摘　　要

随着我国互联网的高速发展，尤其是电子商务的急速膨胀，信息过载问题急需解决，而推荐系统是解决该问题的有效途径。个性化推荐系统通过收集用户的资料，预测用户可能感兴趣的商品，做出个性化的推荐服务，具有良好的发展前景。

本文分析了当前广泛使用的基于内容的推荐系统、协同过滤推荐系统、基于关联规则的推荐系统和混合推荐系统等系统，讨论了这几种系统研究的内容及其应用现状，并针对其中存在的数据稀疏，冷启动等问题提出改进方法。

于是，本文以 Movielens 数据集为载体，实现了基于二部分图网络推荐算法的改进。以往关于二部分图网络的推荐算法，在处理用户与项目之间关系时候，大多只是简单判断用户是否选择过该项目，非是即否。这样处理的结果不能很好地考虑用户对项目的喜欢程度，无法提高推荐结果的准确性。据此，本文提出了改进算法。为区分用户对项目的不同评价分数，引入权值系数λ_1、λ_2，以提高推荐的准确性。再引入项目的度和项目的权值之和的比值 θ 来提高项目推荐的多样性和流行性等推荐性能。

最终应用以上算法处理策略进行实验。实验结果表明，改进的算法在准确性、多样性和流行性等推荐性能上较原著算法有所提高。

关键词：二分图网络、推荐系统、个性化、权值系数、多样性、流行性

第一章　　绪论

1.1 选题背景与研究意义

计算机技术的发展和互联网的不断普及，满足了用户在信息时代对信息的需求，但随着网络的迅速发展而带来的网络信息量的指数增长，使得用户在面对大量信息时无法从中快速获得对自己真正感兴趣或有用的那部分信息，对信息的使用效率反而降低了。在庞大且复杂无比的互联网信息库面前，用户所需的信息量是微乎其微，如何从中快速、准确地找到所需的信息，是许许多多的学者所关注的问题。

个性化推荐系统正是解决这一问题的有效方法，它是根据用户的兴趣特点和历史行为，将用户感兴趣的信息、产品推荐给用户。和搜索引擎相比，个性化推荐系统通过研究用户的兴趣爱好，利用算法计算，能更好地找到用户需求的信息和引导用户发现自己对信息的需求。一个好的推荐系统不仅能为用户提供个性化的信息服务，还能与用户建立密切的联系，使用户对推荐系统产生依赖。

如今个性化推荐系统广泛应用在很多领域中，其中最为典型的就是电子商务领域。凭借个性化推荐系统，商家根据用户的兴趣和购买行为为用户推荐他们可能感兴趣的商品，能更好地激发用户的潜在需求。 目前，一些大型的电子商务网站，比如亚马逊、京东商城、淘宝网、当当网以及一些音乐、电影网站等都应用到了个性化推荐服务，可以说，个性化推荐的研究极具现实意义。

1.2 个性化推荐的研究现状

1.2.1　国内外研究现状

随着 Internet 的普及和电子商务的发展,推荐系统得到了越来越多研究者的关注。遗传算法、神经网络等机器学习技术也在推荐系统中得到应用,涌现出了越来越多的推荐方法,有代表性的如利用神经网络和遗传 K-means 算法通过分

析用户在电子商务网站的浏览路径来获取用户偏好的方法、基于案例式推理协同过滤推荐方法等等。

国外比较著名的推荐系统有 GroupLens、PHOAKS 和 Ringo 等等。

GroupLens 是一个应用于 Usenet 新闻的协同过滤系统,它的目标是让用户一起协作,从大量的 Usenet 新闻中发现他们感兴趣的内容。该系统分为客户端和服务器端两部分。客户端是一个新闻阅读器 NewsReader,服务器端提供协同过滤服务。NewsReader 一般连接到本地 NNTP 服务器,同时也连接到 GroupLens 服务器共享过滤信息,只要用户下载一篇文档,NewsReader 都会向 GroupLens 服务器发送消息请求对该文档内容的预测。此外,用户也可以评价文档,NewsReader 会将该用户评价发送到 GroupLens 服务器上进行处理,以提供给其他用户浏览,GroupLens 会利用这些信息调整该用户和其他用户的相关性[1]。

Terveenet 等人开发出 PHOAKS(people Helping One Another Know stuff)系统,将大家都认为值得看的网站推荐给用户。其运作的方式就是分析用户在 Usenet 中所张贴的布告,找出文章内所推荐的网站URL,并统计每个URL有多少人推荐,藉此来将相关的网站U既推荐给需要的人。实验结果证实PHOAKS是有效的,可以达到 90%的准确性[2]。

Ringo 是由麻省理工学院所设计的一个音乐推荐系统。这个系统会先要求使用者针对音乐家做评比,再依评比的结果计算使用者相似度,然后将使用者分群,最后再由同一族群的使用者互相推荐音乐给彼此[2]。

尽管目前我国在 Internet 领域取得了很大的发展,但是和西方发达国家比起来仍然存在着不小的差距,Internet 的发展落后严重影响我国电子商务的发展,从而使得推荐技术的发展失去了应用背景和基础。目前我国在电子商务推荐方面所使用的主要是查找或检索技术。这种推荐策略的优点是技术比较成熟,实现比较简单,然而其在推荐策略个性化、自动化、持久化三个方面与世界先进推

荐系统仍然存在着很大的差距,严格来说,这种查找策略并不具备主动提供个性化服务的功能,其与真正意义上的推荐策略仍存在差别。

目前我国电子商务的推荐功能相对国外存在较大的差距,主要表现在[3]:

(1)缺乏个性化的推荐。很多的推荐结果是对所有用户的,而非个性化的推荐,可能很多的推荐与某一用户的兴趣并不相符,这是我国电子商务推荐与国外推荐最主要的差别。

(2)推荐的自动化程度低。大多数的推荐功能都需要用户经过一段时间与计算机的交互,输入自己的兴趣信息,然后才能得到推荐结果,而系统不能保存用户每次的输入信息。总体说来,所有的推荐策略基本上停留在查找这一层次上,不能实现自动推荐。

(3)推荐的持久性程度低。目前我国的推荐技术都是建立在当前用户会话基础上,不能利用用户以前的会话信息,推荐的持久性程度非常低,这也是我国推荐技术和国外的推荐技术的一个重要差别。

(4)推荐方法单一。大多数所用的推荐策略基本就是分类浏览和基于内容的检索,缺乏多种推荐策略的混合使用,尤其缺少个性化与非个性化推荐策略的混合使用。

(5)不能在线推荐。有的推荐不能做到在线推荐,推荐不能及时反馈用户。然而随着我国电子商务事业的不断发展,对相应技术的迫切需求必将推动电子商务推荐技术研究不断深入,应用更加广泛。

1.2.2 传统个性化推荐系统的比较

个性化推荐研究于 20 世纪 90 年代被作为一个独立的概念提出来[4]，其目的是根据用户的兴趣爱好为用户推荐感兴趣的商品或信息, Web2.0 技术的成熟，使得该项研究迅猛发展，用户不再是被动的网页浏览者，而是成为主动参与者。推荐算法的关键是如何提高推荐精度和效率，对此研究者提出了多种改进策略和

算法。目前,个性化推荐方法主要有基于内容过滤的推荐、基于关联规则的推荐、协同过滤推荐、混合推荐系统等[5]。

基于内容过滤[6]的推荐系统是信息过滤技术的延续和发展，具有较为广泛的应用，主要是其具有简单、高效的特点。该算法根据用户已经选择的对象，从推荐对象中选择其他特征相似的对象作为推荐结果。这一推荐算法首先提取推荐对象的内容特征，和用户模型中的用户兴趣爱好匹配，最终把匹配度高的推荐对象作为推荐结果推荐给用户。其缺点在于很难出现新的推荐结果和新用户出现时具有冷启动问题。

基于关联规则[7]的推荐系统，从用户的历史消费数据中挖掘出事务之间一些关联规则，然后根据规则向用户做出推荐。例如，“尿布和啤酒”的故事，在美国一家超市里，把尿布和啤酒摆在一起出售，尿布和啤酒的销售额竟然同时增加了。沃尔玛根据所有门店的交易数据利用数据挖掘的方法进行分析，发现与尿布一起被购买最多的是啤酒，建立起两者的联系。关联规则挖掘可以发现不同商品在销售过程中的相关性，在零售业已经得到了成功的应用。但其规则抽取难、耗时，存在产品同名性问题

协同过滤推荐[8]是应用最广泛、最成功的个性化推荐算法，它于 20 世纪 90 年代开始研究并促进了整个推荐系统研究的繁荣。它主要分为基于用户的协同过滤、基于项目的协同过滤、基于模型的协同过滤。协同过滤的优点在于其克服了基于内容的推荐算法中无法为用户推荐新增兴趣的缺点,并且善于发现用户潜在的但自身暂时未意识到的新兴趣。然而,协同过滤算法仍面临着许多问题要解决,最典型的是冷启动问题、稀疏性问题和可扩展性问题。

各种推荐系统都有它的优缺点,在实际应用中可以根据具体问题进行混合推荐。混合推荐的目的在于通过组合不同的推荐算法，扬长避短，给出更符合用户需求的推荐结果。目前研究和应用最多的混合推荐是把基于内容的推荐和协同过

滤推荐的组合。主要混合思路有两种：推荐结果的混合、推荐算法的混合。尽管从理论上讲可以有很多种推荐组合方式，但在某一具体问题上并不一定见效，混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缺点。

1.2.3 个性化推荐系统面临的问题及挑战

尽管协同过滤推荐算法是最成功的、应用最广泛的推荐系统，但这并不能表示现阶段个性化推荐技术的研究已经进入一个成熟的时期，随着网络技术的大范围普及，海量的用户和商品的信息快速呈现，当前个性化推荐系统面临严峻的问题与挑战[9]：

(1)数据稀疏性问题

数据稀疏性是当前限制协同过滤推荐系统推荐质量不高的首要原因。随着推荐系统规模越来越庞大，用户和商品项目的数量呈指数形式增加，而用户对项目的评价信息却非常稀缺，使得大部分基于关联分析的算法（例如协同过滤）都无法建立起准确的推荐模型，给用户返回准确地推荐结果。一个大型的电子商务网站陈列的资源项目往往达成百上千万,一个积极的用户所给出的评价信息对系统的成百上千万的项目资源来说,简直是杯水车薪，这个本质上由高维数据引起的评价数据稀疏性的问题是协同过滤算法的一个经典问题。

(2)冷启动问题

冷启动问题也是限制协同过滤推荐结果准确性的又一个重要因素。所谓的冷启动就是新的用户(新项目)在他们刚刚进入系统时没有选择过任何物品(没有被任何用户选择过),导致他们可以被利用的行为信息非常罕有。因此,无法建立用户（项目）间的关系，分析他们的特征,算法也就不能对这些孤立的新用户(新项目)产生准确的推荐结果。如果我们能够获得用户（项目）充分的文本信息并据此计算用户（项目）之间的相似性，就可以很好解决冷启动的问题，不幸的是，目前的电子商务推荐系统,每天都不断有大量新的项目加入,各种各样的新产品

涌现,因此无法指望构造大量的用户项目评价信息来避免冷启动问题。因此，怎么样才能将新产品推荐给有可能喜欢上它的用户,这是一个极具意义的挑战,同时提高新产品的销售量。

(3)推荐算法的可扩展性

在协同过滤推荐算法中,全局数值算法能及时利用最新的信息为用户产生相对准确的用户兴趣度预测或进行推荐，但是面对日益增多的用户，数据量的急剧增加,算法的扩展性问题(即适应系统规模不断扩大的问题)成为制约推荐系统实施的重要因素。在一个具体的推荐系统中,用户和产品的信息是动态改变的,新用户和新产品不断地加入，如果每次改变都需要完全重新计算,这个计算量是巨大的。虽然与基于模型的算法相比，全局数值算法节约了为建立模型而花费的训练时间,但是用于识别“最近邻居”算法的计算量随着用户和项的增加而大大增加,对于上百万的数目,通常的算法会遇到严重的扩展性瓶颈问题。该问题解决不好,直接影响着基于协同过滤技术的推荐系统实时向用户提供推荐问题的解决,而推荐系统的实时性越好，精确度越高，该系统才会被用户所接受。

1.3 本文的主要工作和组织结构

本文在对不同个性化推荐系统进行对比研究的基础之上,结合二部分图网络的推荐算法并对其做了改进，把用户和项目抽象为节点，结合用户对项目评分情况赋予每一条用户-项目边不同的权值，另外引进项目的度与权值之和的比值θ,使得推荐结果能够一定程度上抑制热门项商品目的推荐,并在准确性和多样性上有所提高。

本文共分五章，具体安排如下：

第一章　绪论。介绍了本论文选题背景与研究意义,以及国内外推荐系统技术的研究现状,主要是介绍了目前国内外主流的商务推荐系统及发展前景。另外总结了目前推荐系统所面临的问题和挑战。

第二章　个性化推荐系统。主要介绍目前使用最广泛的推荐系统——协同过滤推荐系统，分别阐述了基于用户的协同过滤、基于项目的协同过滤以及基于模型的协同过滤等推荐算法，以及算法的原理、算法的优缺点等。

第三章　基于二部分图网络推荐系统的研究。分析基于二部分图网络结构的推荐算法对比于协同过滤技术存在哪些方面的优点，重点描述了基于二部分图网络的算法的具体实现步骤。

第四章　二部分图网络结构推荐算法的改进。在基于二部分图网络结构推荐系统的基础之上，通过权值参数λ_1、λ_2，及项目度与权值之和的比值θ的引入，改进推荐算法的准确性、多样性及流行性，解决冷启动问题，并通过实验数据验证了改进后的算法是可行的，取得良好的推荐效果。

第五章　总结与展望。主要是概述了整个论文的工作、取得的成绩，以及进一步地研究方向。

第二章　基于协同过滤的推荐技术

2.1 协同过滤技术基本概念原理

协同过滤算法是根据用户对物品或者其他信息的评价，发现目标用户和已知用户对物品的相关性，或者物品本身的相关性，然后基于相关性程度的高低来对用户进行预测推荐。通常所指的协同过滤推荐技术是指基于用户的推荐技术，后来有研究提出了基于项目的协同过滤推荐技术和基于模型的协同过滤技术。下面分别介绍这三种协同过滤技术的原理和机制[10]。

2.2 基于用户的协同过滤技术

该技术算法的实现过程要经过三个步骤：利用用户-项目矩阵获取用户评价信息，计算用户间的相似性并产生相似用户集，及推荐结果的生成。

步骤 1：用户信息获取

在个性化推荐中引入用户-项目矩阵 $R_{i,j}$ 来描述用户对商品项目的评价信息，该矩阵中就包含了 m 个用户对 n 个商品项目的评价信息。

步骤 2：产生相似用户集

可以选择相似度 Sim 值降序排列的前 K 个用户作为目标用户的相似用户。或者将预先设定的相似度阈值作为标准，从这群用户里面选出相似度大于阈值的用户为相似用户群。

关于用户相似度的计算，常用的 2 种途径都是基于向量的相似度计算[11]：

a)余弦相似性

利用行向量 $\vec{i}$ 和 $\vec{j}$ 之间的余弦夹角 $Sim(i, j)$ 来表示两用户间的相似性，其计算公式为：

$$Sim(i, j) = \cos(\vec{i}, \vec{j}) = \frac{\vec{i} \square \vec{j}}{\|\vec{i}\| \times \|\vec{j}\|}$$

b)修正的余弦相似性

修正的余弦相似性是对余弦相似性的一个完善和补充。其计算公式为：

$$Sim(i,j)=\frac{\sum_{c\in I_{ij}}(r_{i,c}-\overline{r_i})(r_{j,c}-\overline{r_j})}{\sqrt{\sum_{c\in I_i}(r_{i,c}-\overline{r_i})^2}\sqrt{\sum_{c\in I_j}(r_{j,c}-\overline{r_j})^2}}$$

公式中，I_i 和 I_j 分别被看成是用户 i 和 j 历史评价过的商品项目的集合，I_{ij} 被看作是用户 i 和 j 共同评价过的商品项目的集合，$\overline{r_i}$ 和 $\overline{r_j}$ 分别被用来表示用户 i 和用户 j 对项目的平均评价值，$r_{i,c}$ 表示用户 i 对商品 c 的评价值。

步骤 3：推荐列表生成

在搜索到相似用户群之后，就可以根据推荐方法推荐结果了，下面介绍 2 种常用的方法：

a) 计算目标用户 $User$ 对相似用户 N_i 购买或评价过的商品项目 $Item_j$ 喜爱程度的预测值，根据预测值的情况生成推荐结果[12]。其计算公式为：

$$\mathrm{Pr}ediction_{User,Item_j}=\frac{\sum_{N_i\in N}Sim(User,N_i)\times(r_{N_i,Item_j})}{\sum_{N_i\in N}(|Sim(User,N_i)|)}$$

公式中，$r_{N_{i,Item_i}}$ 表示相似用户 N_i 对 $Item_j$ 的评价值信息。

b) 第二种推荐方法是预测目标用户对所有未评分项目的评价值，然后根据预测评价值的大小，从中选取预测值较大的前 N 项推荐给用户。只针对项目间的相似性评分，其计算公式为：

$$\mathrm{Pr}ediction_{User,Item_j}=\overline{r_{User}}+\frac{\sum_{N_i\in N}Sim(User,N_i)\times(r_{N_i,Item_j}-\overline{r_{N_i}})}{\sum_{N_i\in N}(|Sim(User,N_i)|)}$$

公式中，$\overline{r_{Ni}}$代表相似用户集中的N_i用户对所有商品项目的平均评分，$\overline{r_{user}}$代表目标用户*User*对所有商品项目的平均评分。

2.3 基于项目的协同过滤技术

基于项目的协同过滤利用用户对商品的偏好，来发现商品和商品之间的相似度，然后根据目标用户对相关物品的历史偏好，将与其偏爱物品相似度高的物品推荐给用户[13]。其实现过程与基于用户类似，将用户与商品进行替换计算。这里不再赘述。下面介绍生成推荐结果，同样介绍两种办法：

a)考虑对商品项目的平均评分情况

计算公式如下：

$$\Pr ediction_{User,i} = \overline{r_i} + \frac{\sum_{j \in N} Sim(i,j) \times (r_{User,j} - \overline{r_j})}{\sum_{j \in N} (|Sim(i,j)|)}$$

计算公式中，$\overline{r_i}$和$\overline{r_j}$分别代表目标项目*i*和相似项目*j*的平均评分值，r_{Userj}代表用户*User*对相似项目*j*的评分值

b)不考虑对商品项目的平均评分情况

用户*User*对目标商品项目*i*的预测评分值，可以根据用户*User*对目标商品项目*i*的相似项目集*N*中任一项目*j*的评分计算出，再可根据预测评分值情况决定是否推荐，计算公式为：

$$\Pr ediction_{User,i} = \frac{\sum_{j \in N} Sim(i,j) \times (r_{User,j})}{\sum_{j \in N} (|Sim(i,j)|)}$$

r_{Userj}代表用户*User*对商品项目*j*的评分值。

2.4 基于模型的协同过滤技术

以用户为基础的协同过滤和以项目为基础的协同过滤统称为以记忆为基础的协同过滤技术。它们存在着数据稀疏性问题，所以研究者提出了以模型为基础

的协同过滤技术。以模型为基础的协同过滤是先利用历史数据，对历史数据进行数据挖掘分析，得到一个模型，当对目标用户做出相应推荐时，利用此模型对推荐结果进行预测。

2.5 协同过滤技术中存在的问题

基于用户的协同过滤技术存在的致命性问题就是数据稀疏性问题。当在前面介绍到用户-项目矩阵数量巨大时，如达到百万数量级。每个用户大多数情况只会对极有限的商品（低于 1%）进行评价，或者只有极少数用户参与评价工作，这就使得用户-项目矩阵的数据非常稀疏，以致于很难准确建立相似用户群和一些隐藏的用户间信息得不到发现，从而严重导致了推荐系统质量急剧下降。

基于项目提出的初衷是为了解决数据稀疏，另外，商品之间的联系相对稳定，即相似度计算可以离线进行，这为系统的可扩展性和推荐精度提供了良好的条件。但是，只根据项目之间的相似度来推荐，无法为用户提供新类别商品的推荐。

2.6 本章小结

本章首先讨论现金广泛使用的个性化推荐系统技术，对推荐系统的发展趋势和研究有了较好的了解和把握。其次，本章着重讨论了当前运用最广泛的协同过滤技术，包括协同过滤技术实现的具体步骤，分类，还讨论了该技术当前存在的问题和可能的解决办法，从而为下一步研究个性化推荐系统算法奠定了基础。

第三章 二部分图网络推荐系统的研究

随着个性化推荐技术的发展,关注和研究该技术的人们也越来越多,大家都试着去寻找更高效的推荐算法。最近，基于用户-项目二部分图网络结构的推荐算法得到了研究者的关注,该算法是对原有的协同过滤和基于内容推荐技术的完善和发展，利用二部分图上的物质扩散[14]、热传导[15]等复杂网络动力学过程来对用户进行个性化推荐。

3.1 二部分图网络推荐技术概述

基于二部分图网络的推荐算法与个性化推荐系统中常用的协同过滤的推荐技术相比,其区别主要体现在忽略用户和商品项目的实际物理属性，对海量的用户和商品项目以及彼此之间的关联进行抽象和规约成网络拓扑的表现形式，采取从统计的角度出发研究网络中节点及其彼此连接间存在的关联性质[16]。前者在解决算法复杂度及数据稀疏性问题上都有明显的效果。以下是两者在算法复杂度和数据稀疏性的比较[9]：

(1)算法复杂度上。协同过滤推荐技术在为目标用户寻找相似邻居时,是通过比较他们使用过的项目之间的相似性,按照相似度的大小从而确定邻居用户。在这个过程中,选用合适的计算相似度指标是一个非常重要的过程,而且在计算相似度时,要通过提取项目的特征进行计算,牵涉到人工智能,数据挖掘及智能算法方面的知识,整个计算过程的复杂度很大。但是基于用户项目的二部分网络结构图算法,不是用传统的相似性度量方法去计算,而是通过用户和项目之间的这种通过边连接的节点信息的传递来衡量相关用户之间的相似性。这在计算量上大大地减小了算法的复杂度,而且也很容易编程实现。

(2)冷启动。传统的冷启动问题一直困扰着协同过滤技术的发展,但是基于用户-项目的二部分网络结构图算法可以在一定程度上解决这个问题。因为可以通过网络节点和边的引入,我们可以很方便的添加一些难以分析的资源信息,可以

通过系统初始化时,对冷启动问题采用各种赋初始值的办法进行解决，对于那些新项目可以进行基于内容的标签推荐,对于新用户可以采用初始推荐系统流行项目的办法解决。

3.2 基本二部分图网络推荐算法的实现

3.2.1 二部分图网络的系统模型

在复杂网络研究中，把网络视为描述现实世界中对象之间所存在的某种关系的一种数学模型。每个对象被抽象成节点来表示，而对象之间的关系则是通过存在于属于不同集合的两个节点的连边来表示。节点属于同一种类型的网络，称为单模式网络，节点不属于同一种类型的网络，即称为二部分图网络。在二部分图网络中，同一种类型的节点彼此不能连接，不同类型的节点之间才能连接。

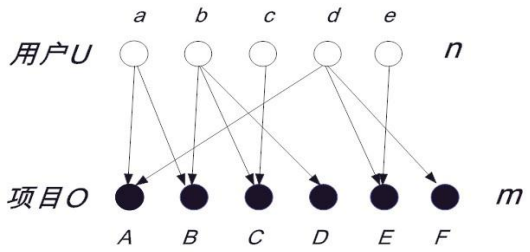


图1　二部分图网络示意图

例如：用户和他们所观看的电影就构成一个用户-电影二部分网络图，如果某位用户观看了某部电影，则用户和电影这两个不同类型的节点便会存在一条连接，从而将用户电影之间联系了起来。这就是所谓的二部分图网络。

3.2.2 算法推荐过程

（1）用户-项目二部分图的构造

对于二部分图网络推荐算法,基本思想是关于资源分配的问题,其推荐处理方法视为是一个资源分配的过程，所以该方法也称为资源的动力扩散算法。基于用户项目相关性的二部分网络结构图的推荐算法,不考虑用户和项目的内容特征,而仅仅把它们抽象成节点，并把每个用户所选择过的项目与用户之间用边相连接,

这样就构成了用户和项目节点之间的关系图，而算法利用的信息都隐藏在用户和项目之间的选择关系之中，通过寻找这种隐藏的信息便可实现对项目的推荐。

由n个用户和m个项目构成的二部分图网络的推荐系统中，定义用户集$U=\{u_1,u_2,u_3,...,u_n\}$，项目集$O=\{o_1,o_2,o_3,...,o_n\}$，因此可以用$(n+m)$个节点表示系统。若用户$u_a$选择过项目$o_i$，则$a_{ai}=1$，否则$a_{ai}=0$，从而得到用户与项目构成的邻接矩阵$A$。

（2）项目的资源配额w_{ij}

周涛等[14]根据用户-项目二部分图提出了一种基于资源分配的算法：假设用户选择过的项目都具有某种向该用户推荐其他产品的能力，这个抽象的能力看作位于相关产品上的某种可分的资源配额。

假设项目o_i具有的初始资源$f(o_i)\geq 0$，当系统开始进行资源分配时，项目节点的资源流向用户节点，其中用户u_a所获的资源：

$$f(u_a)=\sum_{i=1}^{m}\frac{a_{ai}f(o_i)}{k(o_i)} \tag{1}$$

其中：$k(o_i)=\sum_{a=1}^{n}a_{ai}$，表示项目$o_i$的度（该项目被多少个用户选择过）。

同样，当用户节点的资源流向项目节点时，项目节点o_i的资源为：

$$f'(o_i)=\sum_{a=1}^{n}\frac{a_{ai}f(u_a)}{k(u_a)} \tag{2}$$

其中：$k(u_a)=\sum_{i=1}^{m}a_{ai}$，表示用户$u_a$的度（该用户选择过多少项目）。

则联立公式（1）、（2），可得：

$$f'(o_i)=\sum_{a=1}^{n}\frac{a_{ai}f(u_a)}{k(u_a)}=\sum_{a=1}^{n}\frac{a_{ai}}{k(u_a)}\sum_{j=1}^{m}\frac{a_{aj}f(o_j)}{k(o_j)} \tag{3}$$

令

$$w_{ij} = \frac{1}{k(o_j)} \sum_{a=1}^{n} \frac{a_{ai} a_{aj}}{k(u_a)} \tag{4}$$

可得：

$$f'(o_i) = w_{ij} f(o_j) \tag{5}$$

公式（5）揭示两个不同项目节点之间的资源联系，w_{ij} 为项目 o_j 愿意分配给项目 o_i 的资源配额，也就是两个项目之间的相关程度。

(3)用户相似度的计算

在计算用户 u_a 和 u_b 的相关程度，即相似度，我们可以参考不同项目之间相似度 w_{ij} 的计算方法，即公式（4），从而可以定义两个用户的相似度：

$$s_{ab} = \frac{1}{k(u_b)} \sum_{i=1}^{m} \frac{a_a a_b}{k(o_i)} \tag{6}$$

其中：$k(u_b) = \sum_{i=1}^{m} a_{bi}$，表示用户节点 u_b 的度（即该用户选择过多少项目）；

$k(o_i) = \sum_{l=1}^{n} a_{li}$，表示项目节点 o_i 的度（即该项目被多少用户选择过）。

(5)生成基于用户相似度的项目推荐列表

根据用户相似度，可以预测用户对于未选择过的项目的偏好程度，即预测用户对某项目的评分。评分越高，则表示用户对该项目的偏好程度越高，项目被推荐的程度也就越高。用户 u_a 对未选择过的项目 o_i 的预测评分为 v_{ai}：

$$v_{ai} = \frac{\sum_{b=1, b \neq a}^{n} s_{ab} \times a_{bi}}{\sum_{b=1, b \neq a}^{n} s_{ab}} \tag{7}$$

推荐系统对用户未选择过的项目分别利用公式（7）进行评分预测后，就可以根据评分值对项目进行降序排列，生成 $top-N$ 个项目推荐列表。

3.3 本章小结

本章主要介绍二部分图网络结构相比较于传统的协同过滤算法在处理数据稀疏性和复杂度等经典问题上有一定的优越性。重点介绍了基于二部分图网络结构的推荐算法的原理，系统模型的生成以及算法的具体步骤。

第四章　二部分图网络结构的推荐算法的改进

王茜[17]等人针对基于二部分图网络结构的推荐算法做了改进，考虑了数据集中用户的评分特性，提出用带有评分值的有权二部分图替代无权二部分图，并引入项目的度和权值之和的比值θ，从而提高算法的推荐性能。

4.1 项目评分多等级划分

在对评分矩阵的进行数据预处理时，同样区分项目评分的高低，并按照评分等级分别赋予每一条用户-项目边不同的权值w，但本文对跟原文却有所不同，本文所提出的算法把项目评分划分成更多个等级，分别是：当u_i选择过项目o_j，且评分为5（非常喜欢），则$w_{ij}=1$；评分为3或4（$0.5<\lambda_2<1$）（喜欢），则$w_{ij}=\lambda_2$；评分为1或2，则$w_{ij}=\lambda_1$（$0<\lambda_1<0.5$）（基本不喜欢）；当用户未选择过o_j，则$w_{ij}=0$。评分多等级的划分，能够更精确地考虑了用户对项目的喜欢程度，把3、4与5分项目的喜欢程度区分出来，使得推荐结果能够更进一步地、精确地符合用户的兴趣需求，提高推荐结果的准确性。

项目评分多等级划分后，基于用户相似度的计算公式：

$$s_{ab}=\frac{1}{d(u_b)}\sum_{i=1}^{m}\frac{w_{ai}w_{bi}}{d(o_i)} \tag{8}$$

其中：$d(u_b)=\sum_{i=1}^{m}w_{bi}$，表示用户$u_b$所有连边的权值之和；$d(o_i)=\sum_{l=1}^{n}w_{li}$，表示项目$o_i$所有连边的权值之和；$w=0,\lambda_1,\lambda_2$或1。

而用户u_a对未选择过的项目o_i的预测评分为v_{ai}：

$$v_{ai}=\frac{\sum_{b=1,b\neq a}^{n}s_{ab}\times w_{bi}}{\sum_{b=1,b\neq a}^{n}s_{ab}} \tag{9}$$

4.2 引入项目的度和权值之和的比值θ

基于二部分网络结构的算法与传统的协同过滤算法相比,虽然精度有所提高,但是依然受到新用户,新项目等问题的制约(冷启动问题)。另外前面所述推荐算法比较倾向推荐热门商品,这是因为经过资源配置后所得的资源量最多的电影仍是那些热门电影,但是热门商品的推荐可以通过一些更为简单地方式去实现,因此推荐算法给用户推荐冷门商品比推荐热门商品将显得更加有意义。

如果两个用户同时选择一个度较大的项目(热门商品),这不能反映这两个用户的兴趣有较大的相似性;相反,如果两个用户同时选择一个度较小的项目(冷门商品),这更有可能反映这两个用户兴趣存在较大的相似性,所以在改进的推荐算法中应该关注项目的度,提高小度项目的推荐能力,而适当降低大度项目的推荐能力。

另外对于两个相同度的项目o_i和o_j,如果大多数选择过o_i的用户对其评分都较低,即项目o_i的权值之和比较低,说明选择过它的用户基本不怎么喜欢该项目;相反,项目o_j的评分基本都较高,即项目o_j的权值之和相对较高,说明选择过它的用户基本都喜欢它。因此推荐算法应该降低这种选择过它的用户基本不怎么喜欢的项目的推荐能力,而提高用户大多数都喜欢的项目的推荐能力。

考虑冷启动问题,以及项目度的大小、权值之和对推荐结果的意义、准确性的影响,定义了新的参数θ,$\theta(o_i)=\frac{k(o_i)}{d(o_i)}$,即项目$o_i$的度与权值之和的比值。当选择过项目$o_i$的用户都喜欢该项目时,$\theta(o_i)$=1,取得最小值;当选择过项目$o_i$的用户基本都不喜欢该项目时,$\theta(o_i)>1$。也就是说,当$\theta(o_i)$越大,项目越不受用户喜欢,因此算法中就应该减弱其推荐能力。

考虑了θ能够准确地反映项目受喜欢的程度,引进了一个新的函数:

$$f(\theta)=\delta+e^{0.5(\theta-1)} \qquad (10)$$

引进函数$f(\theta)$后,用户之间相似度的计算公式最终为:

$$s_{ab}=\frac{1}{d(u_b)}\sum_{i=1}^{m}\frac{w_{ai}w_{bi}}{d^{f(\theta)}(o_i)} \tag{11}$$

用户u_a对未选择过得项目o_i的最终预测评分为v_{ai}同公式（9）一样：

$$v_{ai}=\frac{\sum_{b=1,b\neq a}^{n}s_{ab}\times w_{bi}}{\sum_{b=1,b\neq a}^{n}s_{ab}} \tag{12}$$

公式（10）中：δ为可调参数。当δ>0时，大度项目的推荐能力会被压制；当δ<0时，大度项目的推荐能力会得到提高。因此，通过调整δ的大小，商家能够结合自己的意愿并根据用户实际消费情况以及热门商品项目的盈利情况选择是否为用户推荐较为热门的商品，还是推荐用户感兴趣的、但却不热门的商品。因此该算法不仅解决了推荐冷门商品项目的问题，同时也给商家更多的自由度，不仅在推荐商品给用户的时候体现了个性化，同时也为商家提供了个性化的服务，这正是给电子商务推荐系统追求个性化提供了一个较好的支持。

4.3 算法的步骤、流程图

4.3.1 算法的详细步骤

改进的基于加权二部分图网络结构的推荐算法的详细步骤如下：

输入：用户和项目的评分矩阵R，目标用户u_a；

输出：目标用户u_a的推荐列表。

（a）根据用户和项目的评分矩阵构造用户-项目的权值矩阵。其中边的权值w通过用户对项目的评分而确定，当评分为5，则w=1；评分为3或4，则$w=\lambda_2$；评分为1或2，则$w=\lambda_1$；当用户未选择过o_i，则$w=0$。以此来生成用户-项目的权值矩阵。

（b）调节参数δ的值，以调节项目度对推荐质量的影响，最终确定函数$f(\theta)$的形式。

（c）根据步骤（b）确定δ的值和函数$f(\theta)$的形式后，利用用户相似性计算公式（11），计算用户u_a和其他用户u_b之间的相似性s_{ab}。

（d）利用式（12）计算目标用户u_a对未评分项目o_i的预测评分v_{ai}。

（e）将评分值进行降序排列，并把列表中的评分最高的N个项目推荐给目标用户u_a，完成推荐。

4.3.2 算法的流程图

算法的流程图如图 2 所示：

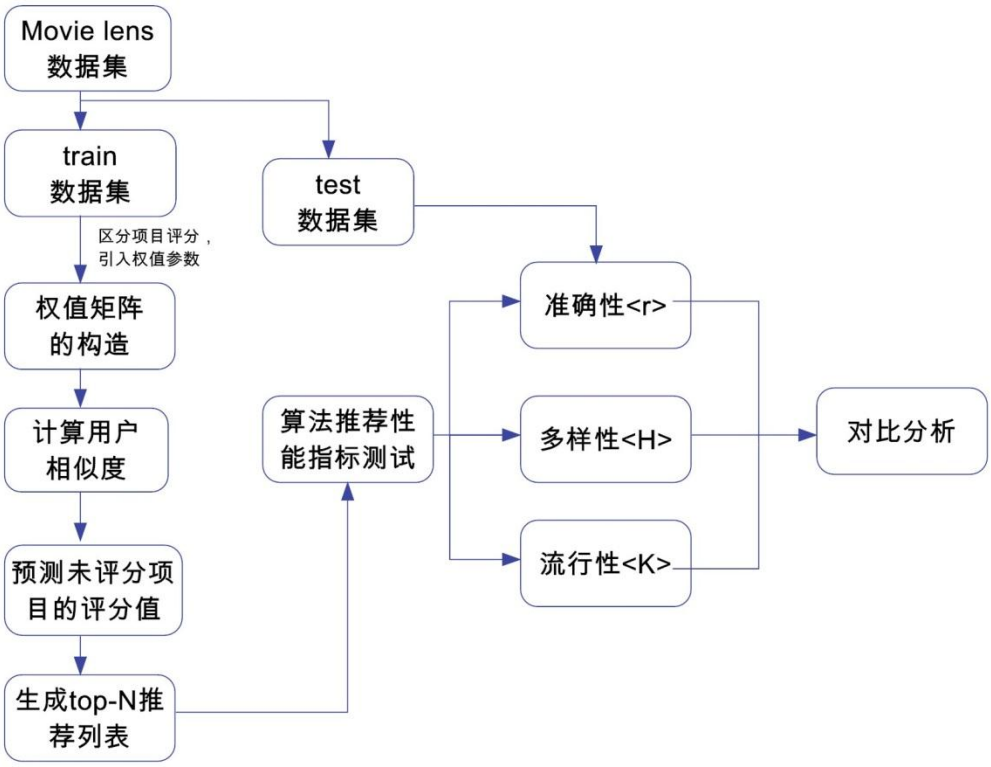


图 2　　算法流程图

4.4 实验环境、数据的选取和算法评价指标

4.4.1 实验环境

为了验证本文中提出的的改进型的基于二部分图网络推荐算法其可行性和有效性，在 MATLAB 环境中进行了实验，并针对推荐准确性、多样性和流行性等方面进行了相关测试和分析。实验机器的主要配置参数为：CPU Intel(R) Core(TM) i3-2350M 2.30GHz；内存 4GB；操作系统 Windows 7。

4.4.2 实验数据选取

实验采用当前广泛使用的 Movielens 数据集，该数据集来自于明尼苏大学的 GroupLens 小组。MovieLens 数据集包括三个数据集，本文采用其中最小的数据集，共包含了 943 名用户对 1683 部电影的 10 万条评分记录，评分值从 1 到 5，评分值越高代表用户越喜欢该部电影。我们采用数据集提供方划分的测试集及训练集。训练集用于构造用户关联网络，生成推荐列表，测试集用于验证推荐系统准确性。

在实验中，我们将训练集存放于 test.mat 文件中，测试集存放于 train.mat 文件中。

4.4.3 算法评价指标

我们用平均排名分数（rank score）来衡量算法推荐算法准确性。此外，用汉明距离（Hamming distance）评价多样性，用推荐项目的平均度（K）评价流行性。

(1)平均排名位置$<r>$

系统的准确性可以用项目平均排名位置来衡量[18]。针对用户u_i，推荐算法会给其一个长度为L的推荐列表。根据测试集，如果用户u_i选择了项目o_j，而o_j在推荐列表中的位置为R_i，则认为项目o_j的相对位置为

$$r_{i,j} = \frac{R_{i,j}}{L}$$

因为测试集中的项目是用户实际选择过的，所以准确度越高的算法，测试集中的项目在其推荐列表中应占据越靠前的位置，即$r_{i,j}$越小。将测试集中所有用

户-项目数据的相对位置求平均，求得平均值<r>，即平均排名位置，<r>越小算法准确性越好。

(2)多样性H

对于系统的多样性，可以用平均汉明距离来衡量[19]。对于任意两个用户u_i和u_j，其推荐列表的汉明距离H定义为

$$H_{i,j} = 1 - \frac{Q_{i,j}}{L}$$

其中，L表示推荐列表长度；$Q_{i,j}$代表用户u_i和u_j长度为L的推荐列表中相同的项目数目，不分位置次序。计算出任意两个用户之间的汉明距离，然后计算其平均值，用该平均值H来衡量算法的多样性。如果H为1，此时其值最大，表示所有用户项目推荐列表全部相同，推荐多样性最好，即个性化程度最好；如果H为0时，表示用户推荐列表完全相同，此时多样性最差。

（3）流行性K

用推荐列表中的L个项目的平均度K来评价算法所推荐项目的流行性[20]。平均度越小，说明不是非常流行的项目也能被推荐。

4.5 实验结果及分析

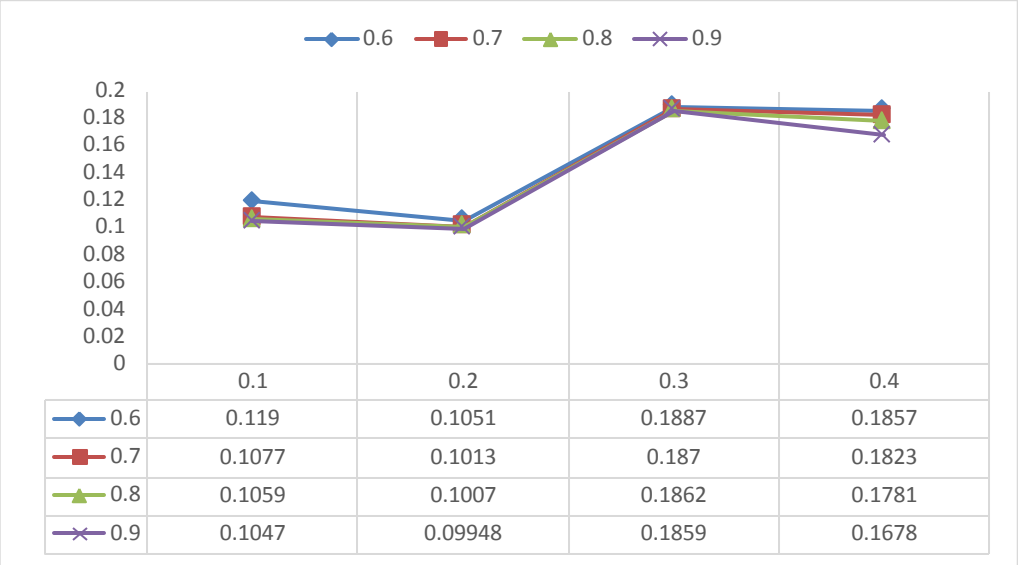
本文实验分设了十六组，分别对应不同的参数λ_1和λ_2，取值如下表所示：

λ_1	0.1	0.2	0.3	0.4
λ_2	0.6	0.7	0.8	0.9

实验中，当λ_1取0.1时，λ_2分别取0.6、0.7、0.8、0.9进行推荐计算，以此类推，一共进行了十六次实验，再对实验中推荐算法给出的推荐结果进行准确性、多样性和流行性等指标的验证。综合各次实验结果，可得：

（1）准确性<r>的分析：

图3显示了在不同参数λ_1、λ_2下，平均排名位置<r>的值：



	0.1	0.2	0.3	0.4
0.6	0.119	0.1051	0.1887	0.1857
0.7	0.1077	0.1013	0.187	0.1823
0.8	0.1059	0.1007	0.1862	0.1781
0.9	0.1047	0.09948	0.1859	0.1678

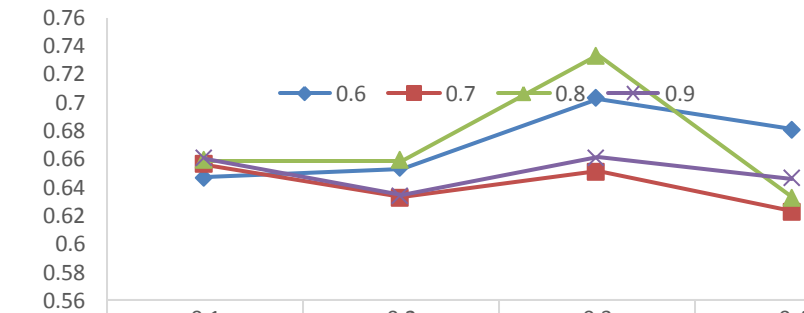
图 3　参数 λ_1、λ_2 对平均排名位置 $<r>$ 的影响

由上图可见，推荐结果的准确性主要受到 λ_1 的影响，即用户低评价分数电影的影响。当 λ_1 取 0.1、0.2 时，系统能得到较好的准确度。当 λ_1 取 0.3 以上时，曲线出现一个跳跃，系统的准确性下降。这大致符合本文最初的猜想，通过引入权值参数来调整系统的准确性。但从图中也可以看出，参数 λ_2 对系统的准确性影响不明显。综上所述，可以得出结论，当 λ_1=0.2，λ_2 =0.9 时，推荐算法的准确性最优，此时 $<r>$=0.0995。

通过查找其他文献中的资料，得到：基于 Pearson 系数的协同过滤推荐算法的平均相对位置< r >=0.120，全局排序算法(GRM)的平均相对位置值< r >=0.136，而基本二部分图网络推荐算法的平均相对位置值< r>=0.106，而本文所提出的基于二部分图网络推荐算法的改进算法，在准确性上有所突破，其平均相对位置值<r>=0.0995,且对比于未改进的算法提高了 6.13%，这验证了本文提出的提高推荐结果准确性的猜想是正确的。

（2）多样性 H 的分析：

图 4 显示了在不同参数 λ_1、λ_2 下，汉明距离 H 的值：



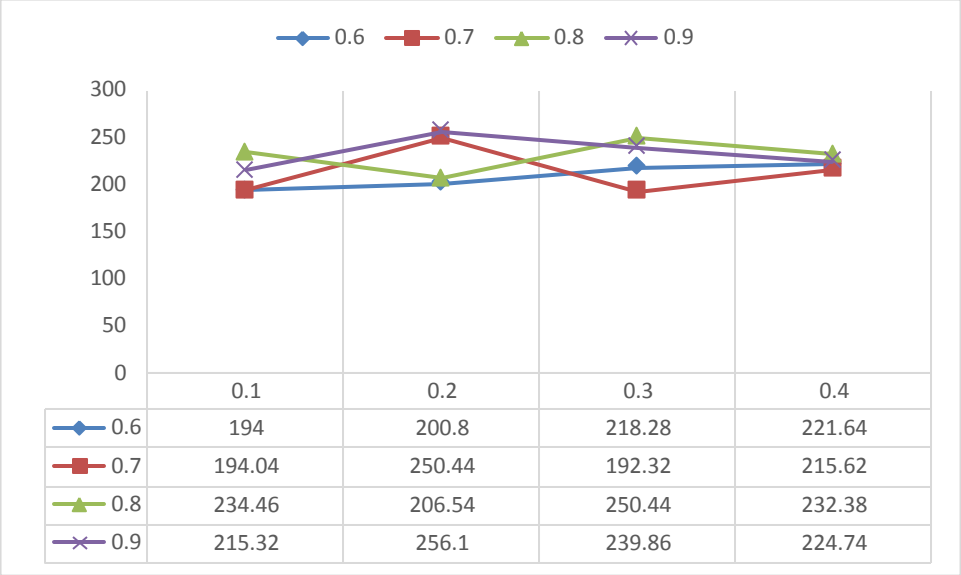
	0.1	0.2	0.3	0.4
0.6	0.647	0.653	0.703	0.681
0.7	0.656	0.633	0.651	0.623
0.8	0.659	0.659	0.733	0.633
0.9	0.66	0.634	0.661	0.646

图 4　　参数 λ_1、λ_2 对汉明距离 H 的影响

从图 4 中可以观察得到，当 λ_1=0.1、0.2 时，λ_2 对系统多样性的影响不大，H 基本相等且较低，即此时系统多样性较差；当 λ_1=0.3、0.4 时，λ_2 对系统多样性有较大影响，且随着 λ_2 的升高，H 的值减小，系统多样性下降；另外由图 4 发现，当 λ_1 取 0.3 的时候，折线图存在一个峰顶。最后可以得出结论，当 λ_1 为 0.3，λ_2 =0.8 时，系统的多样性取得最优，此时 $<H>$=0.733。

通过查找其他文献中的资料，将本文改进算法得到的结果与 NBI 和 SA-CF 两个算法在流行性上做比较。当推荐列表长度 L =50 时，NBI 算法的多样性 $<H>$ =0.54，基于 SA-CF 算法的多样性 $<H>$=0.635，而本文提出的改进算法多样性达到最优时，$<H>$=0.733。这验证了改进算法提出的提高算法推荐结果多样性的性能指标的这一猜想是正确。

（3）流行性 K 的分析：



	0.1	0.2	0.3	0.4
0.6	194	200.8	218.28	221.64
0.7	194.04	250.44	192.32	215.62
0.8	234.46	206.54	250.44	232.38
0.9	215.32	256.1	239.86	224.74

图 5　参数 λ_1、λ_2 对平均度 $<K>$ 的影响

通过查找其他文献中的资料，将本文改进算法得到的结果与 NBI 和 SA-CF 两个算法在流行性上做比较，当推荐列表长度 L =50 时，NBI 的平均度 K 约为 230，SA-CF 算法的平均度约 K 为 240，而本文算法那通过权值系数 λ_1、λ_2 调整，可将平均度 K 的值控制在 220 以内，在保证准确性的同时，也能够降低推荐项目的流行性。这说明本改进算法提出的抑制热门项目推荐能力，提高冷门商品项目推荐能力的猜想验证正确，有利于解决传统推荐系统面临的冷启动问题。

4.6 本章小结

本章着重研究了基于二部分图网络结构的推荐算法的改进，分析了改进算法的具体思想，分别从区分项目评分的高低、引入项目的度与权值之和的比值θ等方面来阐述算法的改进内容，并且通过实验，验证了改进的二部分图网络推荐算法在提高系统推荐的准确性、多样性及提高冷门商品项目推荐能力等方面有所改善。

第五章　总结与展望

基于网络结构的推荐是一种较新的推荐算法，由于该算法中不用考虑用户和推荐对象的内容没有区分用户对项目评分的高低以及大多数推荐算法都存在缺乏推荐冷门商品功能的问题，本文展现了一种改进的基于二分图网络结构的推荐算法。本文通过分段赋予权值的方法合理地区分各个评分所具的价值，构成加权网络。本文在计算用户相似度时综合考虑项目的度和项目权值之和的比值对推荐系统的影响并引入可调参数δ，在提高推荐质量的同时具备推荐相对冷门商品的能力。通过系统分组实验分析发现，适当降低低分评分的权值，削弱其在推荐系统中的作用，能显著提高推荐结果的准确性，虽然系统的准确性和多样性无法同时满足最优，但是合理分配评分权值，能得到较高的准确性、多样性。同样，选择合适的参数，推荐系统也能推荐较为冷门的商品，这在一定程度上解决了冷启动问题。

本文算法在一些性能指标上具备一定的优越性，但仍有许多需要改进的地方，我们将继续对以下几个方面进行探讨研究：

（1） 推荐算法的可扩展性问题。在大量增加了用户和项目的节点数后，推荐算法的准确性和多样性是否能够保持，需要我们进一步试验探索。

（2） 系统性能最优问题。如何通过设定λ_1、λ_2合理地给用户-项目边赋权值，找到系统性能最优的工作点，是我们下阶段探索方向之一。

（3） 推荐系统的实时响应速度问题。在使算法推荐准确率和多样性保持在高水平的同时，提高系统的响应速度，也是我们下阶段的努力方向之一。

（4） 基于网络结构的推荐算法与协同过滤推荐算法存在契合点，下阶段会尝试结合两种算法进行混合推荐，提高推荐质量。

参考文献

[1] 黎星星,黄小琴.朱庆生.《电子商务推荐系统研究》[J].《计算机工程与科学》,2004年第26卷第5期.

[2] 邓爱林等,《基于项目评分预测的协同过滤推荐算法》[J].《软件学报》,2003,第14卷第9期,1621一1628页.

[3] 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年,第39卷第8期,986-989页.

[4] 刘建国,周涛,汪秉宏.《个性化推荐系统的研究进展》[J].自然科学进展;2009,19(1):1-15.

[5] 王国霞,刘贺平.《个性化推荐系统综述》[J].计算机工程及应用;2012,48(7):66-76.

[6] BALABANOVC M,SHOHAM Y.Fab: content-based collaborative recommendation[J]. Communications Of the ACM ,1997,40(3):66-72.

[7] 赵超. 数据挖掘关联规则的研究[J]. 网友世界, 2012 (3): 43-45.

[8] 郭艳红.推荐系统的协同过滤算法与应用研究7D],大连理工大学,2008年6月

[9] 刘友林. 基于网络结构的个性化推荐系统的研究[D]. 东华大学, 2012.

[10] 赵帆. 基于复杂网络数据挖掘的个性化电子商务推荐系统研究[D]. 中南民族大学,2011..

[11] 何安.协同过滤技术在电子商务推荐系统中的应用研究[D]. 浙江大学,2007:17-19.

[12] Schafer,J.B.,Konstan,J.A, and Riedl,J. E-Commerce Recommendation Applications [J]. Data Mining and Knowledge Discovery, 2001, 5(1-2):115-153.

[13] 陈玲.协同过滤推荐算法的研究[D].中山大学,2007:13-16.

[14] ZHOU Tao，REN jie,MEDO M,et al.Bipartite network projection and Personal recommendation [J].Physical Review E,2007,76(4);046115.

[15] SHANG Ming-sheng, LV Lin-yuan, ZHANG Yi-cheng, et al. Empirical analysis of Web-based user-object bipartite networks [J]. Euro physics Letters, 2010,90(4:48006)

[16] 李德毅, 刘常昱, 杜益, 韩旭. 不确定性人工智能[J]. 软件学报, 2004, 15(11):

[17] 王茜, 段双艳. 一种改进的基于二部图网络结构的推荐算法[J]. 计算机应用研究, 2013, 30(3): 771-774.

[18] M.S. Shang, L. Y. Lu, Y.C .Zhang and T. Zhou, Relevance is more significant than correlation: Information filtering on sparse data, Euro physics Letters, 2009,27(8):68-83

[19] J, Ren, T. Zhou and Y. C. Zhang, Information filtering via self-consistent refinement, Euro physicsLetters,2008,82(13):58-67.

[20] 刘建国, 周涛, 郭强, 等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学, 2009, 6(3): 1-10.

[21] 全佳妮. 基于二分网络的协同推荐研究[D]. 苏州大学, 2012.