文章编号：2096-1472(2018)-03-16-03
DOI:10.19644/j.cnki.issn2096-1472.2018.03.004

基于消费者购物记录的商品推荐去重方案

张鹏程，马佳琳
(沈阳师范大学软件学院，辽宁 沈阳 110000)

摘　要：伴随着网购的不断发展，各大电商网站均引入了商品推荐系统。但消费者却常常对这类系统的有效性抱有疑问，因为他们发现自己前一天刚买的商品依然出现在今天的推荐列表中，而自己已经不再需要这类商品了。针对这样的情况，需要给推荐系统增加一个过滤模块，将一些在当前明显不会被目标用户所需要的商品去除。本文在前人提出的回购周期去重方案的基础上进行优化，提出了一套综合去重方案。

关键词：推荐方法；推荐去重；推荐过滤；大数据
中图分类号：TP391　　**文献标识码**：A

Commodity Recommendation and Filtering Scheme Based on Consumer Shopping Records

ZHANG Pengcheng,MA Jialin
(*College of Software,Shenyang Normal University,Shenyang* 110000,*China*)

Abstract: With the continuous development of online shopping,the main e-commerce websites have all introduced commodity recommendation systems.However,consumers usually have doubts on the effectiveness of such kind of system,as they find that the commodities they purchased a day before still appear on the recommendation list today,which they do not need any more.Aimed at such cases,a filtering module needs to be added to the recommendation system to eliminate the commodities that are obviously undesired by the target users at present.This paper conducted optimization based on the filtering schemes put forward on buy-back cycle previously and put forward a set of comprehensive filtering scheme.

Keywords: Recommendation method;repetition removal;recommendation filtering;big data

1 引言(Introduction)

互联网的出现使人们足不出户就能获得各类信息。而多年以来互联网的发展也呈现出爆炸性趋势，加上物流运输产业的的发展，越来越多的人选择通过网上商城进行购物。商家们为了获取更多的利润，几乎都在网站上使用了商品推荐系统，以期获得更大的利润。

Resnick和Varian在1997年给出了推荐系统的定义：“它利用电子商务网站向客户提供商品信息和建议，帮助用户决定应该购买什么产品，模拟销售人员帮助客户完成购买过程”[1]。推荐系统的核心是推荐方法，经过多年的发展，当前较为成熟的推荐方法包括协同过滤，基于内容推荐，基于图结构推荐和混合推荐[2]。

然而当前几乎所有的购物网站使用的推荐系统在对消费者进行商品推荐时都没有考虑推荐去重的问题。例如目标用户经过一个月的筛选，最终买下了一台几千元的手机，而当用户第二天打开网站的时候，网站的推荐系统却仍然在卖力的向其推荐手机。这样的情况不仅浪费了推荐的机会，更让消费者对网站的推荐系统能力产生怀疑，用户对系统推荐性能持否定态度的结果将导致其不再关注系统推荐的任何商品。这样的局面不仅是消费者的损失，更是商家的损失。

而对购买过的商品采取“一刀切”，不再推荐的做法也显然不行，如果消费者购买的商品属于消耗类商品，例如牙膏，那么显然很有必要再次进行推荐，但推荐的时机是一个很重要的问题。

针对以上的问题，张志清[3]等人提出了一种考虑商品重复购买周期[4]的协同过滤推荐方案。本文则受其启发，在其基础上提出一种独立的去重模块，以期能与各类推荐方法组合使用在推荐系统中，提升推荐系统的有效性。

2 去重方案(Filtering scheme)

2.1 考虑商品重复购买周期的协同过滤推荐方案

该方案主要包括如下步骤：第一是建立商品数据库和顾客交易数据库，并将所有商品分成三类，即长重复购买周

期、短重复购买周期和零重复购买周期的商品；第二步是计算短重复购买周期商品的重复购买周期，生成重复购买周期表；第三步则是计算目标推荐用户已购商品的回购状态，找出处于消费周期，失效状态和长重复购买周期的商品；最后一步是将传统的协同过滤方法得到的商品推荐列表与以上三步获得的商品列表进行比对，将两个列表中重复的商品从协同过滤结果的列表中去除，最后将过滤后列表内的商品推荐给用户。

2.2 考虑商品重复购买周期方案的局限性

首先是分类的局限，原方案中对商品进行了事先人工分类，但这样的分类缺乏个性化，因为一些商品对于不同的人来说情况不同，存在一部分人几年内只购买一次，而另一部分人会隔一段时间就购买一次的商品，这样的情况下人工分类吃力不讨好。第二是重复购买周期计算方法的局限。原方案对顾客遇到特价活动时的提前重复购买行为在计算时仅简单的去除，没有考虑提前重复购买对于后续购买周期的影响。第三是对重复购买周期离散值确定上的局限。原方案将离散值设定为固定的30天，但由于商品特性和消费者复杂的情况，该离散值在实际使用上很难取得较好的效果。最后是对失效期商品去除的局限。当超出回购周期后，原方案会认为用户已经在其他地方购买了商品，于是不再对其推荐该商品，这其中没有考虑消费者不购买的原因而是做简单抛弃。

2.3 基于消费者购物记录的去重方案

针对以上问题，本文提出了一套综合的去重方案，主要改动包括：首先是不再进行人工三段式分类，改为先按商品品种[5,6]简单分类(这样的分类每个电商网站本来就有，相当于不需要做额外分类)，再通过计算来划分出真正的针对目标用户的个性化分类的方式。其次，充分考虑优惠打折活动对消费者带来的影响，修改计算方式，使得到的回购周期更加合理。第三，通过考虑用户多次购买同类产品的回购周期离散值来决定商品的黄金推荐周期，消费周期和超期失效周期。第四，对于超过回购周期的商品，通过参考用户在最后一次购买后的浏览记录，购物车记录和商品售价情况综合判断用户对商品的态度后，在有必要的情况下会根据实际情况再次计算回购周期来决定待推荐的商品的去留，不再是一超期就无条件丢弃。

3 方案实现(Scheme realization)

3.1 数据准备

在进行去重前，需要在数据库中先建立如下表格：

表1：用户历史购买记录表。表中记录项目包括“商品ID，购买日期，购买数量，单件容量，单价，购买时是否享有折扣”。

表2：商品分类表。表中记录项目包括“商品ID，商品所属类别ID，全体用户日均消耗离散值”，其中商品的分类按照商品品种进行划分。

表3：用户浏览记录表。表中记录项目包括“商品ID，浏览日期，单件容量，单价，用户浏览时是否有折扣”。

表4：用户购物车记录表。表中记录项目包括“商品ID，放入购物车时间，单件容量，单价，放入购物车时是否有折扣”。

以上表中的“单件容量”属性均指代消耗类商品的容量，非消耗类商品此值留空。考虑到节省存储空间问题，一个用户对一个商品的浏览或购物车放入记录都只保留最后一次。

3.2 去重方案具体步骤

在正式开始去重前还需要准备以下列表用于计算：一个针对目标用户的推荐列表I，该列表通过任意一种推荐算法得出；一个待处理列表I^*，I^*的内容与I一致；一个排除列表I_k，I_k初始为空，用以保存待清除的推荐商品。以上列表均只保存商品的ID。

步骤1：依次检查I^*中的所有商品，若目标用户对待检商品及与待检商品同类商品的购买次数之和仅为1次，则将商品ID从I^*移动到I_k中，而购买记录为0的商品ID则从I^*中删除。I^*经过处理后称为I_2^*。

步骤2：依次检查I_2^*中所有非特价购买次数小于等于1次的商品，若这些商品正在打折，则将其从I_2^*中删除，否则将商品ID从I^*移动到Ik中。I_2^*经过处理后称为I_3^*。

步骤3：尝试从I_3^*中取一个商品，若I_3^*为空，则将I与I_k中均存在的ID值相同的商品从I中删除，I经处理后最终得到I_2，将I_2中的商品推荐给用户，全部流程结束；否则将取到的商品ID记为C，并获得C所属的商品类别，记为U。接着计算目标用户对该类商品的日均消耗量：

$$G_{CDA}=\frac{D_{EnR}-D_{FiR}}{\sum_{t=1}^{n-1}(WN_t\times WC_t)}$$

其中，D_{EnR}代表目标用户最后一次以非打折价格购买U类商品的日期，D_{FiR}代表目标用户第一次以非打折价格购买U类商品的日期，WN代表单次购买U类商品的数量，WC代表单次购买U类商品的容量(若WC为空或0，则WC=1，之后步骤同理)，n代表在D_{EnR}与D_{FiR}之间购买过U类商品的次数。

计算用户对U类商品的单次回购周期内对U的日均消耗量G_{ct}：

$$G_{ct}=\frac{D_{t+1}-D_t}{WN_t\times WC_t}$$

计算用户对商品C的日均消耗量的离散程度SD_c：

$$SD_c=\sqrt{\frac{\sum_{t=1}^{n-1}(G_{ct}-G_{CDA})^2}{n}}$$

其中，D_t为第t次购买U类商品的日期，G_{ct}为第t次和t+1次购买的U类商品的日均消耗量，n代表在D_t与D_{t+1}之间购买

过U类商品的次数。

步骤4：最后一次以非特价购买U类商品到最后一次购买U类商品期间总购买量P_E：

$$P_E = \sum_{i=1}^{N}(WN_i \times WC_i)$$

其中，N为最后一次以非特价购买U类商品到最后一次购买U类商品期间对U类商品的购买次数。

步骤5：计算目标用户对商品C的消耗周期预测范围(即黄金回购周期)：

最短消耗周期：

$$\lfloor T_{\mathrm{HMin}} \rfloor = \frac{P_E}{G_{CDA} + \min(SD_C, SD_A, SD_{CA}{}^*)}$$

最长消耗周期：

$$\lceil T_{\mathrm{HMax}} \rceil = \frac{P_E}{G_{CDA} - \max(SD_C, SD_A, SD_{CA}{}^*)}$$

其中，$SD_{CA}{}^*$代表商品C的全体用户日均消耗离散程度，SD_A为可选参数，默认为0，该值代表与目标用户相似的其他用户对U类商品的离散程度，例如推荐算法中使用到“相似用户”[7]时可以计算相似用户对商品C的离散程度。

另外考虑到电商的物流情况，若计算后$T_{\mathrm{HMax}}<1$，则令$T_{\mathrm{HMax}}=1$。

步骤6：判定当前日期与黄金回购周期的关系：

若$T_{\mathrm{HMin}} \leqslant T_{\mathrm{Act}} \leqslant T_{\mathrm{HMax}}$，说明商品处于黄金回购周期，可以推荐。将C从$I_3{}^*$中删除，完成后执行步骤3。

若$T_{\mathrm{HMin}}>T_{\mathrm{Act}}$，说明商品处于消耗周期，不需要推荐。将C从$I_3{}^*$移动到$I_k$内，完成后执行步骤3。

若$T_{\mathrm{Act}}>T_{\mathrm{HMax}}$，说明商品处于失效周期，需要进一步确定情况，执行下一步骤。

其中，前时间与最后一次非特价购买U类商品的时间差为T_{Act}。

步骤7：检查目标用户在回购期开始后有无将U类商品放入购物车的记录，若无，执行步骤8，否则首先检查购物车记录中商品价格情况，若满足$M_{\mathrm{Now}} \geqslant M_{\mathrm{Se}} \geqslant M_{\mathrm{CM}}$，说明消费者因为价格因素放弃购买，则将C从$I_3{}^*$移动到$I_k$内，完成后执行步骤3；若不满足，说明消费者超期未购买的因素可能与价格无关，将D_{EnR}值设置为购物车记录生成日期，而P_E值通过购物车信息计算得到，之后返回步骤5。

其中，M_{Now}代表当前商品C的最新单价，M_{SC}代表目标用户最后一次将U类商品放入购物车时的商品单价，M_{CM}代表用户购买过的U类商品的平均单价。

步骤8：检查目标用户在回购期开始后有无对U类商品页面的浏览记录，若无，则将商品C从$I_3{}^*$中删除，完成后执行步骤3，否则采用与购物车记录相同的处理步骤，如果需要重新计算黄金回购周期，则D_{EnR}的值设置为目标用户对U类商品最后浏览日期，P_E值设置为：

$$P_E = \frac{\sum_{i=1}^{N_W}(WN_i \times WC_i)}{N_W}$$

其中，N_W代表目标用户对U类商品的购买次数。

4 算法检验(Algorithm inspection)

4.1 去重有效性评价指标

主要从两个指标进行评价：召回率和误识别率[8,9]。

召回率。召回率在本文中定义为正确去除的推荐商品占所有需要去除的推荐商品的百分比，召回率越高越好。其计算公式为：

召回率=(正确识别的多余推荐商品数目/实际包含的多余推荐商品总数)×100%

误识别率。误识别率在本文中定义为被错误去除的推荐商品占所有被识别为待去除商品的百分比，误识别率越低越好。其计算公式如下：

误识别率=(错误识别为待去除商品的数量/被识别为待去除商品的总数)×100%

4.2 评价指标获取方案

步骤1：获取算法所需表格。

步骤2：任意取一个用户P，统计其历史购买记录，获取每个类别商品的总购买次数。

步骤3：对每种购买次数>2的商品，忽略掉其最后一次购买记录。将剩余相关记录导入算法中计算，算法会输出每一个商品的黄金回购周期。

步骤4：将黄金回购周期与用户最后一次购买该商品的日期进行比对，若用户最后一次购买该商品的日期在黄金回购周期内，则判定为成功识别，否则为错误识别。

步骤5：通过前四个步骤获取的参数计算召回率和误识别率。

4.3 实验结果

实验采用的数据集是包含有关网站部分用户购物信息汇总的数据集，实验结果如表1所示。

表1 去重方案实验结果对比

Tab.1 Filtering scheme with comparison of experiment results

方案名	评价指标 %	
	召回率	误识别率
考虑购买周期的去重方案	40.24	23.70
综合考虑的去重方案	55.49	15.74

注：结果四舍五入后保留两位小数。

5 结论(Conclusion)

目前各大电商网站的推荐系统基本都存在重复推荐消费

者当前所不需要的商品的情况。本文在前人的基础上提出了一种以重复购买周期为主，考虑多种情况的推荐商品去重方案，让商家和消费者能得到双赢的结果。该去重方案还可以反向运用，即作为一种推荐方法，定期向用户推荐其可能需要重复购买的商品。

在后续的研究中，将着重处理以下问题：一是对于某些商品来说，消费者重复购买并不是因为周期消耗因素，例如目标用户间隔一段时间就帮不同的朋友购买一块电脑硬盘，这时候计算消耗周期的意义不大，因为即使目标用户也不一定清楚自己会在什么时候需要硬盘；而直接将硬盘归入非消耗类商品从而排除计算周期的方案也不可行，因为存在一些消费者需要大量的存储空间而定期购买硬盘的情况，例如商店内的监控录像存储，因此需要研究一种全新的思路。另一个问题是分类计算周期，在本文的算法中，考虑到很多用户在重复购买时并不局限于购买一个品牌的商品，因此在第一次人工分类时没有细化分类，但部分同类产品高端与低端之分，消费者可能出现种种原因导致高低端产品都有购买，这时在计算周期前可能还需要考虑例如价格等因素来进行更细致的分类，之后才计算回购周期。

参考文献(References)

[1] 王国霞,刘贺平.个性化推荐系统综述[J].计算机工程与应用,2012,48(07):66-76.

[2] 杨博,赵鹏飞.推荐算法综述[J].山西大学学报,2012(3):337-350.

[3] 张志清,李梦,胡竹青.考虑商品重复购买周期的协同过滤推荐方法改进[J].武汉科技大学学报,2017,40(04):307-313.

[4] 罗子明.消费者购买行为的测量指标[J].北京商学院学报,2000(04):30-33.

[5] 陈立平.浅谈商品分类的目的[N].中国商报,2004-05-21.

[6] 联商.便利店商品分类与编码[N].中国商报,2003-11-28.

[7] 高董英,邓新国,肖如良.融合相似用户和信任关系的动态反馈协同过滤推荐算法[J].福州大学学报,2017,45(01):25-31.

[8] 周奕辛.数据清洗算法的研究与应用[D].青岛大学,2005:40-41.

[9] 朱郁筱,吕琳媛.推荐系统评价指标综述[J].电子科技大学学报,2012,41(02):163-175.

作者简介：

张鹏程(1993-)，男，硕士生.研究领域：中小企业信息化.

马佳琳(1972-)，女，硕士，教授.研究领域：中小企业信息化，电子商务.

(上接第22页)

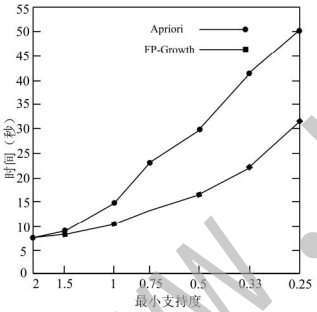

图4 Apriori算法与FP-Growth算法效率比较

Fig.4 Comparison between Apriori algorithm and FP-Growth algorithm

由图3和图4可以看出，FP-Growth算法效率要高于Apriori算法。

5 结论(Conclusion)

本文对Apriori算法及其改进的算法进行了研究，并将Apriori算法应用于游戏销售数据中，挖掘出销售数据中的强关联规则，并对相关规则作出描述，总结出了一套视频游戏的销售规律。

参考文献(References)

[1] Tsai C F,Lin Y C,Chen C P.A new fast a-lgorithms for mining association rules in large databases[C].IEEE International Conference on Systems,Man and Cybernetics.IEEE,2002,7:6.

[2] Khabzaoui M,Dhaenens C,TalbiEG.Fast algorithms for mining association rules[J].Journal of Computer Science & Technology,2008,15(6):619-624.

[3] Yang L,Wang F,Wang T.Analysis of dishonorable behavior on railway online ticketing system based on k-means and FP-growth[C].IEEE International Conference on Information and Automation.IEEE,2017:1173-1177.

[4] Pamba R V,Sherly E,Mohan K.Automated Information Retrieval Model Using FP Growth Based Fuzzy Particle Swarm Optimization[J].International Journal of Information Technology & Computer Science,2017,9(1):105-111.

[5] 林颖华,陈长凤.基于关联规则的企业财务风险评价研究[J].会计之友, 2017(1):32-35.

[6] 高晓佳.Apriori算法优化策略的研究[J].长春理工大学学报(自然科学版),2009,32(4):660-662.

[7] 刘玉锋.数据挖掘中关联规则算法的研究与应用[D].长春理工大学,2010.

作者简介：

闫东明(1989-)，男，硕士生.研究领域：数据库与数据挖掘.

陈占芳(1980-)，男，博士，副教授.研究领域：数据库与数据挖掘.

姜晓明(1988-)，男，硕士生.研究领域：数据库与数据挖掘.