第43卷 第6A期
2016年6月

计 算 机 科 学
Computer Science

Vol. 43 No. 6A
June 2016

# 一种改进的协同过滤推荐算法

黄 涛[1] 黄 仁[1] 张 坤[2]

(重庆大学计算机学院 重庆 400044)[1] (重庆大学自动化学院 重庆 400044)[2]

摘 要 协同过滤推荐算法是电子商务推荐系统中应用最成功的推荐技术之一,而影响协同过滤推荐算法准确率的关键因素是用户相似性度量方法。针对传统相似性度量方法没有考虑共同评分项数量对推荐质量的影响,将用户之间的共同评分项数量作为相似性计算的一个重要指标,从而得到一种改进的相似性度量方法。但这仍然不能解决数据稀疏带来的推荐质量下降的问题,鉴于此,在上述改进的基础上,提出了利用复杂网络中的结构相似性来度量用户之间相似性的方法,使计算结果更具实际意义和准确性。实验表明,通过这些改进能够有效避免传统方法带来的弊端,提高系统的推荐质量。

关键词 协同过滤,推荐系统,共同评分项,结构相似性,电子商务

中图法分类号 TP391 文献标识码 A

## Improved Collaborative Filtering Recommendation Algorithm

HUANG Tao[1] HUANG Ren[1] ZHANG Kun[2]

(College of Computer Science, Chongqing University, Chongqing 400044, China)[1]

(College of Automation, Chongqing University, Chongqing 400044, China)[2]

**Abstract** The collaborative filtering recommendation algorithm is one of the most important recommendation technologies in E-commerce recommendation system, and the similarity measuring method plays a key role for the accuracy of recommendation results. However, the traditional similarity measure methods ignore the influence on recommendation quality resulting from the number of the common grading items between users. Given this situation, a novel approach was firstly proposed based on the number of the common grading items when measuring the similarity between users. Further more, to protect recommendation result from the data sparsity, the structural similarity measure method of complex network was employed to evaluate the similarity between users. The experimental results show that the proposed approaches can avoid the disadvantages of traditional methods effectively and improve the quality of the recommendation.

**Keywords** Collaborative filtering, Recommendation system, Common grading items, Structural similarity, E-commerce

为了解决“信息过载”的问题,像阿里巴巴、亚马逊、京东、Snapdeal等这样的大型电商都应用了各种形式的推荐系统。电子商务推荐系统可以通过多种方式向用户推荐,比如根据用户浏览记录、商品项目销量排行、用户对商品项目的评分等。而用户对商品项目喜好的最直接体现是用户对商品项目的评分,据此给用户的推荐往往更符合用户喜好,达到更好的推荐质量。协同过滤推荐算法正是基于用户对商品项目的评分数据来产生推荐[1]。迄今,协同过滤推荐算法是在电子商务推荐系统中应用最成功的推荐技术之一[2,3],而影响协同过滤推荐算法准确率高低的关键因素是用户相似性度量方法。

传统的用户相似性度量[3]方法主要有:余弦相似性[4]和相关相似性[4],而这些方法在度量用户相似性时,都是先寻找两个用户的共同评分项,然后只在共同评分项的基础上计算相似性。通常情况下用户之间共同评分项越多,相似性应该越高,但上述传统方法无法反映这一事实。例如有评分集合 $a=\{3,3,0,0\}$、$b=\{2,2,0,0\}$ 和 $c=\{3,3,3,3\}$、$d=\{2,2,2,2\}$,$a$、$b$ 的共同评分项数量为 $2$($0$ 表示没有评分),$c$、$d$ 的共同评分项数量为 $4$,使用余弦相似性、相关相似性度量方法,则 $a$、$b$ 和 $c$、$d$ 的相似性结算结果均为 $1$,显然共同评分项数量不同,而它们的相似性计算结果却是相同的,但后者由于共同评分项数量更多,因此相似度应该更高。另外更加普遍的现象是,传统的相似性度量计算结果相差较小,但共同评分项数量相差很大,这时应该更加关注由于共同评分项数量不同对推荐质量所造成的影响,显然传统方法无法对上述情况加以区分。为了解决这一问题,本文在传统相似性度量方法的基础上,考虑共同评分项数量,得到一个改进的相似性度量方法(Comparing Method Based on Number of Common Grade, CGCM),实验结果表明,本方法可以提高系统的推荐质量。

为了缓解数据稀疏带来的推荐质量下降的问题,本文在CGCM的基础上进一步优化,提出基于共同邻居数量的相似性度量方法(Comparing Method Based on Number of Com-

本文受重庆市研究生科研创新项目(CYS15026)资助。

黄 涛(1990—),男,硕士,主要研究方向为数据挖掘,E-mail:240020416@qq.com;黄 仁 男,博士,副教授,主要研究方向为数据挖掘、模式识别、图像处理;张 坤(1989—),男,硕士,主要研究方向为智能控制系统、模式识别。

mon Grade and Neighbor, NCGCM)。共同邻居数量是复杂网络中结构相似性的一个重要指标[5]，Liben-Nowell 和 Kleinberg[6]以及周涛、尧吕琳媛和张翼成[7]研究了大量刻画复杂网络节点接近性的定量指标，发现两个节点的共同邻居数量越多，它们之间存在直接连接的可能性就越大[6,7]，这说明共同邻居数量可以用来衡量两个节点的相似性。本文将用户作为网络中的节点，将用户关系作为网络中的边，基于 CGCM 来判定节点之间是否存在边，从而形成一个用户关系网络。然后在网络中计算用户的共同邻居数量，并以此作为相似性度量的重要指标。实验结果表明，本方法可以进一步提高系统的推荐质量。

## 1 相关工作

推荐系统是数据挖掘的一个热点课题，为了能够进一步提高推荐的准确率，并降低推荐系统的时间复杂度以保证实时性要求，一直以来，研究者从各个方向提出了各种不同的推荐算法，如协同过滤推荐系统、基于内容的推荐系统、Bayesia 网络技术、聚类技术、关联规则技术以及基于网络结构的推荐算法等。

Grundy[2]是最早投入应用的协同过滤推荐系统，它建立用户兴趣模型，并基于这个模型向用户推荐相关书籍。Video 推荐系统[2]和 Ringo 推荐系统[3]利用相同的社会信息过滤方法分别向用户推荐电影和音乐。文献[8]提出一种基于用户间多相似度的协同过滤推荐算法，克服了单一的评分相似度无法反映人们对不同类型事物的喜好程度不同的问题。Breese 等人[9]对各种协同过滤推荐算法及其改进进行了深入分析

协同过滤推荐依赖于用户对项目的评价意见，而基于内容的推荐系统通过用户已经选择的项目的内容信息计算用户之间的相似性，然后通过相似用户对相关项目的评分预测目标用户对项目的评分[10]。

Bayesian 网络技术用决策树来表示相应的模型，树中的每个节点代表一个领域里的产品，节点的状态代表产品的打分值，网络的相关信息都从训练集中学习得到[3]。这个方法的优点是模型非常小，因此应用该方法的速度非常快；缺点是每个用户只属于一个类，但有一些领域的推荐中在用户属于多个类的情况下才能获得更好的推荐效果。

聚类技术假设用户的打分是彼此独立的，将兴趣类似的用户聚集到相同的簇中[10]，之后根据用户聚集簇中相关用户对商品的历史评价数据推测出目标用户对相应商品的评价。由于聚类过程是离线处理的，因此推荐系统在线产生推荐的速度会很快

关联规则技术普遍的应用在零售业，在营销过程中各种商品的潜在相关性可以通过关联规则发掘。基于关联规则的推荐算法使用由用户数据生成的关联规则，建立推荐模型，并基于推荐模型和用户的购买行为向用户产生推荐[10]。由于可以离线构建关联规则模型，关联规则技术能够满足推荐系统的实时性要求。

基于网络结构的推荐算法是一种较新的推荐算法[3]，该算法不用考虑用户和推荐对象的内容，而是把用户和推荐对象抽象为节点，若用户选择了某一推荐对象就会在用户和对象之间存在选择关系，该策略认为信息就隐藏在该选择关系中。

文献[11]针对数据的极端稀疏性问题提出一种基于项目评分预测的协同过滤推荐技术。该技术根据用户对项目的评分信息估测项目之间的相似性，然后利用用户已评分项目来估计未评分项目的得分，并填充用户_项矩阵，进而根据填充了的用户_项矩阵计算用户相似性，这种做法取得了较好的效果。然而，该算法在度量项目之间或用户之间的相似性时依旧使用的是传统的相似性度量方法，在一定程度上影响了算法的效能。

## 2 传统的相似性度量方法及其分析

协同过滤推荐依靠与目标用户最相似的若干用户的喜好来预测目标用户的喜好，这样做的理由是用户之间的相似度越大，他们拥有共同兴趣爱好的可能性也就越大[9]。

那么推荐系统首先要做的就是找到目标用户的最近邻居，而要做到这一点，必须度量用户之间的相似性，然后选择相似性最高的若干用户作为目标用户的最近邻居[15]，最后根据最近邻居的喜好对目标用户进行推荐。显然，准确度量用户相似性是影响协同过滤推荐系统质量的关键因素。

用户评分数据可以用一个 $r\times c$ 阶矩阵 $M(r,c)$ 表示，如表 1 所列，其中 $r$、$c$ 分别代表 $r$ 个用户和 $c$ 个项目，$R_{i,j}$ 代表矩阵的第 $i$ 行第 $j$ 列的元素，即第 $i$ 个用户对第 $j$ 个项目的评分。

表 1 用户评分数据矩阵

| | $Item_1$ | … | $Item_k$ | … | $Item_c$ |
|---|---|---|---|---|---|
| $User_1$ | $R_{1,1}$ | … | $R_{1,k}$ | … | / |
| … | … | … | … | … | |
| $User_j$ | $R_{j,1}$ | … | / | … | $R_{j,c}$ |
| … | … | … | … | … | … |
| $User_r$ | / | … | $R_{r,k}$ | … | $R_{r,c}$ |

设 $sim(i,j)$ 代表用户 $i$ 和用户 $j$ 之间相似性，其传统计算方法主要有余弦相似性和相关相似性[3]。

·余弦相似性(cosine)：设经用户 $i$ 和用户 $j$ 共同评分的项目集合用 $I_{ij}$ 表示，则用户 $i$ 和用户 $j$ 之间的相似性如式(1)所示。

$$sim(i,j)=\frac{\sum_{c\in I_{ij}} R_{i,c}\cdot R_{j,c}}{\sqrt{\sum_{c\in I_{ij}} R_{i,c}^2}\sqrt{\sum_{c\in I_{ij}} R_{j,c}^2}} \tag{1}$$

·相关相似性(correlation)：设经用户 $i$ 和用户 $j$ 共同评分的项目集合用 $I_{ij}$ 表示，则用户 $i$ 和用户 $j$ 之间的相似性 $sim(i,j)$ 通过 Pearson 相关系数度量，如式(2)所示，其中，$\overline{R_i}$ 和 $\overline{R_j}$ 分别表示用户 $i$ 和用户 $j$ 对项目的平均评分。

$$sim(i,j)=\frac{\sum_{c\in I_{ij}} (R_{i,c}-\overline{R_i})(R_{j,c}-\overline{R_j})}{\sqrt{\sum_{c\in I_{ij}} (R_{i,c}-\overline{R_i})^2}\sqrt{\sum_{c\in I_{ij}} (R_{j,c}-\overline{R_j})^2}} \tag{2}$$

显然，以上相似性度量方法都未考虑共同评分项数量对推荐质量的影响，而且在用户评分数据极端稀疏[9]的情况下并不能有效地度量用户之间的相似性，从而使得计算出来的目标用户的最近邻居不准确，导致整个推荐算法的推荐质量下降。针对上述方法的不足，本文将用户之间的共同评分项数量作为相似性计算的一个重要指标，得到一个改进的相似性度量方法，并在此改进的基础上提出了利用复杂网络中的结构相似性来度量用户之间相似性的方法，以提高推荐质量。

## 3 本文算法设计

本文的推荐算法主要分为两步：寻找最近邻居和产生推荐。3.1 节、3.2 节分别介绍了基于 CGCM 和基于 NCGCM 的邻居查找方法，3.3 节描述了根据最近邻居产生推荐的计算方法。

### 3.1 CGCM 算法

传统的相似性度量方法忽略了共同评分项数量对相似性度量的影响，但事实上，共同评分项数量越大，用户的相似性应该越高，尤其是在传统相似性计算结果相差很小的情况下，就更不能忽略共同评分项数量，否则找到的最近邻居可能会是计算相似但实际上相差较大的用户。为了解决这一问题，本文在传统相似性度量方法中加入了共同评分项数量作为影响因子。

$I_i$、$I_j$ 分别表示经用户 $i$ 和用户 $j$ 评分的项目集合，二者的共同评分项数量记为 $|I_i \cap I_j|$，项目的总数记为 $L$，用户 $i$ 和用户 $j$ 之间的传统相似性计算结果用 $sim(i,j)$ 表示，则本文的 CGCM 相似性计算方法如式(3)所示，其中，$\beta$ 是控制因子，用来控制共同评分项数量在相似性度量中所起的作用，当 $\beta=1$ 时相似性完全根据传统相似性度量方法来计算，而当 $\beta=0$ 时相似性完全根据共同评分项数量来计算，$\beta$ 在 $[0,1]$ 区间滑动。

$$cgsim(i,j)=sim(i,j)\times\beta+\frac{|I_i\cap I_j|}{L}\times(1-\beta) \tag{3}$$

### 3.2 NCGCM 算法

基于 CGCM 查找邻居能够提高协同过滤推荐质量，但提高程度有限，这主要是因为 CGCM 依然无法应对过于稀疏的数据。为了缓解这一问题，本文从另一个方面——共同邻居数量[5]来度量用户相似性。

共同邻居数量[5]是复杂网络中度量节点之间相似性的重要指标。而在推荐系统当中，用户以及用户之间的关系就可以构成一个网络。本文将用户作为网络中的节点，将用户之间的关系作为网络中的边，并且基于 CGCM 判定节点之间是否存在边。

节点 $i$ 和节点 $j$ 之间是否存在边的判定方式如式(4)所示，其中 $N_{ij}=1$ 代表节点 $i$ 和节点 $j$ 是相邻的，反之不是。$\eta$ 是一个阈值，只有节点之间的 CGCM 相似性大于阈值时，它们之间才存在边。

$$N_{ij}=\begin{cases}1, & (cgsim(i,j)>\eta)\ \text{and}\ (i\neq j)\\ 0, & (cgsim(i,j)\leqslant\eta)\ \text{or}\ (i=j)\end{cases} \tag{4}$$

对于 $\eta$ 的取值，最好是能够尽量均衡地划分节点的邻居和非邻居，这样可以让共同邻居数量的分布更为均匀。相反，如果 $\eta$ 取值为 0，那么网络中每个节点与其他任何一个节点之间都存在边，即所有节点之间都是邻居，这样就会导致所有节点之间的共同邻居数量都相等，且为最大值，从而也就无法根据共同邻居数量来度量节点之间的相似性。因此，$\eta$ 取节点之间相似性的平均值，具体如式(5)所示，其中，$U$ 代表节点集合。

$$\eta=\frac{\sum_{i<j,i\in U,j\in U,i\neq j} cgsim(i,j)}{|U|\times(|U|-1)}\times 2 \tag{5}$$

设 $N_i$、$N_j$ 分别表示节点 $i$ 和节点 $j$ 的邻居集合，节点 $i$ 和节点 $j$ 的共同邻居数量记为 $|N_i\cap N_j|$，则节点 $i$ 和节点 $j$ 的 NCGCM 相似性的计算方法如式(6)所示。

$$ncgsim(i,j)=\frac{|N_i\cap N_j|}{|N_i\cup N_j|} \tag{6}$$

### 3.3 产生推荐

本文首先使用 CGCM 和 NCGCM 相似性度量方法得到目标用户的最近邻居，然后基于最近邻居预测目标用户对项目的评分，计算方法如式(7)所示[16]。其中，$P_{u,i}$ 表示用户 $u$ 对项目 $i$ 的预测评分，$NBSu$ 表示用户 $u$ 的最近邻居集合，$R_{n,i}$ 表示用户 $u$ 对项目 $i$ 的评分，$\overline{R_u}$和$\overline{R_n}$分别表示用户 $u$ 和用户 $n$ 对项目的平均评分。最后，由式(7)计算得到目标用户对未评分项目的预测评分后，选择预测评分最高的若干项目即可以作为推荐结果推荐给用户，实现推荐系的最终目的。

$$P_{u,i}=\overline{R_u}+\frac{\sum_{n\in NBSu} ncgsim(u,n)\times(R_{n,i}-\overline{R_n})}{\sum_{n\in NBSu}(|ncgsim(u,n)|)} \tag{7}$$

## 4 实验结果及分析

### 4.1 数据集

本实验采用的数据集是目前在衡量推荐算法质量中比较常用的由美国 Minnesota 大学计算机科学与工程学院公布的 MovieLens(http://grouplens.org/datasets/movielens) 数据集。

本文选取其中一部分数据，包括 943 个用户在 1682 部电影上的 100000 条评分记录。评分的范围是 1～5，评分越高，表示用户对电影越感兴趣。在本文的实验中，数据集中 80% 的数据用来作为训练集、20% 的数据作为测试集来对比验证算法的有效性。此外，根据用户评分矩阵中未评分项目在整体数据集中所占的比例，可以观察数据的稀疏程度，比例计算结果如式(8)所示，显然，本文所使用的数据集是相当稀疏的，因此所选用的数据集比较适合于检验所提出的算法在数据稀疏状况下的处理效果。

$$\varphi=1-\frac{100000}{943\times1682}=0.936953 \tag{8}$$

### 4.2 度量标准

评价推荐系统推荐质量的度量标准主要包括统计精度度量方法和决策支持精度度量方法两类[11-13]。本文采用统计精度度量方法中的平均绝对偏差 MAE 作为度量标准。MAE 计算预测的用户评分与实际的用户评分之间的偏差，偏差越小代表推荐质量越高。设 $\{p_1,p_2,\cdots,p_{N-1},p_N\}$ 代表预测的用户评分集合，对应的实际的用户评分集合为 $\{q_1,q_2,\cdots,q_{N-1},q_N\}$，那么平均绝对偏差 MAE 定义如式(9)所示[14]。

$$MAE=\frac{\sum_{i=1}^{N}|p_i-q_i|}{N} \tag{9}$$

### 4.3 实验结果

#### 4.3.1 相似性度量标准比较

本实验首先对传统相似性度量标准(余弦相似性、相关相似性)进行了实验，目的是为了选择最佳的传统相似性度量标准作为下一步实验的基础。以余弦相似性、相关相似性分别对数据集进行实验，计算其平均绝对偏差 MAE，最近邻从 4 个开始，递增到 28 个最近邻，每次递增 4 个，实验结果如图 1 所示。从图 1 中可以看出，在各种条件下，相比于相关相似性度量方式，平均绝对偏差 MAE 在余弦相似性度量方式下的

值均明显较低，由此说明余弦相似性度量方法在衡量用户之间相似性时占据一定优势。因此，在下文的实验中，将使用余弦相似性度量方法作为 CGCM 即式(3)的改进基础。

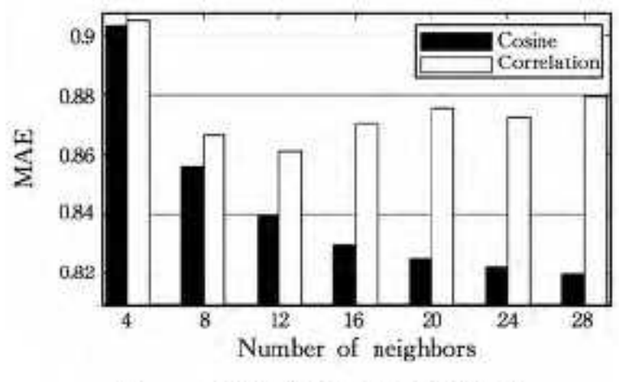


图 1 相似性度量标准比较

### 4.3.2 共同评分项数量和余弦相似性比例实验

这个实验的目的是观测共同评分项数量在推荐预测中所起的作用，以确定共同评分项数量对推荐质量的影响权重，为此，本文在 CGCM 计算式(3)中引入了 $\beta$ 因子。$\beta$ 在[0,1]区间滑动，每次滑动的间隔为 0.1。产生推荐所使用的邻居数量固定为 28。从图 2 中可以看出，共同评分项数量对推荐计算是起着积极作用的。当 $\beta=0.6$ 时，MAE 降到最低。

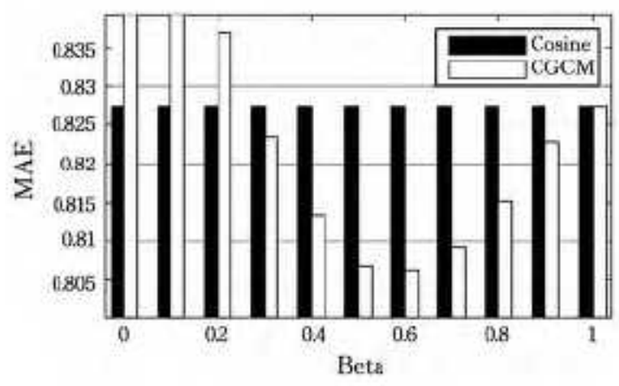


图 2 在不同的 Beta 值上的 MAE 比较(式(3))

### 4.3.3 相似性度量方法比较实验

为了检验本文提出的改进的相似性度量方法的有效性，以传统的相似性度量方法即余弦相似性作为对照，观察 CGCM 相似性度量方法和 NCGCM 相似性度量方法对推荐质量的影响。其中 CGCM 的 $\beta$ 系数确定为 0.6，最近邻数量从 5 开始，递增到 30 个最近邻，每次递增 5 个。实验结果如图 3 所示。

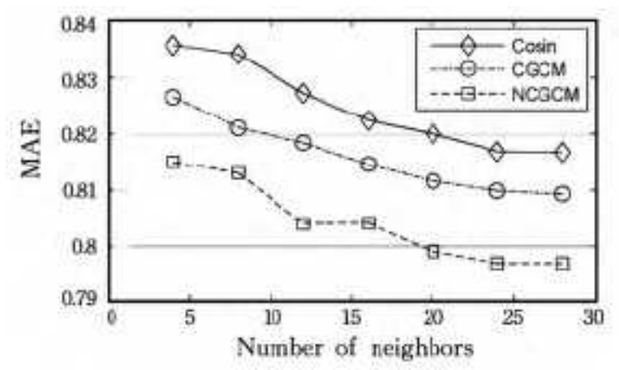


图 3 相似性度量方法的准确率比较

由图 3 可知，在各种实验条件下，本文提出的两个改进的相似性度量方法在协同过滤推荐的框架下均具有较小的 MAE。由此可知，与传统的相似性度量方法相比，本文提出的改进可以显著地提高推荐系统的推荐质量。

结束语 本文首先深入分析了余弦相似性、相关相似性度量方法在计算目标用户的最近邻居时存在的问题。接着，针对这些问题，提出基于 CGCM 和基于 NCGCM 的邻居查找的改进方法，有效地解决了上述度量方法存在的不足，使得计算得到的目标用户的最近邻居比较准确。最后，实验结果表明，本文所提出的改进方法显著提高了推荐系统的推荐质量。

## 参 考 文 献


[1] Herlocker J L, Konstan J A, Riedl J. Explaining collaborative filtering recommendations[C]// Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. ACM, 2000: 241-250

[2] Hu J. Application and research of collaborative filtering in e-commerce recommendation system[C]// 2010 3rd International Conference on Computer Science and Information Technology. 2010: 686-689

[3] Adomavicius G, Tuzhilin A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(6): 734-749

[4] Ahn H J. A New Similarity Measure for Collaborative Filtering to Alleviate the New User Cold-Starting Problem[J]. Information Sciences, 2008, 178(1): 37-51

[5] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报, 2010, 39(5): 651-661

[6] Wang X, Pournaras E, Kooij R E, et al. Improving robustness of complex networks via the effective graph resistance[J]. European Physical Journal B, 2014, 87(9): 1-12

[7] Zhou T, Lü L, Zhang Y. C. Predicting missing links via local information[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2009, 71(4): 623-630

[8] 范波，程久军. 用户间多相似度协同过滤推荐算法[J]. 计算机科学, 2012, 39(1): 23-26

[9] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C]// Fourteenth Conference on Uncertainty in Artificial Intelligence. 1998: 43-52

[10] Sharma L, Gera A. A Survey of Recommendation System: Research Challenges[J]. International Journal of Engineering Trends & Technology, 2013, 4(5): 1989-1992

[11] 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

[12] Herlocker J L, Konstan J A, Terveen L G, et al. Evaluating collaborative filtering recommender systems[J]. ACM Transactions on Information Systems, 2004, 22(1): 5-53

[13] Fukazawa, Yusuke. Intellectual Property DepartmentNTT DOCOMO INC. Sanno Park Tower11-1. Recommendation information evaluation apparatus and recommendation information evaluation method: EP, EP2264624 A1[P]. 2010

[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. ACM, 2001: 285-295

[15] Hurley N, Zhang M. Novelty and Diversity in Top-N Recommendation--Analysis and Evaluation[J]. ACM Transactions on Internet Technology, 2011, 10(4): 63-74

[16] Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering[C]// Fourteenth Conference on Uncertainty in Artificial Intelligence. 1998: 43-52