DOI:10.13471/j.cnki.acta.snus.2016.05.007

# 混合用户和项目协同过滤的电子商务个性化推荐算法*

李清霞[1]，魏文红[2]，蔡昭权[3]
（1. 东莞理工学院城市学院计算机系，广东 东莞 523106；
2. 东莞理工学院计算机学院，广东 东莞 523808；
3. 惠州学院科研处，广东 惠州 516007）

**摘 要**：针对传统的协同过滤算法在电子商务系统中存在数据稀疏性和扩展性方面的问题，提出了一种混合用户和项目协同过滤的电子商务个性化推荐算法。该算法采用聚类技术，将基于用户协同过滤和基于项目的协同过滤结合起来进行双重聚类，结合基于用户协同过滤和基于项目协同过滤两方面的优点，从而获得更好的性能。实验表明，通过与其他推荐算法的比较，文中算法具有较高的推荐质量，更好的准确率和召回率。

**关键词**：协同过滤；电子商务；个性化推荐；聚类
**中图分类号**：TP301 **文献标志码**：A **文章编号**：0529－6579（2016）05－0037－06

# Hybrid user and item based collaborative filtering personalized recommendation algorithm in E-commerce

*LI Qingxia*[1], *WEI Wenhong*[2], *CAI Zhaoquan*[3]
(1. Department of Computer and Information Science, City College of Dongguan University of Technology, Dongguan 523106, China;
2. School of Computer, Dongguan University of Technology, Dongguan 523808, China;
3. Scientific Research Office, Huizhou University, Huizhou 516007, China)

**Abstract**: In view of the traditional collaborative filtering algorithm in E-Commerce system data sparseness and scalability issues, a hybrid user and item based personalized collaborative filtering recommender algorithm in E-Commerce was proposed. Combined with user based collaborative filtering and item based collaborative filtering, the algorithm uses the clustering technology to cluster twice, and can get better performance. Experiments results show that the algorithm is superior to other recommendation algorithms obviously in the aspect of recommender quality, precision and recall rate.

**Key words**: collaborative filtering; E-commerce; personalized recommender; cluster

随着互联网络应用的快速发展，电子商务已经渗入到了人们生活的各个领域。当人们利用电子商务系统进行购物时，想方便快捷地从中挑选自己感兴趣的商品却是一件费时又费力的事情。因此，能够根据用户浏览商品的历史记录，挖掘出用户的消费偏好，并能主动推荐给自己一些可能喜欢的商品，以节省时间和精力。

在这种需求的推动下，推荐系统就慢慢地发展

* **收稿日期**：2016－01－14
**基金项目**：国家自然科学基金资助项目（61103037，61370185）；广东省自然科学基金资助项目（2013010013432）；东莞市科技计划资助项目（2014106101019）；东莞理工学院城市学院青年基金资助项目（2014QJZ002Z）
**作者简介**：李清霞（1973 年生），女；**研究方向**：电子商务；E-mail：lee_ qxia@163.com

起来[1]。推荐系统顾名思义就是向用户推荐信息，具体来说就是抓住用户的个人爱好和习惯，投其所好，推荐一些可能感兴趣的信息给他。因为这些信息经常具有个性化的特点，所以又称为个性化推荐系统[2]。在电子商务系统应用中，个性化推荐系统可以根据用户浏览信息的历史记录，分析其购物特点，并向其推荐可能感兴趣的商品。从而帮助用户从浩瀚的商品列表中购买其所需要的商品，为每个用户提供个性服务[3]。由于个性化推荐系统可以带来显著的商业效益，它已成为当前电子商务发展的研究热点之一[4-6]。

个性化推荐算法是推荐系统的主体部分，研究推荐系统其实就是研究个性化推荐算法，因为个性化推荐算法的性能决定了推荐系统的性能。目前推荐系统中使用的主要推荐算法包括很多，而且各类推荐算法也都有各自的优缺点和适应环境，其中协同过滤推荐是目前研究最多、应用最广的个性化推荐算法[7-11]。协同过滤推荐利用相似用户购买行为存在可能相似的特性进行推荐，而不考虑商品的自身属性，主要依赖于最近邻居用户的意见进行推荐，偏向于个性化推荐[12]。EKSTRAND 等[7]讨论了影响协同过滤推荐算法的各种变化因素，为理解协同过滤推荐算法提供了很大的帮助。蔡观洋[8]利用双重阈值近邻查找思想，提出了两类协同过滤算法：基于用户协同过滤算法 DT－UBCF 和基于项目的协同过滤算法 DT－IBCF。郭艳红等[9]分析了传统协同过滤推荐算法的不足，提出了针对稀疏矩阵改进的个性化推荐算法。蔡强等[10]针对传统协同过滤算法对稀疏数据和新资源的推荐质量下降的情况，结合标签技术，提出了基于标签和协同过滤的推荐算法。朱夏等[11]为了让协同过滤推荐算法能够适用于云计算平台，采用分布式评分管理策略，提出了云计算环境下基于协同过滤的个性化推荐算法。

协同过滤算法可以分为基于用户的协同过滤算法（user-based collaborative filtering）和基于项目的协同过滤算法（item-based collaborative filtering）[13]。其中，前者通过分析用户的历史数据，计算用户之间的相似度，然后依靠近邻用户来提供推荐服务。后者通过分析用户作用于项目的行为数据，计算项目之间的相似度，然后根据计算出的项目相似度与用户的历史兴趣为用户进行推荐。虽然这两类协同过滤推荐算法在推荐系统中获得了广泛的应用，但无论基于用户还是基于项目的协同过滤推荐算法，在实际应用中都存在以下两个主要难题：

1）数据稀疏性问题。

在大数多网站中，用户的历史记录数据相对整个商品集来说，都是很小的一部分。例如在亚马逊网站中，用户购买商品后，对商品的评价占商品的总额不到 1%，这便产生了数据稀疏性问题。数据的稀疏性问题就导致了项目之间会出现没有交集的情况，因此就不能判断用户的喜好是否相似，没有相似的用户集，推荐效果会急剧下降。

2）扩展性问题。

扩展性问题一直是协同过滤推荐算法研究的重点。我们知道，无论是基于用户的协同过滤推荐算法还是基于项目的协同过滤算法，它们的计算量会随着用户和项目的增加而呈线性增长，因此便产生了扩展性问题。另外，即使采用了改进的基于项目的协同过滤推荐算法，在数据量巨大时，计算复杂度依然会成为性能瓶颈。

为了解决以上两个难题，本文结合基于用户协同过滤和基于项目协同过滤两方面的优点，提出了一种混合用户和项目协同过滤的电子商务个性化推荐算法。

## 1 传统协同过滤推荐算法原理

### 1.1 协同过滤推荐算法的输入与输出

在协同过滤中，需要收集所有用户对各种产品的评价作为推荐系统的依据，这些评价数据就是协同过滤推荐算法的输入。假设有 $m$ 个用户 $\{u_1,u_2,\ldots,u_m\}$ 和 $n$ 个项目 $\{i_1,i_2,\cdots,i_n\}$，用 $m\times n$ 矩阵表示原始的数据，矩阵中的元素 $v_{ij}$ 表示用户 $i$ 对项目 $j$ 的评分。另外，$v_{ij}$ 一般用某一区间的整数值表示，如文［1，5］，若 $v_{ij}=0$，则表示用户还没有对项目做出评价。原始数据的来源路径有很多种，既可以要求用户直接对所浏览或购买的商品打分，也可以通过用户的历史购买记录提取。

在协同过滤推荐算法中，当前推荐的对象称为目标用户，用 $u_a$ 表示。协同过滤推荐算法的输出结果一般有两种形式：一种是以预测目标用户 $u_a$ 对项目 $j$ 的评分形式输出；另一种是以目标用户最感兴趣的 $N$ 个产品的推荐列表形式输出，显然，推荐的这个 $N$ 个产品是目标用户尚未购买过的。典型的协同过滤推荐过程如图 1 所示。

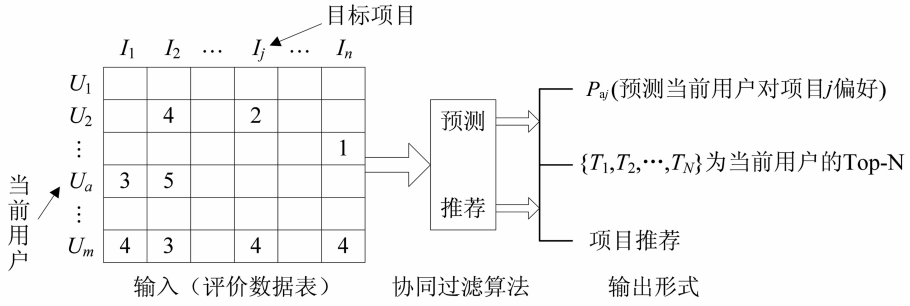


图 1　协同过滤推荐过程

Fig. 1　Processing of collaborative filtering recommendation

### 1.2　基于用户的协同过滤推荐算法

基于用户的协同过滤推荐算法原理是以用户行为的相似性作为基础，这是因为用户对待不同的事物都会存在或多或少的相似性，例如用户在评价某些项目时存在相似性，则他们在评价其他项目时也很可能存在相似性，这称为相似用户行为选择的相似性。系统通过比较目标用户的一系列历史行为选择和其他用户之间的相似性，来识别出一组和目标用户有着相似喜好的用户，称为“最近邻居”[13]。在电子商务系统中，找出目标用户的最近邻居用户后，通过分析最近邻居的行为，挖掘出目标用户喜好的商品，就可以向目标用户进行推荐。

该算法首先要搜索目标用户的“最近邻居”集，具体来说就是针对用户 $u$，要寻找到其“邻居”集合 $N = \{N_1, N_2, \cdots, N_a\}$，且用户 $u$ 不属于集合 $N$，从 $N_1$ 到 $N_a$，相似度 $\mathrm{sim}(u, N_i)$ 从大到小排列。用户之间的相似度计算方法主要有 Pearson 相关系数和余弦值相似度。本文采用 Pearson 相关系数来计算用户之间的相似度，具体的计算公式如下：

$$\mathrm{corr}_{ab} = \frac{\sum_{j\in l}(v_{aj} - \bar{v}_a)(v_{bj} - \bar{v}_b)}{\sqrt{\sum_{j\in l}(v_{aj} - \bar{v}_a)^2}\sqrt{\sum_{j\in l}(v_{bj} - \bar{v}_b)^2}} \tag{1}$$

式中 $v_{ij}$ 为用户 $i$ 对项目 $j$ 的评分，$\bar{v}_a$，$\bar{v}_b$ 分别是用户 a 和用户 b 对所有评过分的项目的平均得分。

计算出了用户间的相似度后，下面就可以确定目标用户的“最近邻居”集了。设目标用户的邻居数为 $l$，通常可以采用以下两种方法寻找目标用户的“最近邻居”集：

1）直接以目标用户为中心，找出与之最相似的 $l$ 个用户。

2）用聚集的方式，首先找出离目标用户最近的 1 个用户，然后依次找出余下的 $l-1$ 个用户。

以方法 2 为例说明目标用户寻找“最近邻居”集的过程，假设目前已经找到 $j$ 个邻居，并把这 $j$ 个邻居归于“最近邻居”集中，此时计算出他们的中心位置 $\boldsymbol{C} = \frac{1}{j}\sum_j V$，然后从目标用户的其他邻居中寻找离该中心位置最近的邻居作为目标用户的第 $j+1$ 个邻居。如此循环，直到 $j = 1$ 结束。

最后统计目标用户“最近邻居”集中的邻居对商品的评分，一般采用加权和的方法，权值系数由实际情况决定。系统可以根据该评分预测目标用户对商品的评分，该评分值的计算公式为

$$P_{aj} = \bar{v}_a + k\sum_{i=1}^{l}\mathrm{corr}_{ai}(v_{ij} - \bar{v}_i) \tag{2}$$

其中，$k$ 为权值因子。

基于用户的协同过滤推荐算法一经提出便取得了很大的成功，但随着系统规模的扩大，计算量成线性增加，巨大开销逐渐成为瓶颈，系统性能越来越差。针对这个问题，基于项目的协同过滤推荐算法便应运而生[14]。

### 1.3　基于项目的协同过滤推荐算法

与基于用户的协同过滤推荐算法比较用户之间的相似度不同，基于项目的协同过滤推荐算法比较的是项目与项目之间的相似度。

该算法首先查找目标用户已经评价项目，计算它们与目标项目 $i$ 之间的相似度，根据相似度选出最相似的 $k$ 个项目 $\{i_1, i_2, \cdots, i_k\}$，设这 $k$ 个项目的相似度为 $\{s_{i1}, s_{i2}, \cdots, s_{ik}\}$。当找到相似的项目后，以相似度为权重，计算目标用户对这些相似项目评分的加权平均值，即可得到考虑推荐项目的评分预测值。算法的实现可分为相似度计算和预测两个阶段。

相似度计算是基于项目的协同过滤推荐算法比较关键的一个阶段，相似度计算的基本思想是首先查找同时评估了项目 $i$ 与项目 $j$ 的所有用户，然后

采用相似度计算公式（如公式（1））计算它们的相似度 $s_{ij}$。

预测阶段则为基于项目的协同推荐算法输出项目结果最为重要的一步，当计算所有项目的相似度后，从中找出 $k$ 个值最大的项目，然后采用权值和的方法计算项目评分的和，即为所需的预测值。具体公式如下：

$$P_{ai} = \frac{\sum_{j \in l} (s_{ij} - v_{aj})}{\sum_{j \in \mathbf{N}} (|s_{ij}|)} \tag{3}$$

与基于用户的协同过滤推荐算法相比，采用基于项目的协同过滤推荐算法的系统计算量要少很多，因而可以获得更好的性。当然，在提高性能的同时，对推荐的精确性也有所牺牲。无论是基于用户的协同过滤推荐算法还是基于项目的协同过滤推荐算法，它们存在各自的优缺点，为了结合两者的优点，摒弃两者的缺点，本文提出了混合基于用户和项目的协同过滤推荐算法。

## 2 混合用户和项目协同过滤推荐算法

为了解决传统协同过滤算法中的用户数据的稀疏问题和扩展性问题，本文综合考虑用户和项目双方的因素，结合基于用户的协同过滤算法和基于项目的协同过滤算法的优点，提出了混合用户和项目的协同过滤推荐算法。该算法能够解决数据的稀疏问题和扩展性问题，它将基于用户协同过滤和基于项目协同过滤结合起来进行双重聚类把具有相似兴趣爱好的用户归到相同的类。系统运行时先根据用户或项目进行聚类，聚类过程可离线进行，因而能大大缩小最近邻居的查找范围，减少实时计算量，加快预测和推荐速度，提升系统性能。

算法实现思路：首先将相似度较高的项目归入一个聚类，其他的项目归入其他的聚类；然后对每个项目进行第一次聚类，再针对用户进行再次聚类，把具有相似兴趣的用户归在一起。

以下为算法的具体步骤。

输入：目标用户 $a$，聚类数目 $k$、$s$ 和用户评分数据库 URDB

输出：Top-N 推荐集

1）检索 URDB 中所有项目，获得 $n$ 个项目集合 $I = \{i_1, i_2, \cdots, i_n\}$；

2）检索 URDB 中所有用户，获得 $m$ 个用户集合 $U = \{u_1, u_2, \cdots, u_m\}$；

3）随机选取 $k$ 个项目，以初始评分作为聚类中心，获得聚类中心集合 $UI = \{ui_1, ui_2, \cdots, ui_k\}$；

4）设 $k$ 个聚类集合 $C = \{c_1, c_2, \cdots, c_k\} = \phi$；

5）repeat

for 每一个项目

for 每一个聚类中心 $ui_i \in UI$

计算项目 $i_i$ 和 $ui_i$ 的相似性 $\mathrm{sim}(i_i, ui_i)$；

endfor

$\mathrm{sim}(i_i, ui_m) = \max\{\mathrm{sim}(i_i, ui_1), \cdots, \mathrm{sim}(i_i, ui_k)\}$；

聚类 $c_m = c_m \cup i_i$；

endfor

until 聚类 $c_1, c_2, \cdots, c_k$ 不再改变；

6）将 $c_1, c_2, \cdots, c_k$ 转换成 $k$ 个子矩阵；

7）随机选取 $s$ 个项目，以初始评分作为聚类中心，获得聚类中心集合 $UC = \{uc_1, uc_2, \cdots, uc_s\}$；

8）设 $s$ 个聚类集合 $T = \{t_1, t_2, \cdots, t_s\} = \phi$；

9）对每个子矩阵

repeat

for 每一个项目

for 每一个聚类中心 $uc_j \in UC$

计算用户 $u_i$ 和 $uc_j$ 的相似性 $\mathrm{corr}(u_i, uc_j)$；

endfor

$\mathrm{corr}(u_i, uc_m) = \max\{\mathrm{corr}(u_i, uc_1), \cdots, \mathrm{corr}(u_i, uc_s)\}$；

聚类 $t_m = t_m \cup u_i$；

endfor

until 聚类 $t_1, t_2, \cdots, t_k$ 不再改变；

10）找出目标用户 a 所在的子矩阵，对这些子矩阵中用户 a 未评分过的项目 $j$，采用公式（2）计算预测值 $P_{aj}$；

11）if 项目 $j$ 属于多个类别 then $P_{aj} = \max(P_{aj1}, P_{aj2}, \cdots, P_{ajk})$；

12）对每一类别的 $P_{aj}$ 值进行计算后降序排列，选取 $P_{aj}$ 最高的前 $n$ 个项目作为 Top-N 推荐集推荐给用户。

## 3 实验结果及分析

本文采用 MovieLens 网站提供的数据集（http://www.grouplens.org）进行测试，并对混合用户和项目协同过滤的推荐算法（简称 HUICF）和其他同过推荐滤算法的性能进行比较，这些协同过滤推荐算法主要包括：DT-UBCF 算法[8]、DT-IBCF 算法[8]、PCF 算法和 TCF 算法[9-10]。我们使用的实验数据是被广泛应用于推荐系统评测中的 MovieLens 数据集，该数据集中包含了 943 个观影者，1 682 部影集以及 100 000 多条的评价。

推荐系统一般有三种质量评价方法：平均绝对偏差（MAE）、准确率（Precision）和召回率

(Recall)。

1）平均绝对偏差：通过计算预测的用户评分与实际的用户评分之间的偏差来度量预测的准确性。假设大小为 $N$ 预测的评分集合和对应的实际用户评分集合分别表示为 $\{p_1,p_2,\cdots,p_n\}$ 和 $\{q_1,q_2,\cdots,q_n\}$，则 MAE 的计算公式如下：

$$\mathrm{MAE}=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{4}$$

从公式（4）可知，MAE 值越小，表示算法的推荐质量越高。

2）准确率：其计算公式为

$$\mathrm{Precision}=\frac{\sum_{u\in Usr}|R(u)\cap T(u)|}{\sum_{u\in Usr}|R(u)|} \tag{5}$$

3）召回率：其计算公式为

$$\mathrm{Recall}=\frac{\sum_{u\in Usr}|R(u)\cap T(u)|}{\sum_{u\in Usr}|T(u)|} \tag{6}$$

在公式（5）和公式（6）中，$R(u)$ 是推荐系统给用户提供的推荐列表，$T(u)$ 是用户在测试集上的行为列表。通过定义可以看出，准确率定义了推荐列表中包含的用户行为占用户在测试集上行为记录的比例，召回率定义了推荐列表中包含的用户行为占推荐列表中项目数的比例。

图 2 显示了 HUICF 算法与其他四种算法在推荐精度方面的比较结果。

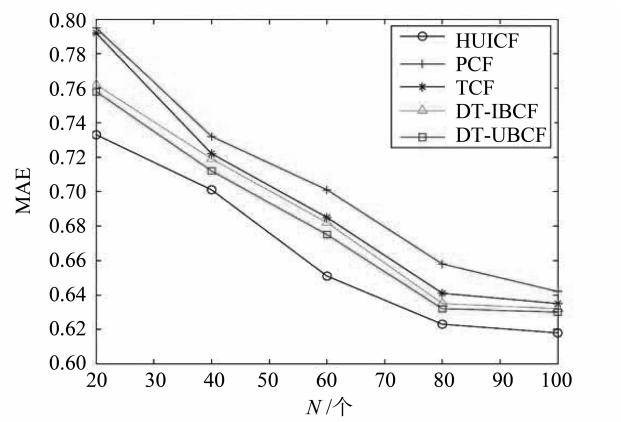


图 2　推荐精度的比较

Fig. 2　Comparison of MAE

从图 2 可以看出，HUICF 算法的推荐精度明显优于 DT-UBCF 算法、DT-IBCF 算法、PCF 算法和 TCF 算法，其中 PCF 算法的推荐质量最低，TCF 算法次之。

图 3 和图 4 显示了五种算法对于推荐准确率和召回率的实验结果，通过实验结果可以看出，在准确率和召回率方面，相对于其他 4 种算法，HUICF 算法具有最好的推荐效果。

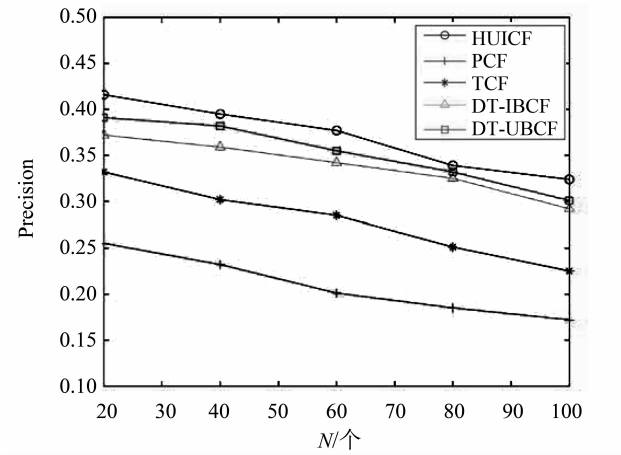


图 3　准确率的比较

Fig. 3　Comparison of precision

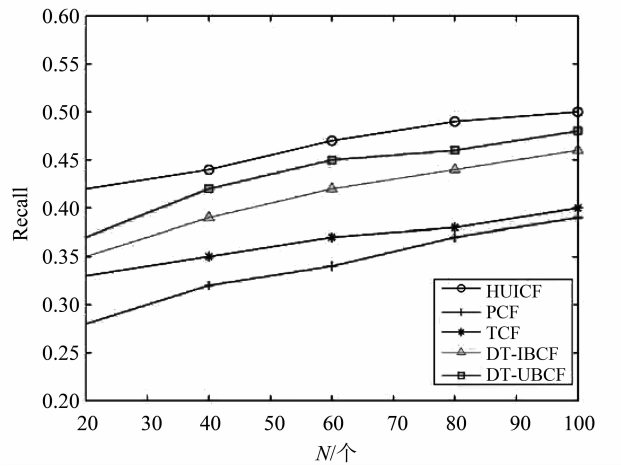


图 4　召回率的比较

Fig. 4　Comparison of recall

综上所述，混合基于用户和项目双重聚类的协同过滤推荐算法在推荐质量、准确率和召回率上超过了其他四种协同过滤推荐算法。

## 4　结　语

在当今电子商务领域中，个性化推荐系统正发挥着越发重要的作用，已成为电子商务网站至关重要的一部分。与其他推荐技术相比，协同过滤推荐算法具有明显的优势，也是当今应用最广、最为成功的推荐算法。本文针对协同过滤算法在应用中存在的主要问题：数据稀疏性和扩展性问题，提出了

混合用户和项目双重聚类的协同过滤推荐算法 HUICF。实验证明，通过与 DT-UBCF 算法、DT-IBCF 算法、PCF 算法和 TCF 算法的比较，HUICF 算法在推荐质量、准确率和召回率方面都要优于这些算法。

**参考文献：**


[1] RESNICK P, VARIAN H R. Recommender systems [J]. Communications of the ACM, 1997, 40(3): 56 – 58.

[2] WANG Y F, CHUANG Y L, HSU M H, et al. A personalized recommender system for the cosmetic business s [J]. Expert Systems with Applications, 2004, 26(3): 427 – 434.

[3] LI S S, KARAHANNA E. Online recommendation systems in a B2C E-commerce context: A review and future directions [J]. Journal of the Association of Information Systems, 2015, 16(2): 72 – 107.

[4] HAN M. The design and implementation of E-commerce personalized services based on collaborative filtering recommendation system [J]. Applied Mechanics and Materials, 2014, 687/688/689/690/691:2039 – 2042.

[5] LIN Z. An empirical investigation of user and system recommendations in e-commerce [J]. Decision Support Systems, 2014, 68: 111 – 124.

[6] ZHAO W, ZHANG H T. E-commerce recommendation system based on mapreduce [J]. Computer Modelling and New Technologies, 2014, 18(12): 264 – 269.

[7] EKSTRAND M D, RIEDL J T, KONSTAN J A. Collaborative filtering recommender systems [J]. Foundations and Trends in Human-Computer Interaction, 2011, 4(2): 81 – 173.

[8] 蔡观洋. 个性化推荐中协同过滤算法的改进研究 [D]. 长春:吉林大学, 2013.

[9] 郭艳红, 邓贵仕. 协同过滤的一种个性化推荐算法研究[J]. 计算机应用研究, 2008, 25(1): 39 – 41.

[10] 蔡强, 韩东梅, 李海生. 基于标签和协同过滤的个性化资源推荐[J]. 计算机科学, 2014, 41(1): 69 – 71.

[11] 朱夏, 宋爱波, 东方,等. 云计算环境下基于协同过滤的个性化推荐机制[J]. 计算机研究与发展, 2014, 51(10): 2255 – 2269.

[12] 李改, 李磊. 基于双向主题模型的协同过滤算法[J]. 中山大学学报(自然科学版), 2013, 52(5): 68 – 72.

[13] WANG J, VRIES A P D, REINDERS M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion [C]// Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006: 501 – 508.

[14] 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, 2001: 285 – 295.