

# 基于用户-项目的混合协同过滤算法

陈彦萍, 王 赛

(西安邮电大学 计算机学院, 陕西 西安 710121)

**摘 要:**针对传统协同过滤方法中存在的冷启动和数据稀疏等问题,结合基于用户的协同过滤和基于项目的协同过滤提出一种混合协同过滤算法。在相似度的计算中提出改进算法来提高相似度计算的精确度;在预测未评分值时引入控制因子、平衡因子进行加权综合预测,最后再进行综合推荐。实验过程中采用 MovieLens 数据集作为测试数据,同时采用平均绝对误差作为实验的测试标准。实验结果表明,基于用户-项目混合协同过滤算法在评分矩阵极度稀疏的环境下提高了推荐的性能,并能有效提高预测的精度。

**关键词:**协同过滤;推荐;未评分值预测;冷启动;数据稀疏

中图分类号:TP311

文献标识码:A

文章编号:1673-629X(2014)12-0088-04

doi:10.3969/j.issn.1673-629X.2014.12.021

## A Hybrid Collaborative Filtering Algorithm Based on User-item

CHEN Yan-ping, WANG Sai

(School of Computer, Xi'an University of Posts and Telecommunications,  
Xi'an 710121, China)

**Abstract:**According to the problems such as cold start, sparse data existed in the traditional collaborative filtering algorithms, a hybrid collaborative filtering algorithm is proposed which combines user-based and item-based collaborative filtering. An improved algorithm is proposed to improve the accuracy of similarity calculation in the similarity algorithm. The control factors and balance factors are introduced in the missing data prediction process for the finally comprehensive recommendation. MovieLens dataset is applied in the experiments, the mean absolute error is used for the experiment as a test standard. Experimental results show that the user-item hybrid collaborative filtering algorithm can improve the recommendation performance and prediction accuracy in the extremely sparse matrix.

**Key words:**collaborative filtering; recommendation; missing data prediction; cold start; sparse data

## 0 引 言

目前推荐系统<sup>[1]</sup>根据不同的推荐对象可分为两种方式:第一种推荐对象为网页,采用数据挖掘<sup>[2]</sup>的方式来分析用户行为并将网页链接推荐给用户;第二种推荐对象为商品,主要是应用在电子商务中并为用户推荐商品。

电子商务没有传统商业模式存在的地域限制,消费者有更大的选择空间<sup>[3]</sup>。但随着互联网上的信息迅速膨胀会出现“信息过载”现象<sup>[4]</sup>,即当用户在网站浏览或购买时,因网站中充斥着各种各样的物品,可能会有多种分类可供用户进行选择。用户如何从海量的资源中快速地找到喜欢的物品是目前急需要解决的问题,而且用户也需要一种能根据自己的需求自动提供其感兴趣物品的技术。

目前协同过滤<sup>[5]</sup>技术已经广泛应用于推荐系统中。协同过滤的基本思想是先从相似用户中收集数据,然后向特定的用户预测其感兴趣的内容<sup>[6]</sup>。根据协同过滤的相关特征,将协同过滤算法分成基于用户的协同过滤算法、基于项目的协同过滤算法。

基于用户的协同过滤算法是根据用户与其具有相似的用户集合之间的相似性来产生推荐的,但是随着用户数量的增加有时会产生不能及时反映出用户的兴趣变化等情况。基于项目的协同过滤的算法是根据项目与其具有相似的项目集合之间的相似性来产生推荐的,但它会存在如用户对项目的评分过少容易忽略项目自身属性的问题,这样也会造成预测的精度不准确。它们存在的这些缺点会产生数据稀疏<sup>[7]</sup>、冷启动<sup>[8]</sup>和可扩展性<sup>[9]</sup>差等问题。鉴于它们各自存在的一些缺

收稿日期:2014-02-26

修回日期:2014-05-30

网络出版时间:2014-10-23

基金项目:陕西省自然科学基金资助项目(2012JQ8029);陕西省教育科研计划资助项目(12JK0938)

作者简介:陈彦萍(1979-),女,博士,副教授,CCF会员,从事服务计算研究;王 赛(1986-),女,硕士研究生,研究方向为面向服务的计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20141023.1124.038.html>

点,文中将基于用户的协同过滤和基于项目的协同过滤两者相结合,提出一种基于用户和项目的混合协同过滤算法。

## 1 用户-项目评分矩阵

一个推荐系统中包含  $M$  个用户  $\{u_1, u_2, \dots, u_m\}$  和  $N$  个项目  $\{i_1, i_2, \dots, i_n\}$ , 它们形成了一个  $M \times N$  的矩阵列表,称为用户-项目矩阵表。矩阵中的每一个  $r_{x,y}$  ( $1 \leq x \leq m, 1 \leq y \leq n$ ) 表示用户  $x$  对项目  $y$  的评分值,这个评分值可以设置为 0-5 之间的整数。如果用户  $m$  对项目  $n$  未评分,则  $r_{x,y} = 0$ 。评分值越高,表示用户对项目越认可。

## 2 协同过滤算法

### 2.1 相似性计算

文中主要利用皮尔逊相关系数<sup>[10]</sup> (Pearson Correlation Coefficient, PCC) 计算用户或项目之间的相似度, PCC 的值介于 -1 到 1 之间。相比其他相似度的计算方法,在推荐系统中使用 PCC 可以达到更高的推荐精度。在基于用户的协同过滤中, PCC 可以用来计算基于项目时的用户  $u$  和用户  $v$  的相似度。公式如下所示:

$$\text{sim}(u, v) = \frac{\sum_{i \in I(u) \cap I(v)} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in I(u) \cap I(v)} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in I(u) \cap I(v)} (r_{v,i} - \bar{r}_v)^2}}$$

基于项目的协同过滤算法<sup>[11]</sup> 与上述方法类似。基于项目 PCC 计算方法如下:

$$\text{sim}(i, j) = \frac{\sum_{v \in V(i) \cap V(j)} (r_{v,i} - \bar{r}_i)(r_{v,j} - \bar{r}_j)}{\sqrt{\sum_{v \in V(i) \cap V(j)} (r_{v,i} - \bar{r}_i)^2} \sqrt{\sum_{v \in V(i) \cap V(j)} (r_{v,j} - \bar{r}_j)^2}}$$

### 2.2 相似度的改进算法

PCC 一般适用于用户对项目的评分数据较多的情况。但是实际应用中,用户评分的数据都非常稀疏, PCC 就不能及时反映出用户或项目之间的相似度了。文献[12]为了解决出现的数据稀疏问题,提出了如下的相似度的改进算法:

$$\text{sim}'(u, v) = \frac{\text{Max}(|I_u \cap I_v|, \theta)}{\theta}$$

上述改进算法虽然可以有效解决用户评分项目过少出现的数据稀疏问题,但如果  $\theta$  在足够小而  $|I_u \cap I_v|$  比  $\theta$  大的多的情况下,  $\text{sim}'(u, v)$  的值有可能会超过 1。

为了有效解决这个问题,提出一种有效的相似度

度量 (Effective Similarity Measurement, ESM) 方法,它可以有效地避免参数  $\theta$  的影响。基于用户的 ESM 的相似度定义如下:

$$\text{sim}'(u, v) = \frac{|I_u \cap I_v|}{|I_u \cup I_v|} \cdot \text{sim}(u, v)$$

与上述方法类似,基于项目之间的 ESM 的相似度定义如下:

$$\text{sim}'(i, j) = \frac{|U_i \cap U_j|}{|U_i \cup U_j|} \cdot \text{sim}(i, j)$$

### 2.3 最近邻集合的选择

最近邻集合的选择是预测未评分值中最重要的一步。将相似度小于 0 或等于 0 的情况排除可有效寻找最近邻的集合。

所以,基于用户最近邻的集合表示如下:

$$N(u) = \{u_m \mid \text{sim}'(u, v) > 0, u_m \neq u\}$$

与基于用户的方法类似,基于项目的最近邻集合表示如下:

$$N(i) = \{i_n \mid \text{sim}'(i_n, i) > 0, i_n \neq i\}$$

### 2.4 未评分值预测

用户-项目矩阵中的数据一般比较稀疏,文献[13]提出了在矩阵中预测未评分值的方法来改善矩阵存在的稀疏问题。

在基于用户的协同过滤中,假设用户的最近邻集合为  $N(u)$ ,则用户  $u$  对项目  $i$  的预测评分  $P_u(r_{u,i})$  的计算方法如下:

$$P_u(r_{u,i}) = \frac{\sum_{u_m \in N(u)} \text{sim}'(u_m, u)(r_{u_m,i} - \bar{u}_m)}{\sum_{u_m \in N(u)} \text{sim}'(u_m, u)}$$

与基于用户的方法类似,基于项目的协同过滤未评分值预测的方法如下:

$$P_i(r_{u,i}) = \frac{\sum_{i_n \in N(i)} \text{sim}'(i_n, i)(r_{u,i_n} - \bar{i}_n)}{\sum_{i_n \in N(i)} \text{sim}'(i_n, i)}$$

在计算未评分值的时候,如果只使用基于用户的预测方法或是只使用基于项目的预测方法,这样有可能忽略有用的信息。为了尽可能精确地预测未评分值,文献[13]使用了系数  $\lambda$  将基于用户的方法与基于项目的方法进行有效结合,具体方法如下:

$$P(r_{u,i}) = \lambda \cdot P_u(r_{u,i}) + (1 - \lambda) \cdot P_i(r_{u,i})$$

为了精确达到预测结果,文中提出了平衡因子  $m_u$  与  $m_i$  和控制因子  $\lambda$  ( $0 \leq \lambda \leq 1$ ) 相结合的方法来平衡两种预测方法的精度。

平衡因子  $m_u$  的计算方法如下:

$$m_u = \frac{\sum_{u_m \in N(u)} (\text{sim}'(u_m, u))^2}{\sum_{u_m \in N(u)} \text{sim}'(u_m, u)}$$

平衡因子  $m_i$  的计算方法如下:

$$m_i = \frac{\sum_{i_n \in N(i)} (\text{sim}(i_n, i))^2}{\sum_{i_n \in N(i)} \text{sim}(i_n, i)}$$

在平衡因子  $m_u, m_i$  和控制因子  $\lambda$  二者结合的基础上提出参数  $t_u$  和  $t_i$ 。其中参数  $t_u, t_i$  分别定义如下:

$$t_u = \frac{m_u \times \lambda}{m_u \times \lambda + m_i \times (1 - \lambda)}$$

$$t_i = \frac{m_i \times (1 - \lambda)}{m_u \times \lambda + m_i \times (1 - \lambda)}$$

从  $t_u, t_i$  的计算公式中可以得出:

$$t_u + t_i = 1$$

当  $N(u) \neq \emptyset \cap N(i) \neq \emptyset$  时, 未评分值  $P(r_{u,i})$  则计算如下:

$$P(r_{u,i}) = t_u \times P_u(r_{u,i}) + t_i \times P_i(r_{u,i})$$

### 3 实验结果及其分析

文中采用 MovieLens 数据集作为实验所需要的测试数据, 同时采用平均绝对误差 (Mean Absolute Error, MAE) 作为实验的测试标准。并通过相关实验来验证文中算法的可行性。

#### 3.1 实验所用的数据集

实验采用的是 MovieLens (<http://www.grouplens.org/>)<sup>[14]</sup> 数据集, 在实验中使用训练数据集 (u1. base) 来实现用户对未评分的电影进行分值预测, 预测的评分与测试数据集 (u1. test) 中的分值进行比较。训练数据集包括了 943 个用户对 1 682 部电影的 8 万条评分在 1~5 分之间的评分记录, 测试数据集包括了 462 个用户的 2 万条评分在 1~5 分之间的评分记录。

该实验中用户评分的密度为:

$$\frac{100\ 000}{943 \times 1\ 682} \times 100\% = 6.30\%$$

因此用户评分的稀疏度为 93.70%, 由此看出评分数据非常稀疏。

#### 3.2 实验度量标准

文中采用 MAE 作为衡量推荐质量的标准, MAE 是通过用户的预测评分与实际评分之间的差值来判断预测的精确度。例如, 训练数据集中预测用户对项目的评分集合为  $\{r_1, r_2, \dots, r_i, \dots, r_n\}$ , 而实际的测试数据集中用户对项目的评分集合为  $\{t_1, t_2, \dots, t_i, \dots, t_n\}$ , 则 MAE 可定义如下:

$$\text{MAE} = \frac{\sum_{i=1}^n |r_i - t_i|}{n}$$

从 MAE 的计算公式中可以看出, MAE 越小代表预测的性能越高。

#### 3.3 实验方案

文中主要通过三个实验分别验证文中提出算法的

可行性。

实验一: 控制因子  $\lambda$  的最佳取值。

控制因子  $\lambda$  在用户对项目的未评分值预测的过程中起到了十分重要的作用, 同时最近邻居数目的选择也很重要, 邻居数目如果选取的过多会造成系统运行比较慢, 如果邻居数目选取的过少则会存在冷启动问题。

由文献[15]可知在数据集极端稀疏时, 项目之间的相似性会比用户之间的相似性更稳定一些。从项目在整体上优于用户方面分析, 文中的控制因子  $\lambda$  在  $[0, 0.6]$  之间取值, 间隔为 0.1。为了更有效地验证  $\lambda$  的最佳取值范围, 将邻居数 (用  $N$  表示) 从 10 取值到 50, 当用户数为 200 和 300 时对应的 MAE 曲线变化图分别如图 1 和图 2 所示。

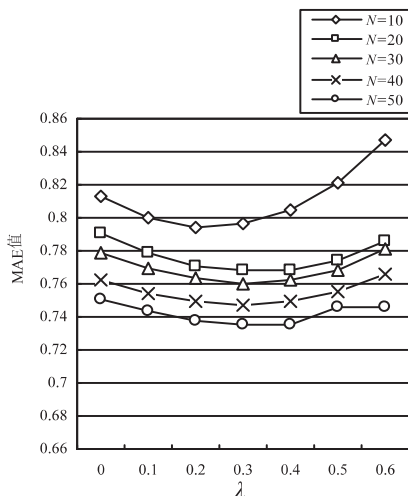


图 1 用户数为 300 时不同  $\lambda$  取值下的 MAE 变化值

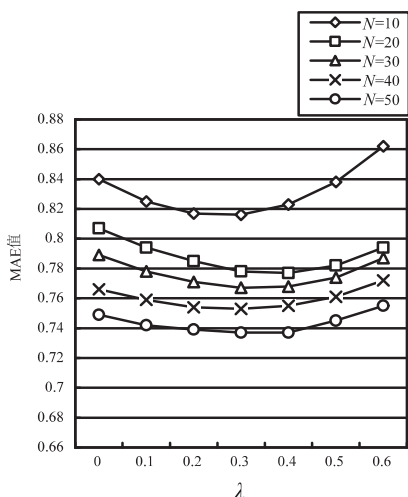


图 2 用户数为 200 时不同  $\lambda$  取值下的 MAE 变化值

MAE 的值越小, 表明预测越精确。从图 1 和图 2 中可以观察到,  $\lambda$  的值在  $[0.2, 0.4]$  之间时 MAE 的值逐渐趋于平滑。当  $\lambda$  取值为 0.3 时 MAE 的值最小; 同时邻居数目为 50 时性能最优。

实验二: 算法性能最优时邻居数目取值。

从图 1 和图 2 中可以得出  $\lambda = 0.3$  时性能最优,为了分析在最近邻居数目(用  $N$  表示)不同取值时对 MAE 的影响,将最近邻居数目的取值分别设置为  $[5, 50]$  之间的整数,区间为 5。在用户数(用  $U$  表示)取值为 200 和 300 时,实验结果如图 3 和图 4 所示。

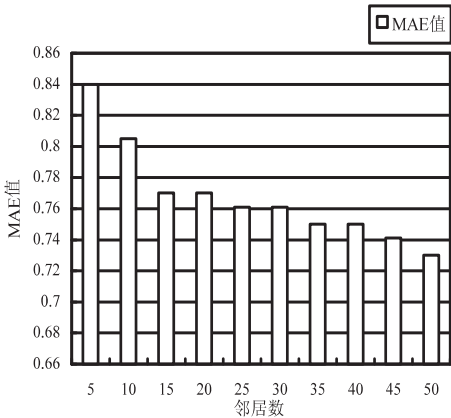


图 3 用户数为 300 时不同邻居数取值下的 MAE 变化值

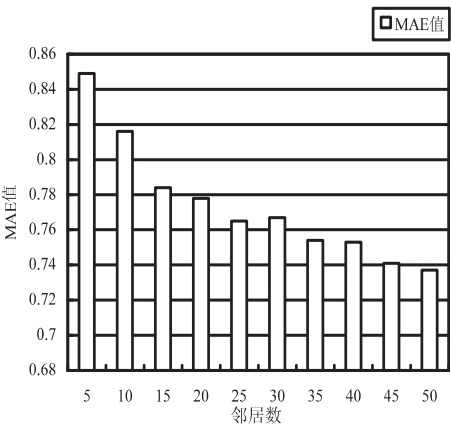


图 4 用户数为 200 时不同邻居数取值下的 MAE 变化值

从图 3 和图 4 中可以看出在用户数为 200 和 300 时,随着邻居数目的取值的增大 MAE 值逐渐变小,在邻居数目取值在  $[35, 50]$  之间时逐渐趋于平滑,且用户数为 300 时优于用户数为 200 时的系统性能。

实验三:不同算法的 MAE 值比较。

为了进一步验证文中的基于用户-项目(用 Proposed 表示)算法的推荐效果,分别计算了基于用户的协同过滤算法(用 User-Based 表示)和基于项目的协同过滤算法(用 Item-Based 表示),邻居数目同样从 10 增加到 50。将文中的算法与这两种算法的 MAE 值分别进行比较,实验结果如图 5 所示。

由图 5 可以清晰地看出,文中提出的基于用户-项目的协同过滤算法的 MAE 值明显小于基于用户的协同过滤算法和基于项目的协同过滤算法,所以文中提出的算法有更好的推荐效果。

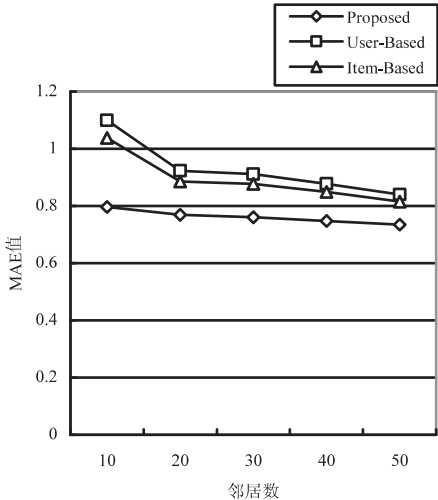


图 5 三种不同算法的 MAE 值比较

4 结束语

针对目前协同过滤中存在的冷启动和数据稀疏等问题,文中提出一种基于用户-项目的混合协同过滤算法。在算法中引入控制因子  $\lambda$ 、平衡因子  $m_u$  与  $m_i$  以及参数  $t_u$  与  $t_i$ ,将基于用户的预测未评分值的方法与基于项目的预测未评分值的方法有效地结合起来,并与传统的基于用户的协同过滤和基于项目的协同过滤算法分别比较之后,验证了文中提出算法的可行性。

从文中可以看出,两种算法相结合可以产生出更好的性能,因此下一步工作的研究重点是在增强算法的可扩展性分析的基础上做出更多有关用户-项目算法的研究。

参考文献:

[1] Gong Songjie. The collaborative filtering recommendation based on similar-priority and fuzzy clustering [C]//Proceeding of 2008 workshop on power electronics and intelligent transportation system. [s. l.]:Inst of Elec and Elec Eng Computer Society,2008:248-251.

[2] 冯海涛,谷文星. 一种洞察客户的“价值-行为”数据挖掘方法及应用[J]. 西安邮电学院学报,2012,17(5):116-121.

[3] 朱 岩,林泽楠. 电子商务中的个性化推荐方法评述[J]. 中国软科学,2009(2):183-192.

[4] 李 勇,徐振宁,张维明. Internet 个性化信息服务研究综述[J]. 计算机工程与应用,2002,38(19):183-188.

[5] Candillier L, Meyer F, Boulle M. Comparing state-of-the-art collaborative filtering systems[C]//Proceedings of the 5th international conference on machine learning and data mining in pattern recognition. [s. l.]:Springer-Verlag,2007:548-562.

[6] 王 霞. 协同过滤在电子商务推荐系统中的应用研究[D]. 南京:河海大学,2003.

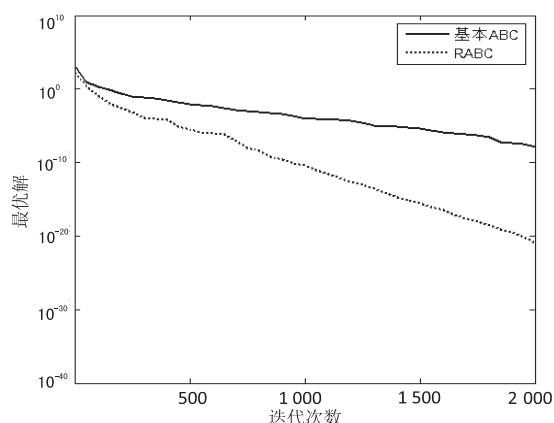


图 2 函数 Schwefel2.22 的进化过程图

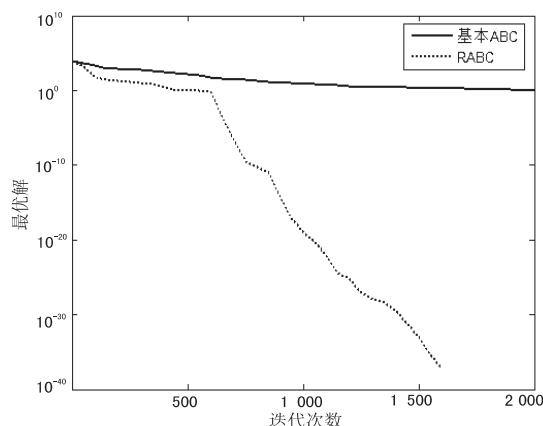


图 3 函数 Rastrigin 的进化过程图

## 5 结束语

针对基本 ABC 算法局部寻优能力较差、收敛精度不高的问题,利用混沌与逆向学习混合算子有效改善了全局分布的均匀性,利用具有最佳引导个体的检索方程提高局部寻优能力,利用迭代次数线性减小检索空间以加快算法的收敛速度,并提高其全局寻优能力<sup>[13]</sup>。通过对单峰函数和复杂的多峰函数的实验结

果表明,与基本 ABC 算法相比,RABC 算法在检索效率、局部寻优能力、收敛精度、稳定性等方面都优于基本 ABC 算法,是一种较好的函数优化算法。

## 参考文献:

- [1] 袁亚杰. 一种改进的人工蜂群算法[J]. 中国科技信息, 2011(24):102-103.
- [2] Karaboga D, Basturk B. A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm[J]. Journal of Global Optimization, 2007, 39(3): 459-471.
- [3] Gao Weifeng, Liu Sanyang. Improved artificial bee colony algorithm for global optimization[J]. Information Processing Letters, 2011, 111(17): 871-882.
- [4] Banharnakun A, Achalakul T, Sirinaovakul B. The best-so-far selection in artificial bee colony algorithm[J]. Applied Soft Computing, 2011, 11(2): 2888-2901.
- [5] 郭超峰, 李梅莲. 基于自适应搜索的人工蜂群算法[J]. 信阳师范学院学报: 自然科学版, 2013, 26(3): 446-449.
- [6] 张银雪, 田学民, 曹玉苹. 改进搜索策略的人工蜂群算法[J]. 计算机应用, 2012, 32(12): 3326-3330.
- [7] 李海生. 蜂群算法及其在垂直 Web 搜索中的应用[D]. 广州: 广州大学, 2010.
- [8] 银建霞. 人工蜂群算法的研究及其应用[D]. 西安: 西安电子科技大学, 2012.
- [9] 于君, 刘弘. 基于人工蜂群算法的群体动画研究与应用[J]. 计算机仿真, 2012, 29(1): 180-183.
- [10] 张超群, 郑建国, 王翔. 蜂群算法研究综述[J]. 计算机应用研究, 2011, 28(9): 3201-3205.
- [11] 毕晓君, 王艳娇. 加速收敛的人工蜂群算法[J]. 系统工程与电子技术, 2011, 33(12): 2755-2761.
- [12] 王珊, 顾幸生. 基于混沌优化的双种群人工蜂群算法[J]. 上海电子学院学报, 2012, 15(1): 11-17.
- [13] 李林菲, 马苗. 基于 ABC 算法的逻辑推理题快速求解方法[J]. 计算机技术与发展, 2011, 21(6): 125-127.


(上接第 91 页)

- [7] 田伟, 彭玉青. 基于电子商务应用的协同过滤技术改进综述[J]. 计算机工程与科学, 2008, 30(10): 61-63.
- [8] 邓爱林, 朱扬勇, 施伯乐. 基于项目评分预测的协同过滤推荐算法[J]. 软件学报, 2003, 14(9): 1621-1628.
- [9] 周军锋, 汤显, 郭景峰. 一种优化的协同过滤推荐算法[J]. 计算机研究与发展, 2004, 41(10): 1842-1847.
- [10] 朱锐, 王怀民, 冯大为. 基于偏好推荐的可信服务选择[J]. 软件学报, 2011, 22(5): 852-864.
- [11] Deshpande M, Karypis G. Item-based top-n recommendation algorithms[J]. ACM Transactions on Information Systems, 2004, 22(1): 143-177.
- [12] McLaughlin M R, Herlocker J L. A collaborative filtering algorithm and evaluation metric that accurately model the user ex-

perience[C]//Proceedings of SIGIR. Sheffield: Association for Computing Machinery, 2004: 329-336.

- [13] Ma Hao, King I, Lyu M R. Effective missing data prediction for collaborative filtering[C]//Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. Amsterdam, The Netherlands: [s. n.], 2007: 39-46.
- [14] Kim B M, Li Q, Park C S, et al. A new approach for combining content-based and collaborative filters[J]. Journal of Intelligent Information System, 2006, 27(1): 79-91.
- [15] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international World Wide Web conference. [s. l.]: [s. n.], 2001: 285-295.

## 基于用户-项目的混合协同过滤算法

作者: [陈彦萍](#), [王赛](#), [CHEN Yan-ping](#), [WANG Sai](#)  
作者单位: [西安邮电大学 计算机学院, 陕西 西安, 710121](#)  
刊名: [计算机技术与发展](#)   
英文刊名: [Computer Technology and Development](#)  
年, 卷(期): 2014(12)

本文链接: [http://d.g.wanfangdata.com.cn/Periodical\\_wjz201412021.aspx](http://d.g.wanfangdata.com.cn/Periodical_wjz201412021.aspx)