

优化稀疏数据集提高协同过滤推荐系统质量的方法

刘庆鹏, 陈明锐*

(海南大学 信息科学技术学院, 海口 570228)

(* 通信作者电子邮箱 mrchen@hainu.edu.cn)

摘要:协同过滤是目前个性化推荐系统中效果较好的一种推荐技术。由于用户和项目数量的急剧增加,使得反映用户喜好信息的评分矩阵非常稀疏,严重影响了协同过滤技术的推荐质量。针对这一问题提出了综合均值优化填充方法,该方法相比较于缺省值法和众数法,考虑到了用户评分尺度问题,同时也不存在众数法中的“多众数”和“无众数”问题。在同一数据集上,通过使用传统的基于用户的协同过滤算法进行验证,表明此方法可以有效提高推荐系统的推荐质量。

关键词:推荐系统; 协同过滤; 均值; 众数; 信息过载

中图分类号: TP301.6; TP391; TP311 **文献标志码:** A

Optimization of sparse data sets to improve quality of collaborative filtering systems

LIU Qing-peng, CHEN Ming-rui*

(College of Information Science and Technology, Hainan University, Haikou Hainan 570228, China)

Abstract: Currently, the collaborative filtering is one of the successful and better personalized recommendation technologies that have been applied to the personalized recommendation systems. As the number of users and items increase dramatically, the score matrix which reflects the users' preference information is very sparse. The sparse matrix seriously affects the recommendation quality of collaborative filtering. To solve this problem, this paper presented a comprehensive mean optimal filling method. Compared to the default method and the mode method, this method has two advantages. First, the method takes account of user rating scale issues. Second, the method does not have the "multiple mode" and the "no mode" problems. On the same data set, using traditional user-based collaborative filtering to test the effectiveness of the method, and the results prove that the new method can improve the recommendation quality of recommendation systems.

Key words: recommendation system; collaborative filtering; mean value; mode; information overload

0 引言

由于“信息过载”^[1],人们在享受足不出户购买商品乐趣和方便的同时,也被电子商务网站成千上万的商品信息所困扰,在电子商务网站寻找自己想要购买的商品已经不是一件简单的事情,这也成为困扰电子商务发展的一大难题。如何留住客户,如何为每个客户提供个性化的商品信息,电子商务个性化推荐系统应运而生。但是目前推荐系统的推荐精度不高,可信度较差,很难满足现代电子商务发展的需求。导致推荐精度低、可信度差等问题的原因之一是应用于推荐算法的数据集过于稀疏。

在一般的大型电子商务网站上,商品信息成千上万,但每个用户真正购买的商品却很少,而用户对已购买商品的评价就更少,通常在1%以下^{[2]285-286}。为此本文针对数据集的稀疏问题提出了一种数据预处理方法,增加了数据集的稠密度,提高了推荐算法的推荐质量。

1 电子商务推荐系统及其主要推荐技术

电子商务推荐系统的出现为解决“信息过载”问题提供了一种方案^{[3]158},为电子商务网站实现“一对一营销”^[4]的战略提供了技术支持。电子商务推荐系统的正式定义^[5]为:

“它是依据电子商务网站向客户提供商品信息,帮助用户决定应该购买什么商品,模拟销售人员帮助客户完成购买过程”。在电子商务网站中推荐系统被当作虚拟店员向用户提供商品信息,根据用户的兴趣爱好帮助用户找到其感兴趣的商品信息和服务,同时也提高了网站的销售额,因此逐渐成为电子商务网站的一种重要工具^{[2]288}。

电子商务推荐系统的最大特点是能够收集用户感兴趣的资料,并根据用户的兴趣偏好提供个性化的服务。也就是说当商品信息和用户的兴趣资料发生变化时,推荐系统给出的推荐结果也会随之发生变化,这样大大方便了用户对商品信息的浏览,同时也提高了企业的服务水平。总之,电子商务推荐系统的作用主要体现在以下几个方面:1)将电子商务网站的浏览者转变为购买者;2)提高电子商务网站的交叉销售能力;3)提高客户对电子商务网站的忠诚度^[6]。

整个电子商务推荐系统由三个部分组成:输入模块(Input Module)、推荐方法(Recommendation Method)和输出模块(Output Module)。输入模块用于接收用户的兴趣偏好信息,包括显式兴趣信息和隐式兴趣信息。输出模块将按照输入的用户偏好信息计算出的结果推荐给用户,推荐的形式包括Top-N推荐和预测评分。推荐方法模块是推荐系统的核心部分,决定着推荐系统的优劣,它的主要功能是按照推荐算法

收稿日期:2011-09-14;修回日期:2011-12-19。 基金项目:海南慧人公司项目(HNHR2011-1)。

作者简介:刘庆鹏(1986-),男,山东临沂人,硕士研究生,主要研究方向:软件工程; 陈明锐(1960-),男,海南澄迈人,教授,主要研究方向:软件工程。

计算推荐结果。

电子商务网站为了满足用户个性化需求的推荐技术主要有:信息检索技术(Information Retrieval)、关联规则(Association Rule)、基于内容的过滤(Content-Based Filtering, CBF)和协同过滤(Collaborative Filtering, CF)。信息检索是基于“人找信息”的被动推荐模式,其他都是基于“信息找人”的主动推荐模式。下面从自动化程度、持久性程度以及个性化程度三个方面,比较各种推荐技术的优缺点。自动化程度(Degree of Automation)^{[3]160}从完全自动化到完全手工,取决于用户为得到推荐结果是否需要显示的输入信息,以及输入信息的多少等。持久性程度(Degree of Persistence)^{[3]161}从暂时性推荐到永久性推荐,暂时性推荐全部基于用户的单一会话,而不基于用户先前的任何会话信息。永久性推荐则是基于用户先前的多个会话来决定用户的喜好和厌恶等兴趣信息来进行推荐。个性化程度(Degree of Personalization)^[4]反映了算法的推荐结果符合用户兴趣喜好的程度。表1列出来主要推荐技术的比较。

表1 主要推荐技术比较

评价指标	信息检索	关联规则	基于内容	协同过滤
自动化程度	低	高	高	高
持久性程度	低	低	高	高
个性化程度	低	低	高	高
推荐模式	被动	主动	主动	主动
用户是否参与	参与	不参与	参与	参与
是否有新颖推荐	没有	有	没有	有
缺点	查准率低	规则质量难以保证	受推荐对象限制	数据稀疏

2 协同过滤算法

基于内容的过滤技术首先为每个用户建立用户描述,然后将用户描述与项目的内容进行比较,将最相似的项目推荐给用户,但是基于内容的推荐技术存在以下两个主要缺点,限制了该技术的推广和应用^[7]。1)有限的内容分析(Limited Content Analysis)。在提取被推荐对象的特征值进行对象描述时,要求对象具有良好的结构性,所以基于内容的过滤技术主要用于文本的过滤,对于音乐、视频、图像等非结构化的、难以对其提取特征值的对象无法进行推荐。即使是对于文本信息,基于内容的推荐也只能描述出对象的内容信息特征而无法去判断出对象资源的质量特征。2)不能够向用户提供新颖的推荐^[8],即发现除了用户描述信息所表现出来的用户兴趣之外的用户潜在兴趣并推荐相关的产品。基于内容的推荐往往存在“受描述所限”(Circumscribed by the Profile)^[9]和“过度专门化”(Over-Specialization)^[10]的现象,即基于内容的推荐系统根据用户的描述信息和对象信息向用户做出的推荐往往被限制在与用户以往熟悉内容相似的项目上,不能够发掘用户潜在的兴趣并将推荐推广到更广的范围。而在电子商务领域发现用户的潜在兴趣并向用户推荐其潜在需要的产品是促进销售和提高利润的重要途径。

为了克服基于内容的推荐所存在的缺点,在1992年由Goldberg等首先提出了“Collaborative Filtering”即协同过滤。协同过滤基于这样的假设:人与人之间存在兴趣和偏好上的相似,而且人对事务的偏好具有一定的稳定性,可以根据过去的偏好来预测未来的选择。它与基于内容过滤(CBF)的不同

在于,协同过滤是基于与目标用户有相似兴趣偏好的其他用户对信息的态度和行为来判断该信息是否对目标用户有价值,进一步决定是否应该将该信息推荐给目标用户。其原理如图1^[12]所示。

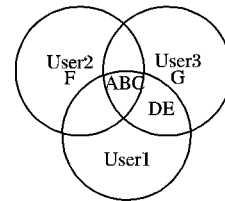


图1 协同过滤原理

对于协同过滤的研究最早始于20世纪90年代美国施乐公司的PARC研究中心。主要是为了解决员工无法在大量的Email文档中获取自己感兴趣资料的问题,PARC开发出了最早的协同过滤系统Tapestry^[13]。Tapestry系统采用的是C/S架构,该系统的推荐要求用户输入文档评价以及使用特定的查询语言;同时要求用户确定有与自己兴趣爱好相似的用户,然后系统根据用户的兴趣给出推荐,因此Tapestry只适合应用在小规模的用户环境当中。针对Tapestry系统的不足,研究人员提出了自动协同过滤的概念,并设计实现了第一个自动协同过滤系统GroupLens^[14]。GroupLens为了帮助用户从大量文章中找到符合自己兴趣偏好的文章。它采用了开放式的架构,并第一次把协同过滤技术应用于Internet环境,推动了协同过滤技术的进一步发展。目前协同过滤技术已经成为个性化推荐系统中应用比较成功的技术。但这一技术仍然存在一些需要改善和解决的问题,如稀疏问题、冷启动问题、可扩展性问题等,这些都严重影响协同过滤推荐算法的推荐质量。本文针对稀疏性问题,提出一种数据集优化方法,减少了数据集的稀疏性,并利用传统的基于用户的协同过滤算法在预处理的数据集和未经处理的数据集上进行实验,比较数据集处理前后的推荐质量。

3 协同过滤算法的步骤

一般说来,协同过滤算法可以分为构建用户档案、寻找最近邻居、预测三步^[11]。

1)构建用户档案。收集用户的评分、评价行为等,对数据进行清理、转换等,最终形成用户对项目的评价矩阵。如表2所示,其中 R_{ij} 表示用户 U_i 对项目 I_j 的评分。 R_{ij} 的取值一般是 $[0,5]$ 上的整数,分数越高表示用户对该项目的兴趣度越高。

表2 用户—项目评分

用户	商品							
	Item ₁	Item ₂	Item _j	...	Item _{n-1}	Item _n
User ₁	4	5	2
User ₂	...	1	3	...	4	...
...
User _i	2	2	R_{ij}	...	2	3
...
User _m	4	...	5	...	2	1	3	...

2)寻找最近邻居。利用相似性计算方法寻找数据库中与目标用户兴趣偏好最相似的用户作为邻居用户。计算相似性的方法主要有Pearson相似度、Cosine相似度和修正的Cosine相似度,它们的计算公式分别如下:

Pearson相似度:

$$\text{sim}(u, k) =$$

$$\frac{\sum_{i \in I_{u,k}} (R_{u,i} - \bar{R}_u) \cdot (R_{k,i} - \bar{R}_k)}{\sqrt{\sum_{i \in I_{u,k}} (R_{u,i} - \bar{R}_u)^2} \cdot \sqrt{\sum_{i \in I_{u,k}} (R_{k,i} - \bar{R}_k)^2}} \quad (1)$$

Cosine 相似度:

$$\text{sim}(u, k) = \frac{\vec{u} \cdot \vec{k}}{|\vec{u}| \cdot |\vec{k}|} \quad (2)$$

修正的 Cosine 相似度:

$$\text{sim}(u, k) = \frac{\sum_{i \in I_{u,k}} (R_{u,i} - \bar{R}_u) \cdot (R_{k,i} - \bar{R}_k)}{\sqrt{\sum_{i \in I_u} (R_{u,i} - \bar{R}_u)^2} \cdot \sqrt{\sum_{i \in I_k} (R_{k,i} - \bar{R}_k)^2}} \quad (3)$$

其中: u 和 k 分别表示用户 u 和用户 k , $I_{u,k}$ 表示用户 u 和用户 k 共同评价过的项目集合; $R_{u,i}$ 表示用户 u 对项目 i 的评分; $R_{k,i}$ 表示用户 k 对项目 i 的评价; \bar{R}_u 表示用户 u 的平均评分; \bar{R}_k 表示用户 k 的平均评分; I_u 表示用户 u 评价过的项目集合; I_k 表示用户 k 评价过的项目集合。

3) 预测。通过最近邻居集产生推荐,常用的推荐方法如下:

$$P_{u,i} = \frac{\sum_{k \in N_u} \text{sim}(u, k) \cdot R_{k,i}}{\sum_{k \in N_u} |\text{sim}(u, k)|} \quad (4)$$

$$P_{u,i} = \bar{R}_u + \frac{\sum_{k \in N_u} \text{sim}(u, k) \cdot (R_{k,i} - \bar{R}_k)}{\sum_{k \in N_u} |\text{sim}(u, k)|} \quad (5)$$

其中: $P_{u,i}$ 表示目标用户对未评分项目 i 的预测评分, $R_{k,i}$ 表示目标用户 u 的邻居用户 k 对项目 i 的评分, N_u 表示目标用户 u 的邻居用户集。式(5)与式(4)相比考虑到了用户评价尺度的问题。

典型的协同过滤推荐流程如图 2^[15] 所示。

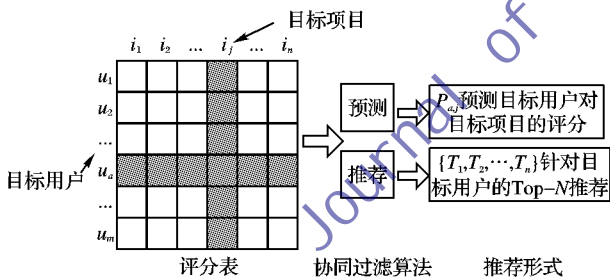


图2 协同过滤流程

4 综合均值优化填充方法

协同过滤技术完全依赖于用户的兴趣偏好数据,一般来说是用户的评分。通过构建用户—项目评分矩阵 $R_{(m,n)}$ (m 为用户数, n 为项目数), 使用统计技术寻找与目标用户具有相同或相似兴趣偏好的邻居用户, 然后再根据邻居用户对项目的评分来预测目标用户对其未评分项目的评分值, 最后选择预测评分值最高的前 N 项作为推荐的结果推荐给目标用户。基于用户的协同过滤的基本思想是用户会对与其有相似兴趣偏好的邻居用户所感兴趣的商品感兴趣。因此, 用户的偏好数据收集得越多, 那么协同过滤的推荐质量也就越高。但是由于电子商务站点的用户数目和商品数目都很庞大, 而且增长速度很快, 这使 $R_{(m,n)}$ 成为高维矩阵, 同时由于用户购买的商品很少, 而给予评分的商品数量就更少, 通常在 1% 以下。从而导致了 $R_{(m,n)}$ 中的数据极端稀疏。协同过滤技术的稀疏问题,

也是造成推荐质量下降的主要原因。

对于降低评分数据集的稀疏性, 研究人员提出了很多方法, 其中最简单的方法就是对未评分的项目设定一个缺省值, 增加评分的项目数。设定缺省值的方法主要有以下两种方式。

1) 大多数情况下, 缺省值设置为中值或稍低的值, 也可以设置为用户的评分均值或者项目的评分均值。但是该填充方法存在的问题是用户对项目的评分不可能完全相同, 以这种方法填充的评分矩阵的可信度不高。

2) 众数法。众数是指一组数据中出现频率最高的数, 采用众数法对未评分项目进行赋值, 即采用目标用户所有评分的众数作为未评分项目的预测值, 但是众数法存在“多数数”(有两个或两个以上的评分值出现的次数都是最多)和“无众数”(所有评分值出现的次数都一样)的问题, 导致这种方法的应用局限性很大。

根据以上缺省值设置方法存在的缺陷和不足, 本文给出了综合均值优化填充方法, 这种方法考虑到了用户评分尺度差异的问题, 而且也没有“多数数”和“无众数”的问题。综合均值优化评分由优化了的用户平均评分和优化了的项目平均评分两部分组成。设 $r_{a,t}$ 为计算得到的用户 a 对未评分项目 t 的评分值。首先从评分矩阵的行和列的角度对评分矩阵的未评分项进行估计, 然后再综合处理得到未评分项的最终评分值。具体计算步骤如下:

步骤 1 计算目标项目 t 的用户优化评分值 r_a 。

$$r_a = \bar{r}_a + \frac{\sum_{m=1}^M (r_{m,t} - \bar{r}_t)}{M} \quad (6)$$

其中: M 为用户空间中对项目 t 有评分的用户数目, $r_{m,t}$ 为用户 m 对未评分项目 t 的评分, \bar{r}_t 为项目 t 的平均评分。

步骤 2 计算目标项目 t 的项目优化评分值 r_t 。

$$r_t = \bar{r}_t + \frac{\sum_{c=1}^C (r_{a,c} - \bar{r}_a)}{C} \quad (7)$$

其中: C 为用户 a 在项目空间中评分过的项目总数, $r_{a,c}$ 为用户 a 对项目 c 的评分, \bar{r}_a 为用户 a 的平均评分。

步骤 3 计算用户 a 对项目 t 的评分。

$$r_{a,t} = \sqrt{r_a \cdot r_t} = \sqrt{\left[\bar{r}_a + \frac{\sum_{m=1}^M (r_{m,t} - \bar{r}_t)}{M} \right] \cdot \left[\bar{r}_t + \frac{\sum_{c=1}^C (r_{a,c} - \bar{r}_a)}{C} \right]} \quad (8)$$

$$r_{a,t} = \frac{r_a + r_t}{2} = \frac{\left[\bar{r}_a + \frac{\sum_{m=1}^M (r_{m,t} - \bar{r}_t)}{M} \right] + \left[\bar{r}_t + \frac{\sum_{c=1}^C (r_{a,c} - \bar{r}_a)}{C} \right]}{2} \quad (9)$$

通过式(8)或式(9)计算出评分矩阵中未评分项的评分值, 使得稀疏的评分矩阵变得稠密, 再在这样稠密的评分矩阵上搜寻目标用户的最近邻居集, 然后根据邻居的项目评价情况向目标用户做出推荐。

5 实验与分析

5.1 数据集

实验数据集使用由美国明尼苏达大学的 GroupLens 项目

组提供的 MovieLens 数据集。MovieLens (<http://www.movielens.org/>或者<http://movielens.umn.edu/>)是一个基于Web的研究型推荐系统,用于接收用户对电影的评分并提供电影的推荐列表。该系统使用用户打分数据,分值为1到5,分值越高表示用户对电影的兴趣也越高。美国明尼苏达大学公布的一个数据集中包含943名用户对1682部电影100000条评分记录,每个用户至少对20部电影做出评价。整个数据集的稀疏等级,即评分矩阵中未评分数据在整个数据集中所占的比例。实验选用的数据集的稀疏度为:

$$1 - \frac{100000}{943 \times 1682} = 93.7\% \quad (10)$$

目前 MovieLens 数据集在协同过滤研究领域得到了广泛的应用,也是使用最多的数据集之一。实验用到的三张数据表 USERS、MOVIES、RATINGS 中的内容如下所示。

用户表 (USERS) 中,UserID 为用户编号;Gender 为用户性别;Age 为用户年龄;Occupation 为用户职业。

电影表 (MOVIES) 中,MovieID 为影编号;Title 为影名称;Genres 为影种类。

评分表 (RATINGS) 中,UserID 为用户编号;MovieID 为影编号;Rating 为用户对电影的评分。

5.2 评价标准

目前评价推荐系统质量的标准主要有两种:统计精度度量方法和决策支持精度度量方法^{[2]293}。统计精度度量方法通过计算预测数据与真实数据之间的差别来衡量推荐效果的好坏。最常用的就是平均绝对误差 (Mean Absolute Error, MAE) 法。MAE 的值越小,表明算法的评分预测越准确,推荐质量越高。设目标客户的预测评分值为 $\{p_1, p_2, \dots, p_n\}$, 目标客户的真实评分值为 $\{q_1, q_2, \dots, q_n\}$ 。则 MAE 值为:

$$MAE = \frac{\sum_{i=1}^n |p_i - q_i|}{n} \quad (11)$$

由于 MAE 方法简单、易于理解和操作,本实验采用 MAE 的方法度量预测的效果。

5.3 实验结果与分析

实验中使用传统的基于用户的协同过滤算法,验证稀疏数据集使用本文给出的填充方法填充后对预测精度的影响。实验采用 All But One 协议,即将实验数据集中的一个用户的实际评分中的一个评分隐藏,用其他的评分数据来预测被隐藏的评分数据。相似度量方法采用余弦相似度量方法。邻居个数分别取 5、10、15、20、25、30、35 时在数据集填充前后运行协同过滤算法,得到预测值,进而计算出在数据集填充前后的 MAE 值。如表 3 所示。

表 3 数据集填充前后 MAE 值对比

邻居数	MAE 值	
	填充前	填充后
5	0.830	0.826
10	0.819	0.807
15	0.802	0.792
20	0.809	0.790
25	0.813	0.788
30	0.796	0.775
35	0.756	0.749

由表 3 可以看出,采用本文给出的数据集填充方法对原始数据集进行填充优化后,协同过滤算法的 MAE 值减小,说

明采用综合均值优化填充方法后,提高了协同过滤推荐方法的推荐质量。同时也说明了数据的稀疏性,是降低协同过滤算法推荐质量的主要原因之一。

6 结语

本文针对协同过滤算法的稀疏性问题,给出了一种数据集填充方法,并通过实验验证了该方法的有效性。对稀疏的数据集进行合理的填充后,可以提高协同过滤算法的推荐质量。下一步的工作将进一步对协同过滤算法进行改进,以提高算法的推荐质量。

参考文献:

- [1] BORCHERS A, HERLOCKER J, KONSTAN J, *et al.* Ganging up on information overload [J]. *Computer*, 1998, 31(4):106-108.
- [2] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Item-based collaborative filtering recommendation algorithms [C]// *Proceedings of the 10th International Conference on World Wide Web*. New York: ACM Press, 2001:285-295.
- [3] SCHAFER J B, KONSTAN J A, RIEDL J. Recommender systems in e-commerce[C]// *Proceedings of the 1st ACM Conference on Electronic Commerce*. New York: ACM Press, 1999:158-166.
- [4] SCHAFER J B, KONSTAN J A, RIEDL J. E-commerce recommendation applications [J]. *Data Mining and Knowledge Discovery*, 2001, 5(1/2):115-153.
- [5] VARIAN H. Recommender systems [J]. *Communications of the ACM*, 1997, 40(3):56-58.
- [6] LEE K C, KWON S. Online shopping recommendation mechanism and its influence on consumer decisions and behaviors: A causal map approach [J]. *Expert Systems with Applications: An International Journal*, 2008, 35(4):1567-1574.
- [7] SHARDANAND U. Social information filtering for music recommendation TP-94-04 [R]. Cambridge: MIT Media Laboratory, 1994.
- [8] SHARDANAND U, MAES P. Social information filtering: Algorithms for automating "word of mouth" [C]// *Proceedings of the 1995 ACM SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 1995:210-217.
- [9] MALTZ D, EHRLICH K. Pointing the way: active collaborative filtering [C]// *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York: ACM Press, 1995:202-209.
- [10] ADOMAVICIUS G, TUZILIN A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2005, 17(6):734-749.
- [11] GOLDBERG D, NICHOLS D, OKI B M, *et al.* Using collaborative filtering to weave an information Tapestry [J]. *Communication of the ACM*, 1992, 35(12):61-70.
- [12] HERLOCKER J, KONSTAN J, BORCHERS A, *et al.* An algorithmic framework for performing collaborative filtering [C]// *Proceedings of Conference on Research and Development in Information Retrieval*. New York: Springer, 1999:263-266.
- [13] BORCHERS A, HERLOCKER J, KONSTAN J, *et al.* Ganging up on information overload [J]. *Computer*, 1998, 31(4):106-108.
- [14] RESNICK P, LACOVU N, SUCHAK M, *et al.* GroupLens: An open architecture for collaborative filtering of net news [C]// *Proceedings of the 1994 ACM on Computer Supported Cooperative Work*. New York: ACM Press, 1994:175-186.
- [15] SARWAR B. Sparsity, scalability, and distribution in recommender systems [D]. Minneapolis: University of Minnesota, 2001.