. Agents that reduce work and information overload[M]. ACM, 1994.

[11] 孙光福, 吴乐, 刘淇,等. 基于时序行为的协同过滤推荐算法[J]. 软件学报, 2013(11):2721-2733.

[12] Goldberg D. Using collaborative filtering to weave an information tapestry[J]. Communications of the Acm, 1992, 35(12):61-70.

[13] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]// ACM Conference on Computer Supported Cooperative Work. ACM, 1994:175-186.

[14] Resnick P, Varian H R. Recommender systems[M]. ACM, 1997.

[15] 项亮. 推荐系统实践[M]. 人民邮电出版社, 2012.

[16] Das A S, Datar M, Garg A, et al. Google news personalization:scalable online collaborative filtering[C]// International Conference on World Wide Web. ACM, 2007:271-280.

[17] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]// International Conference on World Wide Web. ACM, 2001:285-295.

[18] Salakhutdinov R, Mnih A. Probabilistic Matrix Factorization[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2007:1257-1264.

[19]Stern D H, Herbrich R, Graepel T. Matchbox:large scale online bayesian recommendations[C]// International Conference on World Wide Web. ACM, 2009:111-120.

[20]Christakopoulou K, Radlinski F, Hofmann K. Towards Conversational Recommender Systems[C]// The, ACM SIGKDD International Conference. ACM, 2016:815-824.

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

[22]Song Y, Zhang L, Giles C L. Automatic tag recommendation algorithms for social recommender systems[M]. ACM, 2011.

[23]Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithm for Collaborative Filtering[J]. 2013, 7(7):43-52.

[24]Karypis G. Evaluation of Item-Based Top-N Recommendation Algorithms[C]// Tenth International Conference on Information and Knowledge Management. ACM, 2001:247-254.

[25]Silverman B W. Density estimation for statistics and data analysis[M]. CRC press, 1986.

[26]Givens G H, Hoeting J A. Computational statistics[M]. John Wiley & Sons, 2012.

[27]陈希孺. 非参数统计教程[M]. 华东师范大学出版社, 1993.

[28]Kullback S, Leibler R A. On Information and Sufficiency[J]. Annals of Mathematical Statistics, 1951, 22(1):79-86.

[29]Hazewinkel M. Encyclopaedia of Mathematics[J]. Reference Reviews, 1995, 17(7):49-50.

[30]Olkin I, Pukelsheim F. The distance between two random vectors with given dispersion matrices[J]. Linear Algebra & Its Applications, 1982, 48(82):257-263.

[31]Kong W, Liu Q, Yang Z, et al. Collaborative filtering algorithm incorporated with cluster-based expert selection[J]. 2012, 9(12):3421-3429.

[32]胡江溢, 陈西颖. 对大用户直购电交易的探讨[J]. 电网技术, 2007, 31(24):40-45.

[33]胡清智. 关于发电企业向用户直供电情况的探讨[J]. 广西电业, 2014(1):93-96.

[34]Kong W, Liu Q, Wang S, et al. Relation-based collaborative filtering algorithm[J]. Journal of Computational Information Systems, 2012, 8(15):6257-6265.

[35]D. J. Field. What is the goal of sensory coding? Neural Computation, 6:559–601, 1994.

[36]Hoyer P O. Non-negative matrix factorization with sparseness constraints[C]// Jour. of. 2004:1457--1469.

[37]Huang Z. Applying associative retrieval techniques to alleviate the sparsity problem in

collaborative filtering[J]. Acm Transactions on Information Systems, 2015, 22(1):116-142.

[38]Ren J, Zhou T, Zhang Y C. Information Filtering via Self-Consistent Refinement[J]. Epl, 2008, 82(5):1-4.

[39]Sun D, Zhou T, Liu J G, et al. Information filtering based on transferring similarity.[J]. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2009, 80(1 Pt 2):017101.

[40]郭云飞, 方耀宁, 扈红超. 基于 Logistic 函数的社会化矩阵分解推荐算法[J]. 北京理工大学学报, 2016, 36(1):70-74.

[41]燕彩蓉, 张青龙, 赵雪,等. 基于广义高斯分布的贝叶斯概率矩阵分解方法[J]. 计算机研究与发展, 2016, 52(12):2793-2800.

[42]赵长伟, 彭勤科, 张志勇. 混合因子矩阵分解推荐算法[J]. 西安交通大学学报, 2016, 50(12):87-91.

[43]梅忠, 肖如良, 张桂刚. 基于受约束偏置的概率矩阵分解算法[J]. 计算机系统应用, 2016, 25(5):113-117.

[44]Jamali M, Ester M. A matrix factorization technique with trust propagation for recommendation in social networks[C]// ACM Conference on Recommender Systems. ACM, 2010:135-142.

[45]Billsus D, Pazzani M J. Learning Collaborative Information Filters[C]// Machine Learning. In: Proceedings of the Fifteenth International Conference. 1998.

[46]Deshpande M, Karypis G. Item-based top- N recommendation algorithms[J]. Acm Transactions on Information Systems, 2004, 22(1):143-177.

[47]Ruchansky N, Crovella M, Terzi E. Matrix Completion with Queries[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015:1025-1034.

[48]孙晓宇. Android 手机界面管理系统的设计与实现[D]. 北京邮电大学, 2009.

[49]刘昭. 基于 MVC 模式在重构 Android 开发的应用[J]. 科技致富向导, 2014(36):243-243.

致 谢

韶华易逝，毕业将期。东北大学的一花一木让人眷恋颇深。正值毕业之际，借此文向曾经在学习和生活中给予我指导、帮助和关爱的老师、同学和朋友致以衷心的感谢!

感谢我的恩师焦明海副教授！不论是在学习上，还是生活上，焦老师给予我的指导、关心和爱护都是无微不至的。在这短短的两年半时间里，焦老师经常与我讨论问题，指导我解决问题，更是教会我如何寻找解决问题的思路。焦老师的生活态度积极乐观、工作认真负责、治学严谨、求真务实。令人由衷敬佩。感谢焦老师给我树立的榜样，让我时刻思进取，不断图进步！

感谢于戈教授、张岩峰教授等计算中心的每一位老师。他们不仅在学术上给了我很多启发，开拓了我的研究视野，而且都像朋友一样给予我很大的帮助和关心。

感谢左鹏、朱明浩、于明乐、马祥振、许珊珊、威丰霞、李普、田申申等同学的帮助和关心。能与这些同学一同学习交流，让我倍感荣幸。感谢朝夕相处的信息楼 409 实验室的同伴、师弟师妹们，以及已经毕业的师兄师姐，是你们为我的生活带来了欢乐，增添了色彩。实验室浓郁的学习气氛和互相帮助的氛围，将成为我永生难忘的回忆!

感谢室友辛秉哲、战照昕、王泽众和岳春成，计算机学院硕士 21 班的全体同学，这些同学不仅给我学习和生活上的支持和鼓励，还陪伴我度过了整个研究生生涯。

感谢父母含辛茹苦的抚育，感谢我的家人，我的每一分成就都凝聚了你们无限的关爱、理解和支持!正是你们无言的付出与鼓舞，点燃了我前进的激情，使我能安心完成学业并收获一点一滴的进步。

最后，再次感谢所有关心、支持和帮助过我的人。衷心祝愿我的母校东北大学拥有更加美好，更加辉煌灿烂的明天!

攻读硕士学位期间的论文和项目情况

1. 2014.12-2017.12，教育部留学回国人员科研启动基金项目：基于服务的云数据资源优化关键理论与算法研究(项目编号：49-1)。

2. 2015.01-2016.12，企事业单位委托科技项目：电力交易信息采集与数据挖掘技术研究与开发，(项目编号： 2014-0-1-16489)。