第37卷 第14期 计 算 机 工 程 2011年7月
Vol.37 No.14 Computer Engineering July 2011

·软件技术与数据库· 文章编号：1000—3428(2011)14—0034—02 文献标识码：A 中图分类号：TP311

基于Tag和协同过滤的混合推荐方法

王卫平，王金辉

(中国科学技术大学管理学院，合肥 230026)

摘 要：针对传统协同过滤方法的稀疏性问题，提出基于标签(Tag)和协同过滤的混合推荐方法TAG-CF。通过Tag分类信息获取项目的最近邻居，采用基于项目的最近邻方法预测用户评分值，并利用该预测值填充用户评分矩阵，构造密集的伪矩阵，运用基于用户的的协同过滤方法在伪矩阵上计算用户对项目的预测评分值。实验结果表明，TAG-CF能有效降低推荐系统的平均绝对误差，提高推荐质量。

关键词：推荐系统；协同过滤；稀疏性；大众分类法；标签

Hybrid Recommendation Method Based on Tag and Collaborative Filtering

WANG Wei-ping, WANG Jin-hui

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

【Abstract】Aiming at the problem of sparse user ratings facing tradition recommendation system, this paper proposes a hybrid recommendation method based on Tag and collaborative filtering(TAG-CF) to provide a solution to this problem. The neighbors set for the target item can be gained based on tagging information. It uses item-based collaborative filtering to generate the predictive ratings. By filing these predictive ratings into the sparse user-item rating matrix, it constructs a full pseudo ratings matrix. It computes the predictions based on the pseudo ratings matrix by using user-based collaborative filtering. Experimental results show that the proposed method performs significantly better than the traditional CF method.

【Key words】recommendation system; collaborative filtering; sparsity; mass taxonomy; Tag

DOI: 10.3969/j.issn.1000-3428.2011.14.009

1 概述

电子商务中存在信息过载问题，用户必须花费大量的时间和精力寻找所需的信息。推荐系统是解决该问题的方法之一。推荐系统能模拟商品销售人员向用户提供商品推荐，帮助用户找到所需商品，从而顺利完成购买过程。根据推荐方法不同，推荐系统可分为：(1)协同过滤推荐系统；(2)基于内容的推荐系统；(3)混合推荐系统。

协同过滤根据不同用户兴趣的相似性来推荐资源(项目)[1]，通常使用用户项目评分矩阵，按一定的度量标准在评分矩阵上计算相似度，根据相似度找出最近邻居，然后参考这些最近邻居的评分值，预测当前用户对目标项目的评分值，将具有最高评分值的项目推荐给当前用户。

协同标签(Tag)系统又称为大众分类系统，是一种知识分享机制。它允许用户用随意的单词或短语标记喜爱的资源(URL、电影、图片、音乐等)，这些短语和单词就称为Tag。Tag是一种重要的显式评分技术，用户可保存这些Tag，以便日后检索和查询。此外，用户之间还可共享Tag信息。目前，较多大型网站(如delicious.com、凤凰网、CSDN等)都不同程度地使用Tag来描述资源。与传统的生物分类系统方法不同，Tag不是由领域专家预先设定的，而是用户自由选择的特定单词或短语；分类可以是多层树状结构，而Tag只有一层，不同的Tag之间只有平行关系，不存在父-子节点关系；用户可以同时为项目设定多个Tag，而在目录结构的分类中，一个对象一次只能存放在一个分类目录中。

目前电子商务系统中的用户数目和项目数目呈指数级增长，用户评分数据极端稀疏。在大型电子商务系统中，用户评分的项目一般不会超过项目总数的1%，经2个用户共同评分的项目则更少，如果2个用户没有共同评分项目，那么他们的相似度为0[2]，这将影响最近邻的查找，降低推荐质量。相关研究给用户未评分项目设定缺省值，或引入其他的信息，如产品分类、用户人口统计[3]、项目内容[4]，用户浏览行为[5]等信息来缓解稀疏性对协同过滤方法的负面影响。但产品分类信息需要领域专家预先设定；由于涉及到用户隐私，人口统计信息获取困难；而诸如图片、视频等项目的内容信息较难抽取；给用户未评分项目设定统一的缺省值可信度不高，因为用户对未评分项的评分不可能完全相同；用户浏览行为信息不能精确反映用户偏好，而且，不同的浏览行为组合可能导致相互矛盾的兴趣倾向。

为此，本文将Tag信息加入到最近邻协同过滤方法中，提出一种基于Tag和协同过滤的混合推荐方法(TAG-CF)。

2 基于Tag和协同过滤的混合推荐方法

2.1 相关定义

定义1(用户-项目评分矩阵$\boldsymbol{R}_{m\times n}$) m行代表m个用户；n列代表n个项目；第i行j列的元素r_{ij}为用户i对项目j的评分值。

定义2(项目-Tag评分矩阵$\boldsymbol{T}_{n\times k}$) n行代表n个项目；k列代表k个Tag。其中，

$$t_{ij}=\begin{cases}1 & \text{if } tag_j\text{标记在项目}i\text{上}\\ 0 & \text{otherwise}\end{cases}$$

作者简介：王卫平(1953－)，女，高级工程师，主研方向：信息系统，商务智能；王金辉，硕士研究生

收稿日期：2010-12-07 **E-mail：**jhwang08@mail.ustc.edu.cn

2.2　算法描述

基于 Tag 和协同过滤的混合推荐方法核心算法如下：

(1)基于项目-Tag 评分矩阵 $\boldsymbol{T}_{n\times k}$ 计算项目相似度 *itemSim*；

(2)基于 *itemSim* 确定项目最近邻，按基于项目的协同过滤方法建立预测评分矩阵 $\boldsymbol{C}_{m\times n}$；

(3)利用评分矩阵 $\boldsymbol{R}_{m\times n}$ 和 $\boldsymbol{C}_{m\times n}$ 构造伪矩阵 $\boldsymbol{M}_{m\times n}$；

(4)在伪矩阵 $\boldsymbol{M}_{m\times n}$ 上计算用户相似度 *userSim*；

(5)基于 *userSim* 确定用户最近邻，按基于用户的协同过滤方法计算预测评分值；

(6)将预测评分值最高的 *N* 个项目推荐给用户。

2.2.1　Tag 预处理

用户可以使用任意的单词或短语作为 Tag，这具有很大的灵活性。但同时也带来了同义词、多义词等问题，由此导致 Tag 评分质量下降。相关研究表明，尽管一个项目可能会被添加上百个 Tag,，但只有很少一部分 Tag 是频繁用到的，而且这些少量频繁使用的 Tag 集合也是稳定的[6]。因此，本文从原始 Tag 信息中提取频繁 Tag 来表示项目。以降低 Tag 数据的噪音对推荐结果的影响。

2.2.2　基于 Tag 的项目相似度计算

定义 3(项目 *i* 的 Tag 集 T_i) 用户标记到项目 *i* 上的的所有 Tag 的集合

本文采用 2 个项目共同的 Tag 来度量项目的相似度。当 2 个项目共同的 Tag 越多，相似度就越大。

$$itemSim(i,j)=\frac{\left|T_i\cap T_j\right|}{\max(\left|T_i\right|,\left|T_j\right|)}$$

2.2.3　最近邻确定与项目评分值预测

本文确定相似度阈值 η=0.2。对于当前项目 *i*，确定 $itemSim(i,j)>\eta$，且 $r_{uj}\neq 0$ 的所有项目组成项目 *i* 最近邻居，记为 *itemNeighbor*(*u*, *i*)，从而采用下式预测用户 *u* 对项目 *i* 的评分值：

$$c_{ui}=\frac{\sum_{j\in itemNeighbor(u,i)} itemSim(i,j)\,r_{ui}}{\sum_{j\in itemNeighbor(u,j)} itemSim(i,j)}$$

按上述方法，为每个用户未评分项目计算预测评分值，建立预测评分矩阵 $\boldsymbol{C}_{m\times n}$。

2.2.4　伪矩阵构造

在传统 CF 方法中，由于用户评分数据极端稀疏，如果用户没有对项目进行评分，则该评分值设置为∅，这导致评分矩阵非常稀疏，在该稀疏矩阵上产生推荐，产生较大误差，推荐系统的推荐质量难以保证。TAG-CF 综合利用用户的评分值和基于 Tag 的预测值，创建了伪矩阵 $\boldsymbol{M}_{m\times n}(m_{ui})$。其中，

$$m_{ui}=\begin{cases} r_{ui} & \text{if 用户}u\text{评价过项目}i \\ c_{ui} & \text{otherwise}\end{cases}$$

2.2.5　基于伪矩阵的用户相似度计算

常用的用户相似度度量方法有 Peatson 相关系数、余弦相似度、Speatman 相关系数、基于熵的不确定性等。本文采用 Peatson 相关系数作为相似度度量标准，在伪矩阵上计算用户相似度：

$$userSim(u,n)=\frac{\sum_{i\in CR_{u,n}}(r_{ui}-\bar{r}_u)(r_{ni}-\bar{r}_n)}{\sqrt{\sum_{i\in CR_{u,n}}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i\in CR_{u,n}}(r_{ni}-\bar{r}_n)^2}}$$

其中，$CR_{u,n}$ 表示用户 *u* 和用户 *n* 共同评价过的项目集合；$\bar{r}_u$ 表示用户 *u* 对所有项目的平均评分值。

2.3　推荐产生

用户 *u* 对未评分项目 *i* 的预测评分计算公式如下：

$$p_{ui}=\bar{r}_u+\frac{\sum_{n\in neighbors(u)} userSim(u,n)(r_{ni}-\bar{r}_n)}{\sum_{n\in neighbors(u)} userSim(u,n)}$$

其中，*neighbors*(*u*)是评价过项目 *i*，且与用户 *u* 最相似的 *K* 个用户的集合

以此方法预测用户 *u* 对于所有未评分项目的评分值，取预测评分值最高的 *N* 个项目(*top-N*)推荐给用户 *u*。

3　实验结果与分析

3.1　数据集

本文使用美国明尼苏达大学 GtoupLens 研究项目组提供的 MovieLens 数据集(http://www.gtouplens.otg)，该数据集记录了 71 567 个用户对 10 681 部电影的 10 000 054 个评分和 95 580 个 Tag，每个用户至少对 20 部电影进行了评分。评分值范围从 1~5。

本文从以上数据集中选取 800 个用户对 1 000 个项目的评分记录，以及相关的 122 个频繁 Tag 记录。按 90%/10%的比例分成训练集和测试集，比较 TAG-CF 与传统的协同过滤推荐方法的性能。

3.2　度量标准

平均绝对误差(Mean Absolute Error, MAE)广泛用于协同过滤推荐系统的推荐质量评价。MAE 通过计算项目的预测评分与实际评分之间的偏差来度量预测的准确性。MAE 越小，推荐系统的准确度越高[1]。MAE 定义如下：

$$M_{\mathrm{MAE}}=\frac{\sum_{j=1}^{N}\left|p_{uj}-r_{uj}\right|}{N}$$

其中，p_{uj} 为预测的用户评分；r_{uj} 为实际的用户评分；预测的项目个数为 *N*。

3.3　实验结果

为检验本文提出的 TAG-CF 方法的有效性，本文以传统的协同过滤方法作为基础，对比随着最近邻居个数取值不同时，2 种方法的 MAE 变化情况。实验设置邻居个数从 3 逐步增加到 15，间隔为 3，分别计算平均绝对误差，结果如图 1 所示。

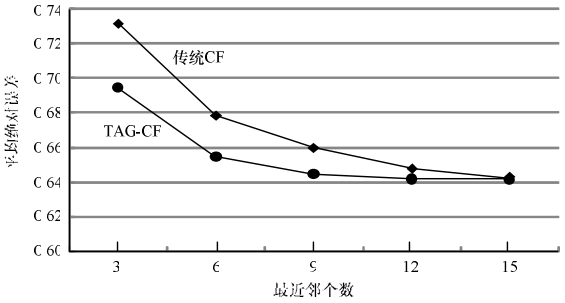


图 1　平均绝对误差比较

由图 1 可知，与传统的协同过滤方法相比，本文提出的 TAG-CF 方法具有较小的 MAE。因此，将 Tag 信息和协同过滤方法相结合，可提高推荐系统的质量。

4　结束语

随着用户和商品数目的不断增加，传统的协同过滤方法面临用户评分稀疏性的挑战。本文利用 Tag 信息构造伪矩阵，从而有效缓解了数据稀疏问题。实验数据表明，本文方法能有效提高推荐质量。下一步将对 Tag 的处理进行研究，引入 Tag 的语义信息，使基于 Tag 的相似度度量更加精确，以进一步提高推荐方法的推荐质量。

(下转第 38 页)