基于动态社会行为和用户背景的协同推荐方法

蒋 胜 王忠群 修 宇 皇苏斌 汪千松
(安徽工程大学计算机与信息学院 芜湖 241000)

摘 要 针对传统协同过滤推荐算法推荐精度低及冷启动的问题,提出了一种基于动态社会行为和用户背景的协同推荐方法。作为用户标注行为的结果,变化的标签体现了用户行为的动态性。该方法首先根据动态社会化标签得出用户的动态兴趣偏好相似度,然后根据用户背景信息计算出用户相似度,最后计算基于时间权重的用户评分相似度,并集成上述3个相似度找出最近邻居集,以为目标用户提供更加准确的个性化推荐。实验结果证明,该方法不仅能较好地解决数据稀疏和冷启动的问题,还能有效提高推荐算法的精确度。

关键词 推荐精度,冷启动,社会化标签,用户背景信息,动态社会行为,时间权重
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.3.052

Collaborative Filtering Recommendation Method Based on Dynamic Social Behavior and Users' Background Information

JIANG Sheng WANG Zhong-qun XIU Yu HUANG Su-bing WANG Qian-song
(School of Computer and Information, Anhui Polytechnic University, Wuhu 241000, China)

Abstract To address the difficulty of data sparsity and lower recommendation precision in the traditional collaborative filtering recommendation algorithm, a new collaborative filtering recommendation method was presented based on dynamic social behavior and users' background information. As the result of user annotation behavior, variable social tags can reflect the changes of user social behavior. Firstly, the similarities of users' dynamic preferences are calculated based on users' social tags. Secondly, the similarities between users are calculated based on users' background information. Finally, the similarities of user rating are calculated based on time weight, and the above three similarities are integrated to get the nearest neighbor set for targeted users to provide more accurate individual recommendation. The experimental results show that the new method can not only improve the accuracy of recommendation, but also solve the problems of data sparsity and cold-start.

Keywords Recommendation precision, Cold-start, Social tags, Users' background information, Dynamic social behavior, Time weight

1 引言

电子商务使得人们淹没于海量的用户和商品信息中。如何有效利用用户信息以方便用户购买到满意的商品给人们提出了挑战。为此,推荐技术应运而生,但仍然存在一些困难,如冷启动、数据稀疏和推荐精度低等。

作为海量信息源之一,用户信息(如社会网络、社会化标签等)具有潜能价值。针对传统协同过滤推荐(Collaborative Filtering Recommend- ation, CF)算法[1]存在的上述问题,人们引入了社交网络技术。文献[2]利用了社会网络分析技术分析用户间的关系,将其量化为信任度以填充用户-项矩阵,并将信任度融入到用户相似性计算中。文献[3]通过用户社交关系提取用户相关性来提升预测精度。文献[4]提出了一种基于用户偏好自动分类的社会媒体数据共享和推荐方法。

除了上述文献使用了社交网络中社交用户间的关系外,文献[5]研究了社交行为信息(如社会化标注)对推荐质量的影响。人们的社交行为是动态的,拥有时间属性,作为用户标注行为结果的社会化标签更能体现用户的兴趣。但更为重要的是,人类的社交行为往往与其背景密切相关,相同背景的人往往具有相似的社交行为。对此,集成动态社会行为和用户背景信息,本文提出了一种基于动态社会行为和用户背景的协同推荐方法。该方法基于动态社会化标签计算用户间的相似性,然后分别计算用户背景信息的相似度和基于时间权重的用户评分的相似度,最后集成上述相似度并找出目标用户的可信最近邻居集实施推荐。

到稿日期:2014-05-16 返修日期:2014-07-12 本文受国家自然科学基金项目(71371012,71171002,61300170),教育部人文社科规划项目(13YJA630098)资助。

蒋 胜(1991—),男,硕士生,主要研究方向为电子商务;王忠群(1965—),男,教授,主要研究方向为信息管理与信息系统;修 宇(1976—),男,讲师,主要研究方向为数据挖掘与机器学习;皇苏斌(1986—),男,主要研究方向为软件工程与电子商务;汪千松(1978—),男,工程师,主要研究方向为网络信息安全、电子商务。

2 理论与方法

2.1 基于动态社会化标签的用户相似度

社会化标签即从用户的角度对社交媒体中的资源添加标注。在社会化标签系统中，标签信息和标签标注时间是两个重要因素。在不同时期，用户所使用的标签在一定程度上能够体现出用户的兴趣偏好及其变化趋势，体现其行为的动态性。

2.1.1 标签信息权重

用户在使用社会化标签来标注资源时，某标签被该用户使用的频率越高，说明该用户对此相关资源的偏好程度越强，反映标签对用户兴趣偏好的影响较大，因此相关标签应具有更高的权重。本文采用 TF-IUF 方法来计算单个标签的权重。计算公式如式(1)所示：

$$w_i = tf \times iuf = f_i \times \lg \frac{N}{n_i} \tag{1}$$

其中，w_i 为用户标签中第 i 个标签的权重，f_i 表示标签 t_i 的频数，N 为测试集中总用户数，n_i 为标签 t_i 出现在不同用户标签集中的次数。

2.1.2 标签时间权重

随着时间的推移，用户动态标注的标签隐含着用户偏好的漂移信息。相对于用户较早标注的标签，最近标注的标签对用户兴趣偏好的影响占有更大的比重。因此，最近标注的标签资源应该比那些很久之前的标签资源具有更高的权重。Cheng 等[6]提出了一种自适应指数衰减函数来计算标签中的时间权重，计算公式如式(2)所示：

$$w_{time}(u,r) = \exp\{-\ln 2 \times time(u,r)/hl_u\} \tag{2}$$

其中，$w_{time}(u,r)$ 为标签的时间权值，表示用户偏好的衰减程度；$time(u,r) \geqslant 0$ 且 $time(u,r) \in N$。当 $time(u,r)=0$ 时，表示用户 u 对资源 r 的最后那次标注时间；当 $time(u,r)=1$ 时，表示用户对资源的倒数第二次标注的时间，以此类推。hl_u 为用户 u 的半衰期，其值随着用户的生活周期而有所不同。对于一个生活周期比较短的用户，其偏好兴趣会下降得比较快，而对于那些有着较长使用期的用户，他们的偏好兴趣会相对变化得比较慢。从式(2)中可以看出，$w_{time}(u,r)$ 是一个对 $time(u,r)$ 单调递减的函数。当 $time(u,r)=hl_u$ 时，$w_{time}(u,r)$ 将下降到原先的 1/2；对于一个用户，标注的时间越是靠近最近标注的时间，其对应的 $time(u,r)$ 值也相对较小。通过式(2)的计算，最近标注的标签资源能够获得相对更大的权重，而较早标注的标签资源的权值则较小。

2.1.3 标签权重

用户-标签伪矩阵 $\boldsymbol{M}_{i\times j} = \begin{pmatrix} M_{11} & \cdots & M_{1j} \\ \vdots & \ddots & \vdots \\ M_{i1} & \cdots & M_{ij} \end{pmatrix}$，其中矩阵的值 $M(u_i,t_j)$ 可以表示用户 u_i 对标签 t_j 的总评分，可以通过标签信息权重和时间权重的加权得出，计算公式如式(3)所示：

$$M_{u,r} = \chi w_i + (1-\chi) w_{time}(u,r) \tag{3}$$

其中，参数 χ 为调和因子且 $\chi \in (0,1)$，标签权重更多的是反映用户偏好，而时间权重则更多的是反映用户偏好的变化。所以 χ 的具体取值可以根据 w_i 和 $w_{time}(u,r)$ 这两个因素的重要程度进行调整。当标注信息对用户偏好有更重要影响时，χ 取较大的值；当标注时间对用户偏好有更重要的影响时，χ 取较小的值。一般情况下，取 $\chi=0.5$，即认为两者具有同样的重要性。

2.1.4 基于动态社会化标签的用户相似度计算

本文采用余弦相似度来计算基于用户标签的相似度 $sim_1(a,b)$，其计算公式如式(4)所示：

$$sim_1(a,b) = \frac{\sum\limits_{i \in I_{a,b}} M_{a,i} \times M_{b,i}}{\sqrt{\sum\limits_{i \in I_{a,b}} M_{a,i}^2} \sqrt{\sum\limits_{i \in I_{a,b}} M_{b,i}^2}} \tag{4}$$

其中，$I_{a,b}$ 表示用户 a 和用户 b 共同标注过的资源集合。

2.2 基于用户背景信息的用户相似度

在较大时间窗口内用户背景会发生变化，但总体来说，用户背景信息相对稳定。计算用户相似性时引入用户背景信息，可以降低相似用户与目标用户中不相关用户的干扰，以找到更加准确的最近邻居用户集。用户背景信息相近的人在兴趣取向上往往有很大的相似性，而不同背景信息类别的人的兴趣偏好差异也较大[7]。

本文将用户背景信息的属性分为二元属性(如性别)、连续的数值属性(如年龄)、分类属性(如地域、职业、文化程度等)。

1)如果 p_k 是二维属性，相似度计算公式如式(5)所示：

$$sim(p_{ak}, p_{bk}) = \frac{1}{|p_{ak} - p_{bk}| + 1} \tag{5}$$

2)如果 p_k 是连续的数值属性，相似度计算公式如式(6)所示：

$$sim(p_{ak}, p_{bk}) = \begin{cases} 1, & \text{当 } p_{ak} = p_{bk} \\ 0, & \text{当 } p_{ak} \neq p_{bk} \end{cases} \tag{6}$$

3)如果 p_k 是分类属性，采用 *Jaccard* 系数计算其之间的相似度，计算公式如式(7)所示：

$$sim(p_{ak}, p_{bk}) = \frac{|p_{ak} \cap p_{bk}|}{|p_{ak} \cup p_{bk}|} \tag{7}$$

其中，p_{ak} 表示用户的分类属性，$sim(p_{ak}, p_{bk})$ 表示对于属性 p_k 用户 a，b 之间的相似度；$p_{ak} \cap p_{bk}$ 表示属性 p_{ak} 与属性 p_{bk} 间所属的类别的交集；$p_{ak} \cup p_{bk}$ 表示属性 p_{ak} 与属性 p_{bk} 之间所属的类别的并集。

集成上述 3 个属性的相似度得出基于用户背景信息的相似度：计算公式如式(8)所示：

$$sim_2(a,b) = \sum_{k=1}^{n} [sim(p_{ak}, p_{bk}) \times w(p_k)] \tag{8}$$

其中，$\sum\limits_{k=1}^{n} w(p_k) = 1$；$n$ 是属性的维度，$sim(p_{ak}, p_{bk})$ 是用户 a 和用户 b 在属性 p_k 上的相似度，$w(p_k)$ 是属性 p_k 所占的权重，该权重可由领域专家提供或者根据统计数据产生。

2.3 基于时间权重的用户评分相似度

2.3.1 用户评分矩阵的建立

本文定义用户集合为 $U=\{u_1, u_2, \cdots, u_n\}$，商品项目集合为 $I=\{i_1, i_2, \cdots, i_n\}$；$r_{u,i}$ 表示用户 u 对项目 i 的行为(如购买或评分行为)，当用户无行为时，$r_{u,i}$ 则记为 0；所有用户对项目的行为可以表示为一个 $m \times n$ 的矩阵，记为 $\boldsymbol{R}_{m\times n} = \begin{pmatrix} R_{11} & \cdots & R_{1n} \\ \vdots & \ddots & \vdots \\ R_{m1} & \cdots & R_{mn} \end{pmatrix}$，其中元素 $R(u_m, i_n)$ 的值表示用户 u_m 对项目 i_n 的评分。

2.3.2 评分时间权重的计算

用户的兴趣偏好可通过用户的评分来体现，但其会随着

时间的推移而变化。而传统的协同过滤推荐算法较少考虑时间因素，本文在计算基于评分相似度的同时考虑了时间因素对用户兴趣的影响，时间权重的计算公式如式(9)所示：

$$f(t_{u,i})=\frac{1}{1+\exp(-t_{u,i})} \tag{9}$$

其中，$-1\leqslant t\leqslant 1$ 且 $0<f(t_{u,i})<1$。$f(t_{u,i})$为单调递增函数，输出值随着时间 t 的增加而一直增加且始终保持在(0,1)的范围内；$t_{u,i}$表示用户 u 在 t 时刻对项目 i 的评分值。本文在实验之前，预先将用户评分的时间变量通过标准化公式转化为[0,1]。时间变量的变化会直接体现在权重的变化上，算法通过权重可准确地追踪用户兴趣的漂移，从而为用户提供更加准确的个性化推荐。

2.3.3 基于时间权重的用户评分相似度计算

采用 *pearson* 相关系数方法计算用户 a 和用户 b 共同评价的项目集合 $I_{a,b}$，则结合时间权重的用户 a 和用户 b 间的相似度为：

$$sim_3(a,b)=\frac{\sum_{i\in I_{a,b}}[r_{a,i}\times f(t_{a,i})-\overline{r_a}][r_{b,i}\times f(t_{b,i})-\overline{r_b}]}{\sqrt{\sum_{i\in I_{a,b}}[r_{a,i}\times f(t_{a,i})-\overline{r_a}]^2}\sqrt{\sum_{i\in I_{a,b}}[r_{b,i}\times f(t_{b,i})-\overline{r_b}]^2}} \tag{10}$$

其中，$r_{a,i}$是用户 a 对项目 i 的评分，$r_{b,i}$是用户 b 对项目 i 的评分，$\overline{r_a}$是用户 a 的平均评分，$\overline{r_b}$是用户 b 的平均评分，$f(t_{a,i})$表示用户 a 对项目 i 评分的时间权重。

2.4 最近邻居集产生及项目推荐

用户总相似度加权计算公式如式(11)所示：

$$sim(a,b)=\alpha sim_1(a,b)+\beta sim_2(a,b)+(1-\alpha-\beta)sim_3(a,b) \tag{11}$$

其中，$sim(a,b)$为集成的总相似度，α、β 分别为 $sim_1(a,b)$、$sim_2(a,b)$相似度的权重，且 $0\leqslant\alpha\leqslant 1$、$0\leqslant\beta\leqslant 1$，$\alpha$、$\beta$ 的取值大小通过多次实验比较并最终找出最佳的相似度权重，即当 α、β取最佳相似度权重时，本文方法可得到较高的推荐效率。通过用户总相似度的计算，找出目标用户的前 k 个最近邻。

在得到目标用户的前 k 个最近邻居集后，对目标用户给出相应的项目推荐。对目标用户 a 按评分差加权平均公式进行预测评分，得到预测项。计算公式如式(12)所示：

$$r_{c_a,i}=\overline{r_a}+\frac{\sum_{b\in U_{c_a},i\in I_b}sim(a,b)\times(r_{a,i}-\overline{r_a})}{\sum_{b\in U_{c_a},i\in I_b}|sim(a,b)|} \tag{12}$$

其中，$r_{c_a,i}$是用户 a 对项目 i 的预测评分，$\overline{r_a}$是用户 a 的平均评分，U_{c_a}是用户 a 的近邻集，$sim(a,b)$是相似用户 a 和 b 的相似性，且：

$$\overline{r_a}=\frac{1}{|I_a|}\sum_{i\in I_a}r_{c_a,i} \tag{13}$$

其中，$I_a=\{i\in I|r_{c_a,i}\neq 0\}$。最终选取出预测目标用户的 Top-N 个推荐项目。

3 算法描述及复杂度分析

3.1 算法描述

输入：目标用户 u_T，标签数据集，用户-评分项目矩阵 $\boldsymbol{R}_{m\times n}$，用户-标签伪矩阵 $\boldsymbol{M}_{i\times j}$，用户背景信息矩阵 $\boldsymbol{B}$，邻居数 k；

输出：Top-N 推荐结果。

推荐步骤如下：

Step1 对每个用户 u，找到同目标用户 u_T 共同评分过的项目，将其记录在 *items* 中；

Step2 根据标签数据集，由式(1)－式(3)计算得出标签权重 $M_{u,r}$并将其值保存在矩阵 $\boldsymbol{M}_{i\times j}$中；

Step3 基于矩阵 $\boldsymbol{M}_{i\times j}$，使用式(4)计算用户标签间的相似度 $sim_1(a,b)$；

Step4 基于矩阵 $\boldsymbol{B}$，使用式(5)－式(8)计算用户背景信息间的相似度 $sim_2(a,b)$；

Step5 根据矩阵 $\boldsymbol{R}_{m\times n}$，使用式(10)计算用户评分间的相似度 $sim_3(a,b)$；

Step6 由式(11)集成得出目标用户 u_T 与其他用户间的总相似度 $sim(a,b)$；

Step7 根据式(12)计算得出目标用户 u_T 的所有项目的预测评分；

Step8 为目标用户 u_T 生成 Top-N 推荐。

3.2 复杂度分析

根据 3.1 节算法的描述，假设用户数为 m，项目数为 n，用户标签数为 l；则在 Step1 系统初始化阶段，遍历目标用户与其他用户间共同评分过的项目，因此其复杂度为 $O(n)$；Step2 中计算用户标签两两之间的标签权重，因此复杂为 $O(l^2)$；Step3－Step5 均是由矩阵计算目标用户与其他用户之间的相似度，复杂度分别为 $O(l^2)$、$O(m^2)$、$O(n^2)$；Step6－Step8 阶段为用户推荐阶段，主要的时间是在项目的排序上，因此复杂度最差为 $O(m^2)$。虽然推荐计算量会随着用户评分数据和标签数据的增加而有所增加，但是相比较传统的协同过滤算法，新方法在提高推荐准确度的同时还缩短了算法的推荐时间。

4 实验及结果分析

4.1 数据集及实验环境

由于目前尚未见到同时具有社交网络和电商交易信息的标准数据集，故本实验采用的数据集是通过集成 MovieLens 站点(http://movielens.umn.edu)提供的 1M 的公开数据集(2000 年加入 MovieLens 6040 个用户所形成的数据集)和 10M 的公开数据集(包括：71567 个用户，10681 部电影，给出对影片的 10000054 次的评级和用户所做的 95580 次的标签总数，其中也包括了重复的标签)实现的。在数据集中，用户对自己看过的电影进行评价，评分范围为 15，“1”表示“不喜欢”，“2”表示“不太喜欢”，“3”表示“一般喜欢”，“4”表示“比较喜欢”，“5”表示“非常喜欢”，其中注册用户必须至少对它所拥有的电影中的 20 部进行评价。在 10M 的 MovieLens 数据中，标签是由用户自由添加的，筛选每个用户至少对所拥有的电影中的 5 部进行标注，每个影片至少被两个用户选择过，每个影片至少赋有一个标签。数据集中时间信息是用时间戳来表示的，时间戳是自纪元表示以秒为返回的时间，两次标签的时间戳差值表示这两次标签标注的时间差，时间戳越小，说明标注的时间越早。

实验数据集首先由 1M 数据集和 10M 数据集中的用户-项目评分通过改进的余弦相似度计算方法分别得出两个数据集用户间的相似度；然后，基于用户的相似度将两个数据集中

的用户进行聚类；继而将1M数据中的用户背景信息整合到10M数据集中，得到一个全新的数据集（包含用户-项目评分数据、电影类别、用户背景信息、标签数据）。其中用户-项目评分数据集中包含：用户ID、项目、评分值、时间戳；标签数据集中包含：用户ID、项目、标签信息、时间戳。通过此数据集对本文提出的算法进行验证。

数据集被随机分为两个部分：一部分作为训练集，另一部分作为测试集。测试集的数据信息是隐藏的，用来检测算法的准确度。在平均排序分的实验中，可通过改变训练集和测试集在数据集中的比重来计算算法的平均排序分；在平均绝对误差实验中，由于通过验证用户-项目预测评分与真实评分间的偏差大小来验证准确性，因此数据集由用户-项目评分被随机地分为两个部分，如80%训练集对应20%测试集，90%训练集对应10%测试集，或者50%训练集对应50%测试集等。本文采用被分为90%的训练集和10%的测试集的数据集对算法进行验证。

为了检验本文方法在数据稀疏、冷启动数据集下相对于其他推荐算法能否可以提高推荐精度，对上述数据集进行随机稀疏采样，采样密度为40%，得到的数据集为冷启动数据集。在冷启动数据集下通过对算法的推荐精度进行比较以验证本文算法的优势。

本文采用Java实现了基于动态社会行为和用户背景的协同推荐方法，并设计了若干实验对本文方法的推荐效率进行验证。实验环境为Intel Core i5处理器、CPU主频为2.5GHz、内存为4G、Windows7操作系统的PC机，jdk版本为1.6。

4.2 度量标准

可以由很多指标来对推荐算法的优劣进行评价。本文通过评价排序分、平均绝对误差两个指标来分析算法的性能。

1）平均排序分[8]。通过计算项目在推荐列表中的位置来度量推荐系统的排序准确度。其计算公式如式(14)所示：

$$r_i = L_i / N \tag{14}$$

其中，N为训练集中用户未选择的对象个数，L_i为预测集中待预测对象i在推荐列表中的位置。排序分越小，说明该系统趋向于把用户喜欢的产品排在前面。对所有用户-对象对的排序分求平均，可得到平均$\bar{r}$，以此值来量化推荐算法的精确度，$\bar{r}$越小说明算法的推荐效果越好。

2）平均绝对误差[9]（Mean Absolute Error，MAE）。通过计算测试集中用户实际评分和利用推荐算法预测出来的评分间的绝对值来度量推荐系统的推荐准确度。当MAE的值越小，系统的推荐精度越高；反之，其推荐精度越差。

假设预测用户评价集表示为$\{p_1, p_2, \cdots, p_N\}$，而相应的实际的用户评价集为$\{q_1, q_2, \cdots, q_N\}$，则平均绝对误差MAE的计算公式如式(15)所示：

$$MAE = \frac{\sum_{i=1}^{n} |p_i - q_i|}{n} \tag{15}$$

其中，n表示测试算法的评分项目数目，p_i表示测试算法的预测评分，q_i为测试集中用户真实评分。

4.3 实验结果分析

为了验证本文所提出的算法在推荐精度上的优势，将其与传统协同过滤推荐算法（CF）、基于社会化标签的协同过滤算法、基于用户背景信息的协同过滤算法进行比较。当参数α、β取值不同时，本文方法与传统协同过滤算法的MAE见表1。

表1 本文方法与传统协同过滤算法的MAE

邻居数	传统协同过滤推荐算法（CF）	$\alpha=0.8,\beta=0.1$时本文算法（CF1）	$\alpha=0.5,\beta=0.25$时本文算法（CF2）	$\alpha=0.2,\beta=0.3$时本文算法（CF3）
10	0.8406	0.7647	0.7047	0.7423
20	0.8247	0.7571	0.7002	0.7215
30	0.8229	0.7385	0.6748	0.7151
40	0.8258	0.7427	0.6783	0.7137
50	0.8294	0.7371	0.6775	0.7213
60	0.8146	0.7426	0.6827	0.7136
70	0.8099	0.7367	0.6846	0.7125
80	0.8158	0.7403	0.6793	0.7094

由表1可得出，在用户邻居数相同的情况下，本文算法的推荐精度均高于传统协同过滤推荐算法；由于相似度权重的选取不同，相应的推荐效率也不同，且在$\alpha=0.5,\beta=0.25$时本文方法的推荐效率达到最佳。故取$\alpha=0.5,\beta=0.25$作为相似度权重，并分别进行相应如图1－图3所示的推荐性能的验证比较。

平均排序分的实验测试结果如图1所示，此时相似度权值$\alpha=0.5,\beta=0.25$。从图1可看出，平均排序分大小与数据集中训练集所占比例成反比，本文算法的平均排序分要明显低于其他几种方法。由此可见，本文算法在推荐精度上要明显好于其他几种算法。

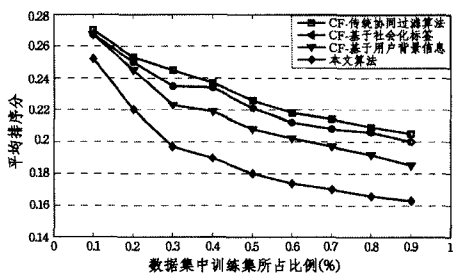

图1 推荐算法的平均排序分比较

当相似度权值$\alpha=0.5,\beta=0.25$时，得出本文算法的MAE变化曲线如图2所示。图2中，当最近邻居数小于30时，随着邻居数量的增加，MAE的值逐渐变小，说明此时算法的推荐精度随着邻居数的增加而得到提高，并且推荐质量受到邻居数大小的影响；当邻居数大于30时，本文算法的MAE变化曲线趋向于平稳，并相对于其他3种推荐算法明显地提高了推荐精度。

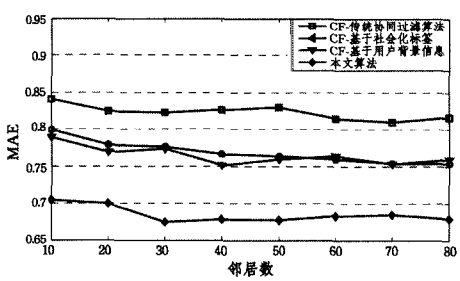

图2 推荐算法的MAE值比较

如图3所示，在冷启动数据集下，相似度权值$\alpha=0.5,\beta=0.25$时，得出推荐算法的MAE值变化曲线。由图3可得出，在相同邻居集下，采用本文算法的MAE值不仅明显低于其它几种推荐算法，而且随着邻居数量的增加其趋向于平稳，其

（下转第265页）

to data mining [M]. Beijing: Post&Telecom Press, 2006
[2] 史忠植. 高级人工智能[M]. 北京：科学出版社，2006
[3] Jain A K. Data clustering: 50 years beyond k-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666
[4] Aggarwal C C, Han J, Wang J, et al. A framework for clustering evolving data streams[C]// Proc of VLDB. 2003: 81-92
[5] Aggarwal C C, Han J, Wang J, et al. A framework for projected clustering of high dimensional data streams [C] // Proc. of VLDB. 2004: 852-863
[6] Cao F, Estery M, Qian W, et al. Density-based clustering overran evolving data stream with noise[C] // Proc of the SIAM Conference on Data Mining (SDM). 2006: 326-337
[7] Huang Z. Extension to K-means algorithm for clustering large datasets with categorical values[J]. Data Mining and Knowledge Discovery II, 1998(2): 283-304
[8] Aggarwal C C, Yu P S. A framework for clustering massive text and categorical data streams[C]// Proc of 6th Siam IntConf on Data Mining. Bethesda, 2006: 477-481
[9] Guha S, Rastogi R, Shim K. ROCK: a robust clustering algorithm for categorical attributes[C]// Proc of ICDE. 1999: 512-521
[10] Barbara D, Couto J, Yi L. COOLCAT: an entropy-based algorithm for categorical clustering[C]// Proc of CIKM. 2002: 582-589
[11] Ralambondrainy H. A conceptual version of the k-means algorithm[J]. Pattern Recognition Letters, 1995: 1147-1157
[12] Huang Z. Clustering large data sets with mixed numeric and categorical values[C]// Pro of 1th Pacific-Asic Conf. 1997: 21-34
[13] Yin Jian, Tan Zhi-fang, Ren Jiang-tao, et al. An efficient clustering algorithm for mixed type attributes in large dataset[J]. IEEE transactions on Machine Learning and Cybernetics, 2005, 8(3): 1611-1614
[14] He Zeng-you, Xu Xiao-fei, Deng Sheng-chun. Scalable algorithms for clustering large datasets with mixed type attributes [J]. International Journal of Intelligent Systems, 2005, 20(10): 1077-1089
[15] 杨春宇，周杰. 一种混合属性数据流聚类算法[J]. 计算机学报，2007, 30(8): 1364-1372
[16] Hsu C C, Huang Y. Incremental clustering of mixed data based on distance hierarchy[J]. Expert Systems with Applications, 2008, 35(3): 1177-1185
[17] 黄德才，沈仙桥，陆亿红. 混合属性数据流的二重 k 近邻聚类算法[J]. 计算机科学，2013, 40(10): 226-230
[18] 陈新泉. 面向混合属性数据集的双重聚类方法[J]. 计算机工程与科学，2013, 35(2): 127-132
[19] Liang Ji-ye, Zhao Xing-wang, Li De-yu, et al. Detemining the number of clusters using information entropy for minxed data [J]. Pattern Recognition, 2012, 45(6): 2251-2265
[20] 王述云，胡运发，范颖捷，等. 基于距离和熵的混合属性数据流聚类算法[J]. 小型微型计算机系统，2012, 12(12): 2365-2371

（上接第 255 页）

他算法的 MAE 值波动较大。说明本文算法在冷启动数据集下，可以较好地提高推荐的准确率。

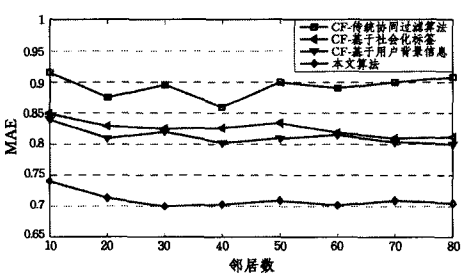

图 3 在冷启动数据集下算法的 MAE 比较

结束语 推荐系统作为解决信息过载的一种有效手段，如何提高其推荐精度已成为非常重要的研究问题。人们对社会标签标注的时间变化体现了其社会行为的动态性。考虑了社会标签的时间属性并集成了用户背景，本文提出的方法相对于传统的协同过滤（CF）等推荐算法能够有效地提高算法的推荐精度和改善数据稀疏及冷启动的问题。目前，社交网络平台与电子商务平台多数是独立运行的，难以获得同时具有电子商务和社交网络的标准数据集，本文验证则另辟途径，先计算两个数据集用户评分间的相似性得出一个用户相似类，再将两个数据集整合成本文所需要的数据集来进行验证，但存在一定的不足。未来工作是集成电商和社交网络平台数据，对本文的工作给予进一步的验证，并考虑其它社交行为对推荐质量的影响。

参 考 文 献

[1] 于洪，李转运. 基于遗忘曲线的协同过滤推荐算法[J]. 南京大学学报：自然科学版，2010, 46(5): 520-527
[2] 冯勇，李军平，徐红艳，等. 基于社会网络分析的协同过滤推荐方法改进[J]. 计算机应用，2013, 33(3): 841-844
[3] Eleftherios T, Yannis M. Product recommendation and rating prediction based on multi-modal social networks[C] // Poceedings of the 5th ACM Conference on Recommender Systems. New York: ACM Press, 2011: 61-68
[4] 贾大文，曾承，彭智勇，等. 一种基于用户偏好自动分类的社会媒体共享和推荐方法[J]. 计算机学报，2012, 35(11): 2381-2391
[5] 顾亦然，陈敏. 一种三部图网络中标签时间加权的推荐方法[J]. 计算机科学，2012, 39(8): 96-98, 129
[6] Cheng Yuan, Qiu Guang, Bu Jia-jun, et al. Model bloggers' interests based on forgetting mechanism[C]// Proceedings of the 17th International Conference on World Wide Web. New York: ACM Press, 2008: 1129- 1130
[7] 庄景明，王明文，叶茂盛. 基于内容过滤的农业信息推荐系统[J]. 计算机工程，2012, 38(11): 38-41
[8] Zhou Tao, Jiang L L, Su R Q, et al. Effect of initial configuration on network-based recommendation[J]. Europhys Lett, 2008, 81: 58004
[9] 赵琴琴，鲁凯，王斌. SPCF: 一种基于内存的传播史协同过滤推荐算法[J]. 计算机学报，2013, 36(3): 671-676