DOI:10.11721/cqnuj20160524

基于时间加权标签的协同过滤推荐算法研究*

宋伟伟[1,2], 杨德刚[1], 郑 敏[1]
(1. 重庆师范大学 计算机与信息科学学院, 重庆 401331; 2. 河南工业与贸易职业学院, 郑州 450053)

摘要:传统的协同过滤推荐算法主要以评分数据为数据源来计算和反映用户的兴趣偏好和资源相似度,并决定是否对潜在用户进行推荐物品。而忽视了用户、资源本身的特征,用户在不同时间对资源的认识和感兴趣程度是会变的。基于这个问题,本研究对传统算法进行了改进,提出了基于时间加权标签的信息推荐算法。该算法的主要思想是标签可被用户依个人偏好进行资源标注,标签代表用户对资源的兴趣特征,以用户集、时间、标签集及物品资源等 4 个量形成的多维关系,可以计算出用户和资源之间的标签特征向量,计算在不同时间段,用户对资源的偏好以及资源相似度,并且依据用户的历史行为来预测用户的偏好,并进行推荐。实验结果显示本算法有效地提高了推荐的准确性,获得了更好的推荐效果。

关键词:时间标签;推荐;个性化;准确性
中图分类号:TP301.6 **文献标志码**:A **文章编号**:1672-6693(2016)05-0113-08

随着时间的推移,人们对于资源事物的认定意义会有所改变。在信息知识迷航时代,网络这个无形之网把能够信息化的事物都做到了信息化,比如人的感情、物品的分类和评价评分等都可以根据自己的喜好来进行标签并分类。标签可被用户依个人偏好自由地进行资源标注。标签是用户对浏览过的信息和数据添加属于用户的个性化标志,所标注的标签格式可以是词汇、句子、标点和符号等[1]。对于已经定义的标签,随着时间的推移和时代的发展,意义和形式总是会有所变化。在当代的大数据信息时代,对于这种变化,如何更精准地从每天以 PB 级单位增加的数据中去推送新事物或者合适的资源到目的用户,是信息推荐领域的一个重点。

有学者提出了基于标签的协同过滤推荐系统,它能有效地解决推荐的冷启动问题,提高推荐的准确性,获得更好的推荐效果。但是标签会随着时间的推移,意义就有所改变,不同时间标记的标签对于用户的意义以及在社会环境中的意义可能不一样,最近时间进行的标签资源可以看作是用户最新感兴趣的资源。本文基于标签意义会随着时间而改变的问题,在基于标签的协同过滤算法基础上加入了时间标签的因素,进行了基于时间加权标签的协同过滤推荐算法的研究。在标注了标签的系统中,由于用户对资源的评价周期长短不同,标签意义有所差别。对于评价资源的时间周期比较长的用户,其较早时间标注的标签对当前的推荐和预测影响较小,而最近时间标注的标签能更全面地反映出用户的喜好,且最近标签特征对推荐预测的准确度影响很大[2]。实验结果表明,加入时间因素对于预测的准确度有所提升。

1 相关工作

当前,推荐系统在学术界以及商业界都取得了较大的进展,特别是在电子商务和网络营销、移动应用、互联网广告和信息检索等领域,在这些领域所应用到的推荐算法中,协同过滤算法应用最广。协同过滤推荐算法的思想是基于用户对物品的评分的共现矩阵,通过评分值来计算用户的相似度或者物品相似度来预测用户对潜在物品的评分,并决定是否对用户推荐相应的物品,且达到一定的推荐效果。在个性化推荐系统的各种算法中,最为广泛采用的推荐技术就是协同过滤算法,协同过滤算法在推荐的过程中没有考虑用户和时间之间的变化因素,时效性欠佳。

传统的推荐算法已经不能满足信息急速增长的步伐。用户使用社会标签对信息进行分类,对于不同的资源

* **收稿日期**:2015-04-06 **修回日期**:2016-01-26 **网络出版时间**:2016-07-13 14:00
资助项目:国家自然科学基金(No. 10971240);重庆市自然科学基金(No. cstc2012jjA40052;No. cstc2013jcyjA0973; No. cstc2013jcyjA80013);重庆市教委科学技术研究项目(No. KJ120615;No. KJ120630;No. KJ130611;No. KJ1400505);重庆师范大学校级项目(No. 13XLZ01;No. 201334;No. xyjg13010);校级研究生科研创新项目(No. YKC15009)
作者简介:宋伟伟,女,研究方向为数据挖掘、推荐系统,E-mail:916548563@qq.com;**通信作者**:杨德刚,教授,E-mail:ydg42@163.com
网络出版地址:http://www.cnki.net/kcms/detail/50.1165.N.20160713.1400.018.html

可以根据自己的喜爱程度和兴趣采用不同的标签来分成不同的种类,可自由组织、管理和搜索所需的资源[3]。由于在推荐的过程中,考虑到用户的兴趣随着时间的推移会改变,在改变兴趣的过程中,用户同样希望推荐的新颖信息和当前兴趣标签吻合。

在传统协同过滤的算法中,为了解决冷启动问题和稀疏问题,提出了基于内容的推荐;为了解决处理复杂的非结构数据,提出了基于内存的协同过滤;为了在缓解数据稀疏性问题的基础上提高预测精确度,并且增强系统的扩张性,提出了基于模型的协同过滤推荐算法;为了解决推荐系统在推荐的过程中依赖于用户的历史行为数据的问题,提出了基于知识的推荐;为了解决推荐过程中的冷启动问题,吴杰等人[4]在传统的协同过滤推荐算法基础上,运用BP神经网络,提出了一种基于优先新品推荐和用户偏好的协同过滤算法,加强了新产品的推荐。单一的协同过滤总是会有不能满足需求的情况,就可以有机地结合不同推荐系统的优点,于是提出了混合式推荐系统,缺点是不能够根据方法的选取来动态调整组合顺序,而且时间复杂度高。为进一步提高协同过滤算法的性能,张莉等人[5]对协同过滤推荐算法进行了改进,主要是充分利用用户历史评分数据并根据邻居用户之间的相似性,通过相应的算法来计算出用户的喜好程度。

用户对于一种事物的认识,随着时间的推移记忆会消退,基于此,郭彩云等人[6]提出了一种改进的基于标签的协同过滤,该算法将用户评分融入到用户对标签权重的计算中,考虑到用户对项目的不同兴趣程度,以及对推荐结果的影响,采用指数渐进遗忘函数和时间窗口相结合的方法来捕捉用户兴趣的变化。易辽宏[7]提出的基于标签张量分块分解的个性化推荐方法(TTBCD),可以把用户设置的标签作为张量分块来进行个性化信息推荐。

在利用数据集计算用户相似度的过程中,往往只考虑到单一的用户矩阵问题,继而赵海燕等人[8]提出一种结合时间权重与信任关系的协同过滤推荐算法TTCF,在推荐的各种途径中,用户倾向于相信好友推荐的物品,因此在个性化推荐的过程中结合用户之间的信任因素,能够提高推荐的精确度[9],从而促进用户的购买行为。孙克等人[10]提出了一种面向社会关系的移动用户好友推荐算法。个性化推荐研究中,垃圾标签不仅会导致数据稀疏性问题,同时影响推荐的实时性和精确性。张明等人[11]提出一种优化标签的矩阵分解推荐算法OTMFR。把矩阵分解技术引入到融合了标签的推荐算法之后,如何将标签信息融合到模型中成为了研究内容之一[12],而对通过优化标签质量提高推荐效果这一方面的研究很少。顾亦然等人[13]提出结合标签使用频率和用户添加标签的时间,有效地提高推荐的精度和多样性。虽然在资源推荐过程中标签的作用很大,但是在推荐系统中,添加标签的随意性和实时性对于标签的质量是有影响的,如标签的一词多义以及时代性意义等,都会影响推荐物品的质量。

在学术界的大多数过滤算法改进中,每一步的改进算法几乎都不太重视用户、资源自身的特征,特别是不同时代下资源所代表的意义,而且对于提高推荐的准确度没有过多涉及,蔡强等人[14]提出了基于标签的个性化资源推荐算法,使推荐结果更具合理化,并且对稀疏数据和新资源的推荐质量明显下降。标签作为体现用户兴趣偏好和资源特征的信息,标签的意义会随着时间[15-16]的推移以及快节奏的信息时代发展而展现出不同的意义,而这个意义有时对于部分用户来说可能截然相反。本文在标签的基础上加入了时间权,根据用户最近的时间段感兴趣的标签作为推荐依据,并决定是否对其进行推荐,基于此,提出了基于时间加权标签的协同过滤思想。

2 时间加权标签和资源相似度计算

2.1 预定义

用户可以对信息资源进行自由的标签标注[16],可以在多个时间段对同一资源进行标注,以最近时间使用的标签作为标签源,从用户、时间和资源标签3个角度分析用户对资源的倾向度。基于此,利用用户使用的标签作为用户偏好模型的基本特征元素。

定义用户集合$U=\{u_1,u_2,\cdots,u_j,\cdots,u_M\}$,其中$M$为用户总数,$j=1,2,\cdots,M$;所有资源的集合记为$I=\{i_1,i_2,\cdots,i_j,\cdots,i_N\}$,其中$N$为资源总数,$j=1,2,\cdots,N$;用户使用的标签集合记为$B=\{B_1,B_2,\cdots,B_j,\cdots,B_L\}$,其中$L$为标签总数,$j=1,2,\cdots,L$;标签标记的时间集合记为$T=\{t_1,t_2,\cdots,t_j,\cdots,t_S\}$,其中$S$为用户对应标签标记的时间。

2.2 时间加权标签

时间加权标签的思想考虑到标签意义随着时间变化的影响,标签意义的变化主要考虑到两个方面,即越早时间标注的标签,其意义对于用户的影响越低;而最近标注的标签,对用户的兴趣和选择影响较大[17]。在推荐的过程中,标签对推荐结果的整体影响比较大,为了降低标签的负影响,本文引入了时间加权标签$f(t)$。结合用户

网络活动信息和标签时间衰减的影响，标签的时间衰减函数如下：$f(t_{ui}) = e^{-t_{ui}}$。其中 t_{ui} 记为在 t 时刻用户 u 对资源 i 的兴趣度。

1）用户的标签特征向量。用户对物品标注的标签，其标签特征向量是利用用户最近或者经常使用的标签来表示用户的兴趣特征[18]，如(1)式所示。

$$\overrightarrow{\boldsymbol{p}_{ub}} = \left(f(t_{ub_1}) \frac{n_{ub_1}}{n_{ub}} \log\left(\frac{M}{n_{b_1 u}}\right), \cdots, f(t_{ub_j}) \frac{n_{ub_j}}{n_{ub}} \log\left(\frac{M}{n_{b_j u}}\right), \cdots, f(t_{ub_L}) \frac{n_{ub_L}}{n_{ub}} \log\left(\frac{M}{n_{b_L u}}\right) \right), \tag{1}$$

其中，$n_{b_j u}$ 表示使用过标签 b_j 的所有用户的数量，M 表示数据集中用户的总数，n_{ub_j} 表示用户 u 使用标签 b_j 的次数，n_{ub} 表示用户 u 所使用的标签数量，$\frac{n_{ub_j}}{n_{ub}}$ 表示 u 使用标签的频率，$\log\left(\frac{M}{n_{b_j u}}\right)$ 表示在用户所有标注的标签中各个标签的重要度，$f(t_{ub_j})$ 表示在 t 时刻用户 u 对标签 b_j 进行了标注，$f(t_{ub_j}) \frac{n_{ub_j}}{n_{ub}} \log\left(\frac{M}{n_{b_j u}}\right)$ 项表示在 t 时刻标签对该用户 u 的重要度。

2）资源的标签特征向量。在 t 时刻对资源 i 标注的标签称为资源的标签特征向量。如(2)式所示。

$$\overrightarrow{\boldsymbol{p}_{ib}} = \left(\frac{n_{ib_1}}{n_{ib}} \log \frac{N}{n_{b_1 i}}, \cdots, \frac{n_{ib_j}}{n_{ib}} \log \frac{N}{n_{b_j i}}, \cdots, \frac{n_{ib_L}}{n_{ib}} \log \frac{N}{n_{b_L i}} \right), \tag{2}$$

其中，N 表示数据集中资源的总数，n_{ib_j} 表示资源 i 被标记为 b_j 标签的数量，$\frac{n_{ib_j}}{n_{ib}}$ 表示标签 b_j 在所有标签中出现的频率，$\log \frac{N}{n_{b_j i}}$ 表示标签 b_j 的重要度，$\frac{n_{ib_j}}{n_{ib}} \log \frac{N}{n_{b_j i}}$ 表示标签对资源 i 的重要度。

2.3 用户对资源的偏好程度

用户 u_j 对资源 i_k 的偏好程度，如(3)式所示。

$$p_{u_j i_k} = \overrightarrow{\boldsymbol{p}_{u_j b}} \cdot \overrightarrow{\boldsymbol{p}_{b_i k}} = \sum_{b=1}^{L} p_{u_j b} \times p_{b_i k}, \tag{3}$$

其中，$u_j \in U, j = 1, 2, \cdots, M; i_k \in I, k = 1, 2, \cdots, N$。

用户 u_j 的资源偏好特征向量，如(4)式所示。

$$\overrightarrow{\boldsymbol{p}_{u_j}} = (p_{u_j i_1}, \cdots, p_{u_j i_k}, \cdots, p_{u_j i_N})。 \tag{4}$$

过滤信息的目的是计算出用户对资源的喜爱程度，并决定是否对其进行推荐。但是随着时间的推移，用户对于事物的认识会改变，推荐算法中对无关的信息进行过滤，过滤目的主要是找到用户对资源的偏好程度，进而决定是否对其进行推荐。但是随着时间的推移，用户对事物的认识是动态变化的。(3)式和(4)式是时间加权标记协同过滤推荐算法的基础公式，在此基础元素上，进行相关的数据分析，达到推荐目的。

2.4 资源相似度计算

用户对于感兴趣的资源进行标注标签，对不同的资源进行了标签分类，以及不同时间对于相同的资源可能会分成不一样的类。计算资源相似度，其实就是计算资源所属标签类型的相似度[19]。基于时间加权标签计算出的资源相似度值，一般以最近时间段的历史行为作为基础数据。

资源的特征信息表示为基于时间标签的资源特征向量 $\overrightarrow{\boldsymbol{I}_k}$，如(5)式所示。

$$\overrightarrow{\boldsymbol{I}_k} = (n_{k1}, n_{k2}, \cdots, n_{ki}, \cdots, n_{kL}), \tag{5}$$

其中，n_{ki} 表示 B_i 被用来标记资源 i_k 归一化后的值。

计算资源相似度的算法有很多方式，文中是基于用户标签进行分类的，采用(6)式表示的 Pearson 相关系数来计算。

$$sim(i,j) = \frac{\sum_{u \in U} (r_{u,i} - \overline{r_u})(r_{u,j} - \overline{r_u})}{\sqrt{\sum_{u \in U} (r_{u,i} - \overline{r_u})^2} \sqrt{\sum_{u \in U} (r_{u,j} - \overline{r_u})^2}}。 \tag{6}$$

通过计算标签之间的相似度，可以构造出相似度矩阵 $\boldsymbol{S}_{n \times n}$，如(7)式所示，代表不同标签之间的相似度。

$$\boldsymbol{S}_{n \times n} = \begin{bmatrix} 1 & \cdots & s_{1j} & \cdots & s_{1n} \\ \vdots & & \vdots & & \vdots \\ s_{j1} & \cdots & 1 & \cdots & s_{jn} \\ \vdots & & \vdots & & \vdots \\ s_{n1} & \cdots & s_{nj} & \cdots & s_{nm} \end{bmatrix}, \tag{7}$$

其中,s_{nm}表示标签代表资源n和m之间的相似度。

2.5 预测评分值

根据最近时间段,用户历史行为的资源标签的相似度,可以预测出其兴趣值并决定是否对用户进行推荐,预测计算如(8)式所示。

$$p_{u_{i_j}}=\sum_{k=1}^{M} f(t_{u_j i_k})\times p_{u_j i_k}\times s_{i_j i_k}, \tag{8}$$

其中,$p_{u_{i_j}}$代表用户u对资源i_j的预测评分,$f(t_{u_j i_k})\times p_{u_j i_k}$代表在$t$时刻用户$u$对资源$i$的感兴趣程度,$s_{i_j i_k}$代表资源之间的相似度。

通过计算用户对资源的兴趣度,可以计算各历史资源与潜在推荐资源的相似度。利用时间因素来更精准个性化地推荐相关资源,提高推荐质量的准确性。

3 时间加权标签算法

3.1 资源标签模型

本文的算法思想是:用户-时间标签兴趣模型的构建脱离了对用户评分的依赖,降低了评分稀疏性对预测用户兴趣的准确性方面影响,并且提高计算精准度。用户的兴趣可能比较广泛,在同一时间可以对多个资源感兴趣,对于每个资源可能属于不同类型标签。同一种资源可以用多个标签来表示,因为同一个资源的意义是多元的。如图1所示,可以为不同为用户的兴趣进行分类[20-22],不同的资源可以标注不同的标签。

图1简单地描述了不同时刻对同一资源的标签变化,以及随着时间的变化对不同标签的衰减程度变化。如果随着时间的推移,标签1和标签n的性质一样,说明用户的喜爱程度比较稳定,可以推荐类型的物品;否则,就以用户最近使用的标签作为爱好特征来进行相似物品的推荐,并能够达到合理的推荐效果。

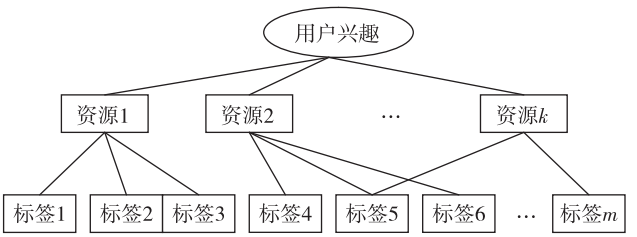

图1 用户-时间标签兴趣模型

Fig. 1 Users-time tags interest model

3.2 算法描述

在相同时间段内,计算哪个标签对用户U_1的重要度大,对于重要度大的标签B_1,找出同时间段也用了此标签B_1的用户U_j,用相似度计算B_1和B_j的资源标签相似度,根据U_j的偏好资源向用户U_1进行推荐。

步骤1,分别对用户数据集、资源数据集、时间记录集和标签集进行信息初始化;

步骤2,统计在同一时刻或者时间段,计算相关标签B对用户U的重要度,并选择Top N的重要标签;

步骤3,对资源I进行标注标签,计算各标签B对资源I的重要度;

步骤4,结合步骤2和步骤3,计算用户U对各标签对应的资源I的偏好程度;

步骤5,计算新资源与用户偏好程度高的资源的相似度,对于相似度越高的新资源,可以对用户进行推荐,可以取得很好的效果;

步骤6,一旦推荐成功,用户就会对新资源进行学习并标注标签,达到推荐的目的,否则,重新学习用户偏好;

步骤7,推荐完成。

4 实验及其分析

本文采用的数据集是来自于数据堂官网网站中的豆瓣读书内容标签数据集,下载地址是http:// www.datatang.com/。数据集中包含有3 315条豆瓣读书页面数据。每条数据包括用户ID、时间戳、图书的豆瓣ID、图书的内容简介、图书的常用标签和次数。

4.1 实验过程

输入:用户-时间-标签-资源的特征数据矩阵;

输出:对于标签相似度高的用户数据集,推荐用户没有浏览的相关资源。

步骤1,数据初始化。

文中数据以典型的Web2.0网站-豆瓣网为数据源,选取其中标注标签最多的前30名“豆瓣读书”用户的标签作为样本数据来进行试验研究。

表 1　豆瓣图书资源数据

Tab. 1　Douban reading data resource

样本资源总数/个	样本资源标签数/个	标注/次	总资源数/个	所有资源标注的标签数/个
30	165	1 549 659	3 315	8 827 523

从表 1 可以得出，每个资源平均被标注 5 个标签，每个用户平均最多标注了 9 000 个标签。数据的标注和图 1 兴趣模型很吻合。表 2 是对用户数据、资源数据、时间记录以及资源标签集合初始化之后的信息。

步骤 2，通过筛选标注量多的前 30 名用户，计算每个用户标注过的标签对其的重要性，对于重要性大的标签 B_i，统计曾经标注过该标签 B_i 的其他用户。通过标签的相似或者一致性，可以预测出兴趣相同的用户，进而对潜在用户推荐类似标签代表的信息资源。根据(1)式计算用户的标签特征向量，计算结果公式如下。

$$p_{165,1\,040\,771}=(0.126,0.248,0.24,0.115,0.122\,4),$$

这里 $p_{165,1\,040\,771}$ 代表用户 165 对资源 ID 为 1 040 771 的所有标签的倾向程度值。如此循环计算其他资源对于 UserID 为 165 的用户重要性。标签对用户的重要度值在 0～1 之间，对于 UserID 为 165 的用户来说，“悬疑”对用户的重要度为 0.126，“达·芬奇密码”对用户的重要度为 0.248，“小说”对用户的重要度为 0.24，“宗教”对用户的重要度为 0.115，“推理”对用户的重要度为 0.122 4。得知，小说和达芬奇标签对用户的重要度比较大。

步骤 3，计算各标签对该资源的重要度。

通过(2)式计算资源的标签特征向量，计算结果如下所示。

$$p_{1\,040\,771,\text{标签}}=(0.319,0.264,0.234,0.183,0.174)。$$

对应的标签是悬疑、达芬奇、小说、宗教和推理。可以算出悬疑和达·芬奇密码这两个标签对图书 ID 为 1 040 771 的重要度比较高。

步骤 4，结合步骤 2 和步骤 3 的计算，可以计算出 ID 为 165 的用户对哪些标签比较喜欢，就可以通过喜欢的标签来推出标签所指的一个或多个资源。步骤 2 中，得出“小说”和“达芬奇”标签对用户的重要度比较大。步骤 3 中，悬疑和达·芬奇密码这两个标签对图书 ID 为 1 040 771 的重要度比较高。可以得出用户 ID 为 165 的用户对标签为小说、悬疑和达·芬奇密码的书目比较感兴趣。

步骤 5，计算新资源与用户偏好程度高的资源的相似度，对于相似度越高的新资源，可以对用户进行推荐，取得很好的推荐效果。

对于新出版的《风之影》，附有“悬疑”、“西班牙小说”、“文学”的标签，风之影的标签与达·芬奇密码标签的相似度为 0.75(相似度最大值为 1)，通过(8)式计算预测值，计算出 ID 为 165 的用户对其比较感兴趣，然后决定对其进行推荐。

步骤 6，单用户单资源的推荐任务完成，转到步骤 2，计算其他的用户或者资源，直到计算完成为止。

步骤 7，结束。

4.2 推荐质量评测指标

对于本文中算法的推荐结果质量进行验证和评测，采用准确率、召回率和覆盖率这 3 个评测指标为依据来进行评测，计算时间加权标签(TBCF)和普通的基于标签的协同过滤算法(TCF)之间的推荐精度的优越性。

以豆瓣图书数据集为数据源，实验选取的数据源中共有 3 313 条数据，是 943 个用户对 3 313 个书目进行标注的标签记录。实验数据中，每个用户对书目标注的标签至少有 5 个。在文中，用 300 个测试数据来进行实验。

首先，采用准确率(Precision)[19-20]来进行测评算法的准确率可行性。

$$Precision=\frac{\text{推荐的正确信息条数}}{\text{推荐的信息条数}}。\tag{9}$$

图 2 是在豆瓣图书数据集下，Precision 值对比的实验结果。

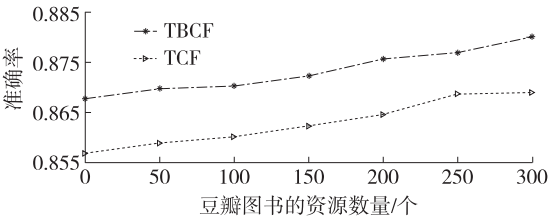

图 2　豆瓣图书数据集下，Precision 值的对比

Fig. 2　In the Douban books data-set, the Precision value contrast

其次，采用召回率(Recall)来测评本文提出的推荐算法的召回率可行性。召回率是指推荐对了的数据占全集的多少。

$$Recall=\frac{\text{推荐的正确信息条数}}{\text{样本中的信息条数}}。\tag{10}$$

表2 豆瓣图书的用户标签数据集

Tab.2 The users tags dataset of Douban books

时间戳	userID	书ID号	书名	标签及其标注数量									
888 551 648	165	1 040 771	达·芬奇密码	悬疑	15 752	达·芬奇密码	13 058	小说	11 553	宗教	9 048	推理	8 602
892 002 353	216	1 090 043	倾城之恋	张爱玲	20 521	倾城之恋	9 988	小说	6 051	爱情	4 874	中国文学	3 240
875 636 389	224	3 211 779	放学后	东野圭吾	15 501	推理	10 073	日本	6 998	小说	3 484	推理小说	3 373
875 637 536	230	1 008 145	围城	钱钟书	31 264	围城	20 758	小说	13 741	中国文学	10 952	经典	9 712
883 599 914	348	1 082 154	活着	余华	23 498	活着	14 578	小说	10 327	中国文学	7 105	人生	5 832
883 268 170	349	1 786 670	百年孤独	百年孤独	16 861	西亚·马尔克斯	10 381	魔幻现实主义	8 778	外国文学	6 521	小说	4 491
892 133 057	463	1 041 482	万历十五年	历史	14 257	黄仁宇	10 405	万历十五年	5 881	明朝	4 455	中国历史	2 551
891 353 950	565	1 770 782	追风筝的人	追风筝的人	27 081	小说	16 162	阿富汗	16 134	卡勒德·胡赛尼	12 996	人性	12 915
891 352 261	576	1 017 143	不能承受的生命之轻	米兰·昆德拉	32 140	不能承受的生命之轻	19 397	外国文学	10 837	小说	9 854	捷克	6 739
882 140 166	138	1 059 406	幻城	郭敬明	12 264	幻城	6 407	小说	5 852	奇幻	3 442	青春	2 221
882 140 318	292	1 084 336	小王子	小王子	32 394	童话	25 325	圣埃克苏佩里	12 710	法国	10 896	经典	9 195
880 131 728	536	2 567 698	三体	科幻	16 019	刘慈欣	12 491	三体	6 625	中国	3 837	小说	3 606
880 129 305	583	1 461 903	何以笙箫默	顾漫	12 961	何以笙箫默	11 212	小说	7 693	爱情	6 543	言情	5 631
880 129 489	628	2 256 039	杜拉拉升职记	职场	14 032	杜拉拉升职记	8 950	小说	7 753	外企	3 610	成长	3 343
878 887 062	851	3 813 669	民主的细节	刘瑜	12 194	民主	10 971	政治	9 420	美国	6 954	随笔	4 554
874 787 017	906	20 427 187	看见	柴静	20 362	纪实	11 302	随笔	6 803	看见	5 591	中国	3 974
885 852 817	72	1 200 840	平凡的世界(全三部)	路遥	16 070	平凡的世界	15 282	小说	8 063	中国文学	6 556	人生	5 349
891 964 326	313	1 045 818	苏菲的世界	哲学	14 524	苏菲的世界	8 229	哲学入门	6 703	乔斯坦·贾德	3 184	外国文学	2 650
888 602 808	324	1 873 231	明朝那些事儿(壹)	历史	12 665	明朝那些事儿	8 796	当年明月	4 705	明朝	3 294	小说	2 605
890 687 473	496	104 265	挪威的森林	村上春树	38 902	挪威的森林	18 486	小说	11 717	日本	10 683	日本文学	10 092
881 105 817	560	3 995 526	目送	龙应台	17 380	散文	9 852	亲情	6 307	随笔	5 456	台湾	3 450
875 975 520	631	1 775 691	少有人走的路	心理学	16 439	少有人走的路	9 449	成长	7 903	心理	5 900	励志	5 173
879 434 688	889	4 242 172	天才在左疯子在右	心理学	15 002	精神病学	10 174	心理	4 605	哲学	4 131	精神病	3 256
888 067 096	20	1 858 513	月亮与六便士	手姆	12 641	月亮与六便士	6 760	小说	6 220	英国	4 258	外国文学	3 909
877 052 400	59	3 259 440	白夜行	东野圭吾	19 552	推理	14 756	日本	13 774	东野圭吾	12 488	小说	5 780
884 304 936	340	1 007 305	红楼梦	红楼梦	15 882	古典文学	12 344	曹雪芹	8 736	经典	5 580	小说	4 945
884 305 165	367	4 742 918	村上春树	树上春树	22 449	日本	7 672	小说	6 590	1Q84	6 469	日本文学	4 587
888 206 254	426	1 400 705	情人	杜拉斯	15 833	情人	10 971	法国	6 701	小说	6 620	外国文学	4 764

图 3 是在豆瓣图书数据集下，Recall 值对比的实验结果。

结合覆盖率（Coverage）来测评推荐算法的覆盖率可行性。覆盖率表示推荐的物品占了物品全集空间的比例，用（11）式表示。Coverage 值对比的实验结果如图 4 所示。

$$Coverage=\frac{\text{推荐的物品}}{\text{物品资源全集}}。\tag{11}$$

从以上实验结果可以得出，采用准确率、召回率和覆盖率等 3 个测量标准，对于不同数量的数据源，TBCF 与 TCF 的实验对比结果有以下 3 点：

1）TBCF 比 TCF 的精确度要高；

2）根据数据量的递增，召回率也成正比增长；

3）根据资源数量的增加，覆盖范围更加精确更加全面。

通过以上实验分析，根据本文中提出的算法在准确率、召回率和覆盖率等 3 方面的测量标准，基于时间加权标签的协同过滤推荐算法对于推荐资源的精确度有所提高，并且能够降低算法的时间复杂度。

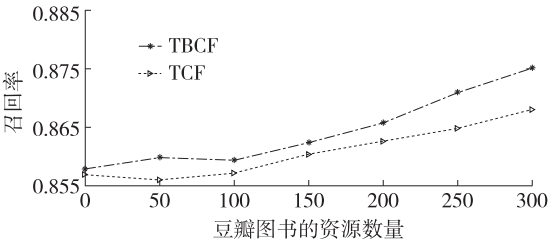

图 3　豆瓣图书数据集下，Recall 值的对比

Fig. 3　In the Douban books data-set, the Recall value contrast

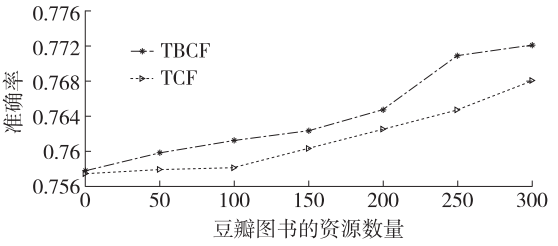

图 4　豆瓣图书数据集下，Coverage 值的对比

Fig. 4　In the Douban books data-set, the Coverage value contrast

5 结语

文中在基于标签和协同过滤的基础上加入了时间权重，时间对于资源的意义变化是很重要的因素。因为时间对于用户的永久兴趣负面影响少些，而对于用户的临时兴趣的负面影响多些。文中以“豆瓣读书”用户的标签数据集为实验数据源，对基于标签的协同过滤推荐算法进行了改进，改进的结果能够提高推荐的准确性，使时间复杂度有所降低。标签对于兴趣固定的用户来说是丰富了资源，对于临时兴趣的用户来说是浏览下新的知识并不对过去的资源有太多的兴趣。基于此情况，加入了时间权重，能够更加全面提高对用户兴趣的预测。有效地提高了推荐的准确性，获得了更好的推荐效果。

参考文献：

[1] 高连花. 基于社会化标签的个性化信息服务研究[D]. 武汉：华中师范大学，2012.
Gao L H. The personal information services based on social labels[D]. Wuhan: Central China Normal University, 2012.

[2] 王岚，翟正军. 基于时间加权的协同过滤算法[J]. 计算机应用，2007，27(9)：2302-2303.
Wang L, Zhai Z J. The collaborative filtering algorithm based on time-weighted[J]. Journal of Computer Application, 2007, 27(9): 2302-2303.

[3] Kohi A, Ebrahimi S J, Jalali M. Improving the accuracy and efficiency of tag recommendation system by applying hybrid methods[C]//International conference on computer and knowledge engineering. Mashhad: IEEE, 2011: 242-248.

[4] 吴杰，冯锋. 综合用户偏好和优先新品推荐的协同过滤算法[J]. 计算机应用与软件，2014，10(31)：285-287.
Wu J, Feng F. A collaborative filtering algorithm incorporating users preference and recommending new products first [J]. Computer Application and Software, 2014, 10(31): 285-287.

[5] 张莉，薛羽青. 结合用户判断力和相似性的协同推荐算法[J]. 计算机科学，2014，41(11A)：320-322.
Zhang L, Xue Y Q. The collaborative recommnendation algorithm combining user's judging power and similarity[J]. Computer Science, 2014, 41(11A): 320-322.

[6] 郭彩云，王会进. 改进的基于标签的协同过滤算法[J]. 计算机工程与应用，2015，12(1)：1-8.
Guo C Y, Wang H J. The improved collaborative filtering algorithm based on tags[J]. Computer Engineering and Applications, 2015, 12(1): 1-8.

[7] 易辽宏. 基于标签张量的个性化推荐方法研究[D]. 辽宁：辽宁大学出版社，2014.
Yi L H. The research of the tag tensor block component decompositions[D]. Liaoning: Liaoning University Press, 2014.

[8] 赵海燕，侯景德，陈庆奎. 结合时间权重与信任关系的协同过滤推荐算法[J]. 计算机应用研究，2015，12：3565-3568.
Zhao H Y, Hou J D, Chen Q K. The collaborative filtering recommendation algorithm combining time weight and trust relationship[J]. The Journal of Computer and Application Research, 2015, 12: 3565-3568.

[9] Ozsoy M G, Polat F. Trust based recommendation systems [C]//Proc of IEEE/ACM international conference on ad-

vances in social networks analysis and mining. US:US7991841, 2013:1267-1274.

[10] 孙克,王新华,窦羚源. 面向社会关系的移动用户好友推荐算法[J]. 山东师范大学学报:自然科学版,2015,30(2):11-15.
Sun K, Wang X H, Dou L Y. The recommended algorithms based on mobile friends who faced social relationships[J]. Journal of Shandong Normal University: Natual Science,2015,30(2):11-13.

[11] 张明,郭娣. 一种优化标签的矩阵分解推荐算法[J]. 计算机工程与应用,2015,51(32):119-124.
Zhang M,Guo D. Matrix factorization recommendation algorithm of optimizing tags[J]. Computer Engineering and Applications,2015,51(23):119-124.

[12] 王海雷,牟雁超,俞学宁. 基于协同矩阵分解的社会化标签系统的资源推荐[J]. 计算机应用研究,2013,30(6):1739-1741.
Wang H L,Mou Y C,Yu X N. Resource recommendation in social tagging system based on collaborative matrix factorization[J]. Computer and Application Research,2013,30(6):1739-1741.

[13] 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):96-99.
Gu Y R,Chen M. One tag time-weighted recommend approach on tripartite graphs networks[J]. Computer Science,2012,39(8):96-99.

[14] 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.
Cai Q,Han D M,Li H S,et al. Personalized resource recommendation based on tags and collaborative filtering[J]. Computer Science,2014,1:69-71.

[15] 李月臣,何志明. 重庆市 MODIS/NDVI 时间序列数据优化研究[J]. 重庆师范大学学报:自然科学版,2015,32(2):32-37.
Li Y C,He Z M. The research on Chongqing MODIS/NDVI time series and data optimization[J]. Journal of Chongqing Normal University: Natural Science,2015,32(2):32-37.

[16] 张斌,张引,高克宁,等. 融合关系与内容分析的社会标签推荐[J]. 软件学报,2012,23(3):476-488.
Zhang B,Zhang Y,Gao K N,et al. Combining relation and content analysis for social tagging recommendation[J]. Journal of Software,2012,23(3):476-488.

[17] Mondal A,Trestian I,Qin Z,et al. P2P as a CDN:a new service model for file sharing[J]. Computer Networks,2012,56(9):3233-3246.

[18] Xia X,Zhang S,Li X. A personalized recommendation model based on social tags[C]// International workshop on database technology & applications. Wuhan:IEEE,2010:1-5.

[19] Rachev S T. The Monge-kantorovich mass transference problem and its stochastic applications[J]. Theory of Probability & Its Applicatons,1985,29(4):647-676.

[20] Hitchcock F L. The distribution of a product from several sources to numerous localities[J]. J Math Phys,1941,20(2):224-230.

[21] 陶新民,郝思媛,张冬雪,等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报:自然科学版,2013,25(1):101-110.
Tao X M,Hao S Y,Zhang D X,et al. The review on unbalanced data classification algorithm[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science,2013,25 (1):101-110.

[22] Du W H,Rau J W,Huang J W,et al. Improving the quality of tags using state transition on progressive image search and recommendation system[C]//IEEE international conference on systems,man,and cybernetics. Seoul: IEEE,2012:3233-3238.

The Research in Collaborative Filtering Recommendation Algorithm Based on Time-tags

SONG Weiwei[1,2], YANG Degang[1], ZHENG Min[1]

(1. College of Computer and Information Science, Chongqing Normal University, Chongqing 401331;
2. Henan Industry and Trade Vocational College, Zhengzhou 450053, China)

Abstract: The traditional collaborative-filtering algorithm is based on users- rating data resource to reflect user's interests and calculate the resources similarity, the aim is to decide whether to recommend items to potential users or not. But it ignored the characteristics of users and resources, the users' understanding and interests for the resources may be changed at different times. To solve this problem, this paper puts forward the recommendation algorithm which is based on time-tag information. The main idea is that labels can be tagged by the users according to their preferences for free resources, it reflect the users' interests and the resources characteristics information. According to the multidimensional relationship of the users, time, tags, and sources, which can generate the tags-feature-vectors of users and resources, then calculate the user's interests and the resources' similarity in different time, the aim is to predict users' preferences for the further prediction and appropriate recommendation. The experimental results show that it can improve the recommendation accuracy, and obtain better recommendation results.

Key words: time-tags; recommendation; individuality; accuracy

(责任编辑 游中胜)