基于用户行为的个性化推荐系统的设计与应用①

王 义 马尚才 (山西财经大学 山西 太原 030006)

摘 要: 目前电子商务网站提供的推荐服务很难满足用户的个性化需求，协同过滤算法作为应用最成功的推荐算法，依然存在数据稀疏性、用户评分真实性等问题，制约着推荐系统的质量。设计和实现了一个基于用户行为的个性化商品推荐系统，主要采用前融合组合推荐策略，避免了单纯使用协同过滤算法的弱点。阐述了基于用户行为的个性化推荐系统的设计思想和实现过程，最终通过实验验证了本推荐系统具有良好的推荐效果。

关键词: 电子商务网站；个性化推荐；协同过滤；用户行为；信息超载

Design and Application of Personalized Recommendation System Based on Users Behavior

WANG Yi, MA Shang-Cai

(Shanxi University of Finance & Economics, Taiyuan 030006, China)

Abstract: It is difficult for the e-commerce system to meet users individual requirements. As one of the most successful algorithms, the Collaborative Filtering Algorithm still has problems like sparsity of data and lack of authenticity in user rating. Based on the related work, a recommendation algorithm based on users actions is designed and implemented. It avoids the weakness of collaborative filtering techniques. This paper describes the recommended system's ideas and the implementation process, and proves that the recommendation system performed well through experimtnts.

Keywords: e-commerce site; personalized recommendation; user behavior; collaborative filtering; information overload

目前电子商务网站数量众多，信息资源总量庞杂、信息增长飞快。推荐系统能够针对不同的用户提供满足个性化需求的服务[1]，提高用户从信息中寻找知识的效率，从而有效保留用户，使得网站能够立于不败之地。根据 Haubl 和 Trifts (2000)研究调查发现，有推荐系统时，93%的消费者选择了更优的产品，而没有的情况下，做出这样选择只有 65%[2]。

推荐算法作为推荐系统中最重要的部分,很大程度上决定了推荐效果的好坏[3]。协同过滤推荐是应用最成功、最广泛的推荐技术，但依然存在数据极端稀疏性、新用户和冷启动等问题。然而，组合推荐可以通过不同算法之间的相互组合，降低或避免单纯推荐技术的弱点。组合策略可分为后融合、中融合和前融合技术。前融合技术是将各种推荐算法有层次的融合在一起，这种组合技术经研究表明具有更好效果。

针对上述因素，本文采用前融合技术，将 WEB 日志挖掘和用户-商品类协同过滤技术相结合的机制，设计实现了基于用户行为的个性化推荐系统。本文详细介绍了算法的三个组成模块：用户聚类了模块、个性化推荐了模块、推荐反馈子模块的实现过程。并根据推荐系统的思想，设计了以“零售中国网”为对象的个性化商品推荐系统。最终，用实验证明了本推荐新系统具有良好的推荐效果。

① 收稿时间:2009-11-25;收到修改稿时间:2010-01-01

1　基于用户行为的个性化推荐系统算法思想

基于用户行为的个性化推荐系统主要使用于面向用户的电子商务网站，列如传统的 B2C、C2C 等。主要以网站销售的商品为推荐对象，如食品、衣服、电影和新闻等。

本系统以用户浏览网站时经常带有明确目的性，用户很多浏览行为都能很好地反映用户兴趣爱好为依据，从服务器日志中挖掘出代表用户兴趣的模型，利用路经聚类方法进行聚类，将大规模用户集合转变为具有近似爱好的用户簇。

用户行为聚类后，在每个聚类簇中应用改进的协同过滤技术。这种前向融合技术，能够通过缩小协同过滤的输入规模，减少用户评分矩阵纵向的深度，将评分矩阵中用户之间没有任何关系，转变为具有一定相似性的用户集合，提高了相似性度量结果的准确性。改进的协同过滤技术是将用户对某个商品的评分转变为对某类商品的评分，减小了用户评分横向的宽度，并通过公式计算出用户的评分值，避免了用户显示评分难获得和真实性问题，从而进一度降低了数据稀疏性，提高了推荐精度。

2　基于用户行为的个性化推荐系统算法

基于用户行为的推荐系统流程如图 1 所示，包括三个子系统:用户聚类子系统、推荐子系统和推荐反馈子系统。

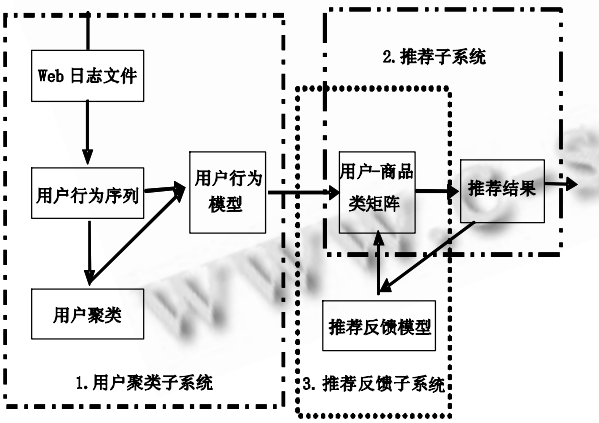

图 1　基于用户行为的个性化推荐系统流程图

2.1 户聚类子系统

主要从 Web 服务器日志中提取出用户频繁访问路径序列，利用聚类技术进行聚类，生成用户行为聚类模型库。

①挖掘用户频繁访问路径

对用户日志进行数据清洗，以得到的用户事务集为输入，网站树状拓扑结构模型为依托，Apriori 算法思想为指导，挖掘出能体现用户兴趣偏好的频繁访问路径序列，建立用户行为模型。

②用户行为聚类

主要利用 FCC 路径聚类方法，对用户频繁访问路径进行聚类，聚类后的用户对象在同一个簇中具有较大的相似性，而不同簇之间相似性最小。

③建立用户行为模型库

提取每个簇的聚类中心，建立用户行为模型库。

2.2 推荐子系统

利用改进的用户–商品类协同过滤推荐算法，进行个性化推荐。用户–商品类协同过滤推荐[4]是根据用户–商品类评分矩阵计算出用户的最近邻，并根据最近邻居信息对目标用户生成推荐。

①建立用户–商品类评分矩阵

对网站商品按概念分层进行归类，从服务器日志和销售记录两部分数据中，按照用户对某类商品的浏览、购买情况计算出用户对某类商品的评分值，填入用户评分矩阵。

②计算用户之间相似度

采用修正的余弦相似性度量[5]方法，计算出目标用户与其他用户之间的相似度。设由 U_i 和 U_j 共同评分的商品集合用 I_{ij} 表示，I_i 和 I_j 分别表示 U_i 和 U_j 各自评分的产品集合，则 U_i 和 U_j 之间的相似性，可由下面公式计算:

$$sim(i,j)=\frac{\sum_{c\in I_{ij}}(R_{i,c}-\bar{R}_i)(R_{j,c}-\bar{R}_j)}{\sqrt{\sum_{c\in I_i}(R_{i,c}-\bar{R}_i)^2}\sqrt{\sum_{c\in Ij}(R_{j,c}-\bar{R}_j)^2}} \quad (1)$$

③生成推荐结果

在相似度计算的结果的基础上，在目标用户最近邻居空间中，应用预测项目的评分公式，计算出目标用户对预测项目的评分，从中选出评分较高的项目推荐给用户。预测项目的评分公式:

$$P_{u,i}=\bar{R}_u+\frac{\sum_{n\in NBS_U}sim(u,n)\times(R_{n,i}-\bar{R}_n)}{\sum_{n\in NBS_U}(|sim(u,n)|)} \quad (2)$$

用户的最近邻居集合用 NBSu 表示，用户对项目 i 的预测评分为 $P_{u,i}$。

2.3 推荐反馈子系统

通过动态调整用户的评分值，实现最大限度的为用户提供个性化的服务。

①建立推荐结果库

对目标用户建立推荐结果库，用于记录对用户产生的推荐信息。

②更新推荐结果库

根据用户对推荐商品的浏览和购买情况，调整浏览和购买动作对应权值的大小，从而动态调整用户的评分数据，进而更新推荐结果库。

3 基于用户行为个性化推荐系统设计与应用

选择零售中国网站（www.lingshoucn.com）中移动电话单元为对象，设计了基于用户行为的个性化推荐系统。

3.1 用户行为聚类

①数据预处理

网络服务器日志并非是专门用来做数据挖掘，必须对其进行数据预处理以获得“纯净”数据。数据预处理的流程，通过线下进行，将日志文件导入数据预处理程序，得到包含时间(time)、ip 地址(ip)、用户浏览页面(u-stem)、用户请求动作(u-query)、状态(status)的用户事务集对象。对用户事务集利用改进的 Apriori 算法[6]挖掘出频繁项集，从而生成每个用户对应的频繁访问路径。

②用户行为聚类

采用 FCC 路径聚类[7]方法，设用户频繁的访问路径为 U_i={V1,V2,…,Vn},Vn 代表用户 i 频繁访问过的页面 n。通过计算用户之间的 CM 系数 S_{ij}，将 S_{ij} 值相近的用户归为一类。根据实验 CM 系数 S_{ij}=0.3 时聚类效果最佳。CM 系数 S_{ij} 公式为：

$$S_{ij}=\left|\frac{comm\left(U_i,U_j\right)}{max\left(|U_i|,|U_j|\right)}\right| \quad (3)$$

$comm(U_i,U_j)$表示用户 i 和用户 j 的频繁访问路径中最长公共路径长度,$max(U_i,U_j)$表示用户 i、用户 j 的频繁访问路径中包含的最长节点数。

对移动手机单元的日志数据进行聚类，聚类结果以二维数据表的形式存储，其中 user_id 属性列记录的是用户编号，class 属性中记录了用户所属类别，每个聚类簇中包含了大量具有相似行为的用户。

③建立用户行为模型库

建立模型库之前，需要对主要网站页面进行编号，例如零售中国网站数码、电脑类主页为... cpxxfl.aspx?id=32 其页面编号为 32，子类中移动手机类页面编号为 1107。以用户行为聚类结果和用户购买商品类别为依据，建立用户行为模型库。用户行为模型库见表 1，其中 buyhistory 字段记录用户购买商品编号，path 字段记录聚类中心频繁访问路径编号。

表 1　用户行为模型库

user_id	class	buyhisitory	path
300	4	8031 1002 506 8037 762 9031	22 10 9 29 30 33 21
301	2	462 10179 840 726 10116 27 301	2 3 9 37 33 25 4
302	6	127 8001 8024 272 98 10116 101	23 11 2 22 13 27 5
303	6	123 107 318 10134 8041 8013 60	23 11 2 22 13 27 5
304	3	637 10157 10101 784 8013 8003	32 7 11 25 20 27 5

例如编号为 304 的用户属于第三类，聚类中心为 32、7、11、25、20、27、5 分别代表电脑、服装、家用电器、礼品、通信产品、电脑配件、电器类的页面，通过分析此类用户较关注家电类产品。用户主要购买的商品编号存放在 buyhisitory 属性列。

3.2 推荐子系统建立

①用户-商品类矩阵的建立

对移动电话类商品进行概念分层可分为 5 大类，分类数据取到第 2 层，这样既可以减少评分矩阵的规模，又不会使分析精度降低太多。商品分层数据如图 2 所示。

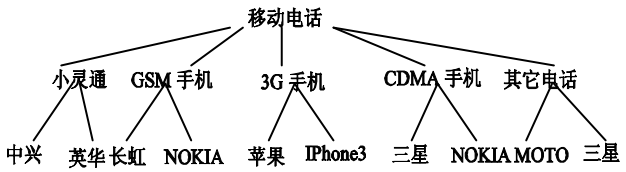

图 2　商品概念分层

用户评分值记作 R_{ij}，表示用户 i 对第 j 类商品的评分。

$$Rij=\begin{cases}0.2m & \text{浏览某类商品 m 次，0 <m< 20}\\ 4 & \text{浏览某类商品 m 次，m>= 20}\\ 1n & \text{购买某类商品 n 个，0 <n< 10}\\ 10 & \text{购买某类商品 n 个，n>= 10}\\ \sum_{i=1}^{N} R_{ij} & \text{用户未购买或浏览某类商品}\end{cases}$$

根据用户对商品类的评分公式，建立用户–商品类评分矩阵。本文选取了属于第一类用户集合的 5 名用户的评分情况，建立了评分矩阵见表 2。

表 2　第一类用户的商品类评分矩阵

user_id	class_id	小灵通	GSM 手机	3G 手机	CDMA 手机	其它类
2	1	4	4	3.8	4	3.7
8	1	3.6	4	3.2	4	3.6
9	1	3.2	3	4	3.5	4
13	1	4	2.7	4	3.6	3.2
16	1	2.8	3.6	3.6	4	3

②用户之间相似度计算

通过修正的余弦相似性度量公式(1)，来查找目标用户 304 的最近邻集合，按相似性由高到低排列得到{87, 93,132,367,1064}，统计这 5 个最近邻居所购买的商品并除去 304 用户本身购买的商品，组成预测项目集合。

③为目标用户生成推荐

在目标用户的预测项目集合中，按预测项目的评分公式(2)，计算出目标用户对预测项目的评分，再根据评分的高低进行排列，将预测评分在 3.5 以上的项目推荐给用户。目标用户的推荐表见表 3。

表 3　目标用户推荐表

user_id	tuijian
300	171 10103 11 9014 800 1029
301	10175 731 200 304 303
302	8027 9019 8030 10099 10148 9016
303	310 21 621 8043 10175
304	429 1017 1175 151 1297
305	8093 10103 612 151 618 8040

从表中得出为 304 用户推荐的商品编号为 429、1017、1175、151、1297 等。

3.3 推荐反馈子系统的建立

根据用户对推荐商品的动作，动态调整用户–商品类评分矩阵，从而提高推荐的精度。如编号为 304 的用户浏览了编号为 429 的推荐类产品，通过查找对应 429 产品属于 GSM 手机类，则对 GSM 手机类评分值设定时的购买次数加 1，以达到更新用户–商品类评分矩阵的目的。

4　实验结果分析

①推荐质量度量标准

采用统计度量方法中平均绝对偏差 MAE(mean absolute error)，作为度量推荐系统好坏的标准。MAE 通过计算预测用户评分与用户显示评分之间的偏差，来度量系统质量。MAE 值越小，推荐质量越高。对于评分数据，预测的评分集合表示为(p1, p2,...,pn)，用户显示评分集合为{q1,q2,...,qn}，则平均绝对偏差 MAE 定义为[8]:

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{4}$$

本文选用零售中国网站(www.lingshoucn.com)为试验对象，考虑到该公司规模以及为避免因各类促销而引起销售量的变化等因素，选取 2009 年 6 月到 2009 年 9 月促销较少月份日志文件作为试验对象。前 3 个月的日志数据和用户的购买记录作为训练集；9 月份的数据作为测试集。随机抽取了 2000 个用户进行跟踪调查，以邮件的形式采集到 1207 个用户显示评分。选取基于余弦和基于项目评分预测的协同过滤推荐算法与本文进行对比。

②实验结果分析

通过平均绝对偏差 MAE 的计算，推荐效果分析如图 3 所示。

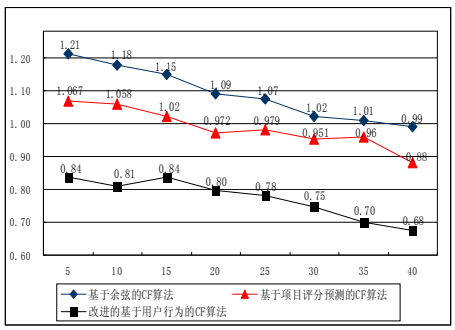

图 3　MAE 随最近邻居数变化图

由图 3 可知，在这种实验条件下，本文基于用户行为的个性化推荐系统均具有最小的 MAE。由此可知，与传统的协同过滤推荐系统相比，本文的基于用户行为的个性化推荐系统可以提高推荐系统的推荐质量。

本文基于用户行为的个性化推荐系统与传统的协同过滤推荐系统的最大不同，就是用户评分矩阵的设定方式不一样。在传统的协同过滤推荐系统中，矩阵中的用户之间没有任何关系，而且评分值是用户显示给出的，造成评分数据难获得和真实性问题。本文采用前向融合技术，首先根据用户频繁访问路径进行聚类，在每个聚类簇中应用改进的协同过滤技术，使得矩阵中用户之间具有相似的访问路径，并且建立推荐反馈系统动态调整推荐结果，不断提高评分的精度，降低的数据稀疏性。

5 结语

随着我国电子商务信息化建设快速发展，如何为用户提供更加准确、高效、便捷的信息化服务成为电子商务建设中的一个重要任务。个性化推荐服务能够为电子商务用户提供满足其个性化需求的商品和特定功能服务，受到了人们的广泛关注，是目前电子商务领域中非常有意义的研究内容。本文提出了一种基于用户行为的个性化推荐系统的具体实现方案，并详细论述了系统模块的具体实现，为电子商务个性化服务的研究和实践提供了一种新的思路。

参考文献

1 Lin WY, Alvarez SA, Ruiz C. Collaborative recommendation via adaptive association rule mining. 2000.

2 Haubl G, Trifts V.Consumer Decision Making in On line Shopping Environments: The Effects of Interactive Decision Aids Marketing Science, 2000.

3 许海玲,吴潇,李晓东.互联网推荐系统比较研究.软件学报, 2009:352－358.

4 崔亚洲,段刚.基于 Web 日志和商品分类的协同过滤推荐系统.电子科技大学学报, 2006:39－41.

5 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法.计算机应用研究, 2008:1622－1623.

6 Jiawei Han, Micheline K.Data Mining Concepts and Techniques,Second Edition.北京:机械工业出版社, 2006:146－173

7 业宁,李戚,梁作鹏,董逸生.一种Web用户行为聚类算法.小型微型计算机系统, 2004:1364－1367.

8 Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. Proc. of the 10th International World Wide Web Conference. 2001.