

文章编号:1001-9081(2009)06-1590-04

一种有效缓解数据稀疏性的混合协同过滤算法

郁雪,李敏强

(天津大学 管理学院,天津 300072)

(yuki@tju.edu.cn)

摘要:目前协同过滤技术已经被成功运用到各种推荐系统中,但是随着资源种类的不断膨胀与用户日益的增加,用来评判的数据矩阵越来越稀疏,严重影响了推荐质量。为此设计了一种混合新算法,对基于项目的协同过滤算法提出两个改进方法:首先根据网站的层次结构信息改进了传统的相似度计算方法;其次增加了预测缺失兴趣值的算法,使用户的交叉兴趣点增多,有效缓解了稀疏性的问题。实验结果证明了新算法具有较高的推荐精度,能够找到用户潜在的兴趣页面。

关键词:推荐系统;协同过滤;数据预测;数据稀疏性

中图分类号: TP311.13 **文献标志码:** A

Effective hybrid collaborative filtering algorithm for alleviating data sparsity

YU Xue, LI Min-qiang

(School of Management, Tianjin University, Tianjin 300072, China)

Abstract: Collaborative filtering has been successfully applied to various recommendation systems. Unfortunately, with the tremendous growth in the amount of items and users, the lack of original rating poses some key challenges for recommendation quality. To address this problem, the paper explored a new hybrid CF approach which improved the traditional similarity coefficient computation combining the portal website natural structure, then the missing preference values were predicted with new similarity based on the item-based CF. The improvements made an increasing intercross of the rating matrix for alleviating sparsity. The experimental results show the proposed algorithm outperforms the traditional CF, and it can recommend potential preference pages for visitors.

Key words: recommendation system; collaborative filtering; data prediction; data sparsity

0 引言

目前个性化推荐技术已经被广泛地应用电子商务的商品推荐、新闻推荐和 E-learning 智能学习推荐等系统中。其中协同过滤技术是目前讨论的热点,其主要优势是对其推荐的对象没有特殊限制,并且能够发现用户的新兴趣,实现兴趣转移预测。但是现有算法也存在一些弊端^[1]:如可扩展性问题,冷启动问题,评分矩阵的稀疏性带来的推荐质量差问题,新用户问题等。

针对上述问题许多学者提出改进的办法,文献[2]提出用 BP 神经网络预测用户对项的缺失评分,平滑了原始矩阵,但算法需要预先建立复杂的神经网络模型。文献[3]利用奇异值分解(Singular Value Decomposition, SVD)有效地减少了输入矩阵的维数,但是分解算法复杂度高,实验证明当评分矩阵极度稀疏情况下,该算法的预测精度不如传统的协同过滤算法。文献[4]结合了奇异值分解与最近邻方法,先采用 SVD 方法预测未打分项目,平滑了原始矩阵,然后用最近邻方法求当前用户的评分。对于聚类的方法目前存在两种思路,一是针对项目聚类,通过用户对项目评分的相似性对项目进行聚类,把所有用户对聚类中心的平均评分作为新的搜索空间^[5]。文献[6]采用新的项目聚类算法,确保在一类项目中用户的评分最相似。文献[7-8]提出通过项目语义进行自然分类,用户对项目的兴趣值转换成用户对某一类项目的兴趣值。二是对用户进行聚类,文献[9]提出 k-means 用户聚类

方法,利用同类中的平均偏移来近似预测评分矩阵中的未知评分,解决了矩阵稀疏性问题,通过对标准数据集的测试证明了该算法有效性。

本文提出一种新的协同过滤推荐算法框架,首先改进了传统相似度算法,引入网站的结构层次特征,增加了计算页面之间的主题相似性,与传统的用户访问相似度进行线性组合,形成资源的综合相似度。然后在新的相似度的基础上采用基于项目的协同过滤算法预测原始兴趣矩阵的缺失兴趣,缓解矩阵的稀疏性,增加用户的交叉分数,填充后的兴趣矩阵作为输入数据,通过最近邻预测用户对所有页面的最终兴趣度。新算法既增加了页面之际的内在关联,平滑了原始矩阵,又保留了传统协同过滤算法的优势。最后实验通过真实数据证明了算法的有效性。

1 经典的协同过滤算法

协同过滤算法是目前应用的最成功的推荐技术,主要是通过分析人们之间的兴趣相似性来进行项目(item)的推荐。它的基本假设是用户可以通过评分来反映各自对项目的兴趣,并且用户对未知项目的评分可由与其兴趣相似的若干用户(最近邻)推导出来。经典的协同过滤算法是基于用户的推荐^[10],通过分析用户历史的评分数据,发现当前用户的最近邻,用邻居的评分来加权平均预测当前用户 x 对项目 i 的兴趣度,预测评分公式为:

收稿日期:2008-12-16;修回日期:2009-02-20。 **基金项目:**高等学校博士学科点专项科研基金资助项目(20020056047)。

作者简介:郁雪(1977-),女,天津人,讲师,博士研究生,主要研究方向:信息系统、Web 智能;李敏强(1965-),男,河北无极人,教授,博士生导师,主要研究方向:系统工程、信息系统,人工智能。

$$p_{x,i} = \overline{R_x} + \frac{\sum_{y \in NBS} \text{sim}(x,y) \times (R_{y,i} - \overline{R_y})}{\sum_{y \in NBS} (1 - \text{sim}(x,y))} \quad (1)$$

其中: $\text{sim}(x,y)$ 是用户 x 与用户 y 的相似性,可以用余弦相似度、Pearson 相关系数和修正的余弦相似度来计算; NBS 是用户 x 的最近邻集合。下面给出常用的 Pearson 相关系数的计算公式^[11]:

$$\text{sim}(u_a, u_b) = \frac{\sum_{s \in S_{a,b}} (r_{u_a,s} - m_{u_a})(r_{u_b,s} - m_{u_b})}{\sqrt{\sum_{s \in S_{a,b}} (r_{u_a,s} - m_{u_a})^2 \sum_{s \in S_{a,b}} (r_{u_b,s} - m_{u_b})^2}} \quad (2)$$

其中: S_{ab} 是用户 a, b 共同评分的项目集合, m_{u_a} 是用户 a 的平均评分分; m_{u_b} 是用户 b 的平均评价分,因此可以看出此公式的计算仅建立在用户曾共同评估过的项目上。

随着用户和项目的不断快速增长,传统算法的可扩展性成为了瓶颈,另外原始评价矩阵的超高维、稀疏的特性给在线实时推荐也提出了新的挑战,因此文献[11]提出了基于项目的协同过滤算法,该算法通过计算项目之间的评分相似性作为权重来预测当前用户对某一项目的未知评分,项目之间的相似性可以参考上述用户相似性的三种计算方法。预测公式为:

$$p_{x,i} = \frac{\sum_{j \in NBS} \text{sim}(i,j) \times R_{x,j}}{\sum_{j \in NBS} (1 - \text{sim}(i,j))} \quad (3)$$

由于项目之间的相似性相对稳定,并且可以离线计算,通过文献的实验证明该模型的可扩展性大大提高,预测精度比基于用户的协同过滤算法高,但还是存在新项目的推荐和数据的高稀疏性问题。

2 算法的改进

2.1 页面之间综合相似性的计算

用户访问矩阵是可以反映匿名用户在一段时间范围内对网页的兴趣的数据结构,大部分的研究者都利用此数据结构为用户兴趣建模,该模型是否能反映用户的真实兴趣,对以后的 Web 挖掘,访问模式发现,推荐质量等都起着至关重要的作用。有些文献则不仅考虑到用户是否访问了某个页面,还把该页面的访问频率、页面驻留时间等因素综合在内,有的文献引入时间窗获取最近最有影响力的访问兴趣,来改进用户兴趣模型质量。由于考虑到日志数据的噪声较大,影响用户停留时间的因素比较多,因此本文在分析用户对页面的兴趣度方面采用了考虑页面的访问频率,通过对 Web 日志的清洗、预处理、会话识别得到用户的兴趣矩阵:

$$PM_{m \times n} = \begin{bmatrix} p_{1,1} & \cdots & p_{1,j} & \cdots & p_{1,n} \\ p_{2,1} & \cdots & p_{2,j} & \cdots & p_{2,n} \\ p_{3,1} & \cdots & p_{3,j} & \cdots & p_{3,n} \\ \vdots & & \vdots & & \vdots \\ p_{m,1} & \cdots & p_{m,j} & \cdots & p_{m,n} \end{bmatrix} \quad (4)$$

用户的兴趣矩阵可以用 $M \times N$ 的矩阵来表示,其中的元素 $p_{u,p}$ 表示用户 u 对页面 p 的兴趣值,计算公式为:

$$p(u,p) = \text{freq}(u,p) \mid_{si} \quad (5)$$

其中 $\text{freq}(u,p) \mid_{si}$ 是指用户 u 在会话 si 期间对页面 p 的访问次数和。

对 Web 门户网站,页面的数量会急剧增加,形成的用户

兴趣矩阵也面临严重稀疏问题。如果采用传统的协同过滤技术,会导致具有相似兴趣的用户难以建立联系,使推荐质量和精度逐渐下降。本文提出将页面使用情况相似性与页面主题内容相似性综合考虑,改进传统的基于项目的协同过滤算法中的相似度计算,然后利用式(3)预测原始兴趣矩阵中缺失的兴趣度(missing preference),平滑原始兴趣矩阵。

1) 页面使用相似性(page usage similarity)。

页面的使用情况来源于用户的主动访问,传统的基于项目的协同过滤算法仅考虑到资源使用的相似性,认为两个被多数用户频繁访问的页面之间具有较高的相似度,如果其中一个页面被当前用户访问,则另一个页面被赋予一个较高的权重进行预测。资源使用相似性可采用 Pearson 相关系数(式(2))进行相似性计算。当两个页面较少地被访问,或没有共同的用户访问时,那么页面之间相似性会非常小甚至为零。但是实际情况是页面之间的相关性除了体现在用户的访问模式外,页面本身的主题相关性也至关重要。

2) 页面的主题相似性(page subject similarity)。

Web 页面的主题相关性用于衡量页面的内容相似性,文献[12]采用了基于本体的语义相似性进行计算,但创建领域本体目前还缺乏自动化的方法,必须手工或半自动化获取,因此对于大规模网站,构造本体是非常困难的事。文献[13]采用信息检索中的向量空间模型,计算网页之间的内容相似度,增加了算法的额外开销。对于开发有序、组织结构严密的网站,自身的层次结构就能够建立底层页面之间的主题关联,形成有效的层次知识。每个 URL 都可以看成是层次结构的底层叶子节点,站点结构的每个分支都能代表一个主题脉络,很显然处于同一个分支的页面具有内容上的相关性,如图1所示,由于篇幅关系本文截取了电子政务门户网站的部分层次结构图。其中包括一级栏目17个,二级栏目68个,三级栏目135个,四级栏目10个。

页面之间的主题相似性可以通过在站点层次结构中的相对距离来计算,本文采用文献[14]中的计算公式获取叶子节点之间的主题相似性,公式如下:

$$\text{SubjectSim}(p1, p2) = \frac{\vec{p1} \cdot \vec{p2}}{2 \times \text{depth}(LCA(p1, p2))} = \frac{\text{depth}(p1) + \text{depth}(p2)}{\text{depth}(p1) + \text{depth}(p2)}; \quad (6)$$

$\text{depth}(p1)$:表示从根目录到叶子节点 $p1$ 的路径长度。

$LCA(p1, p2)$:表示节点 $p1, p2$ 的最低公共祖先。

为了获取更真实的页面相似性,我们集成页面使用相似性和页面主题相似性,引入了 α 权重系数, $\alpha \in [0, 1]$ 通过线性组合来计算页面之间整体的相似度值,通过选用合适的 α 值将两种相似度测量方法的优点结合起来,进一步解决同类主题新页面的推荐问题,公式如下:

$$USSim(p1, p2) = \alpha \times \text{UsageSim}(p1, p2) + (1 - \alpha) \times \text{SubjectSim}(p1, p2) \quad (7)$$

需要说明的是当 $\alpha = 0$ 时,页面的相似度完全由主题相似性来决定,即同类型主题的页面会得到较高的兴趣值;当 $\alpha = 1$ 时,页面的相似性计算即为传统的 Pearson 相关系数。随着参数 α 的不断修正,最终能够找到一个满意的平衡点,优化页面相似度。

2.2 缺失兴趣度的预测

用户兴趣矩阵的稀疏性是影响推荐质量的关键问题,常用的解决办法是将用户的平均兴趣值填充在空白的兴趣项上,但是往往不能正确表达用户的真实兴趣值,导致推荐效果不满意。本文结合了电子政务门户网站的应用背景,在 2.1

节算法基础上,先填充未知兴趣值,平滑了原始矩阵,解决了矩阵极度稀疏性问题,使交叉评分增加,平滑后的兴趣矩阵如图 2(b)所示,作为为下一步推荐算法的输入数据。

图 2(b)中的预测值为:

$$\hat{Pref}(U, p) = \frac{\sum_{q \in NBS} USSim(p, q) \times P_{U, q}}{\sum_{q \in NBS} (1 \mid USSim(p, q) \mid)} \quad (8)$$

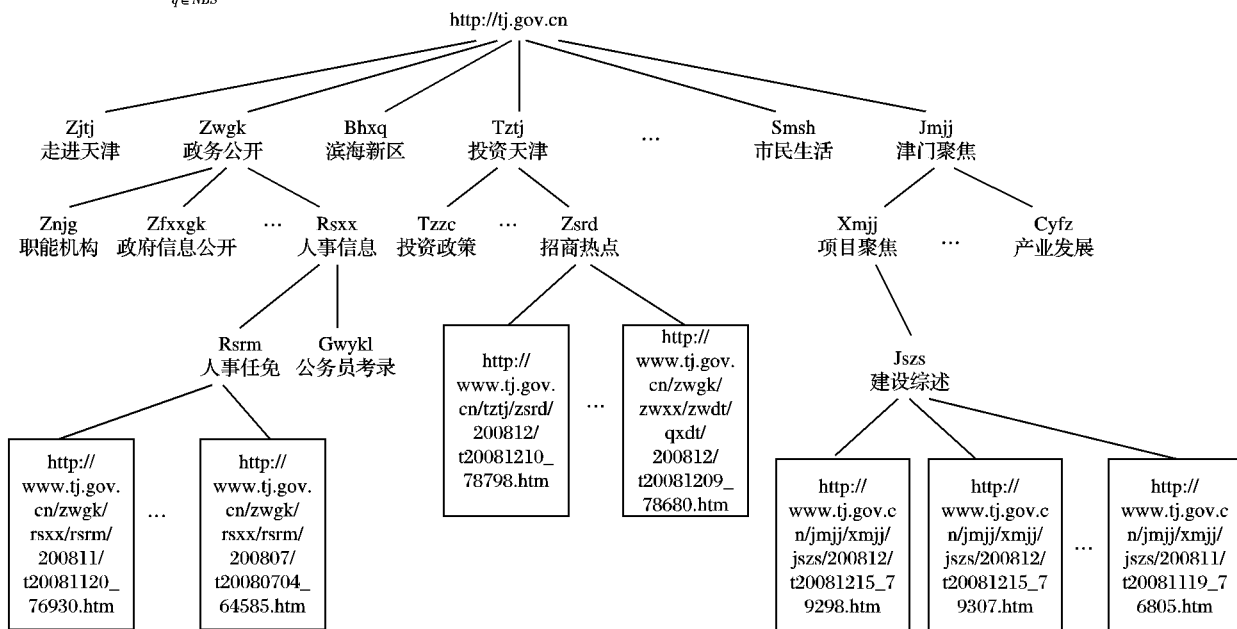


图 1 Web 网站的层次结构

	P_1	P_2	P_3	P_4	P_5	P_6
U_1	0	3	0	1	2	0
U_2	1	0	0	2	5	0
U_3	0	0	3	6	0	0
U_4	3	0	0	0	0	2
U_5	5	2	0	0	0	1
U_6	0	1	0	2	4	0

(a)原始的兴趣矩阵

	P_1	P_2	P_3	P_4	P_5	P_6
U_1	$\hat{Pref}(1,1)$	3	$\hat{Pref}(1,3)$	1	2	$\hat{Pref}(1,6)$
U_2	1	$\hat{Pref}(2,2)$	$\hat{Pref}(2,3)$	2	5	$\hat{Pref}(2,6)$
U_3	$\hat{Pref}(3,1)$	$\hat{Pref}(3,2)$	3	6	$\hat{Pref}(3,5)$	$\hat{Pref}(3,6)$
U_4	3	$\hat{Pref}(4,2)$	$\hat{Pref}(4,3)$	$\hat{Pref}(4,4)$	$\hat{Pref}(4,5)$	2
U_5	5	2	$\hat{Pref}(5,3)$	$\hat{Pref}(5,4)$	$\hat{Pref}(5,5)$	1
U_6	$\hat{Pref}(6,1)$	1	$\hat{Pref}(6,3)$	2	4	$\hat{Pref}(6,6)$

(b)填充后的兴趣矩阵

图 2 原始的兴趣矩阵和填充后的兴趣矩阵

3 实验与评价

3.1 数据集

本文采用了电子政务门户网(<http://www.tj.gov.cn/>)匿名日志数据,在对数据进行基本预处理和清洗后,考虑到 Web 日志数据噪声大的特点,我们又采用了如下的处理过程:

1)删除管理员的访问记录。管理员访问站点会产生大量的日志记录,由于这些记录的主要目的是维护和测试 Web 系统,因此不在我们研究的目标内。

2)删除点击次数少于 10 次,或浏览页面小于 5 的用户。这些用户的访问记录可以忽略不计,可以认为对政务网的提供的信息没有特别的兴趣,因此作为干扰数据可以过滤掉。

3)删除点击次数小于 10 的页面。这些页面可认为没有提供值得用户关注的重要信息,因此可以不作为推荐的对象。

最终从处理后的数据中选择 127 264 个有效的访问记录,其中包括 1 678 个用户对 1 430 个页面的浏览信息。在此数据集中抽取 1 250 条记录,分成训练集合 1 000 条和测试集 50 条(分成 5 组当前用户)。为了测试对新用户的使用效果,随机抽取每组可见的用户兴趣,兴趣页面数为 5,代表当前新用户对页面表现的兴趣较少,数量为 10,代表正常情况下预测,

在对原始兴趣矩阵(如图 2(a))的空白兴趣进行预测填充后,下一步就是根据当前用户的浏览行为预测网页的兴趣值,由于输入的数据为平滑后的兴趣矩阵,因此用户间的共同评分项目增多,可以使用经典的 Pearson 相关系数(式(2))找到当前用户的 K 个最近邻,通过式(1)计算未被浏览的网页的兴趣值,并选择其中的 TOP-N 推荐给当前用户。

因此分别命名为 Given5, Given10^[10]。其余作为测试集进行实验。

3.2 评价标准

我们采用平均绝对误差(MAE)来衡量用户对未知页面兴趣度预测的准确性,公式如下所示:

$$MAE = \frac{\sum_{u \in T} |Pref(u, p) - \hat{Pref}(u, p)|}{|T|} \quad (9)$$

其中: $Pref(u, p)$ 为用户的真实兴趣值, $\hat{Pref}(u, p)$ 是通过算法预测的页面的兴趣值, T 是测试集, $|T|$ 是测试集的元素个数, MAE 值越小说明越接近于真实兴趣值,预测就越准确。

3.3 实验分析

为验证本文算法预测的质量,设计了两组实验。表 1 为实验所用几种算法的比较。

表 1 实验用算法的比较

算法名	USSim	预测填充
IB CF(基于项目评分的协同过滤算法)	否	否
USIB CF(基于改进相似度的协同过滤算法)	是	否
UB_After_USIBMPP CF(本文算法)	是	是

实验中相似度的组合权重参数 α 从 0 逐渐增加到 1 (以 0.1 为步长递增)。当 $\alpha = 0$ 时,相似度就是页面主题相似度;当 $\alpha = 1$ 时,相似度变成了传统的页面使用相似度。

从实验结果(如图 3 所示)可看出:

1) 在 Given5 条件下,当 $\alpha = 0.6$ 时 MAE 最小,取得最佳的权重组合。在 Given10 条件下,曲线的走势与 Given5 条件下基本一致,当 $\alpha = 0.8$ 时 MAE 最小。Given10 的总体预测精度比 Given5 好,这是因为 Given10 测试集给出了足够的历史浏览信息。

2) 采用 UB_After_USIBMPP CF 算法的推荐效果更好。根据实验结果的统计数据表明,MAE 比 USIB CF 算法(平均 MAE)降低了 0.1578。并且在 Given5 历史信息量较少的情况下,应用 UB_After_USIBMPP CF 算法的预测精度没有明显变差(其 MAE 值没有变大),这是因为在两种预测条件下都进行了平滑兴趣矩阵的步骤,利用改进的 USIB CF 算法预测了空白的兴趣值,令填充后的矩阵所含信息量基本一致。

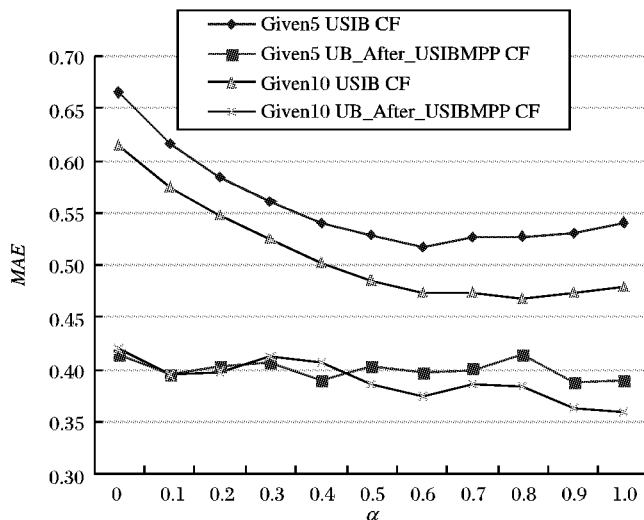


图3 Given5 和 Given10 条件下不同比例的加权值对推荐精度的影响

图 4 为 USIB CF 算法与 IB CF 算法的推荐质量比较(在 Given5 和 Given10 条件下, α 分别取 0.6 和 0.8)。实验中取页面的最近邻个数 $k1 = 10$, 用户的最近邻个数 $k2 = 10$ 。通过实验结果可发现:与传统的协同过滤算法相比,兴趣值信息少的新用户(Given5 实验)采用改进相似度进行预测,比 Given10 条件下的预测精度提高得更显著,MAE 分别下降了 0.0231 和 0.01。

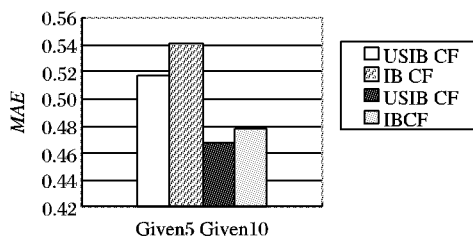


图4 与传统协同过滤算法推荐精度比较

4 结语

随着个性化信息服务的推广,智能化的推荐系统越来越成为人们关注的热点,本文针对传统协同过滤算法中矩阵稀疏性问题,提出一种基于综合相似性的混合协同过滤算法。首先结合页面之间的用户访问相似度和页面主题相似度,计

算出最优的组合权重比例;然后预测用户的空白兴趣值,缓解了原兴趣矩阵的稀疏性问题,最后利用基于用户的协同过滤算法进行页面推荐。本文采用真实的电子政务门户网站日志作实验数据,实验结果证明新算法不仅能够使预测精度明显提高,而且在新用户的经典问题上也有较好的表现。本文着力于解决页面兴趣的预测质量问题,未来的研究重点是结合有效降维的思路,在预测精度不损失的前提下,改善系统的可扩展性。

参考文献:

- [1] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Analysis of recommendation algorithms for E-commerce [C]// Proceedings of the 2nd ACM Conference on Electronic Commerce. New York: ACM Press, 2001: 158-167.
- [2] 张锋, 常会友. 使用 BP 神经网络缓解协同过滤推荐算法的稀疏性问题[J]. 计算机研究与发展, 2006, 43(4): 667-672.
- [3] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Application of dimensionality reduction in recommender system-A case study [C]// Proceedings of the ACM WebKDD 2000 Web Mining for E-Commerce Workshop. New York: ACM Press, 2000: 82-90.
- [4] 孙小华, 陈洪, 孔繁胜. 在协同过滤中结合奇异值分解与最近邻方法[J]. 计算机应用研究, 2006, 23(9): 206-208.
- [5] 邓爱林, 左子叶, 朱扬勇. 基于项目聚类的协同过滤推荐算法[J]. 小型微型计算机系统, 2004, 25(9): 1665-1670.
- [6] QUAN T, FUYUKI I, SHINICHI H. Improving accuracy of recommender system by clustering items based on stability of user similarity [C]// Proceedings of International Conference on IAWTIC. Washington, DC: IEEE Computer Society, 2006: 61.
- [7] 朱征宇, 张小林, 熊茜, 等. 基于用户兴趣子类的协作推荐算法[J]. 计算机科学, 2005, 32(10): 176-180.
- [8] GAO F, XING C, ZHAO Y. An effective algorithm for dimensional reduction in collaborative filtering [C]// ICADL 2007: The 10th International Conference on Asian Digital Libraries, LNCS 4822. Berlin: Springer, 2007: 75-84.
- [9] XUE GUI-RONG, LIN CHEN-XI, YANG QIANG, *et al.* Scalable collaborative filtering using cluster-based smoothing [C]// Proceedings of the 28th Annual International ACM SIGIR conference on Research and Development in Information Retrieval. Brazil: ACM Press, 2005: 114-121.
- [10] BREESE J S, HECHERMAN D, KADIE C. Empirical analysis of predictive algorithms for collaborative filtering [C]// Proceedings of 14th Conference Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann, 1998: 43-52.
- [11] SARWAR B, KARYPIS G, KONSTAN J, *et al.* Item-based collaborative filtering recommendation algorithms [C]// Proceedings of the 10th International Conference on World Wide Web. New York: ACM Press, 2001: 285-295.
- [12] 罗耀明, 聂规划. 语义相似性与协同过滤集成推荐算法研究[J]. 武汉理工大学学报, 2007, 29(1): 85-88.
- [13] FLESCA S, GRECO S, TAGARELLI A, *et al.* Non-invasive support for personalized navigation of Websites [C]// Proceedings of the International Database Engineering and Applications Symposium. Washington, DC: IEEE Computer Society, 2004: 183-192.
- [14] GANESAN P, GARCIA-MOLINA H, WIDOM J. Exploiting hierarchical domain structure to compute similarity [J]. ACM Transactions on Information System, 2003, 21(1): 64-93.