基于项目因子分析的 Web 客户需求协同推荐算法①

赵宏霞[1]，王新海[2]，杨皎平[2]

[1](辽宁工程技术大学 营销管理学院，葫芦岛 125105)
[2](辽宁工程技术大学 工商管理学院，葫芦岛 125105)

摘　要：为解决协同过滤推荐算法中的数据量过大和数据稀疏性的问题，提出了基于项目因子分析的协同推荐算法。该算法通过采用因子分析将项目向量降维为几个具有代表性的项目因子，然后用这些项目因子对目标项目进行回归分析，进而预测目标客户对待评项目的评分。最后通过实验证明了算法的有效性，为以后研究推荐算法提供了一种新的途径。

关键词：电子商务；推荐系统；协同过滤；项目；因子分析

Collaborative Recommendation Algorithm of Web Customer Demand Based on Item Factor Analysis

ZHAO Hong-Xia[1], WANG Xin-Hai[2], YANG Jiao-Ping[2]

[1](School of Marketing Management, Liaoning Technical University, Huludao 125105, China)
[2](College of Business Administration, Liaoning Technical University, Huludao 125105, China)

Abstract: In order to solve the problem that data overload and data sparsity in collaborative filtering recommendation algorithm, the collaborative recommendation algorithm based on item factor analysis is proposed in this paper. The algorithm reduces the dimensions of item vector by use of factor analysis and gets some representative item factors, which are used to regression analysis of target items to forecast the evaluated items. Finally, experiments show that the algorithm is effective, which provides a new way for future recommendation algorithm research.

Key words: E-business; recommendation system; collaborative filtering; item; factor analysis

协同过滤是目前电子商务推荐系统中应用最广泛、最成功的推荐技术[1]。协同推荐技术主要分为 3 类：基于用户的协同过滤[2,3]、基于项目的协同过滤[4-6]、前两类技术的综合应用[7]。其中基于项目的协同推荐在 Amazon.com 为代表的大型电子商务网站中得到了广泛的应用[8]。

基于项目的协同过滤推荐基于假设：如果大部分用户对一些项目的评分比较相似，则当前用户对这些项目的评分也比较相似。基于项目的协同过滤推荐使用统计技术找到目标项的若干最近邻居。由于当前用户对最近邻居的评分与目标项目的评分比较类似，所以可以根据当前用户对最近邻居的评分预测当前用户对目标项目的评分，然后选择预测评分最高的前若干项作为推荐结果反馈给用户[4,5]。

Sarwar 等学者[5]通过实验认为基于项目的协同过滤推荐方法比传统的基于用户的协同过滤推荐方法具有更高的推荐质量，这是因为项目之间的相似性较用户之间的相似性更为稳定[6]，并且基于项目的协同过滤方法可以采用离线计算的方法，从而节省了系统开销。

但基于项目的协同过滤推荐算法同样面临数据的高稀疏性和预测精度问题，很多学者提出了改进的策略。张海鹏等[9]提出了基于项目分类预测的协同过滤推荐算法；龚瑞君等[10]提出了复合项目的概念来解决数据的稀疏性；李聪等[11]和邵伟等[12]通过考虑项目类别进而改善了项目相似性计算的准确度。

① 基金项目:辽宁省教育厅科学技术研究项目(W2010212);教育部博士点基金项目(200801470004);教育部人文社科基金(10YJC630407)
收稿时间:2010-11-08;收到修改稿时间:2010-12-05

以上学者对项目的分类往往基于先验知识和基于距离的方法，实际上决定项目分类的还有消费行为和消费兴趣，如赵宏霞等[13]将商品分为时尚型商品、实用性商品、廉价型商品等，这些分类的依据往往是隐式的，需要采用一定的方法进行显性化处理。

为此，本文提出了项目因子分析的协同推荐算法，该算法假设不同商品项目是不同商品隐含属性的外在表现，纷纭复杂的商品其实是若干商品属性的表现形式。该算法首先对所有商品项目进行因子分析，得到若干(K 个)公共因子(决定商品项目的内在属性)，然后用该 K 个因子和拟推荐给用户的商品进行多元线性回归，从而预测拟推荐商品的评分，最后将预测评分最高的前若干商品作为推荐结果。

1 基于项目的协同过滤算法

1.1 问题描述

在基于协同过滤推荐算法的推荐系统中，用户评分数据库中包括 m 个用户的集合 $U=\{u_1,u_2,\cdots,u_m\}$和 n 个项目的集合 $I=\{I_1,I_2,\cdots,I_n\}$。用户对项目的评分数据可以采用一个 m×n 阶的用户-项目评分矩阵 $R=(r_{ij})_{m\times n}$ 来表示，如表 1 所示。

表 1 用户-项目评分矩阵

	I_1	…	I_t	…	I_n
u_1	r_{11}	…	r_{1t}	…	r_{1n}
…	…	…	…	…	…
u_s	r_{s1}	…	r_{st}	…	r_{sn}
…	…	…	…	…	…
u_m	r_{m1}	…	r_{mt}	…	r_{mn}

其中，评分表示用户对项目感兴趣的程度，评分的级别越高，说明用户越感兴趣。

1.2 项目的相似性

项目的相似性表示用户对项目同时感兴趣的程度。选择不同用户对项目 i 和项目 j 的共同评分数据来计算项目 i 和项目 j 之间的相似性 sim(i,j)，表示如下：

$$sim(i,j)=\frac{\sum_{u\in\Pi_{ij}}(r_{ui}-\bar{r}_i)(r_{uj}-\bar{r}_j)}{\sqrt{\sum_{u\in\Pi_{ij}}(r_{ui}-\bar{r}_i)^2}\sqrt{\sum_{u\in\Pi_{ij}}(r_{uj}-\bar{r}_j)^2}} \quad (1)$$

其中，U(i)表示对项目 i 评分的用户集合，U(j)表示对项目 j 评分的用户集合，u 属于 U(i)和 U(j)的交集，表示对项目 i 和项目 j 共同评分的用户，$\bar{r}_i$ 和 $\bar{r}_j$ 表示项目的评分均值。分别表示如下：

$$\bar{r}_i=\frac{\sum_{u\in\Pi_{ij}}r_{ui}}{|\Pi_{ij}|},\quad \bar{r}_j=\frac{\sum_{u\in\Pi_{ij}}r_{uj}}{|\Pi_{ij}|}$$

其中，$|\Pi_{ij}|=|U(i)\cap U(j)|$表示对项目 i 和项目 j 共同评分的用户个数。

1.3 项目的评分预测

对目标用户 g 的待评分项目 s，选择项目 s 的最近邻居项目集合 Nei(s)，Nei(s)中的项目既可以是与项目 s 相似性排在前几位的项目，也可以是与项目 s 相似性大于某个阈值的项目。

根据 s 的邻居项目集合 Nei(s)，预测目标用户 g 对 s 的评分。预测公式如(2)所示。

$$\hat{r}_{gs}=\bar{r}_s+\frac{\sum_{i\in Nei(s)}sim(s,i)\times(r_{gi}-\bar{r}_i)}{\sum_{i\in Nei(s)}sim(s,i)} \quad (2)$$

2 因子分析相关理论

2.1 因子分析模型

对于标准化的数据矩阵 X_{mn}:

$$\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&\ddots&\vdots\\x_{m1}&x_{m2}&\cdots&x_{mn}\end{bmatrix}$$

其中，$\boldsymbol{X}_n=(x_{1n},x_{2n},\cdots,x_{mn})^{\mathrm{T}}$，该数据面临的因子分析的数学模型为[14]：

$$\begin{cases}\boldsymbol{X}_1=a_{11}\boldsymbol{F}_1+a_{12}\boldsymbol{F}_2+\cdots+a_{1k}\boldsymbol{F}_k+e_1\\\cdots\\\boldsymbol{X}_n=a_{n1}\boldsymbol{F}_1+a_{n2}\boldsymbol{F}_2+\cdots+a_{nk}\boldsymbol{F}_k+e_n\end{cases} \quad (3)$$

其中，k<n,从而达到降维的目的；F_j 是公共因子，它们之间是两两正交的，e_i 是特殊因子，a_{ij} 是公共因子的负载。

2.2 因子值的求法

在因子分析中，常常需要利用公共因子来进一步的研究，例如用公共因子做回归分析等，这样需要计算因子值。假设第 j 个公共因子的因子值 F_j 可以由 X_1，X_2，…，X_n 的样本值计算出来，即：

$$\boldsymbol{F}_j=\begin{bmatrix}f_{1j}\\f_{2j}\\\vdots\\f_{mj}\end{bmatrix}=\begin{bmatrix}x_{11}&x_{12}&\cdots&x_{1n}\\x_{21}&x_{22}&\cdots&x_{2n}\\\vdots&\vdots&\ddots&\vdots\\x_{m1}&x_{m2}&\cdots&x_{mn}\end{bmatrix}\begin{bmatrix}\beta_{1j}\\\beta_{2j}\\\vdots\\\beta_{nj}\end{bmatrix}$$

该式可以表述为 $F_j=X\beta_j$，两边同乘以 X^T，即得到 $X^TF_j=X^TX\beta_j$。容易证明 $X^TF_j=a_j=(a_{1j},a_{2j},\cdots,a_{nj})T$, X^TX 正好是相关系数矩阵 ρ。因此 $a_j=\rho\beta_j$，进而 $\beta_j=\rho^{-1}a_j$。至此得到公式(4)[14]。

$$\boldsymbol{F}_j=\boldsymbol{X}\rho^{-1}a_j \tag{4}$$

3 基于项目因子分析的协同推荐

对于网络数据库中 m 个客户对 n 个商品的评分矩阵 $R=(r_{ij})_{m\times n}$，以及目标客户 g(不妨设为第 m+1 个客户)对其中 s 件商品(不妨设为前 s 件)的评分。

此时将 R 与目标客户 g 的评分作为一个增广矩阵 $\underline{\boldsymbol{R}}$，$\underline{\boldsymbol{R}}$ 可看作 n 个列向量，其中前 s 个列向量为 m+1 维列向量，后 n-s 个向量为 m 维列向量。

3.1 项目因子分析

对前 s 个列向量进行因子分析，得到 K 个项目因子，视为 K 个项目隐分类。

首先，计算前 s 个列向量中每个项目向量 $\underline{\boldsymbol{R}}_j=(r_{1j},r_{2j},\cdots,r_{m+1j})^T$ 在第 k(k≤K)个因子上的负载向量 a_{jk}。

然后，计算前 s 个列向量中任意两个项目向量 $\underline{\boldsymbol{R}}_i$ 和 $\underline{\boldsymbol{R}}_j$ 的相关系数 ρ_{ij}。

最后，根据式(4)得到第 k(k=1,2, …,K)个项目因子为：

$$\boldsymbol{F}_k=\underline{\boldsymbol{R}}^T\rho^{-1}a_k \tag{5}$$

3.2 目标客户评分值预测

以 m 个客户对第 t 项商品(s＜t≤n)的评分 $r_t=(r_{1t},r_{2t},\cdots,r_{mt})$ 为因变量，以第 k(k=1,2, …,K)个公共因子向量的前 m 个元素构成的向量 $\boldsymbol{F}_{1-m,k}=(f_{1k},f_{2k},\cdots,f_{mk})$ 为自变量，进行多元线性回归：

$$r_t=w_1\boldsymbol{F}_{1-m,1}+w_2\boldsymbol{F}_{1-m,2}+\cdots+w_K\boldsymbol{F}_{1-m,K}+C+\varepsilon \tag{6}$$

首先，采用最小二乘法或其它估计方法，得到回归系数 $\hat{w}_k(k=1,2,\cdots,K)$，进而得到：

$$\hat{r}_t=\hat{w}_1\boldsymbol{F}_{1-m,1}+\hat{w}_2\boldsymbol{F}_{1-m,2}+\cdots+\hat{w}_K\boldsymbol{F}_{1-m,K}+\hat{C} \tag{7}$$

然后，将第 k(k=1,2, …,K)个公共因子的第 m+1 项 $f_{m+1,k}$ 代入式(7)得到第 g=m+1 个客户对第 t 项商品的评分 $\hat{r}_{g,t}$。

类似得到客户 g 对后 n-s 件商品的评分：$\hat{r}_{g,s+1},\hat{r}_{g,s+2},\cdots,\hat{r}_{g,n}$。

最后对于预测评分值($\hat{r}_{g,s+1},\hat{r}_{g,s+2},\cdots,\hat{r}_{g,n}$)，满足 $\hat{r}_{gj}>\xi$ (j=s+1,s+2, ,…n)所对应的商品为客户 g 感兴趣的商品。

3.3 算法描述

该方法对应的算法描述为：

1) 输入

① 客户-商品矩阵，m 个客户对 n 个商品的评分 $R_{m\times n}$；

② 目标客户 g，该客户已评价商品 I_g，商品评分阈值 ξ。

2) 输出

目标客户的感兴趣的商品。

算法过程：

预测客户 g 对第 k 个商品的评价分。

① 在线下根据式(3)对|Ig|个项目评分向量(m+1 维向量)进行因子分析，并根据式(5)求出 K 个公共因子 f_1，f_2，…，f_K(每个因子均为 m+1 维向量)。

② 以 m 个用户对第 k 个商品评分值为因变量，以前 m 个用户对应的公共因子值为自变量，根据式(6)进行多元线性回归，进而得到式(7)的表达式。

③ 将客户 g 已评分的商品对应的公共因子值代入式(7)，得到客户 g 待评价商品 k 的评分预测值 $\hat{r}_{gk}$。

④ 对于预测评分值 $\hat{r}_{gk}>\xi$ 所对应的商品为客户 g 感兴趣的商品。

该算法只需要实时地计算多次多元线性回归运算和相应的预测，其余的有关因子分析等运算均可以离线进行，从而节省了系统开销；另外将 s 个客户降维为 K 个项目因子，可以减少数据的稀疏性。

4 算例和实验

为了清晰说明本算法，下面提供如表 2 所示的客户对不同电影的评价数据，其中客户对不同电影的评价值为 1-5 分。现在拟将电影 9，10，11 三部电影中部分电影推荐给客户 7。

表 2　客户-电影评分数据

电影	1	2	3	4	5	6	7	8	9	10	11
客户 1	5	2	2	4	3	2	4	2	2	5	4
客户 2	4	2	2	5	3	1	3	1	1	4	4
客户 3	5	1	3	5	3	2	5	2	1	5	5
客户 4	1	4	5	1	4	4	2	4	4	2	2
客户 5	2	5	5	2	4	3	2	5	5	3	1
客户 6	3	4	4	1	5	5	1	4	5	3	3
客户 7	4	3	3	4	4	3	3	3	?	?	?

对前 8 列数据对应 7 个行向量进行因子分析，得到

两个因子 F_1，F_2，F_3，并得到 3 个因子值如表 3 所示。

表 3 项目因子值

电影	F_1	F_2	F_3	9	10	11
客户 1	-0.72	-0.30	0.50	2	5	4
客户 2	-1.05	-1.42	-1.13	1	4	4
客户 3	-0.03	-0.20	1.93	1	5	5
客户 4	1.30	-0.01	-0.36	4	2	2
客户 5	1.49	-0.37	-0.39	5	3	1
客户 6	-0.51	1.88	-0.68	5	3	3
客户 7	-0.47	0.43	0.13	?	?	?

分别以项目 9，10，11 的评分为因变量，以 F_1，F_2，F_3 为自变量，得到如下三个回归方程：

$$\hat{r}_{79} = 2.98 + 1.04\times F_1 + 1.08\times F_2 - 0.77\times F_3$$

$$\hat{r}_{7,10} = 3.72 - 0.72\times F_1 + 0.33\times F_2 + 0.70\times F_3$$

$$\hat{r}_{7,10} = 3.26 - 1.08\times F_1 - 0.17\times F_2 + 0.75\times F_3$$

区间内预测平均误差为 0.4280；区间外预测得到：$\hat{r}_{79} = 2.85$，$\hat{r}_{710} = 4.29$，$\hat{r}_{711} = 3.79$。

采用传统的协同过滤推荐方法，区间内评价误差为：0.5325，区间外预测得到：$\hat{r}_{79} = 3.20$，$\hat{r}_{710} = 3.76$，$\hat{r}_{711} = 3.69$。

从该算例可以看出，基于因子分析的协同推荐算法比传统的协同过滤推荐算法更精确，这是因为，基于因子分析的方法中，这些因子总体代表了几乎所有已评项目的信息，而传统方法选择少数邻居项目，必然丢掉了很多项目的信息。这与算法的平均绝对偏差 MAE 大体随邻居项目的增加而减少是一致的。

为了说明这一点，下面采用 GroupLens 研究项目组搜集的公共数据集 MovieLens[12]进行实验，数据集中包括由 943 个用户对 1682 部电影的 100000 个评分，评分范围为 1-5 分，每个用户至少评过 20 部电影。选取 430 个客户的 3 万条评分数据，并按照 4:1 的比例划分训练集和测试集。

为了比较传统的基于项目的协同过滤推荐(item-based CFR)算法和本文提出的基于项目因子协同推荐(item-Factor Analysis CR)算法的精确度，采用平均绝对偏差 MAE 作为度量标准。其中传统算法将目标项目最近邻居个数从 5 增加到 40(间隔为 5)，基于因子分析的方法则将因子个数从 5 增加到 40(间隔同样为 5)。将预测结果如图 1 所示。

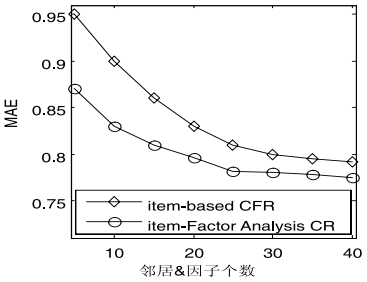

图 1 推荐算法的平均绝对误差 MAE 比较

从图 1 看出，基于项目因子分析的协同预测方法比传统的基于项目的协同过滤算法精确度更好。

5 结语

文章提出了基于项目因子分析的协同推荐算法，算法将项目向量利用因子分析进行降维得到几个具有代表性的项目因子，然后用这些项目因子对目标项目进行回归分析，进而预测目标客户对待评项目的评分。

算法除了线性回归运算外，其余运算均可离线进行，从而大大节省了系统开销；而且，该算法通过项目因子保证了信息的最小丢失，从而具有较高的预测精度；此外，该方法通过降维处理，大大减少了评分矩阵的稀疏性。

最后文章通过实验证明了该算法的有效性，为以后研究推荐算法提供了一种新的途径。

参考文献

1 吴湖,王永吉,王哲,等.两阶段联合聚类协同过滤算法.软件学报,2010,21(5):1042−1054.

2 Sarwar B, Kaypis G, Konstan J, Riedl J. Analysis of Recommendation Algorithms for E-Commerce. Proceedings of the 2nd ACM conference on Electronic commerce. New York: ACM Press, 2000.158−167.

3 李聪,梁昌勇,马丽.基于邻域最近邻的协同过滤推荐算法.计算机研究与发展,2008,45(9):1532−1538.

4 Kaypis G. Evaluation of item-based top-n recommendation algorithms. Proc. of the 10th International Conference on Information and Knowledge Management. New York: ACM Press, 2001.247−254.

5 Sarwar B, Kaypis G, Konstan J, Riedl J. Item-based collaborative filtering recommendation algorithms. Proc. Of

(下转第 210 页)

码就可以独立于对方而修改，做到了客户端模型可以和服务器端模型可以配合工作。

但是，该方法的主要局限是它将解析数据的负担完全交给了客户端。因此客户层的代码会变得更加复杂，如果在大型的应用中采用这种方法，则可以通过重用解析器代码或将一些功能抽象为库的方法来解决。

2.4 用 Ajax 与服务器交互

当需要保存用户在 wiki 页面上的一次修改或者提交新的请求时候触发 Ajax 请求，从客户端 JavaScript 代码角度来看，读取和更新的区别是细微的，我们仅仅需要指定使用 POST 方法，并且传入模型所对应的 XML 字符串就可以了。

Ajax 请求发生在需要和数据库(因此也跟服务器)通信的时候。如果 Controller 监听到绑定在 View 上的事件已触发，Controller 就开始与服务器交互，但是根据 MVC 法则，Controller 是通过 Model 与服务器通信的。如代码 4 所示：

```
Wiki.Controller.prototype = {
callServer:function(xmlData){this.model.callServer(
xmlData);},
}
```

代码 4 Controller 通过 Model 向服务器发送请求

Model 传递给服务器端的是一个 XML 字符串，这样在服务器接受到这个字符串之后可以直接根据定义的 ORM 模板反序列为服务器端的业务模型对象，实现了两个模型之间的同步通信。在服务器跟数据库通信完成之后，再次返回一个 XML 字符串用来更新客户端的界面。Ajax 让我们能够在发送请求的时候就指定回调函数，这样在服务器端进行处理的时候，不会阻止用户的下一步操作。当服务器返回处理之后的结果时，指定的回调函数会自动处理数据，更新用户界面。

3 结论

在上面的文章中，我们利用了 MVC 模式在客户端把用户界面、业务模型和事件处理分离成 3 个不同的部分，以降低应用程序模块之间的耦合性。但在某些情况下，分离可能是不需要的，甚至某些情况下，分离会造成很多不必要的程序冗长。而当我们的应用程序变得越来越复杂，需要 JavaScript 在客服端的多数部分的交互操作的时候，我们把 JavaScript 分离进入 MVC 模式能够产生出更多元化，更重复利用的代码。

参考文献

1 孙卫琴.精通 Structs 基于 MVC 的 Web 设计与开发.北京:电子工业出版社, 2004.

2 黎永良,崔杜武.MVC 设计模式的改进与应用.计算机工程,2005,31(9):96–98.

3 李园,陈世平.MVC 设计模式在 ASP.NET 平台中的应用.计算机工程与设计, 2009,30(13):3180–3184.

4 李学俊,李龙澍,徐怡.新一代网络语言 Wiki.计算机技术与发展, 2007,17(1):85–87.

5 Resion J.陈贤安,江疆译.精通 JavaScript.北京:人民邮件出版社, 2008.

(上接第 191 页)

the 10th International Conference on World Wide Web. New York: ACM Press, 2001.285–295.

6 Sarwar B. Sparsity, scalability and distribution in recommender systems [Ph.D. Thesis]. Minneapolis: University of Minnesota, 2001.

7 汪静,印鉴,郑利荣,黄创光.基于共同评分和相似性权重的协同过滤推荐算法.计算机科学,2010,37(2):99–103.

8 Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing, 2003,7(1):76–80.

9 张海鹏,李列彪,李仙等. 基于项目分类预测的协同过滤推荐算法.情报学报,2008,27(2):218–223.

10 龚瑞君,王佳,戴珺等.基于两阶段聚类的协作过滤推荐算法.郑州大学学报(理学版),2010,42(1):14–16.

11 李聪,梁昌勇,董珂.基于项目类别相似性的协同过滤推荐算法.合肥工业大学学报(自然科学版),2008,31(3):360–363.

12 邵伟,袁方,张瑜.融入项目类别信息的协同过滤推荐算法.数学的实践与认识,2010,40(6):108–112.

13 赵宏霞,杨皎平,陈宗娇.面向客户需求的神经网络挖掘方法.管理评论,2005,17(11):53–57.

14 马庆国.管理统计:数据获取、统计原理.SPSS 工具与应用研究.北京:科学出版社,2002.315–326.