TEFRCF:标签熵特征表示的协同过滤个性化推荐算法

何 明 杨 芃 要凯升 张久伶
(北京工业大学信息学部 北京 100124)

摘 要 标签作为 Web 2.0 时代信息分类和检索的有效方式,已经成为近年的热点研究对象。标签推荐系统旨在利用标签数据为用户提供个性化推荐。现有的基于标签的推荐方法在预测用户对物品的兴趣度时往往倾向于赋予热门标签及其对应的热门物品较大的权重,导致权重偏差,降低了推荐结果的新颖性,未能充分反映用户个性化的兴趣。针对上述问题,定义了标签熵的概念来度量标签的不确定性,提出了标签熵特征表示的协同过滤个性化推荐算法。该算法通过引入标签熵来解决权重偏差问题,利用三分图形式描述用户-标签-项目之间的关系;构建基于标签熵特征表示的用户和项目特征表示,并通过特征相似性度量方法计算项目的相似性;最后利用用户标签行为和项目的相似性线性组合预测用户对项目的偏好值,并根据预测偏好值排序生成最终的推荐列表。在 Last.fm 数据集上的实验结果表明,该方法能够提高推荐准确性和新颖性,满足用户的个性化需求。

关键词 协同过滤,标签,熵,推荐系统
中图法分类号 TP391 **文献标识码** A

TEFRCF:Collaborative Filtering Personalized Recommendation Algorithm Based on Tag Entropy Feature Representation

HE Ming YANG Peng YAO Kai-sheng ZHANG Jiu-ling
(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract Tags are served as an effective way for information classification and information retrieval at the age of Web2.0. Tag recommendation systems aim to provide personalized recommendation for users by using tag data. The existing tag-based recommendation methods tend to assign the popular tags and their corresponding items more larger weight in predicting users' interest on the items, resulting in weight deviations, reducing the novelty of the results and being unable to fully reflect users' personalized interest. In order to solve the problems above, the concept of tag entropy was defined to measure the uncertainty of tags, and the collaborative filtering personalized recommendation algorithm based on tags entropy feature representation was proposed. This method solves the problem of weight deviation by introducing tag entropy, and then the tripartite graphs are used to describe the relationship among users, tags and items. The representation of users and items is constructed based on tag entropy feature representation, and the similarity of items is calculated by the feature similarity measure method. Finally, the user preferences for items are predicted by the linear combination of tags behaviors and similarity of items, and then the recommended list is generated according to the rank of preferences. The experimental results on Last.fm show that the proposed algorithm can improve recommendation accuracy and novelty, and satisfy the requirement for users.

Keywords Collaborative filtering, Tag, Entropy, Recommendation systems

1 引言

Web 2.0 作为互联网时代的标志性技术,不断推动着互联网技术的变革。2017 年 6 月在 Code 大会上发布的互联网趋势报告"Internet Trends"调查显示,全球互联网用户数已达到 34 亿,用户生成内容 UGC 使得互联网上的信息量爆增。在大数据时代,用户的个性化需求不断提高,面对海量数据信息,如何帮助用户有效获取满足其自身需要的信息,以及有力解决"信息过载"问题是数据科研工作者所面临的主要挑战之一。推荐系统(Recommender Systems, RS)[1-2]作为有效缓解该问题的方法,通过分析用户的历史行为数据为用户推荐个性化的内容。在信息智能时代,推荐系统已经成为互联网以及数据服务公司的核心技术模块之一,对于推进推荐系统技术的发展进程具有重要应用意义。

协同过滤(Collaborative Filtering, CF)[3-4]是推荐系统中应用得最为广泛的推荐技术之一,其基本思想是基于用户(Users)对项目(Items)的评分或其他行为模式来为目标用户提供个性化的推荐,且不需要项目的显式特征表示。然而,在

本文受国家自然科学基金项目(91646201,91546111),北京市教委科研计划一般项目(KM201710005023)资助。

何 明(1975-),男,博士,副教授,主要研究方向为推荐系统、数据挖掘、机器学习,E-mail:heming@bjut.edu.cn;**杨 芃**(1994-),男,硕士生,主要研究方向为推荐系统、机器学习;**要凯升**(1994-),男,硕士生,主要研究方向为推荐系统、数据挖掘;**张久伶**(1990-),男,硕士生,主要研究方向为推荐系统、迁移学习。

海量个性化需求的驱动下，协同过滤技术仍然面临一些挑战：1)冷启动[5-6]问题。当新用户或新项目出现时，因缺乏它们的偏好信息而无法生成推荐。2)数据稀疏性问题。当评分数据比较稀疏时，根据传统计算方法很难找到相似用户，导致推荐质量下降。上述问题的主要原因是数据不够充分，为了提供有效的推荐，还需要更多合适且容易获取的数据来丰富用户或项目的特征表示方式。

标签(Tags)作为 Web 2.0 时代在社会化网络中的重要应用，体现了用户对资源的理解，既表达了信息资源的主要特征，又涵盖了用户与资源之间以及用户与用户之间的关系，兼具内容与关联的特征[7]，在 Web2.0 时代其可以更好地实现对各类信息的归类以及处理。Delicious 共享书签、Flickr 共享照片、CiteULike 共享学术文献等都应用了社会标签来描述和共享资源。将标签作为推荐技术的数据来源，充分利用自发标签直接反映用户兴趣和资源内容的特点，便有可能开发出同时具备内容过滤和协同过滤优越性的推荐技术，提高推荐系统的准确性和交互性[8]。Zhang 等[9-11]介绍了基于用户-项目-标签三元组的个性化推荐算法，该算法将标签信息作为一个重要特征应用到推荐算法中，并以三分图的形式来描述用户、项目与标签三者之间的关系，最终实现个性化推荐。Jomsri 等[12]提出了基于标签的研究论文推荐系统的架构，利用标签集来表达用户的偏好，并应用此偏好为用户推荐适合的研究论文；实验结果表明，用户自定义的标签能用于表达每个个体用户的偏好，提高了推荐准确度。蔡强等[13]提出了一种基于标签和协同过滤的推荐算法，以满足用户个性化资源推荐。李慧等[14]综合了用户的资源标签与标签概率模型，提出了一种个性化标签推荐方法。叶剑虹等[15]提出了一种基于混合模式的流媒体缓存调度算法，可用张量表示用户、标签、项目等数据，同时进行高阶奇异分解，在有效缓解数据稀疏性的同时提高了推荐质量。Kideok 等[16]基于标签信息计算用户的兴趣相似度，使用随机游走算法求得 Top-N 推荐。

由此可见，基于标签的推荐系统通过使用用户标注信息，更好地针对资源以及用户自身的特征信息为用户做出个性化推荐，提高推荐质量。然而，现有的标签推荐系统中的推荐方法在预测用户对物品的兴趣度时往往倾向于给热门标签及其对应的热门物品较大的权重，从而降低了推荐结果的新颖性，未能充分反映用户个性化的兴趣。例如，文献[13]提出的基于标签和协同过滤推荐的算法是通过建立用户的标签向量从而对用户兴趣建模，其中每个标签的权重用该标签被用户所使用频率的形式表示。这种建模方式的缺点是在推荐过程中往往倾向于给热门标签对应的项目较大的权重，导致权重偏差，从而降低了推荐结果的新颖性。目前，在基于标签的推荐过程中针对权重偏差问题的研究还较少。针对上述问题，本文借鉴 TF-IDF 思想，提出基于标签信息熵表示的协同过滤个性化推荐算法来提高个性化推荐的准确性和新颖性。本文的主要贡献和创新性如下：

(1)首次提出了标签熵的概念来度量标签的不确定性，描述了基于标签的推荐系统中标签的概率分布。

(2)基于用户-标签和标签-项目特征表示，提出了基于标签熵特征表示的协同过滤个性化推荐算法。

(3)通过仿真实验验证了方法的有效性。实验结果表明，标签熵特征表示的协同过滤推荐算法可以提高推荐的准确性和新颖性。

2 基于标签信息熵的用户偏好和项目相似性计算

2.1 标签信息熵

TF-IDF 是一种用于信息检索与文本相似性度量的方法，其核心思想是通过词频(Term Frequency)和逆向文件频率(Inverse Document Frequency)来求得某一词语对主题的预测能力。本文借鉴了 TF-IDF 思想，将其用于基于标签的协同过滤推荐。

基于 Delicious 数据集 2007 年某一随机月份的数据对标签使用频率及其热度间的关系进行了研究与分析，通过标签热度计算(见算法 1)发现标签热度呈长尾分布，如图 1 所示。Delicious 数据集由德国研究人员公布，其涵盖了自 2003 年 9 月到 2007 年 12 月间该公司用户的 4.2 亿条标签标记行为记录。

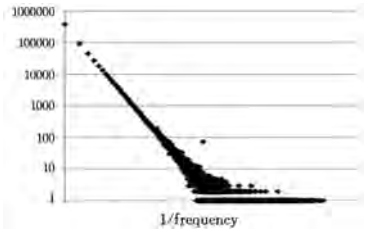

图 1 标签热度的长尾分布

算法 1 标签热度计算

```
Input:Delicious tag data set
Define TData=Delicious tag data set
1. for user,tag,item in TData do
2. if tag not in TData
3.     Pop[tag]=1
4. else
5.     Pop[tag]+=1
6.  end
7. end
```

图 1 中纵坐标是热度 p，横坐标是数据集中相应标签热度所对应的标签使用频率的倒数。此外，其双对数曲线如式(1)、式(2)所示，近似是一条直线。

$$\log^{f(p)}=x\log^{p}+y=\log^{f^{x}\cdot e^{y}} \tag{1}$$

$$f(p)=e^{y}\cdot p^{x}=\gamma\cdot p^{x} \tag{2}$$

其中，f 为标签热度 p 所对应的标签总数。

因此，TF-IDF 思想并不完全适用于标签推荐过程中，可能导致权重偏差，即在推荐过程中往往倾向于给热门标签对应的项目赋予较大的权重，降低了推荐结果的新颖性，无法充分反映用户个性化的兴趣。

例如，“大米”和“小米”虽是两个不同的标签，但是因为都被用来标记类似的主食，所以它们应具有较高的语义相似性。而“使用频率决定确定权重”所导致的权重偏差问题显然无法有效识别这些项目间的相似性，会造成大量重要信息的丢失。因此，针对这一问题，本文将标签看作一个随机变量，用标签熵的概念来度量标签的不确定性。

定义 1 标签熵(Tag Entropy)定义为：

$$H(t_i)=p(t_i)\cdot\log_2^{\left(1+\frac{1}{p(t_i)}\right)} \tag{3}$$

$$H'(t_i)=q(t_i)\cdot\log_2^{\left(1+\frac{1}{q(t_i)}\right)} \tag{4}$$

其中，$p(t_i)$表示标签 t_i 被用户所使用的概率，$q(t_i)$表示标签 t_i 标记项目的概率。

在标签推荐系统中，标签信息熵可表示为：

$$\begin{pmatrix} T \\ H(T) \end{pmatrix}=\begin{pmatrix} t_1 & t_2 & \cdots & t_i & t_n \\ H(t_1) & H(t2) & \cdots & H(t_i) & H(t_n) \end{pmatrix}$$

$H(T)$用于刻画标签的概率分布，用于标签“不确定性”度量。一个标签被用户用来标记项目的概率越大，其对应熵值越低，表示此标签用于标记该项目的“确定性”越大；反之，熵值越大，表示此标签用于标记该项目的“确定性”越小。

本文首先求得标签推荐系统中各个标签的标签熵，然后根据标签熵的特征及用户使用标签、标签标记项目的信息惩罚热门标签，从而提升推荐准确性和新颖性。

2.2　基于标签熵的用户和项目的特征表示

在基于标签熵特征表示的协同过滤个性化推荐中，通常包含 3 种类型的对象：使用标签标记项目的用户、标签本身以及被标签标记的项目。因此，数据表示可以描述为一个三元组：

$T:=(P,Q,L)$

其中，$P=\{p_1,p_2,\cdots,p_i,\cdots,p_x\}$为用户集合，$X$ 为用户总数，$i=1,2,3,\cdots,X$；$Q=\{q_1,q_2,\cdots,q_i,\cdots,q_y\}$为所有项目集合，$Y$ 为项目总数，$i=1,2,3,\cdots,Y$；$L=\{l_1,l_2,\cdots,l_i,\cdots,l_z\}$为用户使用的标签集合，$L$ 为标签总数，$i=1,2,3,\cdots,Z$。

用户-标签-项目之间的关系可以用三分图来描述，如图 2 所示。其中，P,Q,L 3 种类型的节点之间的边用以表示标注关系。

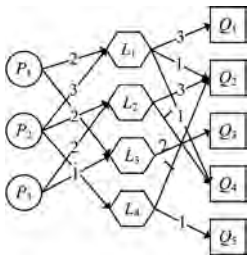

图 2　用户-标签-项目信息

图 2 中，P_i 表示用户，L_j 表示标签，Q_k 表示项目。如果用户 P_i 使用了标签 L_j，则在用户 P_i 和标签 L_j 之间连接一条连线，线上的数字表示用户 P_i 使用标签 L_j 的次数；如果标签 L_j 标记过项目 Q_k，则在标签 L_j 和项目 Q_k 之间连接一条连线，线上的数字表示项目 Q_k 被标签 L_j 标记的次数。

2.2.1　用户的标签熵特征表示

根据用户标签行为建立标签熵特征表示的用户兴趣特征模型，如式(5)所示：

$$\overrightarrow{F_{ul}}=(|H(l_1)-H(\bar{l})|\cdot\frac{T_{ul_1}}{\log^{(1+T_1^u)}}\cdot\log^{\frac{X}{T_{l_1u}}},\cdots,|H(l_j)-H(\bar{l})|\cdot\frac{T_{ul_j}}{\log^{(1+T_j^u)}}\cdot\log^{\frac{X}{T_{l_ju}}},\cdots,|H(l_Z)-H(\bar{l})|\cdot\frac{T_{ul_z}}{\log^{(1+T_z^u)}}\cdot\log^{\frac{X}{T_{l_zu}}}) \tag{5}$$

其中，T_{ul_j} 表示用户 u 使用标签的次数，T_j^u 表示标签 j 被使用过的用户数，X 表示用户总数，T_{l_ju} 表示使用标签 l_j 的用户数；$|H(l_j)-H(\bar{l})|\cdot\frac{T_{ul_j}}{\log^{(1+T_j^u)}}$表示对热门标签削弱后的标签权重，$\log^{\frac{X}{T_{l_ju}}}$ 表示在用户 u 所使用过的所有标签中标签 l_j 的重要程度，$|H(l_j)-H(\bar{l})|\cdot\frac{T_{ul_j}}{\log^{(1+T_j^u)}}\cdot\log^{\frac{X}{T_{l_ju}}}$ 表示标签 l_j 对于用户 u 的重要度。

2.2.2　项目的标签熵特征表示

根据标记项目的标签信息建立标签熵特征表示的项目特征模型，如式(6)所示：

$$\overrightarrow{F_{wl}}=(|H'(l_1)-H'(\bar{l})|\cdot\frac{T_{wl_1}}{\log^{(1+I_1^w)}}\cdot\log^{\frac{Y}{T_{l_12}}},\cdots,|H'(l_j)-H'(\bar{l})|\cdot\frac{T_{wl_j}}{\log^{(1+I_j^w)}}\cdot\log^{\frac{Y}{T_{l_jw}}},\cdots,|H'(l_Y)-H'(\bar{l})|\cdot\frac{T_{wl_z}}{\log^{(1+I_z^w)}}\cdot\log^{\frac{Y}{T_{l_zw}}}) \tag{6}$$

其中，T_{wl_j} 表示项目 w 被标记的标签总数，I_j^u 表示项目 j 被用户打过标签的用户数目，Y 表示项目总量，T_{l_jw} 表示被标签 l_j 标记的项目总数；$|H'(l_j)-H'(\bar{l})|\cdot\frac{T_{wl_j}}{\log^{(1+I_j^w)}}$表示对热门项目削弱后的项目权重，$\log^{\frac{Y}{T_{l_jw}}}$ 表示在项目 w 所使用过的所有标签中标签 l_j 的重要程度，$|H'(l_j)-H'(\bar{l})|\cdot\frac{T_{wl_j}}{\log^{(1+I_j^w)}}\cdot\log^{\frac{Y}{T_{l_jw}}}$ 表示标签 l_j 对于项目 w 的重要度。

2.3　基于标签熵特征表示的用户偏好计算

在确定用户和项目的标签熵特征表示之后，将二者相结合来预测用户 u_i 对项目 w_j 的偏好。本文在计算过程中引入了标签熵的思想，对热门标签、项目进行惩罚，在解决“权重偏差”问题的同时提高了推荐结果的准确性和新颖性，如式(7)所示：

$$F_{u_iw_j}=\overrightarrow{F_{u_il}}\cdot\overrightarrow{F_{lw_j}}=\sum_{l=1}^{Z}\overrightarrow{F_{u_il}}\times\overrightarrow{F_{lw_j}} \tag{7}$$

其中，$F_{u_iw_j}$ 表示 u_i 对项目 w_j 的预测偏好，$u_i\in P$，$i=1,2,3,\cdots,X$，$w_j\in Q$，$j=1,2,3,\cdots,Y$。

用户 u_i 的项目偏好特征模型如式(8)所示：

$$\overrightarrow{F_{u_i}}=(F_{u_iw_1},F_{u_iw_1},F_{u_iw_1},\cdots,F_{u_iw_1}) \tag{8}$$

基于用户对项目的偏好程度构建用户-项目偏好矩阵：

$$F_{U_iW_j}=\begin{bmatrix} F_{U_1W_1} & \cdots & F_{U_1W_j} & \cdots & F_{U_1W_Y} \\ \vdots & \ddots & & & \vdots \\ F_{U_iW_1} & & F_{U_iW_j} & & F_{U_iW_Y} \\ \vdots & & & \ddots & \vdots \\ F_{U_XW_1} & \cdots & F_{U_XW_j} & \cdots & F_{U_XW_Y} \end{bmatrix} \tag{9}$$

其中，$i\in[1,X]$，$j\in[1,Y]$，该矩阵可以反映各个用户对各个项目的偏好程度。

2.4　基于标签特征的项目相似性计算

项目相似性表示两个不同项目之间的相似程度。传统协同过滤推荐方法的相似性计算通常基于用户-项目二维评分矩阵：

$$User_i\begin{bmatrix} R_{1,1} & \cdots & R_{1,j} & \cdots & R_{1,Y} \\ \vdots & \ddots & & & \vdots \\ R_{i,1} & & R_{i,j} & & R_{i,Y} \\ \vdots & & & \ddots & \vdots \\ R_{X,1} & \cdots & R_{X,j} & \cdots & R_{X,Y} \end{bmatrix}\quad Item_j \tag{10}$$

其中，$R_{i,j}$ 表示用户 i 对项目 j 的评分值，$i\in[1,X]$，$j\in[1,Y]$。

考虑到在基于标签的协同过滤推荐算法中用于描述用户-项目-标签之间的关系是三元的，而传统的协同过滤推荐算法无法直接应用，因此本文通过将三元关系映射到低维二维空间，利用用户-标签和标签-项目二维映射矩阵计算项目的相似度，这些映射保留了项目信息。用户-标签-项目三元关系的映射如图 3 所示。

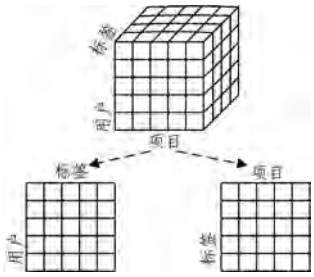

图 3 用户-标签-项目三元关系映射

用户-标签映射矩阵和标签-项目映射矩阵如式(11)和式(12)所示：

$$User_i\begin{bmatrix} & & Tag_k & & \\ G_{1,1} & \cdots & G_{1,k} & \cdots & G_{1,Z} \\ \vdots & \ddots & & & \vdots \\ G_{i,1} & & G_{i,k} & & G_{i,Z} \\ \vdots & & & \ddots & \vdots \\ G_{X,1} & \cdots & G_{X,k} & \cdots & G_{X,Z} \end{bmatrix} \tag{11}$$

其中，$G_{i,k}$ 表示用户 i 使用标签 k 的次数，$i\in[1,X]$，$k\in[1,Z]$。

$$Tag_k\begin{bmatrix} & & Item_j & & \\ B_{1,1} & \cdots & B_{1,j} & \cdots & B_{1,Y} \\ \vdots & \ddots & & & \vdots \\ B_{k,1} & & B_{k,j} & & B_{k,Y} \\ \vdots & & & \ddots & \vdots \\ B_{Z,1} & \cdots & B_{Z,j} & \cdots & B_{Z,Y} \end{bmatrix} \tag{12}$$

其中，$B_{k,j}$ 表示标签 k 标记项目 j 的次数，$k\in[1,Z]$，$j\in[1,Y]$。

项目的特征信息可以用标签特征表示的项目特征向量表示：

$$\vec{W}_Y=(w_{i1},w_{i2},w_{i3},\cdots,w_{iZ})$$

其中，w_{ij} 表示标签 l_j 标记项目 q_i 的次数，$i=1,2,3,\cdots,Y$。

所有项目的特征信息可以通过项目特征矩阵表示：

$$W_{i,j}=\begin{bmatrix} W_{1,1} & \cdots & W_{1,j} & \cdots & W_{1,Y} \\ \vdots & \ddots & & & \vdots \\ W_{i,1} & & W_{i,j} & & W_{i,Y} \\ \vdots & & & \ddots & \vdots \\ W_{X,1} & \cdots & W_{X,j} & \cdots & W_{X,Y} \end{bmatrix} \tag{13}$$

其中，$i\in[1,X]$，$j\in[1,Y]$。

通过项目的特征表示计算不同项目间的相似度：

$$sim(w_i,w_j)=\frac{\sum\limits_{k\in Y_{i,j}}(w_{i,k}-\overline{w_i})\times(w_{j,k}-\overline{w_j})}{\sqrt{\sum\limits_{k\in Y_{i,j}}(w_{i,k}-\overline{w_i})^2}\times\sqrt{\sum\limits_{k\in Y_{i,j}}(w_{j,k}-\overline{w_j})^2}} \tag{14}$$

其中，$Y_{i,j}$ 表示项目 w_i 和项目 w_j 共同使用过的标签集合，$\overline{w_i}$ 和 $\overline{w_j}$ 分别表示项目 w_i 和 w_j 被标签标记的平均次数。

基于项目间的相似性计算，构建项目相似度矩阵：

$$Sim_{w_iw_j}=\begin{bmatrix} 1 & \cdots & Sim_{w_1w_j} & \cdots & Sim_{w_1w_Y} \\ \vdots & \ddots & & & \vdots \\ Sim_{w_iw_1} & & 1 & & Sim_{w_iw_Y} \\ \vdots & & & \ddots & \vdots \\ Sim_{w_Yw_1} & \cdots & Sim_{w_Yw_j} & \cdots & 1 \end{bmatrix} \tag{15}$$

其中，$i,j\in[1,Y]$。

2.5 基于标签特征的用户偏好预测

本文采用 k-近邻方法对用户-项目偏好进行预测，即选择与目标用户 u 最相似的 k 个用户作为最近邻集合来进行计算。设项目集合 $w_u:\{i_1,i_2,i_3,\cdots,i_j\}$ 为用户 u 已标记的项目，则对任意未标记项目 $w_i\notin w_u$ 的预测偏好为：

$$P_{rate(u,w_i)}=\sum_{j=1}^{X}F_{u_iw_j}\times Sim_{w_jw_k} \tag{16}$$

其中，$F_{u_iw_j}$ 表示用户 u_i 对项目 w_z 的喜好程度，$Sim_{w_jw_k}$ 表示不同项目 w_j 和 w_k 之间的相似性。

3 基于标签熵特征表示的协同过滤推荐算法——TEFRCF

3.1 算法描述

本文提出的 TEFRCF 算法主要分为 5 个阶段：

(1)读取数据集 rec_{ij}；

(2)对数据集进行降维处理，将三元关系映射到低维二维空间，得到 $user\_tag$ 和 tag_tem 矩阵；

(3)由式(5)、式(6)求得用户及项目的基于标签熵的特征表示，依据式(7)构建用户-项目偏好模型；

(4)由式(15)可计算出不同项目间的相似度，得到项目相似度矩阵 Sim_{yy}，以此可计算出目标项目的最近邻集合；

(5)根据式(16)求出用户 u 对未评分项目集中各个项目的预测评分。将评分结果按从大到小的顺序排列，取前 N 项作为 Top-N 推荐集。

TEFRCF 算法的具体流程如图 4 所示。

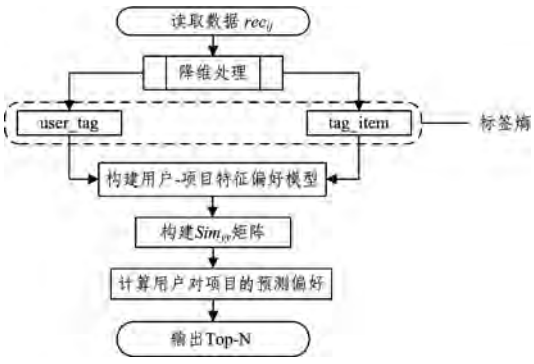

图 4 TEFRCF 算法的流程

TEFRCF 算法如算法 2 所示。

算法 2 TEFRCF 算法

Input:Source data set (user,tag,item)

Output:Top-N Recommended Collections

1. Data set dimensionality reduction:

D=(P,Q,L)→(U−T)、(T−I)

2. Get F_{ul} according to Formula 5
3. Get F_{wl} according to Formula 6
4. for i∈[1,X] do
5. for j∈[1,Y] do
6. for l∈[1,Z] do
7. $F_{u_i w_j} += F_{u_i l} \times F_{l w_j}$
8. end
9. end
10. end
11. Build Matrix $F_{u_i w_x}$
12. for i∈1,Y] do
13. for j∈[1,Y] do
14. compute sim(w_i,w_j)→Formula 14
15. end
16. end
17. Build Matrix Sim_{YY}
18. for i∈[1,X] do
19. for k∈[1,Y] do
20. for j∈[1,Y] do
21. compute P_rate(u,w_i)→Formula 16
22. end
23. end
24. end
25. Sort P_rate by reverse sorting
26. Build Top-N Recommended Collections

3.2　算法复杂度分析

TEFRCF 算法的复杂度分析：假设有 p 个用户，l 个标签，q 个项目，数据规模为 d。TEFRCF 算法首先遍历了所有数据并对其进行了降维处理，计算用户-标签标注行为及标签-项目标注信息，时间复杂度为 $O(n^3)$；其次，计算用户对于项目的偏好以及项目相似性，进而构建用户-项目偏好矩阵和项目相似性矩阵，时间复杂度为 $O(n^2)$；然后，依据上述行为数据预测用户对项目的偏好，时间复杂度为 $O(n^2)$；最后，对推荐项目进行排序，寻找偏好最大的前 N 个项目，时间复杂度为 $O(n\log n)$。因此，综合分析 TEFRCF 算法的整个过程可知，为用户生成推荐结果的时间复杂度为 $O(n)$，总的时间复杂度为 $O(n^3)$。

4　实验及分析

本节通过几组实验来对 TEFRCF 方法及现有一些推荐方法进行比较，并对实验结果进行分析。

4.1　数据集

本文使用 Last.fm 数据集对算法进行测试。Last.fm 是世界上最大的社会音乐[17]平台之一，也是全球最著名的以 Web 2.0 为标签的社会化网络之一。用户在该平台可以用标签随意标记歌手和歌曲。该数据集共有 186479 条标记记录，其中包含 17632 条资源、1946 个标签、1892 位用户。

实验采取 10 折交叉验证的方法，按照各个标签被使用过 10 次以上并且每个用户至少对音乐或歌手标记过 10 次的原则对原始数据进行预处理。数据在预处理后被随机分为训练集和测试集两个部分，其中 80%作为训练数据，剩余 20%作为测试数据。表 1 列出了对 Last.fm 数据集信息的描述。

表 1　数据集信息描述

	标记记录	资源数	标签数	用户数
原始数据集	186479	17632	1946	1892
预处理后的数据集	168437	11982	1493	1260

4.2　方法比较

为了证明 TEFRCF 方法的有效性，本文采用 3 种方法作为基线系统进行对比、评估与分析。

(1)IBTCF[13]：该方法基于项目被标签标记的信息建立标签特征模型，用于基于标签的协同过滤推荐。

(2)UBTCF：该方法基于用户对项目的标记行为建立标签特征模型，用于基于标签的协同过滤推荐。

(3)TEFRCF：本文提出的方法。首先求得标签推荐系统中各个标签的标签熵特征表示，然后根据项目标记信息建立标签熵特征表示模型，预测用户对未使用标签的偏好，最后推荐生成 Top-N 标签。

(4)Benchmark Method(基准方法)：对标签频率进行惩罚是消除流行标签对推荐结果的影响的最简单方法，因此将此方法作为基准方法进行对比。

4.3　评估标准

本文选用如下评估标准：准确率(Precision)、召回率(Recall)、F-Measure 和新颖性(Novelty)。其中，准确率表示在推荐标签集中所有相关标签所占的比例，主要用于衡量推荐系统的查准率；召回率表示推荐结果中的相关标签占所有相关标签的比例，主要用于衡量推荐系统的查全率；新颖性指的是推荐项目的热门程度，其中越不热门的产品越能让用户觉得新颖。

设 $Tr(d)$ 是根据用户在训练集中的行为数据求得的 Top-N 推荐列表；$Te(d)$ 是根据用户在测试集中的行为数据求得的 Top-N 推荐列表。

准确率(Precision)的计算公式如下：

$$Precision=\frac{\sum|Tr(d)\bigcap Te(d)|}{\sum|Tr(d)|} \tag{17}$$

召回率(Recall)的计算式如下：

$$Recall=\frac{\sum|Tr(d)\bigcap Te(d)|}{\sum|Te(d)|} \tag{18}$$

F-Measure(又称为 F-Score)融合了召回率和准确率，是 Precision 和 Recall 的加权调和平均，如式(19)所示：

$$F=\frac{(\alpha^2+1)\times Precision\times Recall}{\alpha^2\times Precision+Recall} \tag{19}$$

其中，α 是参数，当 $\alpha=1$ 时，即为 F1-Measure：

$$F_1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{20}$$

计算新颖性(Novelty)的最常用方法是求得推荐列表中物品的平均热度，如式(21)所示：

$$Novelty(rec_n)=\frac{\sum\limits_{i\in rec_n}p(i)}{n} \tag{21}$$

其中，rec_n 为 Top-N 推荐集，$p(i)$ 为物品 i 的热度。$Novelty$ 值越小说明推荐集合中平均热度越低，推荐结果的新颖性越好。

4.4 实验结果与分析

为了验证和比较本文所提方法的有效性，我们进行了以下几组实验：

实验 1 在生成 Top-N 推荐的过程中，N 的取值对于推荐结果具有较大的影响。因此，本组实验旨在比较本文提出的 TEFRCF 算法在不同 N 值下 Precision，Recall 和 F-measure 的变化。实验中，将 N 分别取值 5，10，15，20，25 和 30 来计算推荐结果。实验结果如图 5 所示。

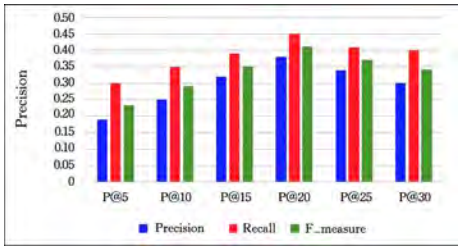

图 5 不同 N 值下 TEFRCF 算法的推荐性能

由图 5 可以看出，随着 N 值的不断增加，算法的 Precision，Recall 和 F-measure 呈递增趋势，当 N 取值为 20 时各项评价指标达到峰值。随后，推荐质量随着 N 值的增大而降低，因此，当 N 的数量为 20 时，推荐效果最好。

实验 2 在生成 Top-N 推荐过程中，*Item* 的取值对于推荐结果同样具有较大的影响。因此本组实验旨在比较文献[13]提出的 IBTCF 方法、基准方法、UBTCF 方法及本文提出的 TEFRCF 方法在不同 Item 下 Precision，Recall，F-measure 的变化。实验将 Item 的数目分别选取为 300，500，800 和 1300 等来计算推荐结果。实验结果如图 6—图 8 所示。

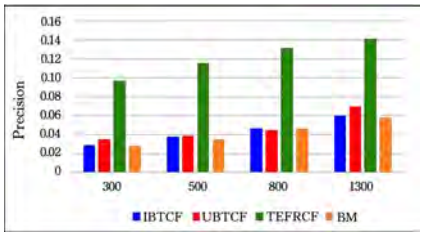

图 6 不同 Item 数目下 Precision 的比较结果

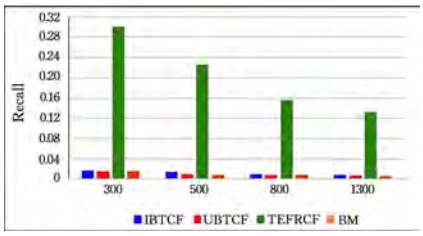

图 7 不同 Item 数目下 Recall 的比较结果

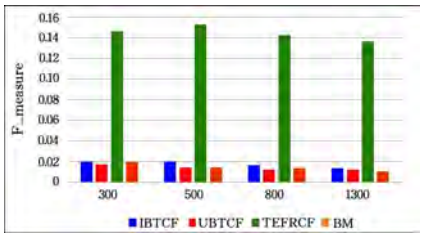

图 8 不同 Item 数目下 F-measure 的比较结果

由图 6 可以看出，Item 数目越多，推荐结果的精确度就越高。实验结果表明，本文提出的推荐算法在 Precision，Recall 和 F-measure 上明显优于 IBTCF、UBTCF 和基准方法，从而验证了本文提出的 TEFRCF 算法可以有效地提高推荐质量。

实验 3 推荐结果的新颖性是评价推荐质量的一个重要指标。因此，本组实验主要比较 TEFRCF 方法、基准方法、UBTCF 方法以及文献[13]提出的 IBTCF 方法在不同 Item 值下推荐结果的新颖性。其中 Item 的取值范围为 300，500，800 和 1300。实验结果如图 9 所示。

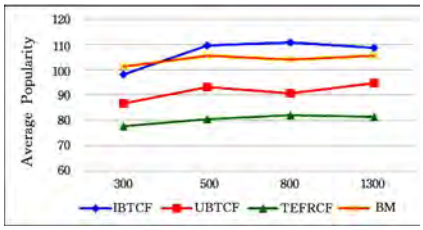

图 9 不同 *Item* 数目下新颖性的比较结果

由图 9 可以看出，TEFRCF 方法推荐项目的平均流行度更低，表示该方法推荐结果的新颖性更高。实验结果表明本文提出的推荐算法在推荐结果新颖性上优于 IBTCF、UBTCF 和基准方法，从而验证了本文提出的 TEFRCF 算法能有效地解决权重偏差现象，提高了推荐结果的新颖性。

实验 4 本次实验对数据集进行了交叉验证，旨在进一步验证本文提出的推荐算法的稳定性。首先将数据集等量地随机分为 10 份，其中 9 份作为训练集，1 份作为测试集；一次实验完成后，在未做过测试集的集合中任意选出 1 份作为新的测试集，将剩下 9 份作为训练集，重复 9 次并观察实验结果中准确率、召回率及 F-measure 的变化。实验结果如图 10 所示。

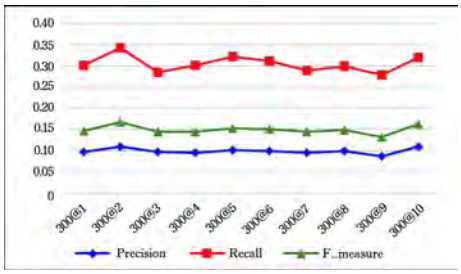

图 10 交叉验证结果

由实验结果可得，随着数据集的随机变化，Precision，Recall 和 F-measure 各项指标并无明显变化，从而证明了本文提出的 TEFRCF 算法具有较好的稳定性。

结束语 标签作为 Web 2.0 时代在社会化网络中的重要应用，体现了用户对资源的理解，既表达了信息资源的主要特征，又涵盖了用户与资源之间以及用户与用户之间的关系。本文根据可自由标注的社会化标签并结合信息熵和协同过滤思想，提出了一种面向用户个体的个性化需求推荐算法。该方法首次融入了标签熵的概念，有效解决了权重偏差问题，使得推荐结果在新颖性和准确率方面得到提升。通过大量仿真实验表明，本文提出的推荐算法在各项性能指标上均优于结合标签和协同过滤的推荐算法，同时也具备了较好的可拓展性。在将来的研究工作中，我们将进一步研究数据增量、用户推荐实时性及时效性等问题。

（下转第 486 页）

参 考 文 献

[1] 王丽珍,周丽华,陈红梅,等.数据仓库与数据挖掘原理及应用(第2版)[M].北京:科学出版社,2009:1-19.

[2] HAN J,KAMBER M,PEI J. Data mining:concept and techniques(Third Edition)[M]. Beijing:China Machine Press,2006:1-23.

[3] HUANG Y,SHEKHAR S,XIONG H. Discovering Co-location Patterns from Spatial Data Sets:A General Approach[C]// IEEE Transactions on knowledge and Data Enginnering (TKDE). 2004:1472-1485.

[4] YOO J S,SHEKHAR S. A partial Join Approach for Mining Co-location Patterns[C]// Proc. of the 12th Annual ACM Int. Workshop on Geographic Information Systems. Washington DC,USA,2004:241-249.

[5] YOO J S,SHEKHAR S,CELIK M. A join-less approach for co-location pattern mining:A summary of results[C]// Proc. of the 5th IEEE Int. Conf. on Data Mining. Washington:IEEE Computer Society,2005:813-816.

[6] WANG L Z,BAO Y Z,LU J,et al. A new join-less approach for co-location pattern mining[C]// IEEE International Conference on Computer and Information Technology (CIT2008). Washington,2008:197-202.

[7] WANG L Z,BAO Y Z,LU Z Y. Efficient discovery of spatial co-location patterns using the iCPI-tree[J]. The Open Information Systems Journal,2009,3(1):69-80.

[8] WANG L Z,ZHOU L H,LU J,et al. An order-clique-based approach for mining maximal co-locations[J]. Information Sciences,2009,179(19):3370-3382.

[9] 欧阳志平,王丽珍,陈红梅.模糊对象的空间 co-location 模式挖掘研究[J].计算机学报,2011,34(10):1947-1955.

[10] 姚华传,王丽珍,陈红梅,等.面向海量数据的空间 co-location 模式挖掘新算法[J].计算机科学与探索,2015,9(1):24-35.

[11] 吴萍萍,王丽珍,周永恒.带模糊属性的空间 co-location 模式挖掘研究[J].计算机科学与探索,2013,7(4):348-358.

[12] 芦俊丽,王丽珍,肖清,等.空间 co-location 模式增量挖掘及演化分析[J].软件学报,2014,12(25):190-199.

[13] 曾新,杨健.带时间约束的 co-location 模式挖掘[J].计算机科学,2016,43(2):293-296.

[14] 杨世晟,王丽珍,芦俊丽,等.空间高效用 co-location 模式挖掘技术初探[J].小型微型计算机系统,2014,35(10):2302-2307.

[15] 江万国,王丽珍,方圆,等.领域驱动的高效用 co-location 模式挖掘方法[J].计算机应用,2017,37(2):322-328.

[16] HAN J W,KAMBER M,PEI J.数据挖掘概念与技术(第3版)[M].北京:机械工业出版社,2014:74-76.

(上接第 470 页)

参 考 文 献

[1] ADOMAVICIUS G,TUZHILIN A. Toward the next generation of recommender systems:A survey of the state-of-the-art and possible extensions[C]// Proceedings of the IEEE Transactions Knowledge and Data Engineering. 2005:734-749.

[2] LÜ L,MEDO M,YEUNG C H,et al. Recommender systems [J]. Physics Reports,2012,519(1):1-49.

[3] SU X,KHOSHGOFTAAR T M. A survey of collaborative filtering techniques [J]. Advances in Artificial Intelligence,2009,2009(12):4.

[4] WEI C,HSU W,LEE M L. A unified framework for recommendationsbased on quaternary semantic analysis[C]// Proceedings of the 34th International ACM SIGIR Conference on Research and Development InInformation Retrieval. Beijing,China,2011:1023-1032.

[5] WANG L C,MENG X W,ZHANG Y J. Context-Aware recommender systems:A survey of the state-of-the-art and possible extensions[J]. Journal of Software,2012,23(1):1-20.

[6] LIN J,SUGIYAMA K,KAN M Y,et al. Addressing cold-start in apprecommendation:latent user models constructed from twitterfollowers[C]// Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval. Dublin,Ireland,2013:283-292.

[7] MISTRY O,SEN S. Tag recommendation for social book marking:Probabilistic approaches [J]. Multiagent and Grid Systems,2012,8(2):143-163.

[8] 于洪,李俊华.一种解决新项目冷启动问题的推荐算法[J].软件学报,2015,26(6):1395-1408.

[9] ZHANG Z K,ZHOU T,ZHANG Y C. Personalized recommendation via integrated diffusion on user-item-tag tripartite graphs [J]. Physica A:Statistical Mechanics and its Applications,2010,389(1):179-186.

[10] ZHANG Z K,LIU C,ZHANG Y C,et al. Solving the cold-start problem in recommender systems with social tags [J]. EPL (Europhysics Letters),2010,92(2):28002.

[11] ZHANG Z K,ZHOU T,ZHANG Y C. Tag-Aware recommender systems:A state-of-the-art survey [J]. Journal of Computer Science and Technology,2011,26(5):767-777.

[12] JOMSRI P,SANGUANSINTUKUL S,CHOOCHAIWATTANA W. A framework for tag-based research paper recommender system:An IR approach[C]// Proceedings of the 2010 IEEE 24th Int'l Conf. on Advanced Information Networking and Applications Workshops. 2010:103-108.

[13] 蔡强,韩东梅,李海生,等.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):69-71,110.

[14] 李慧,马小平,胡云,等.融合主题与语言模型的个性化标签推荐方法研究[J].计算机科学,2015,42(8):70-74.

[15] 叶剑虹,叶双.基于混合模式的流媒体缓存调度算法[J].计算机科学,2013,40(2):61-64.

[16] KIDEOK C,HAKYUNG J,et al. How can an ISP merge with a CDN? [J]. IEEE Communications,2011,49(10):156-162.

[17] 李瑞敏,林鸿飞,闫俊.基于用户-标签-项目语义挖掘的个性化音乐推荐[J].计算机研究与发展,2014(10):2270-2276.