考虑年度日程表事件的协同过滤推荐

夏秀峰,郝仲模,李 磊
XIA Xiu-feng,HAO Zhong-mo,LI Lei

沈阳航空工业学院 计算机学院,沈阳 110136
School of Computer,Shenyang Institute of Aeronautical Engineering,Shenyang 110136,China
E-mail:xiaxiufeng@syiae.edu.cn

XIA Xiu-feng,HAO Zhong-mo,LI Lei.Recommended collaborative filtering with annual schedule events considered.Computer Engineering and Applications,2010,46(11):135-137.

Abstract: Collaborative filtering is the recommendation technology which is applied most widely in the personalized recommendation system.However,existing collaborative filtering algorithms do not consider the relationship between the every year specific event and the user purchase behavior.For this reason,this paper presents a collaborative filtering algorithm considering the year schedule event.It introduces the time weight function,makes recommendation of the commodity which purchased by the consumer of close neighbor and which's purchased time is close to current user access time and belong to the same period to be higher.So the recommendation accuracy of collaborative filtering algorithms can be improved.

Key words: annual schedule events;collaborative filtering;personalized recommendation;time weight function

摘 要:协同过滤是个性化推荐系统中应用最广泛的推荐技术,现有的协同过滤算法不能反映出每年特定的事件与用户购买行为的关联性。针对这个问题,提出了一种考虑年度日程表事件的协同过滤算法,引入了时间权值函数,使得同一时期的越接近当前用户访问时间的近邻用户购买商品的推荐度越高,提高了协同过滤算法的推荐精度。

关键词:日程表事件;协同过滤;个性化推荐;时间权值函数
DOI:10.3778/j.issn.1002-8331.2010.11.041 **文章编号:**1002-8331(2010)11-0135-03 **文献标识码:**A **中图分类号:**TP311

网络所带来的便捷的信息传递和信息服务推动着电子商务的蓬勃发展，人们在逐渐享受由此带来的巨大惊喜的同时，也面临着从传统购物方式向网络虚拟购物方式转变的挑战。推荐系统是一种个性化的信息过滤技术[1],它被用来预测某个特定用户是否会喜欢某个特定的商品(预测问题),或被用来确定N件用户感兴趣的商品（TOP-N推荐问题)。协同过滤(Collaborative Filtering,CF)是建立推荐系统最成功的技术,其优点是能够基于信息的质量和兴趣进行过滤并推荐用户意想不到的有用信息[2],很多电子商务网站都使用该技术,如Amazon,CDNOW和eBay。

协同过滤算法在计算推荐过程中将用户访问过的每个资源同等对待,这显然是不合理的。为了及时反映用户兴趣变化,在文献[3]中提出了一种时间加权协同过滤算法以得到用户兴趣变化的新颖信息。但现有的协同过滤推荐算法中,都忽略了一些销售事件可以和每一年特定的事件关联起来。将年度日程表事件(每一年有特定意义的时间 如:国际劳动节、元旦之类的节日等)对顾客购买行为的影响引入到基于用户的协同过滤算法的推荐过程中,加入新的时间权重,提出一种考虑年度日程表事件的时间加权协同过滤算法。

1 协同过滤技术

协同过滤的出发点是:兴趣相近的用户可能会对同样的东西感兴趣。所以,通过维护关于用户喜好的数据,从中分析得出具有相似兴趣的用户,然后就可以根据相似用户的历史访问记录来向其进行推荐。另一种可能的出发点是:用户可能较偏爱与自己已购买的商品相类似的商品。可以根据用户对各种商品的评价来判断商品之间的相似程度,然后推荐与用户兴趣最接近的那些商品。

在现有的协同过滤推荐系统中,输入数据通常可以表述为一个$m\times n$的用户-资源访问矩阵$\boldsymbol{R}$。其中m是用户数,n是资源数,R_{ij}是第i个用户对第j个资源的兴趣度。对于电子商务网站来说,对于用户对商品的兴趣度的评定通常使用打分的方式。但它有一个明显的缺点,收集数据比较困难,用户通常并不愿意为你贡献这种数据。另外一种被认为更有效的方法是"隐式评分"方法。这种方法不需要用户直接输入评价数据,而是根据用户的行为特征由系统代替用户完成评价。电子商务网站在隐式评分的数据获取上有先天的优势,用户购买的商品记录是非常有用的数据。文中采用的用户兴趣度评定方法是根据用户购买的某个商品的数量来确定用户对商品的兴趣度。具体方法

作者简介:夏秀峰(1964-),男,博士,教授,CCF高级会员,主要研究方向为数据理论与技术;郝仲模(1982-),男,硕士研究生,主要研究方向为点击流数据仓库;李磊(1981-),男,硕士研究生,主要研究方向为点击流数据仓库。

收稿日期:2008-10-14 **修回日期:**2009-01-04

如表1所示。

表1 兴趣度评定方法定义

购买某个商品的数量	兴趣度
1	1
2	2
3	3
4	4
≥5	5

典型的协同过滤算法是基于用户的(user-based),它的基本原理是利用用户访问行为的相似性来互相推荐用户可能感兴趣的资源。对当前用户 u,系统通过其历史访问记录及特定相似度计算公式,计算出与其访问行为(购买的产品集合、访问的网页集等)最相近的 k 个用户作为用户 u 的最近邻居集[4]。在电子商务网站推荐系统中相似性计算公式如下:

设用户 a 和 b 购买过的商品集合用 i 表示,则用户 a 和 b 之间的相似性 $sim(a,b)$ 通过 Peason 相关系数度量。

$$sim(a,b)=\frac{\sum_{i=1}^{m}(R_{a,i}-\overline{R_a})(R_{b,i}-\overline{R_b})}{\sqrt{\sum_{i=1}^{m}(R_{a,i}-\overline{R}_a)^2\sum_{i=1}^{m}(R_{b,i}-\overline{R}_b)^2}} \tag{1}$$

$R_{a,i}$ 和 $R_{b,i}$ 表示用户 a 和 b 对商品 i 的兴趣度,$\overline{R_a}$ 和 $\overline{R_b}$ 表示用户 a 和 b 对商品 i 的平均兴趣度。

然后计算候选推荐集中每个商品 i 对用户 u 的推荐度,取其中 n 个排在最前面的商品作为用户 u 的 top-N 推荐集。推荐度计算公式如下:

$$\mathrm{rec}(u,i)=\overline{R_u}+\frac{\sum_{a=1}^{n}(R_{a,i}-\overline{R_a})\times sim(u,a)}{\sum_{a=1}^{n}|sim(u,a)|} \tag{2}$$

$\overline{R_a}$ 和 $\overline{R_u}$ 分别表示用户 u 和用户 a 对商品的平均兴趣度,$sim(u,a)$ 是用户 u 和用户 a 的相似系数,$R_{a,i}$ 表示用户 a 对商品 i 的兴趣度,n 是相似用户的个数。

为解决传统协同过滤算法的可扩展性问题,文献[5]提出了基于资源(item-based)的协同过滤算法,该算法比较资源与资源之间的相似性,由当前用户已访问的资源集合推荐未访问的资源。由于资源间的相似性比用户相似性稳定,因此可以离线进行计算存储并定期更新,较好地解决了算法的可扩展性问题。

2 算法描述

现有的协同过滤推荐算法都普遍存在一个问题:只注重用户或资源间的相似性,而忽略了一些销售事件可以和每一年特定的事件关联起来。每一年特定的事件是指一些节日之类的对客户购买行为有影响的时间段。例如:在圣诞节时,大家都购买什么样的商品,对别人有很好的借鉴作用。要综合考虑每年的特定时间对顾客购买行为的影响,以接近各个时期顾客关注的焦点。

在传统的协同过滤算法的基础上,提出了考虑年度日程表事件的时间加权协同过滤算法,更好地反映了用户购买行为与每一年特定事件的关联性,提高算法推荐精度。

2.1 年度日程表事件元数据

元数据是"关于数据的数据",它描述了数据仓库的数据和环境,用于存储数据模型和定义数据结构、转换规则、控制信息等,是数据仓库的核心。通常元数据是独立于数据仓库单独存储,它有自己的数据模型和访问方式[6]。

年度日程表事件元数据定义了每个年度日程表事件的时间跨度,通过用户访问时间所处的时间段,来确定用户属于哪个年度日程表事件的时间范畴。元数据的数据结构如表2所示。

表2 元数据的数据结构

序号	属性名称	类型	长度	备注
1	编号	字符型	4	年度日程表事件的编号
2	名称	字符型	20	年度日程表事件的名称
3	开始时间(公)	日期型	8	公历事件的开始时间
4	结束时间(公)	日期型	8	公历事件的结束时间
5	开始时间(农)	字符型	8	农历事件的开始时间
6	结束时间(农)	字符型	8	农历事件的结束时间

例如:定义编号为 $GT1$ 的年度日程表事件"元旦",用户的感兴趣时间设定为从每年的12月12日到12月27日(考虑到用户购买商品后,商品的邮寄时间一般为3天)。历史同期的其他用户在此时间段的购买行为对当前用户在此时间段的购买行为会产生一定的影响。因此,可以定义一系列年度日程表事件,如教师节、母亲节等,通常是指一些对用户购买行为有影响的特殊事件。通过网站以往用户的历史访问记录可以推测每个年度日程表事件的用户兴趣持续时间。通过加入新时间权值函数,改进协同过滤算法,可以提高算法的推荐精度。

具体使用方法如下:

第一步,取出当前用户的访问时间 T_u 和推荐集中相似用户所购买商品的时间 T_a 并计算出它们的农历日期。

第二步,访问年度日程表事件元数据表确定 T_u 和 T_a 及其农历日期属于的年度日程表事件及事件编号。GTi 和 GTj、NTi 和 $NTj(i,j\in N)$ 分别表示 T_u 和 T_a 的公历年度日程表事件和农历年度日程表事件的编号。

第三步,把 T_u 和 T_a 的年度日程表事件的编号运用到下面的时间权重函数的计算中。

2.2 基于时间的数据权重

在时间加权协同过滤算法中,要充分考虑到年度日程表事件对用户购买行为的影响。即最近的、同一时期的购买行为对推荐度的影响最大。引入时间加权函数到推荐度预测中,将商品 i 对目标用户 u 的推荐度公式做了进一步完善。

在计算相似用户 a 购买商品 i 的时间 T_a 与当前目标用户 u 访问时间 T_u 的时间间隔过程中引入了两种度量方式。

(1)年份差值:以年度为单位,计算相似用户 a 购买商品 i 的年份与当前目标用户 u 访问年份的间隔,用 Y_{ua} 表示。

(2)单位年度内日期相对差值:以月和日为单位,计算相似用户 a 购买商品 i 的时间与当前目标用户 u 访问时间的以月和日为单位时间间隔,用 D_{ua} 表示。

定义基于年度日程表事件的两个的时间权重函数 $WD(u,i)$ 和 $WY(u,i)$。

$$WD(u,i)=1-\frac{D_{ua}}{360}a \tag{3}$$

$$WY(u,i)=b^{Y_{ua}} \tag{4}$$

$WD(u,i)$ 主要是为了实现与当前用户登录时间属于同一时期的购买行为推荐度应该最高的策略。如果 T_u 和 T_a 的公历

年度日程表事件编号 GTi 和 GTj、农历年度日程表事件编号 NTi 和 NTj 有一组值相同,以 D_{ua} 为时间变量的权重函数 $WD(u,i)$ 值为 1,如果都不相同就通过计算公式确定时间权重函数 $WD(u,i)$ 的值,取值范围应该保持在(0,1)范围内。

$WY(u,i)$是为了突出近期购买过的商品的重要性,实现同一时期的越接近当前用户访问时间的购买数据推荐度越高。年份差值越小,权重函数 $WY(u,i)$的值越大,其取值的范围在(0,1)内。

$a\in(0,1)$和 $b\in(0,1)$称为权重增长指数,改变 a 或 b 可以调整权重随时间间隔变化的速度。

最终的基于年度日程表事件的时间加权函数表达式:

$$WDY(u,i)=WD(u,i)\times WY(u,i) \tag{5}$$

商品 i 对目标用户 u 的推荐度计算公式改进为:

$$\mathrm{rec}(u,i)'=\overline{R_u}+\frac{\sum_{a=1}^{n}(R_{a,i}-\overline{R_a})\times sim(u,a)\times WDY(u,i)}{\sum_{a=1}^{n}|sim(u,a)|\times WDY(u,i)} \tag{6}$$

2.3　算法的实现

在协同过滤推荐算法中,用户间相似度的计算和推荐的产生是在一个处理过程之中,为了提高推荐的速度,采用文献[7]中的方法把用户之间相似度的计算放在离线处理部分,减少了在线推荐的计算量。下面介绍协同过滤推荐部分的实现,首先获取目标用户 IP,然后判断该用户是第一次访问,还是已经访问过该网站。如果用户是第一次访问,选择访问频率高的前 K 项作为推荐内容。如果用户以前访问过该网站,根据与聚类中心的相似度找出该用户所属类以及类内所有用户,再计算出目标用户和类内其他用户之间的相似系数。最后通过推荐度公式(6)计算出 $\mathrm{rec}(u,i)'$的值。选出推荐度最大的前 N 个商品作为用户 u 的 top-N 推荐集。算法的具体描述如下:

算法:考虑年度日程表事件的时间加权协同过滤推荐算法

输入:目标用户 u,推荐的商品数 n,相似用户数据集 M_u,最相似用户个数 k。

输出:目标用户的 n 个推荐商品。

过程:

(1)开始算法。

(2)计算目标用户 u 和任一用户($a\in M_u$)的相似性 $sim(u,a)$。得到它的 k 个最相似的用户 a_1 a_2 a_3 … a_k,合并相似用户购买过的商品生成资源推荐集 I。

(3)从 I 中删除目标用户 u 已经购买过的商品,得到候选资源推荐集 $candidateI$。

(4)对每个商品 $i\in candidateI$,根据公式(3)、(4)和(5)计算时间权重 $WDY(u,i)$。

(5)对每个商品 $i\in candidateI$,计算 i 对目标用户 u 的加权推荐度 $\mathrm{rec}(u,i)'$。

(6)将 $candidateI$ 中的商品按加权推荐度大小排列,其中最前的 n 个商品作为用户 u 的 top-N 资源推荐集。

(7)结束算法。

3　实验结果及分析

3.1　实验数据

用 SQL Server2005 中,AdventureWorksDW 示例数据仓库中的数据作为测试数据来对提出的考虑年度日程表事件的时间加权协同过滤算法与传统的基于用户的协同过滤算法进行比较。AdventureWorks DW 示例数据仓库所基于的虚构公司 Adventure Works Cycles 是一家大型跨国自行车生产公司。主要商品包括:Adventure Works Cycles 公司生产的自行车;自行车组件(车轮、踏板或刹车部件等);从供应商购买的转售给 Adventure Works Cycles 客户的自行车装饰;从供应商购买的转售给 Adventure Works Cycles 客户的自行车附件(头盔、护膝、运动衫等)。使用的数据见表 3。

表 3　实验数据

FactInternetSales 事实表	60 398 条网上销售记录
DimProduct 维表	606 件商品
DimCustomer 维表	18 484 个用户
DimTime 维表	时间跨度从 2001-07-01 到 2004-08-31

实验是为了比较提出的算法和传统算法的推荐精度,最终选择 1 021 个购买商品数量总数超过 10 的用户共计 14 125 条记录作为训练集。由于测试数据分为训练集和测试集,把训练集中购买商品总数排在前 20 位的 20 个用户在 2004 年的购买记录作为测试集,其余的购买记录作为训练集。训练集用于构建用户-购买矩阵 $\boldsymbol{R}$ 和进行用户相似度计算。

因为商品是面向北美、欧洲和亚洲市场,所以选取一些国际性的节日作为年度日程表事件,如:元旦(1 月 1 日)、西方情人节(2 月 14 日)、国际劳动妇女节(3 月 8 日)、国际劳动节(5 月 1 日)、母亲节(5 月第二个星期日)、国际儿童节(6 月 1 日)、圣诞节(12 月 25 日)等等。

3.2　评价标准

在实验过程中,把用户的访问时间设定为属于 2004 年某个日程表事件的一个时间值,根据每个用户在训练集中的购买记录为其计算 Top-N 推荐集,如果 Top-N 推荐集中某件商品出现在当前用户测试集中的属于这个年度日程表事件时间段购买的商品记录中,则表示生成了一个正确推荐。用信息检索领域中评估系统效果的准确率(Precision)标准作为对比传统算法和提出的算法推荐精度的标准[8]:

$$Precision=\frac{Hits}{N} \tag{7}$$

其中,$Hits$ 表示算法产生的正确推荐数,N 表示算法生成的推荐总数。在文中最终对于某个用户的推荐精度是在各个年度日程表事件中算得的 $Precision$ 值的平均值。

3.3　实验结果

在测试中,实验结果显示 a=0.8、b=0.9 时算法性能达到最优,因此在随后的测试过程中就设定 a=0.8、b=0.9、商品的推荐个数为 10。预测用户对商品的兴趣评分时,参与计算的相似用户的多少影响着算法的 $Precision$ 值。实验中,采取目标用户的相似用户的个数从 5 增加到 40,间隔为 5,查看不同的相似用户个数对预测准确度的影响。

以下的评估数据是根据对实验数据进行统计的结果。用 CF 表示传统的协同过滤算法,用 CF1 表示考虑年度日程表事件的时间加权协同过滤算法。实验结果如图 1 所示。

从实验结果可以看出,在用户的相似用户个数不断变化的过程中,这种考虑年度日程表事件的时间加权协同过滤算法始终要比传统的基于用户的协同过滤算法的准确性高。这种考虑年度日程表事件的时间加权协同过滤算法能够体现出每年特定时期对用户购买行为影响,把用户购买行为存在的相似性利

(下转 154 页)