基于离散评分向量概率分析的CF算法改进研究

田 伟[1] 许 静[1] 彭玉青[2]
(南开大学信息技术科学学院 天津 300071)[1]
(河北工业大学计算机与软件科学学院 天津 300130)[2]

摘 要 协同过滤(CF)个性化推荐算法过程中,用户相似度计算是CF技术的核心问题之一。以用户显式、离散评分的条件为基础,应用概率论分析方法,考察了用户显式离散评分向量,提出了改进方法:全面均值法结合对商品进行分类推荐的改进算法。分析表明这种方案更适合推荐系统实际应用环境。实际数据实验表明,新方法提高了CF推荐的预测精度和推荐质量。

关键词 电子商务,推荐系统,协同过滤,全面均值,个性化
中图法分类号 TP181 **文献标识码** A

Research on CF Algorithm Based on Probabilistic Analysis of Discrete Rating Vector

TIAN Wei[1] XU Jing[1] PENG Yu-qing[2]
(College of Information Technical Science, Nankai University, Tianjin 300071, China)[1]
(Department of Computer and Software Science, Hebei Industrial University, Tianjin 300130, China)[2]

Abstract The users' similarity computation is a key step of Collaborative Filter (CF) algorithm. The All-Average method and classified recommendation improved algorithm based on probabilistic analysis of users' discrete explicit rating vector were proposed to solve the problem of CF sparsity and other practical problems. Experimental result shows that the improved method enhances the precision and quality of CF prediction.

Keywords Electronic commerce, Recommendation system, Collaborative filter, All-average, Personalized

1 引言

到目前为止,在电子商务的个性化推荐方面已有不少的实现技术和算法,其中一般认为协同过滤(Collaborative Filter,CF)是目前研究最多、较为成功的推荐技术。而稀疏性问题是CF技术面临的重要问题[4]。这是由于在实际电子商务使用平台上,一般各个用户只会对很少的项目作出评价,这样会造成整个用户-项目评分矩阵非常的稀疏,即使用户的评价只局限于一类商品(比如电影),评分矩阵的稀疏性也较高。数据实际情况如图1所示。

一些研究提出了将高稀疏的用户-评分矩阵进行奇异值分解(SVD)[1]或矩阵划分的推荐算法,在一定程度上解决了稀疏性问题,但其预测精度较低和训练成本较大。对于CF实际应用,应以简便方法、较低时间复杂度解决稀疏等问题。

本文在研究分析CF算法的基础上,提出了全面均值及分类推荐法的改进算法。在计算用户相似性时,按不同项目类综合考虑用户间已评分与未评分项目,降低了推荐算法计算用户间相似度的时间复杂度,适合实际稀疏数据环境。经实际数据试验验证,本算法提高了CF计算的预测精度和推荐质量。

MovieLens 用户—电影评分矩阵: (具有高稀疏性)

用户	电影:	1	2	3	4	5	6	7	8	9
1	评分:	5	0	0	0	0	0	0	0	0
2	评分:	0	0	0	0	0	0	0	0	0
3	评分:	0	0	0	0	0	0	0	0	0
4	评分:	0	0	0	0	0	0	0	0	0
5	评分:	0	0	0	0	0	2	0	0	0
6	评分:	4	0	0	0	0	0	0	0	0
7	评分:	0	0	0	0	0	4	0	0	0
8	评分:	4	0	0	3	0	0	0	0	0
9	评分:	5	0	0	0	0	0	0	0	0
10	评分:	5	5	0	0	0	0	4	0	0
11	评分:	0	0	0	0	0	0	0	0	0
12	评分:	0	0	0	0	0	0	0	0	0
13	评分:	0	3	0	0	0	0	0	0	0
14	评分:	0	0	0	0	0	0	0	0	0
15	评分:	0	0	0	0	0	4	0	0	0
16	评分:	0	0	0	0	0	0	0	0	0
17	评分:	0	0	0	0	0	0	0	0	0
18	评分:	4	2	0	0	0	0	0	0	0

图1 MovieLens评分矩阵情况展示

2 协同过滤推荐算法的改进——全面均值法和项目分类推荐

2.1 对CF用户显式离散评分向量的概率分析

为简化分析,设定每个用户对于各项目的评分数值等概率出现;各用户评分行为是相互独立的。

设 Rt 为用户对项目的评分等级个数(本文设定包括分值

到稿日期:2009-06-03 返修日期:2009-08-20 本文受天津市科技发展计划项目(08ZCKFGX01100)资助。
田 伟(1982-),男,博士生,主要研究方向为智能计算、数据挖掘等,E-mail:tianwei8202@163.com;许 静(1967-),教授,主要研究方向为软件工程、软件测试、信息安全检测等;彭玉青(1970-),教授。

0)，Is 为用于计算用户间相似度的评分项目的总个数，Us 为参与评分的用户个数，则一个用户可表达的个性化评分向量个数为

$$Vs=Rt^{Is} \tag{1}$$

多个用户一共可能表达的总评分方案状态数为

$$Vss=Rt^{Is*Us} \tag{2}$$

多个用户一共可能表达的总评分向量总个数为

$$Vsum=Us*Rt^{Is} \tag{3}$$

一个用户对各项目的全部评分都相同的概率为

$$P=\frac{Rt}{Rt^{Is}}=Rt^{1-Is} \tag{4}$$

两个独立评分的用户彼此评分向量完全相同的概率为

$$P_s=\frac{Rt^{Is}}{Rt^{2*Is}}=Rt^{-Is} \tag{5}$$

对于 CF 用户相关性计算，上述各个概率指标应当有不同的预期大小值趋势。Vs 和 Vss 的意义是可以表达的个性化状态，个性化评分状态应当多，所以 Vs 和 Vss 的值应当大；而造成相关性计算失误的概率则应当小(式(4))；对式(5)的概率也应当小，因为 CF 的特点是找到和目标用户相似的用户，而一个全部评分范围和目标用户自己完全相同的用户，对于生成推荐无太大帮助。即 CF 方法的特点实际是希望找到(在整个向量长度上)相似但不相同的用户。

为了使 Vs 和 Vss 尽量大，使 P 和 Ps 尽量小，就要对 Rt，Is 取较大的值。但 Rt，Is 并非越大越好。一方面，文献[2]等研究表明，将所有类别的项目集中于一个评分矩阵(Is 此时较大)不利于预测精度。另一方面，Rt 的值受到相似度计算公式的限制：无论是余弦、修正余弦等方法，都只能探查出(评分)向量之间的线性关系。若加大评分等级数(Rt 加大)，则有更大的可能概率错过用户之间的非线性关系。如评分向量(−3，−2，−1，0，1，2，3)和(9，4，1，0，1，4，9)以余弦、修正余弦相关等计算为 0，但它们之间并非“没有关系”，而是非线性关系(平方关系)。

分析说明，对于 Is 的值应当在尽可能大和要求相同类别项目之间做出权衡；而对于 Rt 的值不应当过分大，以免影响相似度计算和探查，虽然加大 Rt 有利于表达更多的个性化状态。

基于以上分析，在此提出全面平均值法(Is 较大、Rt 取适中的五级评分法)比传统的计算方法要更能表达数量多的个性化状态，实际应用中计算相似度应用更安全，在表达个性化程度和减少出错概率上获得了改进。

2.2 在相近类别项目中计算用户相似度的必要

文献[2]等指出，协同过滤推荐较适合用户单一兴趣下的推荐。对于用户多兴趣下的个性化推荐，其准确率会大大降低。即 CF 方法实际上要求用来预测的项目与被预测的项目在内容上具有一定的相似性。因此，应当根据项目的内容(领域)性质分别直接建立各类/子类项目的评分矩阵，如书籍(评分)矩阵、电影矩阵等。文献、实验等表明，在相似类别项目集合中计算用户之间相似度最为可信。比如，用户间分别对于音乐、书籍等领域的评价、喜好相似程度应分别考察。

因此，下节提出直接建立各分类项目评分矩阵的 CF 相似度计算方法。直接建立各分类的项目矩阵可以缩短文献[2]在大综合矩阵中计算某项目的最相似项目(similar items)的运算时间。且通过项目分类别建立各自的评分矩阵，可减轻各评分矩阵的稀疏程度。在实际电子商务系统中，可按商品类别，分别为每一类的商品项目建立用户-此类项目评分矩阵，再对各类商品矩阵使用前述全面均值法。

2.3 处理稀疏用户-项目评分矩阵改进方法——全面均值法

设定全面均值为用户对某一类项目的评分总和除以此类项目的总个数。

全面均值法的协同过滤推荐算法描述如下。

输入：目标用户对一类或几类商品项目显式离散评分；对项目按照类别或子类别分别建立的得分矩阵；

输出：N 种目标用户最可能喜爱的项目；

步骤如下：

A)目标用户对一系列各类项目评分，分数按照商品项目类别记入各用户-分类项目评分矩阵(如用户-电影矩阵等)

B)依次在各个子类项目的评分矩阵中，若目标用户在此分类项目评分矩阵中有非零评分，进行以下步骤 1)－步骤 4)：

1)计算目标用户和本子类矩阵中每个用户对此类项目的全面均值

$$\text{全面均值}=\sum R_i/\text{AllItem}N \tag{6}$$

式中，AllItem$N$ 代表本类的商品总数量，R_i 为用户对此子类的各项目 i 的评分。如上所述，Is=某一子类项目个数。

2)依次计算目标用户与各子类矩阵中各用户间相似性(将用户评分向量中的 0 视为一个用户有效评分数值不仅计算非 0 的项目)，用户相似度计算公式为

$$\text{sim}(i,j)=\frac{\sum_{c\in AI}(R_{i,c}-\bar{R}_i)(R_{j,c}-\bar{R}_j)}{\sqrt{\sum_{c\in AI}(R_{i,c}-\bar{R}_i)^2}\sqrt{\sum_{c\in AI}(R_{j,c}-\bar{R}_j)^2}} \tag{7}$$

式中，AI 表示本矩阵类中的所有项目的集合(包括用户评分为 0 的项目)，C 表示本类每个项目，$\bar{R}_i$ 为用户 i 在本类中的全面均值，$\bar{R}_j$ 为用户 j 在本类中的全面均值。

3)根据目标用户与此类中用户间的相似性 $\text{sim}(i,j)$ 递减排序，在此用 N 值法生成相似邻居集合 SU。

4)根据公式

$$P_{a,j}=\bar{R}_a+\frac{\sum_{n\in SU}\text{sim}(a,n)*(R_{a,j}-\bar{R}_n)}{\sum_{n\in SU}(|\text{sim}(a,n)|)} \tag{8}$$

计算本子类中目标用户未评分项目的预测评分。$\bar{R}_a$ 是用户 a 的全面均值，$\bar{R}_n$ 是相似邻居 n 的全面均值。

C)将各类得分最高的 N 项项目(与来自各种子类型矩阵中项目的预测评分一起不分类排序)推荐给目标用户。

原计算用户相似度的过程中，一般只考虑用户间“共同的非 0 评分”，本方法将用户间共同的 0 评分也考虑进去。以 C 语言伪代码描述，将原先计算用户 i 与 j 之间相似度数值的语句(rating[i][k]代表用户 i 对项目 k 的评分，ItemNum 为某类项目数量)：

```
for (k=0;k<ItemNum;k++)
    {if ((rating[i][k]! =0)&&(rating[j][k]! =0))
        使用此两值计算相似度;}
```

改进为

```
for (k=0;k<ItemNum;k++) {使用此两分值计算相似度;}
```

若在两个用户的评分向量长度为 n，在不计其它计算操作的情况下，上述前者代码时间复杂度为 O($2n$)，后者为 O(n)。可见，本方法降低了时间复杂度，且即使用户间没有共同非 0 项目，评分相似度也可计算。

3 改进算法的效果对比实验

对于以上算法，初步测试各算法对于一类项目电影的推

荐性能。为了恰当分析并与其它试验横向比较，测试数据集选择美国明尼苏达大学的 MovieLens 数据集 million-ml-data 数据文件。文件包含了 6040 个用户对 3952 个电影的 1,000,209 个评分。总体用户-评分矩阵稀疏等级约为 95.8%。使用平均绝对偏差(MAE,Mean Absolute Error)来度量本算法预测的准确性。这里采用 All-But-One 方法，随机取 10 个用户，在其余用户中计算相似用户，以 UBCF,ALL-BUT-ONE 方法对不同计算方式的性能进行评估，得到平均 MAE 值，如表 1 所列。

表 1　各计算方法平均 MAE 值对比

邻居数目	余弦相似	修正余弦	全面均值法
10	0.82847	0.76039	0.69307
20	0.76089	0.70494	0.62195
30	0.77858	0.71745	0.64141
40	0.79219	0.71395	0.60777
50	0.80325	0.73651	0.62469
60	0.80772	0.73853	0.62881

可以看到，本文提出的全面均值法在上述的实验环境下，总体的预测评分精度有所提高；在各个相似邻居数目测试情况下，有较小的 MAE 值。

表 2 为 F1 系数[1]验证的推荐质量(对上述各试验用户推荐 15 部电影。设数据集前 3000 部电影为已知，后 951 部未知，前者作为计算相似邻居依据，后者作为推荐测试集)。

表 2　各计算方法平均 F1 系数对比

邻居数目	余弦相似	修正余弦	全面均值
10	0.185777	0.147855	0.166257
20	0.238812	0.163861	0.246994
30	0.208182	0.160595	0.224860
40	0.226174	0.180726	0.226174
50	0.246846	0.178072	0.255656

由以上数据表明，在相同条件和各相似邻居数目下，全面均值法基本上具有较大的 F1 数值，推荐质量较高。

结束语　本文在研究分析 CF 算法的基础上，提出了全面均值法的改进算法，降低了推荐算法的时间复杂度。经实际数据验证，该算法提高了 CF 计算的预测精度和推荐质量，其所依据的心理理论基础是：一个人对一系列事物不表态本身就是一种表态。以图 1 为例，用户 3,4 对 1－10 号的电影评分都是 0，表明他们对 1－10 号电影都不感兴趣或不屑评价。这也是一种相似：他们虽然未必有多少相同的喜好，但却有范围较一致的“讨厌”。相对于其它的用户而言(如 10 或 18)，他们之间相对更为相似一些。

全面均值在计算时分母取值不宜过大，因此在实际电子商务系统中，分别为每一类(子类)的商品项目建立用户-此类项目评分矩阵，再对各类商品矩阵使用本全面均值法(如本文测试使用的 MovieLens 数据集就是专门对电影一类的 CF 矩阵)，这样做有助于提高 CF 预测精度[2]。对于不同项目类别划分的粒度，可进一步在多类项目环境中试验确定。

对于 CF 算法的性能验证，因数据集不同而存在很多的不确定性，静态试验数据集的 F1 等参数并不一定说明推荐算法的实际效果。但本算法在保持比传统计算方法稍好性能的情况下，降低了计算相似性的时间复杂度。

相关文献研究[1]指出，在实际评分数据稀疏条件下，传统的只考虑用户间共同非 0 评分项目的相似度计算方法会降低精度，而本文提出的全面均值算法考虑了用户间已评分和未评分的项目。在多种类型项目条件下推荐验证问题，可作为下一步的试验方向，以在实际多类型项目推荐数据中测试其效果。

参 考 文 献

[1] Asymeonidis P, Nanopoulos A. Collaborative recommender systems: combining effectiveness and efficiency [J]. Expert Systems with Applications, 2008, 34: 2995-3013

[2] Yu Li, Liu Lub, 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

[3] Papagelis M, Plexousakis D. Qualitative analysis of user-based and item-based prediction algorithms for recommendation agents [J]. Engineering Applications of Artificial Intelligence, 2005, 18: 781-789

[4] 潘红艳，林鸿飞，赵晶. 基于矩阵划分和兴趣方差的协同过滤算法[J]. 情报学报，2006，25(1)：49-54

[5] 吴发青，贺樑，夏薇薇，等. 一种基于用户兴趣局部相似性的推荐算法[J]. 计算机应用，2008，28(8)：1981-1990

[6] 王宏宇. 商务推荐系统的设计研究[D]. 合肥：中国科学技术大学，2007

[7] 王卫平，吴伦. 协同过滤在 CRM 交叉销售中的应用研究[J]. 管理学报，2007，4(4)：436-441

(上接第 129 页)

能力，从而较好地提高了服务选择的成功率。

结束语　本文在常见的基于 OWL-S/UDDI 的 Web 服务发现的研究基础上，对服务质量的描述从单一的服务提供方指定扩展到为由服务提供方、服务使用方、第三方共同评价的模式，增加了服务质量的可信度和可用性。同时，由于服务质量有多个方面的多个指标，对服务质量的多目标决策判定办法在实现中很困难，因此本文尝试将服务质量指标进行规范度量，以便在服务选择时计算机能自动处理。由于服务质量指标类型很多，在服务指标量化时如何对这些指标进行量化，需要较多的经验积累。本文的量化方法还存在不足，应当在以后的研究中予以改进。

参 考 文 献

[1] http://www.w3.org/Submission/OWL-S/. 2009.11

[2] http://www.uddi.org/pubs/uddi_v3.htm. 2009.11

[3] 高亚春，张为群. 基于 QoS 本体的 Web 服务描述和选择机制[J]. 计算机科学，2008，35(12)：273-276

[4] Luo J, Montrose B, Kim A, et al. Adding OWL-S Support to the Existing UDDI Infrastructure[C]// IEEE International Conference on Web Services (ICWS'06). 2006

[5] Ding Zhi-jun, Wang Jun-li, JIANG Chang-jun. Semantic Web Service Composition Based on OWL-[C]// Proceedings of the First International Conference on Semantics, Knowledge, and Grid (SKG 2005). 2005

[6] 赵军. 基于 OWL-S 的 Web 服务发现系统的研究和实现[J]. 计算机技术与发展，2006，16(10)：163-166

[7] 徐利谋，金可音，阳辉，等. 基于 OWL-S 的服务发现算法研究[J]. 计算机工程与科学，2007，29(8)：64-67

[8] 吴健，蔡铭，唐敏，等. 网络制造中 Web Service 的服务质量模糊排序方法[J]. 计算机辅助设计与图形学学报

[9] 牟玉洁，曹健，张申生，等. 扩展的 WebService 服务质量模型研究[J]. 计算机科学，2006，33(1)：5-9

[10] 陈蜀宇，刘刚国. 面向 Web 服务的数字化营区系统构架[J]. 重庆工学院学报：自然科学版，2008，22(9)：103-107