基于客户行为序列的推荐算法

Recommendation Algorithm Based on Customer Behavior Locus

王卫平　刘 颖　（中国科学技术大学 管理学院 安徽合肥 230026）

摘要：客户的兴趣是不断变化的。但是，目前所广泛应用在推荐系统中的协同过滤算法却是静态的，它只是单纯整合客户的历史数据，并未考虑客户的兴趣变化情况，这必然会导致对高信息量客户的低推荐性能。文中将客户的兴趣度变化考虑在内，提出了一种基于客户行为序列的算法，可以在一定程度上提高针对高信息量客户推荐的性能。

关键词：推荐系统 协同过滤 行为序列 关联规则

1 引言

作为一种新型的智能工具，推荐系统在近年来被越来越多的用在电子商务中。它可以识别消费者的偏好并向消费者推荐其真正感兴趣的产品，从而提高厂商的服务质量，降低客户在选择产品时的信息处理负担。在推荐系统中应用最广泛的算法是协同过滤算法，它通过分析客户的历史数据，生成与目标客户兴趣最相近的邻居集，向客户推荐他们最感兴趣的产品[1]。

虽然当给出足够清楚的偏好信息时，该算法可以表现出良好的性能，但由于客户的兴趣是不断变化的，随着时间的增长，由于该算法的单纯整合客户所有历史数据的静态特性，必然会导致低的推荐性能。

本文将客户兴趣度变化考虑在内，提出了一种基于客户行为序列的新的推荐算法，可在一定程度上解决客户兴趣度变化的问题。

2 基于客户兴趣度变化的推荐算法

本文所提出的算法是基于这样一种考虑：一般来讲，客户购买产品的时间系统是有记录的。设用 L 时长的数据来对客户进行分析。将 L 分为 N 段，已知目标客户前 N－1 段时间的数据，求第 N 段时对目标客户推荐的产品。该算法的实现主要有四个阶段：数据预处理，规则挖掘，产品推荐以及反馈。在预处理阶段，对客户历史数据进行分段，通过概念分层及关联规则来降低数据集稀疏性，建立客户偏好档案（customer profile）。规则挖掘阶段，根据客户偏好档案对客户进行聚类，并根据各个时段的聚类情况得到客户行为序列，从而得到行为序列关联规则。推荐阶段，发现目标客户的行为序列，比较目标客户的行为序列与规则的相似度，并根据所选取的规则进行推荐。反馈阶段，对客户是否点击推荐的产品进行跟踪，将结果反馈给系统，作为下一次推荐的参照。

2.1 数据预处理

协同过滤算法有两个主要的问题——稀疏性和扩展性。所谓稀疏性是指在推荐系统中，每个客户所涉及的信息非常有限，使得客户－项矩阵非常稀疏，推荐性能很差。一般来说，在推荐领域中，一个典型数据集的稀疏度大概有 95% 之多[2]。而本文将时间因素考虑在内，对输入数据根据时间重新分配，势必会导致稀疏度的急剧增加。因此，将输入数据根据时间分段后，首先要解决的即稀疏性问题。

本文采用了两种方法来解决稀疏性问题：概念分层和关联规则。

（1）概念分层。概念分层是数据挖掘中的一种数据预处理方法。一个概念分层定义一个映射序列，将低层概念映射到更一般的高层概念[3,4]。对于整个产品集来说，可以分为几个不同的子集，而每个子集又可以分为更小的子集或具体产品的叶结点。比如，一个商店的产品可分为食品，化妆品，生活用品等，而化妆品又可以分为香水和护肤品等，而护肤品又可以再进一步地细分为面霜，眼霜等。将低层的客户－项矩阵上卷到高层，自然可以降低稀疏性。在进行分析时，应

根据经验选择合适的层次进行分析。

(2) 关联规则。此处所使用的关联规则主要用来挖掘产品的联系程度,从而进行有关产品的得分评定,以降低稀疏度。本算法中,关联规则的挖掘主要使用了购物篮数据和购买数据。在概念分层的第 i 层上,对于购物篮数据和购买数据,在时段 N-k(k=0,1,2……,N-1, N≥2)进行如下操作:分别计算频繁 2-项集,并产生规则体(body)和规则首部(head)均只有一项的关联规则,记为 R_b, R_p。最后,合并两个关联规则集,记为 R_{all}。在计算 R_{all} 时,如果 R_p 和 R_b 中出现相同的规则,则采取置信度最高的那条规则[5]。

定义客户偏好档案 C = (c_{ij}) 如下:

$$c_{ij}\begin{cases}1 & j\in PurSet(i)\\ conf(*\Rightarrow j) & j\in AssoSet(i)-PurSet(i)\\ 0 & otherwise\end{cases}$$

i = 1,2,…,m,m 指客户数 j = 1,2,…,n,n 指产品数

PurSet(i) 指目标客户已经购买的产品集,若在概念分层的最低层,是指具体产品;若不是在最低层,则是指产品类。

AssoSet(i) 指将 PurSet(i) 应用在 R_{all} 上时,可以推导出来的产品集。

conf(*⇒j) 指在 R_{all} 中以 j 为首部的关联规则的最大置信度。

对每个时段,根据输入数据,利用以上规则都可以得到相应的客户档案。从而对每个客户,都可以得到其不同时段的客户档案。

2.2 规则挖掘

对每个时段所得到的客户档案进行自组织映射(Self-organizing Map)聚类,可以得到如下有 q 个类的聚类集合:C = {C_1, C_2, ……, C_q},其中每个类代表一组具有相似行为模式的客户。对这些类根据客户和时段进行重新排列,可以得到客户在总时长 L 的行为序列,它是在 N 个时段内某一客户所属类别的变化情况,如下:

$L_i = <C_{i,1},\cdots,C_{i,N-1}, C_{i,N}>$

其中 i = 1,2,…,m $C_{i,N-k}\in C$, k=0,1,2……,N-1, N≥2

单纯的客户行为序列并没有什么意义,必须从所有客户行为序列中发现相同的模式。这个模式发现的过程可以利用关联规则来完成。将客户在这 N 个时段的行为序列分为条件部分和结论部分:关联规则的条件部分由 L_i 中的 $<C_{i,1},\cdots,C_{i,N-1}>$ 组成,而结论部分为 $C_{i,N}$,由客户在最后时段 N 所属类别组成。对所有的客户行为序列进行分析,从而得出行为序列关联规则的支持度和置信度。最后,所得出的行为序列关联规则可表示成如下形式:

R_i: $r_{i,1},\cdots,r_{i,N-1}\Rightarrow r_{i,N}(Sup_i, Conf_i)$ $r_{i,N-k}\in C$ 或 Φ, $r_{i,N}\in C$

该序列关联规则的意思是:如果一个目标客户的行为序列为 $r_{i,N-L+1},\cdots,r_{i,N-1}$,那么其在最后一个时段 N 所属类别是 $r_{i,N}$。

2.3 产品推荐

在此阶段,对于给定的目标客户,推荐与其行为序列最匹配的产品。当给定一个目标客户时,首先要采取与规则挖掘阶段相同的方法找出该客户的行为序列,然后计算与该客户行为序列最相近的规则,并根据该规则进行推荐。

2.3.1 行为序列寻找及规则匹配

对于给定的目标客户,首先要利用与规则挖掘阶段相似的方法挖掘出该客户在前 N-1 个时段的行为序列。然后利用以下度量值判断行为序列与规则的相似程度。

(1) 相似性度量

设目标客户 i 在过去 N-1 时段的行为序列为:$L_i^c = <C_{i,1},\cdots,C_{i,N-1}>$,规则 j 的条件部分为:$R_j^c = <r_{j,1},\cdots,r_{j,N-1}>$。则 L_i^c 和 R_j^c 之间的相似度为:

$$SM_i^j=\sum_{k=1}^{N-1}S_{i,N-k}^j$$

$$S_{i,N-k}^j=\begin{cases}1-(N-1-k)*\alpha & \text{if } C_{i,N-k}=r_{j,N-k}\\ 0 & otherwise\end{cases}$$

其中参数 $0\le\alpha<1/(N-2)$,是由经验决定的调整参数,当客户行为序列在某个时段所属类别与规则在此时段所属类别相等时,该时段离推荐时段越近,给予的权重就越大。

但是,仅仅依靠相似性度量还不足够判定目标客户所属的最终类,因为有不同支持度和置信度的关联规则有着不同的通用性。因此,还必须将规则的通用性考虑在内。

(2) 实用性度量

客户行为序列 i 和规则 j 之间的实用性度量定义如下：

$FM_i^j = SM_i^j * Sup_j * Conf_j$

目标客户在最后时段 N 所属的类别即是有最大 FM_i^j 的关联规则 j 的结论部分：$r_{I,N}$

2.3.2 Top－N 产品推荐

设 C＊为某目标客户在最后时段所属的类。由于上述分析是在概念分层的第 i(i = 0,1,……)层进行，所以 C＊中既可能含具体的产品，也可能是某个产品类。而对客户进行推荐时，只需要推荐具体的产品即可。

对于该目标客户，寻找在第 N 个时段中与其同在 C＊的所有客户，并得到这些客户在这个时段购买的所有产品的集合 P。去掉产品集中目标客户已经购买过的产品，然后对产品集中剩下的产品进行评分。对产品集中的某个产品 p：

$score_p = \sum_{i=1}^{m} C0(i,p)$　　m 表示 C＊中的客户数

C0 是将在第 1.1.2 节得到的第 0 层的关联规则应用于 C＊中的客户时，根据其在第 N 个时段所购买的产品而得到的客户偏好档案，它反应了目标客户及其所有邻居在第 N 个时段对产品的偏好程度。

得出所有需要的产品评分后，对这些产品评分按大小排列，从中选取 N 个对目标客户进行推荐。

2.4 反馈

该部分主要是建立一张产品点击表，对客户是否点击为其推荐的产品进行跟踪。如果客户未点击为其推荐的某种产品，说明他对这种产品可能并无兴趣。因此，在下一次进行推荐时，如果进行评分的产品集 P 中包括这些顾客未点击产品，则在利用上述方法初步计算所有评分后，对这些未点击产品的评分进行修改：

$score_x = score_x * \beta$，$\beta < 1$，由经验决定

3 实验评估

在实验中使用 Movielens 数据集[6]数据集分为两个层次，电影共 19 个类。一部电影属于一个或多个类。

由于只有对高信息量客户才能挖掘出其兴趣度变化情况，因此首先要对数据集中的数据进行筛选。选取从 1997 年 9 月 19 日到 1998 年 4 月 9 日为进行分析的时间段，将其等分为四个时间段。只有在每个时间段中都至少对电影做过一次评分的客户，才能成为分析的对象，对客户在第四个时段的购买进行推荐。

首先将客户的评分转换为二元表。方法如下：对每一评分，如果高于该客户的平均评分，则转换为 1，否则为 0。首先对测试集中的数据进行分析，挖掘出行为序列关联规则。然后找出测试集中目标客户在前三个时段的行为序列，再根据行为序列关联规则，找出该客户在第四个时段所属的类，并根据与其同类的客户的情况对该客户进行推荐。本次实验采取五折交叉确认。在测试中对每一个目标客户隐藏 n 项偏好，在剩余偏好的基础上推荐 N 个产品。如果推荐的某个项目是隐藏的某个项目，则称为一次 hit。本次实验中，n = 3，N = 12。时序关联规则的支持度和置信度阈值分别设为 0.01，0.3。

本次实验以 F1 值作为评价算法性能的依据。其定义如下：

准确率 = hit/推荐产品总数

召回率 = hit/在推荐时段所购买的产品总数

F1 = (召回率 × 准确率)/((召回率 + 准确率)/2)

本次实验的结果如图 1 所示。

图中横线表示协同过滤算法的 F1 值，折线表示对于不同的类别个数，基于客户行为序列的推荐算法的 F1 值。

由图 1 可以得出，当选取合适的类别数目时，对于高信息量客户，本文所提出的算法相对于协同过滤算法而言有着更好的性能。因此，在实际应用中，选取合适的类别数目是该算法成功的一个关键因素。

4 结语

本文提出了一种将用户兴趣度变化考虑在内的推荐算法。它首先使用了概念分层和关联规则来降低数据集的稀疏性，然后根据顾客的行为序列挖掘行为序列关联规则，并根据目标客户的行为序列与行为序列关联规则的匹配程度对客户进行推荐。最后根据客户的反馈对下一次的推荐结果进行了修正。通过实验分析，可见当选取适当的聚类数目时，本文所提出的算法相对协同过滤算法来讲有着更好的推荐性能。

由于本文的实证分析是在一个比较小的数据集上进行的，因此它还有待于在实际中进行检验。另外，本

文的算法是一种基于模型的算法，随着时间的增长，模型的性能必然会下降。如何处理模型性能的下降是今后一个亟待解决的问题。最后，从实验结果可以看出，虽然该算法的性能比协同过滤算法有了一定程度的改进，但是其离满足客户的实际需求还是有一定的差距，必须要进一步提高推荐性能。

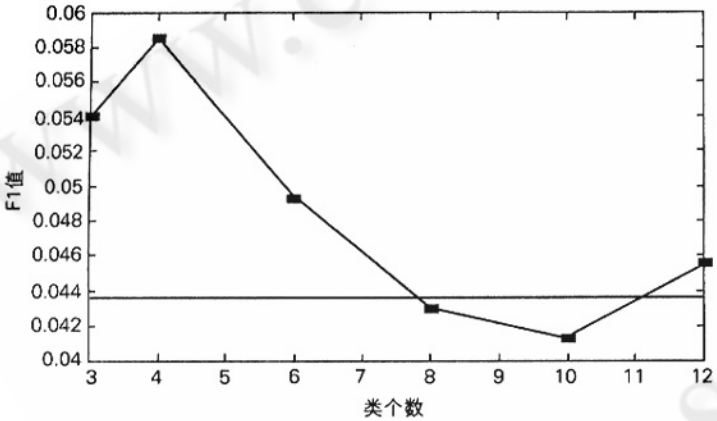

图 1　两种推荐算法在 Movielens 数据集上的性能比较

参考文献

1 YU LI, LIU LU, LI XUEFENG. A hybrid collaborative filtering method for multiple – interests and multiple – content recommendation in E – Commerce [J]. Expert Systems with Applications, 2005(28) 67 – 77.

2 YEONG BIN CHO, YOON HO CHO, SOUNG HIE KIM. Mining changes in customer buying behavior for collaborative recommendation [J]. Expert Systems with Applications, 2005(28) 359 – 369.

3 HAN JW, KAMBER M. Data Mining: Concepts and Techniques [M]. San Mateo, CA: Morgan Kaufmann, 2000.

4 KIM C, KIM J. A recommendation algorithm using multi – Level association rules [A]. Proceedings of the IEEE/WIC International Conference on Web Intelligence (WTO3) [C], 2003.

5 JAE KYEONG KIM, YOON HO CHO, WOO JU KIM, JE RAN KIM, JI HAE SUH. A personalized recommendation procedure for Internet shopping support [J]. Electronic Commerce Research and Applications, 2002(1) 301 – 303.

6 http://www.cs.umn.edu/Research/GroupLens/index.html.