# 一种基于分众分类的协同过滤推荐算法①

吴春旭 李佳俊 石 辉 (中国科学技术大学 管理科学系 安徽 合肥 230026)

**摘 要:** 基于内存的协同过滤算法是推荐系统中使用的最成功的技术之一，但它存在着数据稀疏性和可扩展性的问题。分众分类是一种能使用户发现、组织和理解在线事物的强有力的机制。基于这种机制，提出了一种新的协同过滤算法，来解决该算法中的稀疏性和可扩展性的问题。实验表明，该算法在解决这些问题上是有效的。

**关键词:** 协同过滤；推荐系统；分众分类；算法；稀疏性；可扩展性

## A Collaborative Filtering Recommendation Algorithm Based on Folksonomy

WU Chun-Xu, LI Jia-Jun, SHI Hui

(Department of Management Science, University of Science and Technology of China, Hefei 230026, China)

**Abstract:** The memory-based collaborative filtering algorithm is one of the most successful technologies for recommender systems, although these approaches all suffer from data sparsity and poor scalability problems. Folksonomy has emerged as a powerful mechanism that enables users to find, organize, and understand online entities. This paper focuses on how to address these problems by using tags. The experiment for the tag-based algorithm and other algorithms showed that the novel algorithm achieves better performance compared to the traditional ones, proving the validity of the algorithm.

**Keywords:** collaborative filtering; recommender system; folksonomy; algorithm; sparsity; scalability

## 1 引言

信息时代，客户都会受到信息过载的困扰，推荐系统可以帮助用户在巨大的信息资源中，找到自己所需的信息。

推荐系统中有两种实体及其关系：用户实体与项目实体；用户与项目、用户与用户及项目与项目之间的关系。这三类关系常用相似度来表示，也是各类推荐系统进行推荐的数据基础。充分利用这三种关系，可以更好地解决信息过载的问题，为用户提供更好的个性化的推荐服务。

随着 Web 2.0 技术的成熟，因特网上产生了大量的用户生成内容(UGC)，极大地丰富了客户的偏好信息。标签是一种重要的并引起更多关注的 UGC 的形式。用户用自由的标签标记感兴趣的资源，从而构造了一个社会范围的分类架构，叫做分众分类[1]。

UGC 为协同过滤(collaborative filtering, CF)推荐提供了新的数据。John 等学者将逆用户频率(inverse user frequency, IUF)用于计算用户相似度，以降低那些被很多用户评分的项目的权重[2]。Jörg Diederich 等学者扩展了这一概念，使用了类似 IUF 的概念，将标签向量作为用户概况(User Profile)，使用标签的频率来计算用户之间的相似性，并用 IUF 调整标签频率的权重[3]。本文提出用类似 TFxIDF 的概念，计算标签的 TFxIUF(tag frequency – inverse user frequency)频率和 TFxIIF(tag frequency – inverse item frequency)频率，并利用这两种频率来计算相似度，以提高基于内存的协同过滤算法的相似度计算精度，取得了较好的结果。

## 2 基于内存的协同过滤

基于内存的协同过滤(memory-based collaborative filtering)有两种类型：一种是基于用户(User-

① 收稿时间:2009-09-15;收到修改稿时间:2009-12-08

based)的，一种是基于项目(Item-based)的。

基于用户的协同过滤(UBCF)假设：如果两个用户在一些项目上评分相似，在其它项目上也相似。即要求用户在所涉及的项目上满足：当在用于计算相似性的项目评分上相似时，在用于预测的项目评分上也相似。因此，在 UBCF 中，我们可以寻找目标用户的最近邻居，并根据最近邻居的偏好，来预测目标用户对目标项目的评分。

基于项目的协同过滤(IBCF)假设：其他用户认为相似的项目，目标用户也认为相似。用户处于某个相似群体，他们在认知上是相似的。IBCF 适用于向同一类型的用户推荐，即，评分矩阵中的用户是同一类别的。所以，在 IBCF 中，我们可以寻找目标项目的最近邻居，直接用目标用户对最近邻居项目的评分预测来对目标项目的评分。

使用基于内存的协同过滤有几个前提条件。对于 UBCF，一方面，每个用户的已评分项不能过少，以便计算用户之间的相似度；另一方面，也要有一些未评分项，否则就没有可推荐的项目。IBCF 有类似的要求。概括起来就是要求评分矩阵有适当的密度。事实上的用户评分矩阵非常稀疏[4]，这就造成了所谓的数据稀疏性问题。另外，随着系统中用户和项目的数目的增多，算法的性能随之降低，带来了可扩展性问题。最后还有冷启动问题：如果从来没有一个用户对某一项目加以评价，那么该项目不可能被推荐。

基于内存的协同过滤算法通常将用户的各种浏览、点击、购买、标记等日志数据完全转化成单一的评分，形成评分矩阵，然后基于评分矩阵进行过滤推荐。鉴于 UBCF 和 IBCF 的假设前提，这既不能保证推荐的精度，又不能摆脱数据稀疏性问题。本文充分利用 UGC 环境下大量的标签数据，将标签视为一种独立实体，并将与其相关的频率用于计算相似度，在不增加数据稀疏性的前提下，进一步提高了推荐精度。

## 3　基于分众分类的协同过滤算法

设：用户的集合为 $U=\{u_1,u_2,\cdots,u_i,\cdots,u_m\}$，项目的集合为 $I=\{i_1,i_2,\cdots,i_j,\cdots,i_n\}$，标签的集合为 $T=\{t_1,t_2,\cdots t_k,\cdots,t_l\}$。评分矩阵为 $R=(r_{ij})_{m\times n}$，其中 $r_{ij}$ 为用户 $u_i$ 对项目 ij 的评分。

### 3.1 标签 TFxIDF 矩阵

TFxIDF 是一种权重，常用于信息检索和文本挖掘，使用一种统计的方法，来评估一个文档集合中一个词对一个文档的重要程度。在这里，这一权重被用于相似性计算。如果将用户使用的标签看作文档，得到一个权重 TFxIUF，那么得到的权重就是用户对某个标签的偏好程度，得到的矩阵类似于评分矩阵，也可以计算用户之间的相似度。如果将标记一个项目的所有标签看作文档，得到一个权重 TFxIIF，那么得到的权重就是标签对项目标记的重要程度。以 TFxIIF 形成的矩阵类似于评分矩阵，可以计算项目之间的相似性。

#### 3.1.1 计算 TFxIUF

用户 $u_i$ 的标签的集合为 $T_i^U$，用户 $u_i$ 对所有项目的标记次数记为 $C_i^U(A)$，用户 $u_i$ 用标签 $t_k$ 标记项目的次数记为 $C_i^U(k)$，那么标签 $t_k$ 在用户 $u_i$ 的所有标签中的频率为 $f_i^U(k)$，用(1)式计算。

$$f_i^U(k)=\frac{C_i^U(k)}{C_i^U(A)} \tag{1}$$

定义 $iuf_k$ 为标签 $t_k$ 的逆用户频率，用(2)式计算。

$$iuf_k=\log\frac{C^U(A)}{C^U(k)} \tag{2}$$

其中，$C^U(A)=|U|$ 表示所有用户数，$C^U(k)=|\{u_i|t_k\in T_i^U\}|$ 表示拥有标签 tk 的用户数。定义标签 tk 在用户 $u_i$ 中的 TFxIUF 频率为 $tf\times iuf(k,i)$，且 $tf\times iuf(k,i)=f_i^U(k)\cdot iuf_k$，得到矩阵 $M^U$，表示为(3)式。

$$M^U=\begin{bmatrix} tu(1,1) & \cdots & tu(1,m) \\ \vdots & \ddots & \vdots \\ tu(l,1) & \cdots & tu(l,m) \end{bmatrix} \tag{3}$$

其中，$tu(k,i)=tf\times iuf(k,i)$。

#### 3.1.2 计算 TFxIIF

项目 ij 的标签的集合为 $T_j^I$，所有用户对项目 ij 的标记次数记为 $C_j^I(A)$，所有用户用标签 $t_k$ 标记项目 ij 的次数记为 $C_j^I(k)$，那么标签 $t_k$ 在项目 ij 的所有标签中的频率为 $f_j^I(k)$，用(4)式计算。

$$f_j^I(k)=\frac{C_j^I(k)}{C_j^U(A)} \tag{4}$$

定义 $iif_k$ 为标签 tk 的逆项目频率，用(5)式计算。

$$iif_k=\log\frac{C^I(A)}{C^I(k)} \tag{5}$$

其中，$C^I(A)$表示所有项目数，$C_j^I(k)$表示包含标签 $t_k$ 的项目数。

定义标签 $t_k$ 在项目 ij 中的 TFxIDF 频率为 tf×iif(k, j)，且 $tf \times iif(k,j) = f_j^I(k) \cdot iif_k$ 得到矩阵 $M^I$，用(6)式表示。

$$M^I = \begin{bmatrix} ti(1,1) & \cdots & ti(1,m) \\ \vdots & \ddots & \vdots \\ ti(l,1) & \cdots & ti(l,m) \end{bmatrix} \tag{6}$$

其中，$t_i(k,j) = tf \times iif(k,j)$ 。

很显然，在矩阵 $M^U$ 和 $M^I$ 中，若对应的用户或项目没有该标签，则对应的项为零。与传统的评分矩阵类似，也面临着数据稀疏性的问题。

### 3.2 标签相似度

使用 Pearson 相关系数，基于标签权重矩阵，计算用户之间，项目之间的相似度。为区别于基于评分矩阵计算得到的评分相似度，将基于标签权重矩阵计算得到的相似度称为标签相似度。

用户 $u_x$ 和用户 $u_y$ 的标签相似度，用(7)式计算。

$$sim_U^T(x,y) = \frac{\sum_{k \in T_{xy}} (tu(k,x) - \bar{T}_u(x)) \cdot (tu(k,y) - \bar{T}_u(y))}{\sqrt{\sum_{k \in T_{xy}} (tu(k,x) - \bar{T}_u(x))^2} \sqrt{\sum_{k \in T_{xy}} (tu(k,y) - \bar{T}_u(y))^2}} \tag{7}$$

其中，$T_{xy}$ 为两个用户都使用的标签。其频率不为零。$\bar{T}_u(a)$ 为用户的 TFxIUF 的均值。

项目 $i_x$ 和项目 $i_y$ 的标签相似度 $sim_I^T(x,y)$，用(8)式计算。

$$sim_I^T(x,y) = \frac{\sum_{k \in T_{xy}} (ti(k,x) - \bar{T}_i(x)) \cdot (ti(k,y) - \bar{T}_i(y))}{\sqrt{\sum_{k \in T_{xy}} (ti(k,x) - \bar{T}_i(x))^2} \sqrt{\sum_{k \in T_{xy}} (ti(k,y) - \bar{T}_i(y))^2}} \tag{8}$$

其中，$T_{xy}$ 为两个项目都使用的标签。其频率不为零。$\bar{T}_i(a)$ 为项目 $i_a$ 的 TFxIIF 的均值。

### 3.3 评分相似度

对项目 $i_x$ 和项目 $i_y$ 都已评分的用户集合表示为 $U_{xy}$，那么项目之间的相关相似性用 Pearson 相关系数度量用(9)式表示。

$$sim_I^R(x,y) = \frac{\sum_{c \in U_{xy}} (r_{cx} - \bar{R}_x) \cdot (r_{cy} - \bar{R}_y)}{\sqrt{\sum_{c \in U_{xy}} (r_{cx} - \bar{R}_x)^2} \sqrt{\sum_{c \in U_{xy}} (r_{cy} - \bar{R}_y)^2}} \tag{9}$$

其中，为用户 $u_c$ 对项目的评分，为用户对项目的平均评分。

用户之间的标签相似度将用于对用户的预选，预选步骤中，选择与目标用户相似的最近邻居。这样有助于基于项目的协同过滤得到更准确的推荐结果，因为选出的用户之间更加相似。项目之间的标签相似度将用于计算项目综合相似度并最终用于寻找目标项目的最近邻居。

### 3.4 综合相似度

将项目的评分相似度和标签相似度加权得到综合相似度，用(10)式表示。

$$sim_I(x,y) = \alpha\, sim_I^R(x,y) + (1-\alpha) sim_I^T(x,y) \tag{10}$$

其中，$\alpha$ 的取值范围在 0 到 1 之间。

### 3.5 算法步骤

#### 3.5.1 预选

使用标签相似度计算目标用户与其他用户的相似度，然后用最近邻法对训练集用户进行过滤，得到标签权重矩阵 $M^I$ 和评分矩阵 R。

#### 3.5.2 计算项目评分相似度

使用(9)式，利用预选过滤得到的评分矩阵 R，计算目标项目与其他项目的评分相似性。

#### 3.5.3 计算项目标签相似度

使用(8)式，利用预选过滤得到的标签权重矩阵 $M^I$，计算目标项目与其他项目的相似度。

#### 3.5.4 计算项目综合相似度

使用(10)式，利用上面得到的标签相似度和评分相似度，计算项目的综合相似度。

#### 3.5.5 推荐预测

目标用户 $u_i$ 对目标项目 ij 的预测评分，通过(11)式计算。

$$r_{ij} = \sum_{y \in I_n} r_{iy} sim_I(j,y) \Big/ \sum_{y \in I_n} sim_I(j,y) \tag{11}$$

其中，$I_n$ 为目标项目的最近邻居集合。

## 4 实验结果及分析

### 4.1 数据集

采用 MovieLens 数据集测试算法的效果。MovieLens (http://www.movielens.org)是一个基于 Web 的研究型推荐系统，它接收用户对电影的评分并提供电影推荐。现在该站点已经开始支持标签，并已经接收相当数量的标签。从该数据集中随机抽取

71567 位用户对 10681 个项目的 10M 条评分数据和 100K 条标签作为实验数据集，并将评分文件与标签文件整合得到 100K 的完整数据，然后将数据集分为训练集(占 90% ) 和测试集(占 10% )。

### 4.2 评价标准

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[5]。统计精度度量方法中的平均绝对偏差 MAE(mean absolute error)易于理解，可以直观地对推荐质量进行度量，是常用的一种推荐质量度量方法，本文采用平均绝对偏差 MAE 作为度量标准。平均绝对偏差 MAE 通过计算预测的用户评分与实际的用户评分之间的偏差度量预测的准确性，MAE 越小，推荐的质量越高。

定义平均绝对偏差 MAE 为(12)式。

$$MAE = \frac{\sum_{r_{ij} \in Te} | r_{ij} - \hat{r}_{ij} |}{|Te|} \tag{12}$$

其中，$T_e$ 为测试集，$r_{ij}$ 为预测值，$\hat{r}_{ij}$ 为实际值。

### 4.3 实验结果

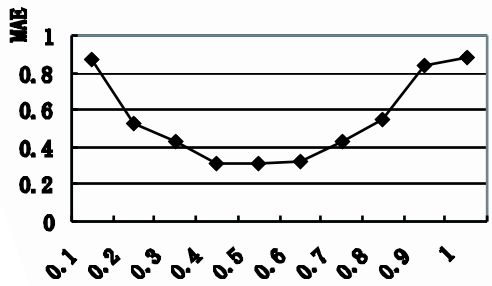


图 1　权重参数对 MAE 的影响

我们首先做了值对 MAE 影响的实验，结果如图 1 所示。实验表明，当的值在 0.4~0.6 之间时，基于分众分类的算法精度较高。

选取=0.5，做基于分众分类的推荐算法(TBCF)与基于内存的协同过滤算法(IBCF)和基于语义相似度的资源协同过滤算法(IBCFUSS)[5]的 MAE 的对比试验，结果如图 2 所示。结果表明：TBCF 与 IBCF 和 IBCFUSS 相比，具有较小的 MAE。这说明基于分众分类的协同过滤算法改善了数据稀疏性造成的推荐不准确的缺点，有效地提高了推荐质量。

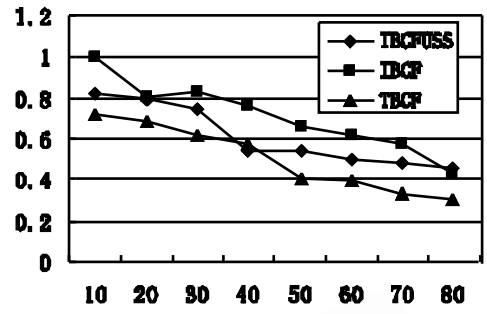


图 2　推荐算法 MAE 比较

## 5　结论

本文分析了基于内存的协同过滤推荐系统面临的主要问题：评分矩阵的稀疏性问题和用户及项目维度过大的扩展性问题，分析了分众分类的特点，提出了基于分众分类的协同过滤推荐算法。实验表明，新算法较好地解决了这些问题。未来的研究工作将进行更广泛地探讨与 Web 2.0 技术的结合，以求进一步缓解数据稀疏性的问题和提高预测精度。

### 参考文献

1 Kim HN, Ji AT, Ha I, Jo GS. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation. Electronic Commerce Research and Applications.2009.

2 Breese JS, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proc. of the 14th Conf. on UAI-98. San Francisco, July 24－26 1998. 43－52.

3 Diederich J, Iofciu T. Finding Communities of Practice from User Profiles Based On Folksonomies. In: Proc. of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs'06).2006.

4 Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proc. of the 10th International World Wide Web Conference. 2001. 285－295.

5 崔林,宋瀚涛,陆玉昌.基于语义相似性的资源协同过滤技术研究.北京理工大学学报, 2005,25(5):402－405.