

# 基于用户层次信息的协同推荐算法<sup>\*</sup>)

秦 国 杜小勇

(中国人民大学信息学院 北京100872)

**摘 要** 在个性化信息推荐服务中,协同推荐作为一种基本的信息过滤方法得到广泛的应用,它根据和目标用户具有相似行为的用户对资源的评价来进行推荐。但是,我们的研究发现,协同推荐算法所获得的相似用户群和实际用户的概念层次没有关系,这和我们的直觉是矛盾的,这驱使我们在协同推荐算法中考虑进用户的分类信息。实验结果表明,这样的方法是有效的,它和传统协同过滤相比具有更高的推荐精度。

**关键词** 个性化推荐,协同过滤,用户层次

## An Efficient Collaborative Filtering Algorithm with User Hierarchy

QIN Guo DU Xiao-Yong

(School of Information, Renmin University of China, Beijing 100872)

**Abstract** As a basic information filtering approach, collaborative filtering (CF) is widely used in personalized information service systems. However, our research reveals that the similar user groups in the CF is independent with that of semantic classification of the users, which is a contrary to our intuition, so we take user's classification into count in our improved algorithm. The experiment result shows that our algorithm is efficient since it has higher accuracy compared with traditional collaborative filtering.

**Keywords** Personalized recommendation, Collaborative filtering, User hierarchy

## 1 引言

随着各种信息处理技术的不断发展,人们可以访问的数字资源越来越多,因此,如何从海量数据中方便、快捷地获取所需的信息便自然而然地成为人们非常关心的问题。

目前的信息服务还处于一种基于关键词的、粗糙的、被动的状况,信息服务系统通过对用户输入的查询关键词进行简单的语法层面的匹配来获取信息,这是一种没有考虑用户个性化需求的通用性信息服务<sup>[1]</sup>;在主动的个性化信息服务中,比较常用的推荐算法有基于内容推荐、协同过滤以及两者相结合的混和方式<sup>[2]</sup>。特别是协同推荐由于不关心资源和用户本身的信息,易于实现,因此受到广泛的关注。然而,这两种推荐方式都不能很好地反映用户对资源“深度”的需求。我们知道,无论是资源还是用户都存在语义上的深度或者层次,例如,大学数字图书馆中的用户可区分为教师、研究生、本科生等,资源也可以区分为学术专著和教材等不同的层次。直觉上,这种层次信息对于提高推荐的质量是有意义的。

我们的研究发现,这种层次信息和协同推荐的用户相似度之间不存在关联性。这个结果驱使我们考虑利用用户的语义分类信息来提高协同推荐的精度。

本文提出一种基于用户分类信息的协同推荐算法,可以更好地向用户推荐需要的数字资源,实现真正的主动向用户推荐最新的满足用户个性化需求的资源。

本算法已经应用在中国人民大学信息学院和图书馆联合开发的数字图书馆个性化推荐系统第二版(DLPers V2.0)中。本文在第2节介绍传统的协同推荐算法,第3节通过一些实

验具体分析传统协同推荐算法存在的问题,然后在第4节提出了结合用户分类层次信息的协同推荐算法,第5节的实验结果表明本文所提出的算法能够改进推荐的精度。

## 2 协同过滤算法

协同过滤技术在推荐系统以及电子商务网站中得到广泛的应用,它是至今为止最成功的信息过滤技术之一。其基本思想是通过比较用户过去的兴趣和行为的相似程度,找出和目标用户具有相同或相似兴趣的用户组,再根据他们对资源的评价来预测目标用户的兴趣,达到向目标用户推荐数字资源的目的<sup>[3]</sup>。

一般来说,协同过滤算法的依据是大量用户对大量数字资源的评分矩阵  $R_{m \times n}$ :

$$R_{m \times n} = \begin{bmatrix} r_{11} & \cdots & r_{1j} & \cdots & r_{1n} \\ \vdots & & \vdots & & \vdots \\ r_{i1} & \cdots & r_{ij} & \cdots & r_{in} \\ \vdots & & \vdots & & \vdots \\ r_{m1} & \cdots & r_{mj} & \cdots & r_{mn} \end{bmatrix}_{m \times n}$$

其中  $r_{ij}$  表示用户  $i$  对资源  $j$  的评分,如果用户  $i$  没有给资源  $j$  评分,则  $r_{ij} = 0$ 。 $m$  表示矩阵中包含的用户数目, $n$  表示矩阵中数字资源的数目,其工作原理大致如下:

首先,计算每个用户对以往评价过的信息资源的平均打分:

$$\bar{r}_i = \frac{1}{|I_i|} \sum_{j \in I_i} r_{ij} \quad (1)$$

其中  $I_i$  为用户  $i$  的评分向量,  $|I_i|$  为  $I_i$  的长度即用户打过分

<sup>\*</sup>)本研究得到中国人民大学211工程项目资助。秦 国 硕士研究生,主要研究方向为智能信息检索。杜小勇 工学博士,教授,博士生导师,主要研究方向为智能信息检索、高性能数据库系统、知识工程等。

的数字资源数目。

其次,计算目标用户对于未评价过的信息资源的预测值:

$$p_{aj} = \bar{r}_a + k \sum_{i=1}^N w(a, i) (r_{ij} - \bar{r}_i) \quad (2)$$

其中  $p_{aj}$  为目标用户  $a$  对信息资源项  $j$  的预测值,协同过滤算法根据和目标用户最相似的  $N$  个用户的评价进行预测,并不是所有用户都参与预测  $p_{aj}$  值,  $w(a, i)$  为用户  $a$  和用户  $i$  之间的

兴趣相似度,  $k$  是归一化因子,等于  $1/\sum_{i=1}^n w(a, i)$ 。算法的核心部分就是计算用户的兴趣相似度,目前较为常用的是用皮尔逊相关系数<sup>[4]</sup> (Pearson Correlation Coefficient) 计算用户的兴趣相似度。其对用户  $a$  和用户  $i$  之间的兴趣相似度  $w(a, i)$  定义为:

$$w(a, i) = \frac{\sum_{I_{ai}} (r_{aj} - \bar{r}_a) (r_{ij} - \bar{r}_i)}{\sqrt{\sum_{I_{ai}} (r_{aj} - \bar{r}_a)^2 \sum_{I_{ai}} (r_{ij} - \bar{r}_i)^2}} \quad (3)$$

其中  $I_{ai} = I_a \cap I_i$ , 即  $j$  为用户  $a$  和用户  $i$  都打过分的数字资源项。

上述协同过滤算法被应用到众多个性化推荐系统中,它能够即时地利用最新的信息来做出预测;但它只是基于用户过去的评分矩阵来预测,而这一评分矩阵不能充分地体现用户的需求,从而影响了推荐结果的精确性。

### 3 问题分析

虽然协同推荐没有考虑资源的“深度”,也没有考虑用户的实际分类,但是,直觉上我们会认为,这些信息应该蕴涵在用户过去对资源的打分中。然而,我们在数据集 MovieLens ([www.grouplens.org/data/ml-data.zip](http://www.grouplens.org/data/ml-data.zip)) 上的实验结果表明,协同推荐计算出的相似用户和实际的用户分类没有关系。

MoiveLens 数据集包含了 943 位用户对 1682 部电影的评分(分值从 1 到 5, 1 代表用户很不喜欢该电影,相对的, 5 代表用户很喜欢该电影),每位用户至少对 20 部电影打分。数据集 MovieLens 提供了七对(训练集,测试集)组合,其中每个训练集里有 80000 条记录(用户对电影的打分),而测试集里有 20000 条记录。MoiveLens 里的用户从事不同的职业,如艺术家、娱乐界人士、作家、教育家、医生、推销员等等。

直观上,不同职业的用户对电影的理解会有很大的差异;因此我们将职业视为用户层次的一个标志,根据具体语义,将用户分为 A 类 = {艺术家、娱乐界人士}, B 类 = {作家、教育家} 以及 C 类 = {其它职业} 三类,通常情况下,对电影的理解,艺术家、娱乐界人士最专业,作家、教育家次之,相比之下从事其它职业的用户最业余,相应地,在层次上也就 A 类最高, B 类次之, C 类最低。

针对目标用户,定义事件  $X$  为“ $a$  用户和目标用户属于同一类别”,事件  $Y$  为“用户  $a$  和目标用户之间(按照协同推荐算法计算的)相似度不小于阈值  $\delta$ ”。设  $C_i$  为和目标用户属于同一类的用户数,  $C_s$  为和目标用户之间的相似度不小于阈值  $\delta$  的用户数,  $C_a$  为和目标用户属于同一类而且和目标用户之间的相似度不小于阈值  $\delta$  的用户数,  $C$  为用户总数。所以事件  $X$  的概率  $P(X) = C_i/C$ , 事件  $Y$  的概率  $P(Y) = C_s/C$ ,  $P(X \cap Y) = C_a/C$ 。我们多次随机地选取若干用户,然后针对不同的阈值,计算事件  $X$  和  $Y$  之间的相关系数  $Cor(X, Y) = \frac{P(X \cap Y)}{P(X)P(Y)}$ , 图 1 表示了其中分别属于三个类别的三个用户的

相关系数随阈值的变化。

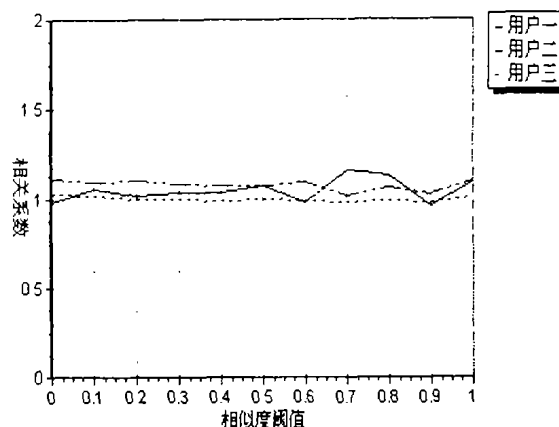


图 1

图 1 中各用户对应的相关系数值都在 1 左右波动,也即是说事件  $X$  和事件  $Y$  几乎是相互独立的。由此可见,MovieLens 中用户兴趣是否具有相似和用户是否属于同一类别几乎没有关系。这和我们的直觉有些矛盾。很显然,这种带有语义的用户层次分类和用户的兴趣之间应该有很重要的关系。这个事实说明,协同推荐没有能够考虑用户和资源的层次,是这个方法的一个重要缺陷,会影响到推荐精度。

本文提出的一种基于用户分类层次的协同推荐算法,即在推荐时考虑到用户层次等因素。

### 4 基于用户分类层次的协同推荐

鉴于上面提出的问题,在推荐时就应该考虑到用户层次因素,结合用户的打分向量  $I$ , 构成一个更完善的用户兴趣模型,以生成更精确的推荐结果。

#### 4.1 层次相关性

**定义 1(层次)** 根据用户对资源  $D$  理解的深浅程度将用户分为若干个类别  $\{C_1, C_2, \dots, C_n\}$ , 并且  $C_i \cap C_j = \emptyset (i \neq j)$ 。如在数字图书馆的应用中,可以根据教授、副教授、讲师等职称来对用户进行分类,也可以按照博士生、硕士生、本科生等学历来对用户进行分类;当然也可以使用一定分类或者聚类算法来分类。

**定义 2(理解度)** 定义函数  $f_D(C_i)$ , 它表示用户组  $C_i$  对资源  $D$  的理解程度,称为用户组  $C_i$  的理解度。

理解度的取值依赖于用户分类方法,对于按照一定语义进行的分类理解度根据具体的逻辑含义来取值,而采用分类或者聚类算法的分类结果理解度的取值取决于具体的算法。

**定义 3(层次)** 若  $f_D(C_1), f_D(C_2), \dots, f_D(C_n)$  构成一个全序关系,则称  $\{C_1, C_2, \dots, C_n\}$  构成一个用户层次;并且定义任意两个用户组  $C_i, C_j$  之间的距离为它们理解度的差值:

$$d(C_i, C_j) = |f_D(C_i) - f_D(C_j)| \quad (4)$$

**定义 4(层次相关性)** 用户  $i$  和用户  $j$  之间的层次相关性系数  $HC(i, j)$  定义为:

$$HC(i, j) = \frac{1}{d(C_i, C_j) + 1} \quad (1 \in C_i, j \in C_j) \quad (5)$$

根据以上定义可见,层次相关性系数  $HC(i, j)$  随着用户  $i$  和用户  $j$  层次属性差别的增大而减少。

#### 4.2 用户兴趣模型

协同推荐的基本思想是通过比较用户间的过去的兴趣和行为的相似程度确定相似用户群,然后通过那些目标用户的相似用户群对数字资源的评价来进行推荐。也即是说,协同

推荐更关心的是用户兴趣的相互关联程度,即用户需求的一个相对定位。

结合前面的分析,用户兴趣模型  $U_{(a)}$  可以重新描述为:

$$U_{(a)} = (C_i, I_a) \quad (a \in C_i, I_a \in R_{m \times n}) \quad (6)$$

其中,  $C_i$  为用户  $a$  相对应的层次,  $I_a$  为用户  $a$  的评分向量。

#### 4.3 层次兴趣相似度

基于上述新的用户兴趣模型,兴趣相似度计算也在原来皮尔逊相关系数(Pearson Correlation Coefficient)的基础上做了相应的改进,用户  $a$  和用户  $i$  之间的层次兴趣相似度定义为:

$$wh(a, i) = \frac{\sum_j (r_{aj} - \bar{r}_a)(r_{ij} - \bar{r}_i)}{\sqrt{\sum_j (r_{aj} - \bar{r}_a)^2 \sum_j (r_{ij} - \bar{r}_i)^2}} * HC(a, i) \quad (7)$$

其中,  $j$  是用户  $a$  和用户  $i$  都打过分的资源项,  $HC(a, i)$  是用户  $a$  和用户  $i$  的层次相关性系数。

#### 4.4 算法步骤

通过以上讨论,基于用户分类层次的推荐算法通过以下步骤完成:

##### (1) 定义层次

根据用户的显性或者隐性的层次属性语义或者采取相关算法对用户进行合理的分组,即定义出层次  $H$ 。

##### (2) 定义层次距离

按各层次代表的逻辑含义或者采用的分组算法得出对应的理解度。

##### (3) 计算层次兴趣相似度

结合用户对资源的评分矩阵以及用户间的层次相关性,计算目标用户和其它用户的层次兴趣相似度并选出与目标用户最相似的  $N$  个用户。

(4) 根据第(3)步选择出来的  $N$  个用户对资源的评价来预测目标用户  $a$  对其没有评价过的资源  $j$  的感兴趣程度  $P_{aj}$ :

$$p_{aj} = \bar{r}_a + k \sum_{i=1}^N wh(a, i) (r_{ij} - \bar{r}_i) \quad (8)$$

这里归一化因子  $k$  的值为  $1 / \sum_{i=1}^N wh(a, i)$ 。

##### (5) 生成结果

将资源按照预测值进行降序排列,并根据实际需要选择前若干个或者全部结果呈现给用户。

## 5 实验分析

基于用户分类层次的协同过滤方法产生于数字图书馆这种特定的应用环境,但是它并不局限于数字图书馆领域,还可以在其它的用户兴趣模型里面含有层次信息的数字资源推荐系统中得以应用。由于 DLPers 系统投入运行不久,系统内用户对书籍或者论文的评价很少,考虑到少量数据的偶然性不足以说明算法的好坏,因此我们在 MovieLens 进行了下面的实验。

和前面一样,将用户分为 A 类{艺术家、娱乐界人士}、B 类{作家、教育家}以及 C 类{其它职业},其中  $f_D(A)=3$ ,  $f_D(B)=2$ ,  $f_D(C)=1$ 。我们选择其中五个训练集中的数据分别

采用传统协同过滤算法和基于用户分类层次的协同算法来预测其中 A 类用户对电影的打分,最后使用平均绝对误差(MAE)来比较两种算法的精度。平均绝对误差是所有用户对资源的打分的预测值与测试集中的实际打分值的偏差绝对值的平均,它越小说明算法的精度越高,效果越好。

实验结果如表1所示。

表1

| 训练集   | 基于用户分类层次协同推荐算法 | 传统协同过滤算法 |
|-------|----------------|----------|
| Test1 | 0.86924        | 0.93296  |
| Test2 | 0.88012        | 0.95021  |
| Test3 | 0.83157        | 0.89517  |
| Test4 | 0.82915        | 0.88639  |
| Test5 | 0.86538        | 0.92372  |
| 平均值   | 0.85509        | 0.91769  |

从表2中可以明显地看出,新算法在推荐精度上优于传统的协同过滤算法。

另外,在 DLPers V2.0 中使用该推荐算法也确实更好地描述了用户的需求,提高了推荐的精度,更好地满足了用户的需要。

**结论** 本文针对传统的协同推荐算法中没有考虑用户和资源层次这一问题,将用户层次因素纳入用户兴趣模型中,提出了基于用户分类信息的协同推荐算法。通过实验分析,新的算法提高了推荐的精确度。将层次这一特征纳入改善了推荐的效果。基于用户分类信息的协同推荐算法已经应用到了 DLPers V2.0,相信随着 DLPers V2.0 用户数目和用户评分的逐渐增加,其良好的推荐效果会得到实践结果的进一步验证。

用户和资源都存在层次或者深度属性,本文提出的算法是从用户的层次属性出发,但由于数字资源建设的历史性原因,目前很少存在资源有深度属性,但我们可以在下一步工作中根据资源的使用情况利用数据挖掘技术来得出这些信息。

算法中我们将两点的层次距离用理解度的差值表示,所以理解度取值对推荐结果有很大的影响,如何更合理地确定理解度也是我们下一步工作的内容。

## 参考文献

- 1 高凤荣,杜小勇,王珊. 数字图书馆环境下一种基于语义分类的个性化推荐算法. A Personalized Recommendation Algorithm Based on Semantic Classification in Digital Library. In NDBC 2003
- 2 Paulson P, Tzanavari A. Combining Collaborative and Content-Based Filtering Using Conceptual Graphs. Modelling with Words, 2003. 168~185
- 3 Prof Dr, Neuhold E J. Personalization and User profiling & Recommender Systems. WI/IM Informationsmanagement Proseminar SS, 2003 LVA-Nr. : 400298 8
- 4 Vozalis E, Margaritis K G. Analysis of Recommender Systems' Algorithms (<http://macedonia.uom.gr/~mans/papiria/hercma2003.pdf>)