第34卷 第23期 计 算 机 工 程 2008年12月
Vol.34 No.23 Computer Engineering December 2008
·博士论文· 文章编号：1000—3428(2008)23—0011—03 文献标识码：A 中图分类号：TP391

协同过滤系统项目冷启动的混合推荐算法

郭艳红，邓贵仕
(大连理工大学系统工程研究所，大连 116023)

摘 要：研究协同过滤推荐系统中的冷启动问题，运用基于内容预测的方法，对系统内未被用户评价过的项目进行评分预测，应用2种优化步骤，过滤掉预测不准确的用户的评分。在此基础上用协同过滤的方法产生推荐，使传统推荐算法中无法推荐给用户的项目得到推荐机会。通过一系列实验证明，该混合推荐算法能保证推荐准确性，提高了新项目的推荐概率。

关键词：协同过滤；冷启动；基于内容的预测；混合推荐

Hybrid Recommendation Algorithm of Item Cold-start in Collaborative Filtering System

GUO Yan-hong, DENG Gui-shi
(Institute of Systems Engineering, Dalian University of Technology, Dalian 116023)

【Abstract】 To address the problem of item cold-start in collaborative filtering systems, this paper advances a new method that using content based prediction before collaborative filtering to get the predictive ratings of items for users. It points out two fine grained parameters to guarantee the accuracy of the predictions. After the content based filtering, collaborative filtering algorithm is used to generate predictions for users. Experimental results show that the method is superior than traditional collaborative filtering algorithm in the coverage which indicates that the item cold-start problem is alleviated.

【Key words】 collaborative filtering; cold-start; content based prediction; hybrid recommendation

协同过滤是一种常用的减少信息过载的技术，已成为个性化推荐系统的主要工具，但大多数协同过滤算法存在一个共性——新项目的冷启动问题。现有方法在解决协同过滤系统中新项目的冷启动问题时，一般采用系统随机推荐给用户的方法。该方法得到的结果并不理想，因为从长期来看，随机方法产生的推荐准确率不会超过50%。而用户一旦对推荐不满意，就可能不再信任推荐系统。

1 协同过滤及冷启动问题分析

协同过滤算法一般可分为3步：构建用户档案，寻找最近邻，产生推荐[1-2]。

(1)构建用户档案(profile)。即收集用户的评分、评价行为等，并进行数据清理、转换和录入，最终形成用户对各种项目的评价表，如表1所示。

表1 用户的评分

用户项目	$Item_1$	$Item_2$	…	$Item_{n-1}$	$Item_n$
User1	3	4		?	3
User2	5	?		5	3
…					
$User_{m-1}$	?	4		4	4
$User_m$	5	5		5	?

(2)寻找最近邻居。在这一阶段，计算目标用户与数据库内各个用户的相似度，寻找相似度最高的作为最近邻居集。一般可采用：1)pearson 相关度公式；2)cosine 相关度公式；3)修正的余弦相关度计算用户之间的相似度。

$$sim(i,a)=\frac{\sum_{j\in I_i\cap I_a}(R_{i,j}-\overline{R_i})(R_{a,j}-\overline{R_a})}{\sqrt{\sum_{j\in I_i\cap I_a}(R_{i,j}-R_i)^2}\sqrt{\sum_{j\in I_i\cap I_a}(R_{a,j}-R_a)^2}} \tag{1}$$

$$sim(i,a)=\cos(i,a)=\frac{\bar{i}\times\bar{a}}{\|i\|\|a\|} \tag{2}$$

$$sim(i,a)=\frac{\sum_{j\in N}(R_{i,j}-\overline{R_i})(R_{a,j}-\overline{R_a})}{\sqrt{\sum_{j\in N}(R_{i,j}-\overline{R_i})}\sqrt{\sum_{j\in N}(R_{a,j}-\overline{R_a})}} \tag{3}$$

(3)预测阶段。一般采用加权平均值的方法，通过最近邻居集的评价产生推荐，经典的推荐算法一般如下：

$$p_{a,y}=\frac{\sum_{u\in NN,y\in N}sim(a,u)R_{u,y}}{\sum_{u\in NN,y\in N}|sim(a,u)|} \tag{4}$$

$$p_{a,y}=\overline{R_a}+\frac{\sum_{u\in NN,y\in N}sim(a,u)(R_{u,y}-\overline{R_u})}{\sum_{u\in NN,y\in N}|sim(a,u)|} \tag{5}$$

其中，$P_{a,y}$代表目标用户对项目y的预测值；$R_{u,y}$代表目标客户a的最近邻居集内的用户u对项目y的评价。这里的目标用户a的最近邻居邻居集用NN(nearest neighbour)表示，因此，$u\in NN$。

可以看出，协同过滤推荐依靠的是用户对项目的评分，才能给出推荐。如果一个新的项目，没有任何用户对它给出评价，那么该项目就永远也没有机会被推荐给用户。这就是新项目的冷启动问题。因此，在协同过滤系统中，为了增加

基金项目：国家自然科学基金资助项目(70671016)；国家自然科学基金资助项目“互联网环境下的关系营销理论与创新”(70532006)
作者简介：郭艳红(1977－)，女，博士研究生，主研方向：电子商务个性化理论与方法；邓贵仕，教授、博士生导师
收稿日期：2008-03-25 **E-mail**：guoyh@dlut.edu.cn

新项目的推荐机会，应增加新项目的评价机会。在一般的协同过滤系统中，会有简单的有关项目内容的分析，可以利用这些项目的相关内容，为用户对这些项目进行预测，然后再利用协同过滤进行算法进行推荐[3]。

2 基于内容预测与协同过滤的混合推荐算法

本文的算法流程为：利用基于内容的过滤方法，对项目的内容进行简单的分析，根据用户对项目的评价和项目内容之间的关联，对用户未评价过的新项目进行初步预测，并利用 2 种优化策略，过滤掉预测不够精确的项目。在此基础上，再应用协同过滤的方法，为用户产生最后的推荐。

2.1 基于特征的项目表示

一般的协同过滤推荐系统会有对项目的简单描述，如以电影推荐，会有电影的类型的相关介绍，是动作片还是爱情片，或是几种的结合。这些对影片的描述，可以看成是有关项目的关键词，这样，每个项目 X_i 就可以由关键词来描述，如下：

$$X_i=\{A_1, A_2,\cdots, A_n\}$$

其中，A_j 表示 X_i 的第 j 个特征，如果这一项为 1，表明 X_i 具备这个特征，如果为 0 则不具备这个特征。

因此，所有的项目就可表示为 1 个关键词的 0, 1 矩阵，如下：

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 1 & 0 & 1 \\ 0 & 0 & 0 & \cdots & 1 & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \end{bmatrix}$$

2.2 基于内容的预测

项目的特征抽取完成后，就可以通过用户已经评价过的项目的内容分析和抽取的特征进行相应的信息过滤[4]。通过分析用户的偏好和项目的特征，判断用户是否对一个项目感兴趣。

winnow 算法在文本分类中效果很好，在 winnow 中，每个单词 x_i 被作为一个布尔特征，winnow 通过设定每个单词的权重构成一个线性的门槛函数，如下：

$$\sum w_i x_i > \theta$$

其中，θ 为门槛值；权值 w_i 初始化为 0.5。本文采用 Winnow 算法，预测用户对电影的评价。

用户参数由分析他以前评价过的项目的关键词的抽取来判断。如果用户评价过的某个电影是 action 类型，则此时这一项的用户参数可设为 1，否则为 0。权值模式被用来区分关键词，如果一个用户喜欢一个电影，则从此项目描述中抽取的单词的权值将增加，否则减少。

在训练的过程中，用户的每一个项目都被计算，找到这个用户的最佳的权值和。如果权重和布尔值乘积之和小于门槛值 θ，但是用户的评价大于 θ，则每个关键词的权重放大，乘以 2；如果权重值与布尔值之和大于门槛值 θ，但是用户的评价小于 θ，则每个单词的权重缩小，除以 2；否则权重合适，不做改变。

在整个训练集中，通过调整权值不断循环，直到所有的项目被正确处理。或者对训练集例子循环 10 次，直到对训练集的例子没有改变为止。经过训练后，每个用户都有一个关于不同类别的电影的权重 $w_{i,k}$，这样对于每一个电影，就会产生一个预测的评价值 $p_{i,j}$，计算如下：

$$p_{i,j}=\sum_{k=1\ldots k} w_{i,k}\times X_{j,k}$$

其中，$w_{i,k}$ 为用户 i 的对第 i 类特性的电影的权重；$X_{j,k}$ 为第 j 个电影在第 k 类特征上的值。

因此，虚拟用户的评价如下：

$$R_{i,j}=\begin{cases} p_{i,j} & r_{i,j}=0 \\ r_{i,j} & r_{i,j}\neq 0 \end{cases}$$

2.3 算法的优化

本文采用基于内容的预测的方法，为用户评价矩阵填上许多空白项，产生虚拟的用户评价矩阵，因此，这个矩阵应该能够正确代表用户的偏好，否则，基于此产生的预测值将误差非常大。为了保证基于内容的预测方法的有效性，要对预测结果进行初步筛选，只有精度达到一定程度，说明经过训练得到的权重能够代表用户的意愿，采用如下优化措施：

(1)为了提高预测的准确率，先对用户进行筛选，即当用户的评价个数很少的时候，认为由于样本个数太少，无法得出准确的预测值。因此，只有当用户的评价个数(RN)达到一定数目以上，才能产生预测，本文选择 $RN\geqslant 90, RN\geqslant 100$。

(2)对基于内容的预测进行初步的评价，评价方法采用精确率来度量，即用户预测值与用户的评价值相符与用户的评价的个数的比例达到一定数目，这个用户的预测值才可以接受。比例 CP 分别为 75%和 80%。

在基于内容预测的基础上，可利用式(1)、式(2)或式(3)，对用户之间的相似度重新进行度量，然后再用式(4)、式(5)对用户作出进一步的项目预测，产生推荐。

3 实验结果与分析

3.1 数据集的筛选

为了验证算法的有效性，本文采用 Grouplens 工作组提供的公开数据集(http://movilens.umn.edu/)。Movilens 是由美国明尼苏达大学 Grouplens 工作组研究人员开发的一个基于 Web 的研究型推荐系统。它用于接受用户对电影的评价，并提供相应的电影推荐列表。目前该系统的用户已经超过 43 000 人，用户评价的项目超过 1 600 个。

本文采用 Movielens 工作组提供的 ml 数据集，它由 943 个用户的 10 000 条 1~5 的评价数据组成。共有 1 682 个电影项目，每个用户至少对 20 个电影项目做出评价。

整个实验数据集划分为训练集和测试集。因此，引入变量 x 作为测试集占整个数据集的百分比。本文选用 x=0.2，在整个数据集中，训练集占 80%，测试集占 20%，即训练集为 100 000×80%＝80 000 条数据，测试集为 100 000×20%＝20 000 条数据。

为保证实验的准确性，重复实验 5 次，每次测试集的数据都各不相同。为了度量整个数据集的稀疏性，引入稀疏度的概念，其定义为用户未评价数据占整个数据集的比例。本文所用数据集的稀疏度为

$$\frac{943\times 1682-100\,000}{943\times 1682}\times 100\%=93.7\%$$

3.2 度量标准

(1)平均绝对误差

设测试集内目标客户的推荐数据集为 $P_a=\{p_{a,j}\mid j=1, 2,\cdots, n\}$，目标客户的真实评价集为 $R_a=\{r_{a,j}\mid j=1,2,\cdots,n\}$。对于每个不为 0 的“预测-评价对”$<p_{a,j}, r_{a,j}>$，都有

$$MAE=\frac{\sum_{i\in N}(p_{i,j}-r_{i,j})}{N}$$

其中，N 为测试集内目标客户 a 的预测值和真实评价值都不为 0 的项目的个数。MAE 越小，推荐精度越高。

(2)覆盖率

覆盖率是所有能够预测的项目占项目总数的比例。因此，设为用户的提供的预测值的集合为 $P_a = \{p_{a,j} \mid j = 1,2,\cdots,n\}$，则所有 $p_{a,j} \neq 0$ 的个数为 K_a，则用户 a 的覆盖率 C_a 为

$$C_a = \frac{K_a}{N}$$

3.3 实验结果与分析

经过实验验证，余弦相似性度量方法产生较好的预测精度，因此，本文的实验都采用余弦相似性度量方法对用户进行相似性度量。

4 种推荐策略如下：(1)Tcfd：传统的推荐算法(考虑评价风格的影响)；(2)tcfs：传统的推荐算法(不考虑评价风格的影响)；(3)cbcfd：基于内容预测与协同过滤的个性化推荐算法(考虑评价风格的影响)；(4)cbcfs：基于内容预测与协同过滤的个性化推荐算法(不考虑评价风格的影响)。

表 2~表 7 是在不同参数下，基于内容预测的协同过滤算法与传统协同过滤推荐算法以及无优化步骤的混合推荐算法的推荐精度和覆盖度。没有经过优化步骤的混合推荐算法，尽管覆盖度有 100%，能为所有的项目产生推荐，但是推荐的准确率却非常低。而经过优化处理的混合推荐算法要优于传统的协同过滤算法。其中本文提出的考虑到不同评价风格的基于内容的预测的算法推荐精度最高，且随着参数 *RN* 和 *CP* 的不断增大，算法的精度逐渐得到提高。

表 2　混合推荐算法(*RN*=90, *CP*=75%)

邻居个数	cbcfd	cbcfs	覆盖率/(%)
5	0.836 0	0.903 9	12.6
10	0.802 0	0.861 5	15.1
15	0.785 6	0.848 4	19.1
20	0.777 2	0.841 6	21.1
25	0.771 1	0.837 6	22.4
30	0.766 3	0.835 3	23.1

表 3　混合推荐算法(*RN*=100, *CP*=75%)

邻居个数	cbcfd	cbcfs	覆盖率/(%)
5	0.837 6	0.905 9	12.0
10	0.803 5	0.862 6	16.2
15	0.783 3	0.848 7	18.8
20	0.773 8	0.839 8	20.7
25	0.767 7	0.838 2	22.2
30	0.765 2	0.835 0	23.4

表 4　混合推荐算法(*RN*=90, *CP*=80%)

邻居个数	cbcfd	cbcfs	覆盖率/(%)
5	0.833 7	0.902 7	11.5
10	0.797 6	0.859 3	15.6
15	0.777 7	0.842 4	18.5
20	0.767 6	0.835 2	20.4
25	0.763 0	0.839 3	22.1
30	0.757 1	0.833 3	23.3

从覆盖度角度来讲，基于内容预测的协同过滤算法的覆盖率远高于传统的协同过滤算法，这说明基于内容的预测，在保证预测的准确性的同时，覆盖率很高。因此，在进行协同过滤算法之前，通过基于内容预测的方法对用户的偏好信息进行优化是可行的。该方法保证了协同过滤算法的精确性，对协同过滤的项目的冷启动问题也有一定改善。

表 5　混合推荐算法(*RN*=100, *CP*=80%)

邻居个数	cbcfd	cbcfs	覆盖率/(%)
5	0.834 0	0.904 1	11.3
10	0.797 7	0.860 9	15.4
15	0.776 1	0.842 2	18.3
20	0.767 0	0.836 1	20.2
25	0.761 5	0.839 0	21.9
30	0.755 9	0.833 1	23.2

表 6　传统的协同过滤算法

邻居个数	cbcfd	cbcfs	覆盖率/(%)
5	0.831 3	0.821 2	9.5
10	0.805 4	0.820 2	13.4
15	0.787 2	0.818 0	15.9
20	0.777 4	0.814 1	17.9
25	0.774 0	0.810 3	19.5
30	0.769 2	0.800 3	20.9

表 7　无优化步骤的混合推荐算法

邻居个数	cbcfd	cbcfs	覆盖率/(%)
5	2.122 5	2.368 5	100
10	2.007 9	2.105 6	100
15	1.972 3	2.001 6	100
20	1.699 1	1.936 4	100
25	1.356 3	1.593 8	100
30	1.231 7	1.361 2	100

4　结束语

本文应用基于内容的预测方法对未评价过的项目产生预测，再应用协同过滤的算法对用户产生推荐，解决过滤系统的冷启动问题。实验证明，该算法在保证预测精度的同时，提高了传统协同过滤推荐算法的覆盖度，在一定程度上解决了协同过滤推荐系统的项目冷启动问题。

参考文献

[1] 曾汇艳, 麦永浩. 基于内容预测和项目评分的协同过滤推荐[J]. 计算机应用, 2004, 24(1): 111-113.

[2] Herlocker J, Konstan J, Borchers A, et al. An Algorithmic Framework for Performing Collaborative Filtering[C]//Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. [S. l.]: ACM Press, 1999.

[3] Sarwar B, Karypis G. Item-based Collaborative Filtering Recommendation Algorithm[C]//Proceedings of the 10th International World Wide Web Conference. Hong Kong, China: [s. n.], 2001.

[4] Balabanovic M, Shohalm Y. Fab: Content Based Collaborative Recommendation[J]. Communication of the ACM, 1997, 40(3): 66-72.

（上接第 10 页）

的信息变得很微弱，环境因素的影响则使其变得更微弱，常常被噪声所淹没。要想重建这些泄漏发射的信息，只能靠后期信息处理算法，而这一切都建立在信息高度同步的基础上，因此，获取同步信息是泄漏发射重建工作中非常重要的一步。本文提出的基于高速 DSP 的相关检测算法较好地解决了提取同步信号的问题，对泄漏发射信息的重建具有理论和实际指导意义。

参考文献

[1] van Eck W. Electromagnetic Radiation from Video Display Units: An Eavesdropping Risk[J]. Computers & Security, 1985, 4(1): 269-286.

[2] 刘　杰, 刘济林. 泄漏发射信息的同步信号提取与信息重建[J]. 浙江大学学报: 理学版, 2005, 32(6): 528-534.