

# 基于用户多兴趣的协同过滤策略改进研究

徐红<sup>1,2</sup>, 彭黎<sup>1</sup>, 郭艾寅<sup>2</sup>, 徐云剑<sup>2</sup>

(1. 湖南大学软件学院, 湖南长沙 410205;

2. 湖南涉外经济学院 计算机科学与技术学部, 湖南长沙 410205)

**摘要:**协同过滤机制利用用户之间的相似性来推荐信息,能够为用户发现新的感兴趣的内容,它作为一种行之有效的技术被应用到很多领域中。但传统的协同过滤算法不能反映用户的多个兴趣及兴趣更新情况。基于此不足,在用户聚类协同过滤算法的基础上进行了改进,其基本思想是在基于用户聚类的基础上研究用户多兴趣的表示。针对用户兴趣偏好及更新情况引入基于时间的数据阈值、兴趣类型和用户项目兴趣权重的概念和公式。在此基础上将它们有机结合,融入基于用户聚类的协同过滤算法的推荐过程中。实验表明,改进后的算法比传统协同过滤算法在推荐准确度上有明显提高。

**关键词:**协同过滤;基于时间的数据阈值;基于兴趣的数据权重;用户多兴趣的表示

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2011)04-0073-04

## User-Based Collaborative Filtering Strategies More Interested in Improvement of Research

XU Hong<sup>1,2</sup>, PENG Li<sup>1</sup>, GUO Ai-yin<sup>2</sup>, XU Yun-jian<sup>2</sup>

(1. Software School of Hunan University, Changsha 410205, China;

2. Department of Computer Science and Technology, Hunan International Economics University, Changsha 410205, China)

**Abstract:** Collaborative filtering using the similarity between users to recommend information to users interested in discovering new content, it acts as an effective technology has been applied to many fields. However, the traditional collaborative filtering algorithms can not reflect the multi-user interest and user interest changes. To address this problem, a collaborative filtering based on user clustering strategies to improve the basic idea is the basis of user-based clustering of users and more interested in that. Time threshold and the introduction of user interest data on the weight of the concepts, definitions and formulas to calculate the user preferences for different project categories and changes of interest. On this basis, they will combine the introduction of user-based collaborative filtering algorithm for clustering the recommended process. Experimental results show that the improved algorithm than the traditional collaborative filtering recommendation algorithm accuracy are dramatically increased.

**Key words:** collaborative filtering; time threshold; weights based on the data of interest; users expressed more interest

## 0 引言

近年来,随着 Internet 的日益普及,各站点纷纷转向“以用户为中心”<sup>[1]</sup>的发展模式。推荐系统应运而生,目前推荐系统中应用最为成功的是协同过滤技术<sup>[2-4]</sup>,广泛地应用于电子商务、数字图书馆等众多领域<sup>[5,6]</sup>。

协同过滤的基本出发点是:①用户是可以按兴趣分类;②用户对不同的信息评价包含了用户的兴趣信

息或潜在需求;③用户对一个未知信息的评价将和其相似(兴趣)用户的评价相似<sup>[7]</sup>。这三条构成了协同过滤系统的基础。它根据群体用户的评价和反馈来向当前活动用户进行推荐,通过发现用户与用户之间、资源项目与项目之间存在着的关系、关联或特征模式来向当前用户推荐可能感兴趣或有价值的资源对象项目。

协同过滤技术在理论和应用上都取得了一定的成功。但是也存在一定的局限性,即要获得满意的效果,需要拥有大量用户信息数据。随着项目或用户的不断增加,协同过滤面临的主要挑战有:①稀疏性问题;②冷启动问题;③可扩展性问题<sup>[8]</sup>;④用户多兴趣的表示问题<sup>[9]</sup>。

文中在分析现有的主流协同过滤算法的基础上,

收稿日期:2010-08-04;修回日期:2010-11-13

基金项目:湖南省教育科学研究项目(09C591,09C600)

作者简介:徐红(1976-),女,湖南长沙人,硕士,讲师,研究方向为数据挖掘和个性化服务;彭黎,博士,副教授,研究方向为数据挖掘、系统分析与建模、软件工程。

重点剖析了协同过滤算法在用户多兴趣下的推荐问题,提出了一种能适应用户多兴趣及兴趣更新的协同过滤推荐改进算法,对改进算法进行了实验仿真,验证了算法的有效性。

## 1 相关工作

### 1.1 基于用户聚类协同过滤算法

传统的协同过滤算法是在所有的用户空间上进行最近邻居,随着项目或用户的不断增加,协同过滤面临的数据稀疏性问题、冷启动问题和扩展性问题越来越突出。为解决协同过滤推荐系统存在的上述问题,一种方案为引入聚类技术<sup>[10]</sup>,即将数据对象分组成为多个类或簇,一旦簇建立,可以使用簇中其它用户(项目)的平均观点对目标用户(项目)进行预测。

基于用户聚类的协同过滤算法<sup>[11]</sup>就是目前应用比较广泛且效率较高的一种个性化推荐算法,其基本思想如图 1 所示。首先,在分析用户特点和习惯等的基础上使用聚类算法将整个用户空间划分为若干个不同的聚类(或簇),不同聚类(簇)之间用户对项目的评分尽可能不同,而聚类(簇)内部用户对项目的评分尽可能相似;其次,搜寻目标用户所在聚类(簇);最后,通过协同过滤算法来预测评分。因此,聚类后的用户空间相对于原始的用户空间要小得多,有效地提高了最近邻的查询效率和推荐算法的准确性。

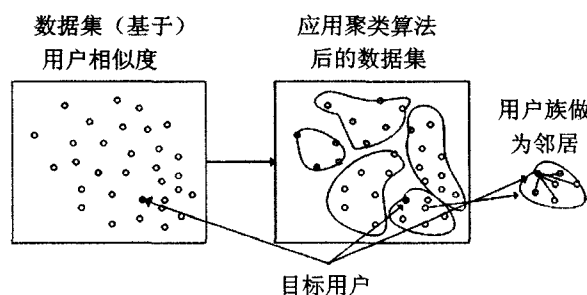


图 1 基于用户聚类的协同过滤示意图

划分的方法很多,K-means 聚类算法是被应用最多的聚类算法之一,其聚类算法过程描述如下:

给定一个包含  $n$  个数据对象的数据库,以及要生成的簇的数据  $k$ ,一个划分的算法将数据对象组织为  $k$  个划分( $k \leq n$ ),其中每个划分代表一个簇。通常会采用一个划分的准则(一般如相似度函数),例如距离,以便在同一个簇的对象是“相似的”,而不同簇中的对象是“相异的”。

采用 K-means 聚类算法对整个用户空间进行聚类的主要步骤如下:

输入:聚类数据  $k$  和用户-项目评分矩阵。

输出: $k$  个簇,使平方误差准则最小。

过程:

(1) 从用户-项目评分矩阵中检索得到所有  $n$  个项,记为集合  $I = \{I_1, I_2, \dots, I_n\}$ 。

(2) 从用户-项目评分矩阵中检索得到所有  $m$  个用户,记为集合  $U = \{U_1, U_2, \dots, U_n\}$ 。

(3) 在用户集  $U$  中,随机选择  $K$  个对象作为初始的聚类中心。

(4) 使用相似性计算公式计算簇中每个对象与聚类中心的相似度;根据计算结果,将每个对象分配到相似性最高的聚类中。

(5) 计算新聚类中所有用户对项的平均评分,生成新的聚类中心。

(6) repeat(3) 到(5)步,until 聚类不再发生改变,得到相似用户群。

(7) 产生推荐,从目标用户的相似用户群中产生  $N$  项推荐。

### 1.2 算法中存在的不足

从实际应用来看,该方法在用户-项目矩阵数据稀疏的情况下仍然存在下列问题:

首先,基于用户聚类的方法没有考虑用户在不同时间段访问某项目时的兴趣差异和变化因素,因此无法反映出用户的兴趣随时间的变化过程,当用户兴趣发生改变的时候,现有的推荐系统无法及时发现,从而导致系统推荐的资源在很大程度上偏离了用户的需求。

第二,用户-项目矩阵没有考虑项目和用户兴趣的实际内容,在数据稀疏的时候极易产生遗漏值。比如用户 X 访问评价了项目 A、B、C,而用户 Y 访问评价了 D、E、F,显然根据用户-项目矩阵,用户 X 和 Y 的相似性为 0,而实际上他们访问的项目在内容上可能是很相似的,两者的兴趣主题也是一致的。

第三,无法处理由于时间推移和数据规模扩大所带来的计算处理矛盾。基于用户-项目矩阵的用户聚类处理时,参与计算的是所有用户所有时间段的项目数据,随着时间的推移,计算规模会越来越大,用户-项目矩阵的稀疏性往往也将加大。

基于上述问题,文中将两种数据策略:基于时间的数据阈值和基于兴趣的数据权重,在此基础上将它们有机结合,引入基于用户背景特点聚类的协同过滤算法的推荐过程中,改进算法流程,以解决此弊端。

## 2 基于用户聚类的协同过滤改进算法描述

### 2.1 基于用户多兴趣问题描述的改进

注意到,随着资源项目的日益增多,用户的评分数据却一般只集中在自己感兴趣的领域,前者的增长速度远远高于后者,这样数据的稀疏性就日益突现。虽然资源项目越来越多,但其可以根据自身的内容,划分

为不同的类别,且类别总数还相对比较固定。

故对传统协同过滤算法提出改进策略:

(1)用户多兴趣的表示,因为实际上用户的兴趣可通过对项目进行选择来进行了解,所以把对用户兴趣的表示转化为用户对多种不同类型项目的选择,即项目类型确定用户兴趣,这样既可以大大降低数据的维度和稀疏程度,又可以扩大资源推荐的范围。

(2)引入一个基于时间的数据阈值和用户兴趣权重的概念,来探讨用户在不同时间段中所表现出来的兴趣差异和变化,进而实现对用户多兴趣的了解。

(3)用户感兴趣程度不一,在推荐集中应考虑从用户对不同项目类别的兴趣权重出发,来确定分配每类项目的推荐数目,以满足用户多兴趣的要求。

在实际的推荐系统中,项目的描述信息中普遍存在着对项目所属类别的描述。当然也可根据具体应用情况对项目进行分类,目前对项目进行分类的方法很多。比如中图分类和主题词法,概念分层和利用分类和聚类技术自动生成项目类别等。

当把整个项目空间划分成若干类别之后,再考虑用户多兴趣模型的表示。用户模型中引入基于时间的数据阈值  $Ttime$  和基于项目类别兴趣的数据权重  $W$ 。

初始的用户兴趣模型用  $I^{(0)} = \{(t_1, W_1, S_1), (t_2, W_2, S_2), \dots, (t_i, W_i, S_i), \dots, (t_n, W_n, S_n)\}$  表示。其中,  $t$  表示某一项目类别;  $W$  表示基于用户访问项目类别的兴趣权重,权重值越大,表示该用户在该项目类别方面的兴趣越大;  $W$  的引入可有效地适应用户兴趣的变化;  $S$  表示是否为长期兴趣。

如用户一年前某项目类别的权值已达到较大值,但一年以来很少或未存在同类项目历史信息,则说明相应的兴趣度有所下降。若此时仍采用原来的向量,则会出现兴趣度量的偏差。以时间作为一个度量因素,给定一个时间阈值  $Ttime$  (可设为两个月,或五个月等),若用户在时间阈值范围内没有相关操作,则应将相应项目类别对应的基本时间的兴趣权重要用公式(1)进行修改。

$$W = W \times (1 - mdate/Ttime) * D \quad (1)$$

$mdate$  为最近访问某资源的时间间隔;  $Ttime$  为最早访问某资源的时间间隔;  $D$  为调节常数,用来控制不同分类兴趣(长期兴趣或非长期兴趣)值减小的速度。

同理,若用户在某一时间阈值  $Ttime$  内对该类别资源的操作比较多,则相应的项目类别的兴趣权重应增加。增加的公式可用公式(2):

$$W = W + r * D * k / Ttime \quad (2)$$

其中,  $r$  为资源的操作方式参数,根据方式的重要性各设定一个值,  $D$  为调节常数,用来控制不同分类兴趣(长期兴趣或非长期兴趣)值减小的速度,  $k$  为项目资源的关联度,如果是项目资源本身  $k$  值为 1。

每隔一个时间阈值  $Ttime$ ,就对用户的兴趣模型进行一次更新,从而不断细化用户的兴趣模型,适应用户兴趣的变化。

## 2.2 改进算法流程

改进算法的处理流程如图 2 所示。

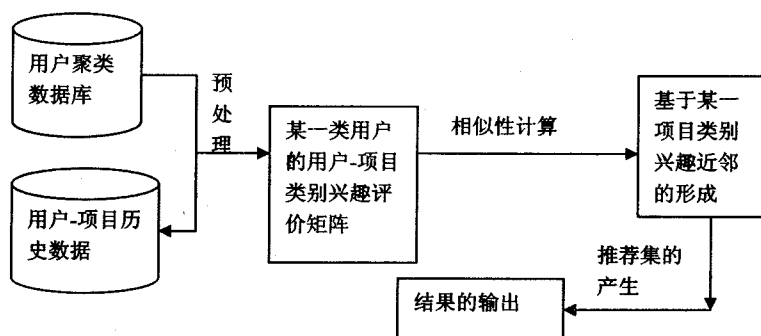


图2 基于用户-项目类别兴趣矩阵协同过滤示意图

该流程主要分为 3 个模块:

(1)数据预处理模块:主要完成用户对某类项目类别的评分数据的初步处理,以形成用户-项目类别兴趣矩阵。将用户对每项资源的评分转换为用户对其所属类别的平均评分用来表示该项目类别的兴趣度,如表 1 所示。

表1 用户-项目类别矩阵

UserID	XL01	XL02	XL03	.....	XL0n
A0162154	0	0	0.422	0	0.242
A1182418	0.266	0	0	0	0
.....	.....	.....	.....	.....	.....
A0910115	0	0.021	0.125	0	0.565

(2)相似性计算模块:在用户-项目类别兴趣矩阵的基础之上,通过向量空间模型计算目标用户和所在群中用户的两两之间的相似性,从而形成基于某一项目类别兴趣的近邻。

(3)推荐集的产生模块:主要完成根据最近邻居的相关数据给活动用户相关建议,即产生推荐集。

## 2.3 相似用户计算

可采用下面的修正的余弦相似性公式<sup>[12]</sup>(3)来计算两两用户之间的相似性:

$$\text{sim}(i, j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \bar{R}_i) (R_{j,c} - \bar{R}_j)}{\sqrt{\sum_{c \in I_i} (R_{i,c} - \bar{R}_i)^2} \sqrt{\sum_{c \in I_j} (R_{j,c} - \bar{R}_j)^2}} \quad (3)$$

输入:基于聚类的用户-项目类别兴趣矩阵  $Q$ , 预设阈值  $S$ 。

输出:用户相似度结果集。

步骤:

- (1) 对于聚类用户集中的任一用户  $u_i$ , 循环进入(2);
- (2) 对于聚类用户集中用户  $u_i$  以后的任一用户  $u_j$ , 循环进入(3);
- (3) 获得用户  $u_i$  和用户  $u_j$  的各组项目类别及其兴趣权重;
- (4) 用相似度计算公式计算用户  $u_i$  和用户  $u_j$  的各组项目类别兴趣的相似度  $S_{ij}$ ;
- (5) 若  $S_{ij}$  大于等于预设阈值  $S$ , 则认为用户  $u_i$  和用户  $u_j$  是近邻相似用户, 存储  $u_i$  和  $u_j$  及其相似度值  $S_{ij}$  入数据库  $R$  中;
- (6) 否则, 转到(2);
- (7) 如果用户集合处理完毕, 进入(8), 否则转到(1);
- (8) 算法结束。

## 2.4 推荐集的产生

根据当前用户的近邻用户生成资源推荐集的具体处理流程如下:

第一步, 对系统中任一用户  $u$ , 从数据库中得到  $u$  的最近邻居集  $R$  (已按相似度大小  $S_{ij}$  降序排列)。对  $u$  取前  $N$  个近邻  $n_i \in R$ , 将  $n_i$  的相应数据信息存储到推荐候选集  $DR_1$  中。

第二步, 把候选集  $DR_1$  中当前用户  $u$  已经访问并有历史记录的资源剔除, 得到推荐候选集  $DR_2$ ;

第三步, 对  $DR_2$  候选集, 先根据资源的项目类别排序; 当项目类别相同时再根据资源的历史评分来对推荐候选集  $DR_2$  排序。

第四步, 将排好序的  $DR_2$  存储到数据库中, 结合当前用户的每个项目类别及其相应的兴趣权重取相应项目类别的  $N \times W$  个资源作为当前用户 Top -  $N$  推荐。

## 3 实验结果与讨论

文中, 像大多数文献中的一样, 采用平均绝对偏差 (mean absolute error) MAE<sup>[13]</sup> 作为度量标准, 其值越小, 推荐质量就越高。MAE 是指通过计算预测值和实际评分来衡量算法的推荐精度。MAE 计算见公式(4):

$$MAE = \frac{\sum_{i=1}^N |p_i - q_i|}{N} \quad (4)$$

其中,  $N$  为测试集大小;  $p_i$  为预测评分;  $q_i$  为实际评分。

实验数据采用 MovieLens (<http://movielens.umn.edu>) 提供的数据集, 并依据电影中已有的分类体系对项目进行了划分。选取了 350 位用户对 1600 部电影的 28655 条评分数据, 每位用户至少对 20 部电影进行了评分。根据实际需要选择了其中包含电影数目在不同数量级上的 12 类项目。

实验方案: 查看最近邻居集大小不同的情况下, 将基于用户聚类的协同过滤算法与文中提出的基于用户多兴趣的协同过滤改进算法的性能进行比较。取最近邻居数分别为 10、20、30、40, 对上述两种算法分别进行实验测试。实验结果如图 3 所示。

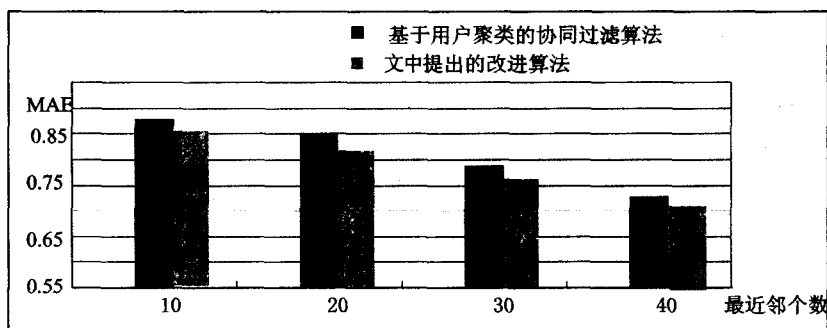


图 3 改进算法与基于用户聚类协同过滤算法对比图

## 4 结束语

文中基于现有用户聚类算法的不足, 提出了一种基于用户聚类的协同过滤改进算法, 首先根据项目自身的类别属性对项目进行分类, 再根据用户评分对所有用户项目类别离线进行基于用户群用户间的相似度计算, 形成基于某一项目类别兴趣的近邻, 最后根据近邻的历史数据产生推荐集。结合离线计算的方式有效提高推荐系统的实时响应速度。实验结果表明, 文中提出的基于用户多兴趣的协同过滤改进算法不仅能反映出用户的多兴趣和兴趣的变化情况, 而且可有效解决推荐系统处理大规模数据面临的实时性问题和推荐质量问题。

### 参考文献:

- [1] Alberto D E, Manuel J. Using linear classifiers in the integration of user modeling and text content analysis in the personalization of a Web-based Spanish News Service[C] // In: Proceedings of the 8th International Conference on User Modeling. [s. l.]: [s. n.], 2001.
- [2] Wu Jianzhong. A dynamic gateway to information; electronic services at the Shanghai library[J]. Information Development, 2004, 20(2): 111-116.
- [3] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet Computing, 2003, 7(1): 76-80.
- [4] Zeng Chun, Xing Chun xiao, Zhou Li zhu, et al. Similarity

(下转第 80 页)

示了在分布式算法下,  $\Delta\epsilon$  对各认知用户公平性的影响。由图可知, 随着  $\Delta\epsilon$  的增加, 各用户的实际速率逐渐偏离相应的期望速率, 即用户的公平性越差。

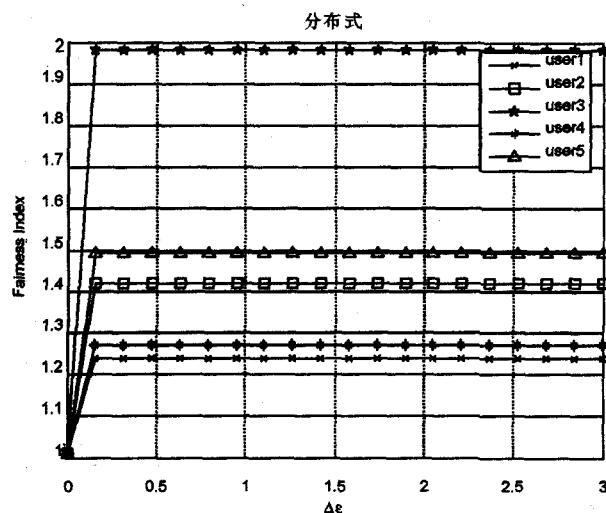


图 6 分布式算法下  $\Delta\epsilon$  对用户公平性的影响

## 5 结束语

根据用户的不同速率需求, 在认知无线网络中提出了基于多用户公平性的功率分配问题。文中以最小化用户需求速率与可达速率为优化目标, 以主用户的干扰限制为约束条件。为求解该问题, 提出了两种集中式分配算法及一种分布式算法。仿真结果表明, 三种求解算法都可收敛于优化问题的最优解, 且集中式算法 2 的收敛速率远大于集中式算法 1, 分布式功率分配算法有利于减少实现的复杂度。

## 参考文献:

- [1] Musavian L, Aissa S. Quality-of-Service Based Power Allocation in Spectrum-Sharing Channels [C]// IEEE GLOBE-COM. [s. l.]: [s. n.], 2008: 1-5.
- [2] Le Long B, Hossain E. Resource allocation for spectrum underlay in cognitive radio networks [J]. IEEE Transaction on Wireless Communications, 2008(7): 5306-5315.
- [3] 谢军辉, 冯平. 认知无线网络中基于凸优化的功率分配研究 [J]. 计算机应用研究, 2010(3): 1161-1166.
- [4] Kang X, Zhang R. Optimal power allocation for cognitive radio under primary user's outage loss constraint [C]// ICC 2009. Dresden: IEEE Press, 2009: 1-5.
- [5] 郭艳艳, 康桂霞, 张宁波, 等. 基于认知无线电系统的协作中继分布式功率分配算法 [J]. 电子与信息学报, 2010, 32(10): 2463-2467.
- [6] Guo Y, Kang G, Zhang N, et al. Outage performance of relay-assisted cognitive-radio system under spectrum-sharing constraints [J]. Electronics Letters, 2010, 46: 182-184.
- [7] Chen Yan, Yu Guanding, Zhang Zhaoyang, et al. On Cognitive Radio Networks with Opportunistic Power Control Strategies in Fading Channels [J]. IEEE Trans. Wirel. Commun, 2008, 7(7): 2752-2761.
- [8] Luo Changqing, Yu F Richard, Ji Hong, et al. Distributed relay selection and power control in cognitive radio networks with cooperative transmission [C]// IEEE communication society subject matter experts for publication in the IEEE ICC 2010 proceedings. [s. l.]: [s. n.], 2010.
- [9] 林琳, 周贤伟, 薛楠, 等. 认知无线网络安全路由问题研究 [J]. 计算机技术与发展, 2010, 20(1): 159-162.
- [10] Musavian L, Aissa S. Outage-Constraint Capacity of Spectrum-Sharing Channels in Fading Environments [J]. IET Communication, 2008, 2(2): 724-732.
- [11] Qin T, Leung C. Fair adaptive resource allocation for multiuser OFDM cognitive radio systems [C]// Proceedings of IEEE ChinaCom. [s. l.]: [s. n.], 2007: 115-119.
- [12] Attar A, Holland O, Nakhai M R, et al. Interference-limited resource allocation for cognitive radio in orthogonal frequency-division multiplexing networks [J]. IET Commun, 2008(2): 806-814.
- [13] Sarwar B, Karypis G, Konstan J. Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International World Wide Web Conference. [s. l.]: [s. n.], 2001: 285-295.
- [9] 潘红艳, 陶剑文, 杨华兵. 基于信息项和用户群的信息推荐机制 [J]. 情报学报, 2006, 25(5): 600-605.
- [10] 张娜, 何建民. 基于项目与客户聚类的协同过滤推荐方法 [J]. 合肥工业大学学报 (自然科学版), 2007, 30(9): 1159-1162.
- [11] 查文琴, 梁昌勇, 曹镭. 基于用户聚类的协同过滤推荐方法 [J]. 计算机技术与发展, 2009, 19(6): 69-71.
- [12] 董祥和, 齐莉丽, 董荣和. 优化的协作过滤推荐算法 [J]. 计算机工程与应用, 2009, 45(8): 229-232.
- [13] Herlocker J L, Konstan J A, Terveen, et al. Evaluating collaborative filtering recommender systems [J]. ACM Trans. on Information Systems, 2004, 22(1): 50-53.

(上接第 76 页)

- measure and instance selection for collaborative filtering international [J]. Journal of Electronic Commerce, 2004, 4(8): 115-129.
- [5] 姜雅倩, 王直杰, 张珏. 基于供求关系及协同过滤的推荐模型研究 [J]. 计算机技术与发展, 2007, 17(6): 18-21.
- [6] 游文, 叶水生. 电子商务推荐系统中的协同过滤推荐 [J]. 计算机技术与发展, 2006, 16(9): 70-72.
- [7] Mobasher B, Jin X, Zhou Y. Semantically enhanced collaborative filtering on the Web [C]// In: Proceedings of the European Web Mining Forum. [s. l.]: [s. n.], 2004.
- [8] Sarwar B, Karypis G, Konstan J. Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International World Wide Web Conference. [s. l.]: [s. n.], 2001: 285-295.