一种改进的协同过滤推荐算法

王 茜 王均波
(重庆大学计算机学院 重庆 400044)

摘 要 传统的协同过滤算法在寻找最近邻居集合时没有考虑时间因素的影响,仅从用户或者项目单方面出发计算用户或者项目的相似性以产生推荐结果,也忽略了用户特征对推荐的影响。针对上述问题,引入时间遗忘函数、黏度函数、用户特征向量,对协同过滤算法寻找用户的最近邻居集合过程进行了改进,体现了时间效应、用户偏好程度和用户特征。采用MovieLens数据集进行了一系列对比实验,结果表明,改进后的算法能够明显提高推荐的准确度。

关键词 协同过滤,时间效应,用户偏好度,用户特征向量

Improved Collaborative Filtering Recommendation Algorithm

WANG Qian WANG Jun-bo
(Department of Computer Science, Chongqing University, Chongqing 400044, China)

Abstract When searching the nearest neighbor set, the traditional collaborative filtering algorithm ignores the impact of the time factor, only from the user or item takes into account the similarity of the user or item unilaterally, and ignores the impact of user characteristics on recommendation. Aiming to the above problems, we introduced the time forgotten function, resources viscosity function and the user feature vector, and improved the process of finding the user's nearest neighbor set, reflected the time effect, degree of user preferences and user characteristics, and tested this new methodology on data set got from MovieLens. The results of experiment show that proposed method can improve the accuracy of the prediction.

Keywords Collaborative filtering, Time effect, User preference degree, User characteristic

1 引言

随着网络的普及,网络资源不断丰富,网络信息量不断膨胀。用户要在众多的选择中挑选出自己真正需要的信息好比大海捞针。个性化推荐系统应运而生,它为不同用户提供不同的服务,以满足不同的需求,带来了由"人找信息"到"信息找人"的转变。常用的推荐方法有基于内容的推荐、基于知识的推荐、协同过滤推荐以及组合推荐等[1],其中协同过滤推荐是迄今为止应用最成功的个性化推荐技术之一[2],其基本思想是:通过计算用户对项目评分之间的相似性,搜索目标用户的最近邻居,然后根据最近邻居的评分向目标用户产生推荐[3]。

现有的协同过滤推荐算法存在下列问题:①只考虑了用户或项目间的相似性,忽略了用户兴趣的动态变化[3-5]。现实生活中,用户对商品的喜好会随着时间的推移发生改变。传统的协同过滤算法利用用户-项目评分矩阵来进行推荐计算,未考虑用户访问项目的具体时间,未反映用户兴趣随时间的变化过程,当用户兴趣发生改变时,现有的推荐系统不能及时反应,从而导致系统推荐的项目偏离了用户的真实喜好。②仅从用户或项目单方面出发,只考虑了用户或者项目单方面的相似性,而实际上用户可能喜欢相似用户所购买的商品,也可能喜欢非相似用户所购买商品的相似商品[1,6,7]。③不同特征的人其兴趣爱好可能不同,而同一类别的人的兴趣爱好则具有一定的相似性。所以,在寻找目标用户的最近邻居时,用户特征也是一个不容忽视的因素,但在传统的协同过滤算法中并没有体现出这一点[2,8,9]。

针对上述问题,在传统协同过滤算法的基础上,论文引入时间遗忘函数、黏度函数、用户特征向量,提出了一种适应用户兴趣变化和基于用户特征的个性化推荐算法,改进了寻找最近邻居集合的过程,综合考虑了时间效应、用户特征和用户偏好程度,提高了推荐的准确性。

2 传统协同过滤算法

协同过滤推荐技术是当前最成功的推荐技术之一[2,10],典型的协同过滤算法是基于用户的。协同过滤推荐算法的实现过程分为3步:建立用户模型、寻找最近邻居和产生推荐项目。

①建立用户模型:协同过滤算法的输入数据通常表示为一个 $m*n$ 的用户-评价矩阵 R,m 是用户数,n 是项目数,$r_{u,i}$ 表示第 u 个用户对第 i 项的评价值。通常采用5级评分的办法,评分越高表明该用户对该项目的兴趣越大。

②寻找最近邻居:在这一阶段,主要完成对目标用户最近邻居的查找。通过计算目标用户与其他用户之间的相似度,找出与目标用户最相似的"最近邻居"集。即:对目标用户 u,

到稿日期:2009-07-14 返修日期:2009-09-27 本文受重庆市自然基金项目(CSTC2009BB2046)资助。

王 茜 女,博士,副教授,主要研究方向为电子商务、网络安全;王均波 男,硕士生,主要研究方向为电子商务、计算机网络与通信。

产生一个以相似度 $\mathrm{sim}(u,v)$ 递减排列的“邻居”集合 $N_u=\{N_1,N_2,\cdots,N_i\}$，$u\notin N_i$。该过程分两步完成：首先计算用户之间的相似度，可采用皮尔森相关系数、余弦相似性和修正的余弦相似性等度量方法[1]，本文采用皮尔森相关系数（式1），其次是根据如下方法选择“最近邻居”：(1)选择相似度大于设定阈值的用户；(2)选择相似度最大的前 k 个用户；(3)选择相似度大于预定阈值的 k 个用户，本文采用第(2)种方法。

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(r_{u,i}-\bar{r}_u)(r_{v,i}-\bar{r}_v)}{\sqrt{\sum_{i\in I_{uv}}(r_{u,i}-\bar{r}_u)^2*\sum_{i\in I_{uv}}(r_{v,i}-\bar{r}_v)^2}} \tag{1}$$

式中，$\mathrm{sim}(u,v)$ 表示目标用户 u 与用户 v 的相似度，$r_{u,i}$ 表示用户 u 对项目 i 的评分，$r_{v,i}$ 表示用户 v 对项目 i 的评分，$\bar{r}_u$ 和 $\bar{r}_v$ 分别表示用户 u 和用户 v 对项目的平均评分，I_{uv} 表示用户 u 和用户 v 共同评价过的项目集合。

③产生推荐项目：设目标用户 u 已评价项目集合为 I_u，则目标用户 u 对任意项目 i 的预测评分 $p_{u,i}$ 可通过邻居用户 v 对项目 $i(i\notin I_u)$ 的评分得到，计算方法如下：

$$p_{u,i}=\bar{r}_u+\frac{\sum_{v\in N_u}\mathrm{sim}(u,v)*(r_{v,i}-\bar{r}_v)}{\sum_{v\in N_u}\mathrm{sim}(u,v)} \tag{2}$$

通过上述方法预测出目标用户对未评价项目的评分，然后选择预测评分最高的 TOP-N 项推荐给目标用户。

3 改进算法

3.1 时间效应

传统协同过滤算法在寻找用户的最近邻居时，将用户不同时间的项目评分同等对待，没有考虑用户对各个项目的评分不是在同一时间段进行的实际情况。一般来说，用户近期访问过的项目对推荐该用户未来可能感兴趣的项目起比较重要的作用，而早期的访问记录对生成推荐结果影响相对较小，这是因为用户的兴趣偏好随时间发生改变，而在较短的一段时间间隔内用户的兴趣相对稳定。因此一个用户感兴趣的项目最可能和他近期访问过的项目相似。也就是说，虽然所有评价都对推荐结果有影响，但最新评价贡献更大，旧数据反映用户以前的喜好，它在推荐的预测上应占较小的权值。

心理学家对遗忘现象的研究表明：人类的遗忘过程是逐步的、非线性的[11]。借鉴人类的遗忘规律，本文引入非线性逐步遗忘策略-遗忘函数。按时间 t 对项目评分进行衰减，改变不同时间内评分对推荐结果的贡献。

遗忘函数 $h(u,t)$ 用来描述用户 u 随时间的遗忘过程。$h(u,t)$ 是一个单调递减函数，以反映用户 u 近期访问过的项目的重要性，其函数值在(0,1]范围内。目前用于描述人类遗忘过程的遗忘函数主要有线性衰减函数[4]和指数函数[3]。

如图1所示，一般指数函数衰减速度过快，函数值迅速趋于0，而线性衰减函数衰减速度太慢。基于上述分析，论文定义遗忘函数为：

$$h(u,t)=e^{[a^b-(t+a)^b]} \tag{3}$$

a,b 为常数。设 $d_{u,i}$ 表示用户 u 访问项目 i 的时间与该用户最晚访问项目的时间间隔，T 为一时间间隔基本单位，$t=\left[\frac{d_{u,i}}{T}\right]$。

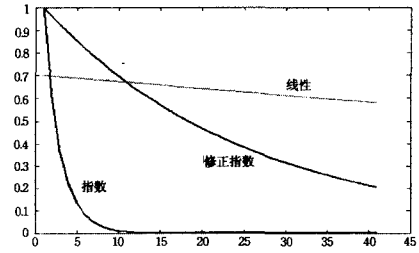

图1 不同函数衰减速度比较

3.2 用户偏好度

传统的协同过滤算法仅考虑了用户或者项目单方面的相似性，忽视了用户和项目两者间的联系。本文引入用户 u 对项目 i 的黏度函数 $y(u,i)$ 来表示用户 u 对项目 i 的偏好度。

设用户 u 已访问过的项目集合为 I_u，对于项目 $i\in I_u$，如果 i 和 I_u 中项目相似度较高，说明项目 i 和用户 u 的当前兴趣较相近，则在未来一段时间内，用户 u 感兴趣的项目很可能也和项目 i 相似，即项目 i 对生成用户 u 的推荐起比较重要的作用。基于此我们定义黏度函数 $y(u,i)$ 来衡量项目 i 和用户 u 当前兴趣的相关程度，它用 i 和 I_u 的平均总体相似度 $\mathrm{sim}(i,I_u)$ 来表示，而 i 和 I_u 平均总体相似度可以通过计算 i 和 I_u 中每个项目的平均相似度获得：

$$y(u,i)=\mathrm{sim}(i,I_u)=\frac{\sum_{j\in I_u}\mathrm{sim}(i,j)}{|I_u|} \tag{4}$$

式中，$|I_u|$ 表示 I_u 中的项目数。为计算 $y(u,i)$，需要计算 i 和 I_u 中每个项目的相似度，若集合 I_u 中项目数为 n，则计算 $y(u,i)$ 的时间复杂度为 $O(n^2)$，一个用户访问过的项目数通常比较小，因此不会对算法的实时性有太大影响。

3.3 用户特征向量

不同特征的人拥有的兴趣爱好可能不同，而同一类别的人兴趣爱好则具有一定的相似性。所以，在寻找目标用户的最近邻居时，用户特征也是一个不容忽视的因素。通过构建用户特征向量，有助于寻找与目标用户更相似的邻居。

年龄层次不同的人，因为其阅历不同，对生活的领悟程度也不同，所以喜欢的物体层次、类别会有些差异。女性多喜欢情感剧，而男性多喜欢警匪片，说明性别差异也会对用户的兴趣取向产生影响。此外，具有相同职业的人对事物更可能有相同的理解角度，往往会喜欢同一类型的东西。因此本文选择年龄、职业和性别3个因素作为表征用户特征的特征因素。

将职业空间中所有职业按其类别描述成一棵倒立的树，称为职业树[1]，如图2所示。

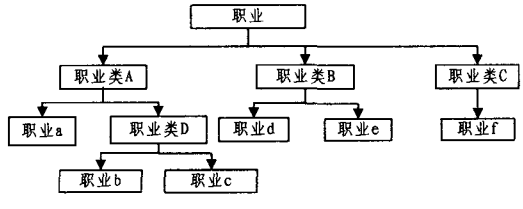

图2 职业树

职业树的总层数称为职业树高度，记为 H_o。职业 a,b 在职业树中最近的共同父类称为 a,b 的最近父类，其位于职业树上的层次称为其高度，记为 $H_{a,b}$，当 a,b 的最近父类为职业树的根节点时，$H_{a,b}=0$。设用户 u 职业为 a，用户 v 职业为 b，定义用户 u,v 在特征向量空间职业维的相似性 $O(u,v)$ 为

职业 a,b 之间的相似度 $\mathrm{sim}_o(a,b)$，即：

$$O(u,v)=\mathrm{sim}_o(a,b)=\begin{cases}\dfrac{H_{a,b}}{H_o}, & (a\neq b)\\ 1, & (a=b)\end{cases} \tag{5}$$

例如在图 2 中，职业树的高度 $H_o=4$，对职业 a 和职业 b 而言，最近共同父类为职业类 A，相应的 $H_{NC(a,b)}=2$，由此职业 a,b 的相似度 $\mathrm{sim}_o(a,b)=0.5$。

设用户 u 的性别为 S_u，用户 v 的性别为 S_v，则用户 u,v 在特征向量空间性别维的相似性 $S(u,v)$表示为：

$$S(u,v)=\begin{cases}0, & S_u\neq S_v\\ 1, & S_u=S_v\end{cases} \tag{6}$$

设用户 u 的年龄为 A_u，用户 v 的年龄为 A_v，则用户 u,v 在特征向量空间年龄维的相似性 $A(u,v)$表示为：

$$A(u,v)=\begin{cases}1, & A_u-A_v\leqslant 5\\ \dfrac{5}{|[A_u-A_v]|}, & A_u-A_v>5\end{cases} \tag{7}$$

用户 u 和 v 在用户特征向量上的相似度 $\mathrm{sim}_v(u,v)$通过如下式来计算：

$$\mathrm{sim}_v(u,v)=\chi S(u,v)+\delta A(u,v)+(1-\chi-\delta)O(u,v) \tag{8}$$

式中，χ,δ 为修正系数，分别影响性别、年龄和职业 3 个维在用户特征向量相似度上所起的作用。一般根据统计信息或者实验结果来调整其大小，不同的推荐系统，可以动态调整其值来优化推荐效果。

3.4 改进算法流程

把上述方法引入到传统的协同过滤推荐算法中，提出了一种改进后的协同过滤算法，通过遗忘函数对原始的用户-评价矩阵 R 按照评价时间进行修正，增加时间效应，突出用户近期评价数据对推荐的贡献，获得修正后的用户-评价矩阵 R'。在传统用户相似度基础上，结合用户在用户特征向量空间上的相似度来构造用户的综合相似度，依据用户综合相似度大小获得用户的最近邻居集。改进后的推荐流程如图 3 所示。

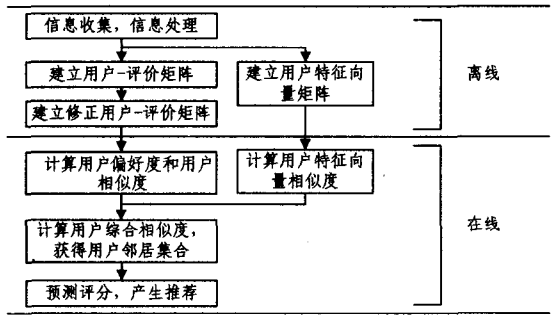

图 3 改进算法流程图

输入：目标用户 u，用户-评价矩阵 R，用户特征向量矩阵 V

输出：向目标用户 u 进行推荐的 TOP-N 推荐集

Setp 1 建立修正用户-项目评分矩阵 R'

对每个评分 $r_{u,i}\in R$，采用式(3)计算 $r'_{u,i}=r_{u,i}\times h(u,t)$，建立修正用户-项目评分矩阵 R'。用户的兴趣在短期内是相对稳定的，修正用户-项目评分矩阵 R' 是可以离线生成的，间隔 T 周期更新。

Step2 计算用户特征向量上的相似度 $\mathrm{sim}_v(u,v)$

在用户特征向量矩阵 V 上，采用式(8)计算目标用户 u 与用户在用户特征向量上的相似度 $\mathrm{sim}_v(u,v)$。

Step3 计算目标用户和候选用户原始相似度 $\mathrm{sim}(u,v)$

在修正用户-项目评分矩阵 R' 上，结合式(4)和皮尔森相关系数法，计算目标用户和候选用户的原始相似度：

$$\mathrm{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(r'_{u,i}-\bar{r}_u)\times(r'_{v,i}-\bar{r}_v)\times y(u,i)}{\sqrt{\sum_{i\in I_{uv}}(r'_{u,i}-\bar{r}_u)^2\times\sum_{i\in I_{uv}}(r'_{v,i}-\bar{r}_v)^2}} \tag{9}$$

Step4 计算用户综合相似度，获取目标用户的邻居集合

在 Step2 和 Step3 基础上，采用式(10)计算目标用户与每个候选用户的复合相似度，同样采用 Top-N 策略获取 n 个用户作为目标用户的邻居用户。

$$\mathrm{sim}'(u,v)=\alpha\mathrm{sim}(u,v)+(1-\alpha)\mathrm{sim}_v(u,v) \tag{10}$$

Step5 进行预测评分，产生推荐

由 step4 计算得到的邻居用户对目标用户未评分的项目进行预测评分，预测评分的计算以及推荐的产生仍然采用第 2 节中介绍的方法。

4 实验结果及分析

4.1 实验数据选取和度量标准

实验数据采用美国 Minnesota 大学 GroupLens 项目组提供的 Movielens 数据集。它包含了 943 名用户对 1682 部电影的评价，每个用户至少评价过 20 部电影，评分值为 1 到 5 之间的整数，数值越高，表明用户对该电影的偏爱程度越高。

本文采用通常使用的平均绝对偏差 MAE 作为准确性度量标准，采用推荐产生的平均消耗时间(MCT)作为算法效率度量标准。MAE 通过计算预测的用户评分与实际的用户评分之间的偏差度量可知预测的准确性，MAE 越小，推荐质量越高。

设目标用户的预测评价集合为 $P_u=(p_{u,i}\mid i=1,2,\cdots,n)$，对应的目标用户实际评价集合为 $R_u=(r_{u,i}\mid i=1,2,\cdots,n\}$。对于每个不为零的预测-评价对 $\langle p_{u,i},r_{u,i}\rangle$，计算

$$MAE=\frac{\sum_{i\in I_u}(p_{u,i}-r_{u,i})}{|I_u|} \tag{11}$$

式中，I_u 为测试集内目标用户 u 的预测值和真实评价值都不为 0 的项目集。

4.2 实验结果及分析

本文设计了 4 组实验，将改进算法与传统的协同过滤算法和基于时间改进型协同过滤算法[3]产生的推荐结果进行对比。式(3)中的 a,b 分别取 20，1/10；式(8)中的 χ,δ 分别取 0.2，0.5。

实验一 测试遗忘函数中时间间隔 T 对推荐结果的影响，T 分别取 10，15，20 进行试验，结果如图 4 所示。

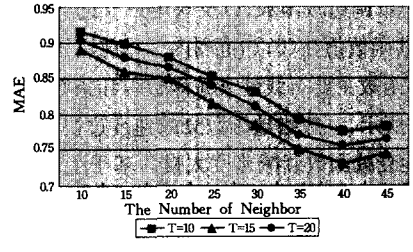

图 4 不同 T 值对推荐结果的影响

实验结果表明，不同 T 值对推荐准确度有一定的影响，过长或者过短都会对算法起到负面作用。当 $T=15$ 时算法

(下转第 243 页)

in Virus-Evolutionary Genetic Algorithm[C]// Proceedings of the IEEE Conference on Evolutionary Computation, ICEC. May 1996:182-187

[10] Kubota N, Arakawa T, Fukuda T, et al. Fuzzy manufacturing scheduling by virus-evolutionary genetic algorithm in self-organizing manufacturing system[C]// Proceedings of the 1997 6th IEEE International Conference on Fussy Systems, FUZZ-IEEE' 97. Part 1 (of 3):1283-1288

[11] Kubota N, Fukuda T. Schema representation in virus-evolutionary genetic algorithm for knapsack problem [C]// Proceedings of the IEEE Conference on Evolutionary Computation, ICEC. May 1998:834-839

[12] 焦李成,杜海峰,刘芳,等. 免疫优化计算学习与识别[M]. 北京:科学出版社,2006

(上接第 228 页)

性能较佳。

实验二　测试式(11)中 α 取不同值对推荐准确度的影响,试验中 α 分别取 0.7,0.8,0.9,如图 5 所示。

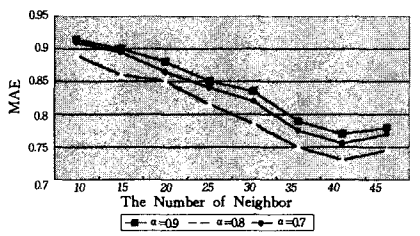

图 5　不同 α 取值对推荐准确度的影响图

从图 5 可以看出,引入时间遗忘函数、用户偏好度、用户特征向量能较大地提高推荐准确度,尤其当邻居数目较大时,准确度提高更为明显。由此可见,综合考虑时间效应、用户偏好有效地突出了用户近期访问数据对生成推荐的重要性,结合用户特征向量,使推荐结果能更好地反映用户的当前兴趣。同时可以看出,α 的取值对推荐准确度也有比较大的影响,α 取值过大或过小都会对推推荐效果产生负面影响,当 $\alpha=0.8$ 时,推荐结果较佳。

实验三　对比本文提出的改进算法、传统的协同过滤算法、基于时间改进型协同过滤算法的推荐准确度。

从图 6 可以看出,本文算法与其他算法相比,推荐精度有一定程度的提高,尤其当邻居数目较大时,准确率提高更为明显。但最近邻居的选取对算法准确率有比较大的影响,其值过大或者过小都会对推推荐效果产生负面影响。本实验取邻居数目为 35 到 45 之间比较合适。

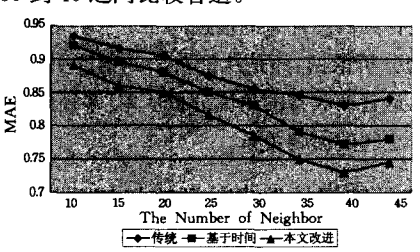

图 6　不同邻居数目对推荐准确度的影响图

实验四　测试本文提出的改进算法与传统的协同过滤算法、基于时间改进型协同过滤算法的时间性能($\alpha=0.8$,邻居数目=35,连续做 50 次实验平均结果)。

表 1　本文改进算法与其他协同过滤在效率的比较

推荐策略	MCT
传统的协同过滤算法	15006.32(ms)
基于时间的协同过滤算法	15194.45(ms)
本文改进协同过滤算法	15206.65(ms)

从表 1 可以看出,本文提出的改进算法在效率上并没有明显的差距。从图 3 可以看到,本文算法采用了离线和在线计算相结合的方式,虽然引入了用户特征向量上的相似度计算和用户偏好度计算,二者的算法复杂度分别为 $O(m)$ 和 $O(n^2)$,但是用户数目 n_1 和用户评价过商品数目 n_2 都相对较小,所以对推荐的效率也没有明显的影响。

结束语　针对协同过滤算法中存在的用户兴趣变化、用户项目间联系单一和缺乏用户特征信息的问题,本文提出了适应用户兴趣变化和基于用户特征的个性化推荐算法,在协同过滤算法的基础上,引入了时间遗忘函数、黏度函数、用户特征向量,改进了寻找最近邻居集合过程,综合考虑了时间效应、用户特征和用户偏好程度,从而提高了寻找最近邻居的准确度。实验表明,本文提出的算法在一定程度上克服了传统协同过滤算法的弊端,能较大提高推荐的准确度,尤其在参数设置得当的情况下,算法性能有明显提高。

参 考 文 献

[1] 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[J]. 合肥工业大学学报:自然科学版,2008(03)

[2] 彭德巍,胡斌. 一种基于用户特征和时间的协同过滤算法[J]. 武汉理工大学学报,2009(03)

[3] 杨怀珍,丛晓琪,刘枚莲. 于时间加权的个性化推荐算法研究[J]. 计算机工程与科学,2009(06)

[4] 郑先荣,曹先彬. 线性逐步遗忘协同过滤算法的研究[J]. 计算机工程,2007(06)

[5] Gong Songjie, Cheng Guanghua. Mining User Interest Change for Improving Collaborative Filtering[C]// Intelligent Information Technology Application 2008. Second International Symposium. Volume 3. Dec. 2008:24-27

[6] 赵智,冯卓楠. 改进的基于相关相似性的协同过滤推荐算法[J]. 长春工业大学学报:自然科学版,2006(04)

[7] Xia Weiwei, He Liang, Ren Lei, et al. A new collaborative filtering approach utilizing item's popularity[C]// Industrial Engineering and Engineering Management. IEEE International Conference, Dec. 2008:1480-1484

[8] Su Xiaoyuan, Khoshgoftaar T M, Greiner R. A Collaborative Filtering Algorithm Based on Variance Analysis of Attributes-Value Preference[C]// IEEE/WIC/ACM International Conference. Volume 1, Dec. 2008:633-639

[9] Dai Y, Ye Hongwu, Gong Songjie. Personalized Recommendation Algorithm Using User Demography Information. Knowledge Discovery and Data Mining[C]// Second International Workshop. Jan. 2009:100-103

[10] Gong Song-jie, Ye Hongwu. Combining Memory - Based and Model-Based Collaborative Filtering in Recommender System [C]// Circuits, Communications and Systems. Pacific-Asia Conference. 2009:690-693

[11] 张国忠. 遗忘规律的探索与运用[J]. 继续教育研究,2002(06)

[12] 王元明,熊伟. 异常数据的检测方法[J]. 重庆工学院学报:自然科学版,2009,23(2):86-89