基于用户阅读时间-频次行为的书籍推荐方法

曹 斌 龚佼蓉 彭宏杰 赵立为 范 菁
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 随着电子阅读在近年来的兴起,通过研究用户对电子书籍的喜好,利用协同过滤推荐算法向用户进行个性化的书籍推荐具有实际应用价值,也成为了推荐系统研究中的重要内容。但当前很多书籍推荐应用中都存在缺少用户评分数据甚至没有用户评分的情况,使得传统协同过滤推荐方法的应用受阻。为解决此问题,通过分析处理用户阅读数据的相关行为数据,将此类行为数据通过时间-频次模型建模并得到用户-书籍评分矩阵,并利用该评分进一步实现基于用户的协同过滤书籍推荐算法。实验结果表明,改进的书籍协同过滤推荐算法的时间-频次模型能够提高书籍的推荐效果,具有实践研究意义。

关键词 协同过滤,推荐系统,用户评分矩阵,用户行为,时间-频次
中图法分类号 TP391 **文献标识码** A

Recommending Books Based on Reading Time and Frequency

CAO Bin GONG Jiao-rong PENG Hong-jie ZHAO Li-wei FAN Jing
(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract With the rise of electronic reading in recent years, the use of collaborative filtering(CF) recommendation algorithm to recommend user personalized books has practical application value, and has become the important research content in the study of recommender systems. But many current e-reading systems for book recommendations lack of users' rating data, which hinders the application of CF. To address this, we analyzed the massive users' reading behavior, and proposed a reading time-frequency(T-F) model to profile the users' interests to the book. Thus, the implicit ratings matrix can be derived from this model and then classical CF algorithm could be used in a natural way. The experimental results show that the user based CF with our proposed T-F rating model can improve the recommendation effectiveness, which is feasible for real scenarios.

Keywords Collaborative filtering, Recommendation system, User ratings matrix, User behavior, Time-frequency

1 引言

近年来,随着移动互联网持续高速的发展,用户能够从终端得到的信息数量趋于海量,推荐系统也在众多行业中得到了广泛的应用,其价值在于能够有效帮助用户获得个性化的信息。电子书籍数量的指数型增长,也使得书籍推荐系统成为推荐系统中一个重要的组成部分。如何将电子书籍准确地推荐给目标用户,让用户能够在书海中及时准确地获得所喜爱的书籍,成为当前一个热门的研究领域。目前书籍推荐系统所采用的主流推荐技术主要有两种,一是基于内容的推荐算法,二是协同过滤推荐算法。

基于内容的推荐是根据书籍信息与用户偏好的相关性对用户做出推荐,并通过比较资源和用户描述文件,推荐与用户兴趣相似的书籍给目标用户。但由于目前并不是所有数据集都含有用来描述书籍的数据,仅利用书籍内容相似度给目标用户推荐的书籍都是统一类型的,过于单一化,因此利用基于内容的推荐算法无法向用户提供多样化的书籍对象。

协同过滤算法则弥补了以上算法的不足。协同过滤算法作为一类经典的推荐算法,主要基于这样一个假设:若某些用户对一些项目的评分比较相似,就认为他们对其他项目的评分也会比较相似[1]。根据这种思想,利用协同过滤进行书籍推荐的主要过程是,基于用户对书籍的评分建立用户评分矩阵,计算用户之间或书籍之间的相似度,找出相似度大的用户群或书籍群,根据这类用户群或书籍群对用户或书籍进行相关的推荐。但协同过滤推荐算法也有它本身的局限性,即如果用户评分缺失,将无法进行准确的推荐。

Yin等人[2]通过观察分析当前的社交应用发现,用户在使用app过程中,对社交平台中的内容或项目进行评分是非常不积极的。此外,即使在有评分的情况下,也存在如用户未能根据自己真实喜好进行评分等问题,因此评分不能完全反映用户的真实喜好,进而无法很好地解决以上描述的用户评分缺失的情形。目前的一些电子阅读软件应用也存在着同样

本文受国家自然科学基金(61173097),浙江省重大科技专项重大工业项目(2013C01112),杭州市重大科技创新专项(20132011A16)资助。
曹 斌(1985—),男,博士,讲师,主要研究方向为大数据、中间件,E-mail:bincao@zjut.edu.cn;龚佼蓉 女,硕士生,主要研究方向为大数据、推荐系统;彭宏杰(1992—),男,硕士生,主要研究方向为大数据;赵立为(1991—),男,硕士生,主要研究方向为空间数据库;范 菁(1969—),女,博士,教授,主要研究方向为服务计算、虚拟现实,E-mail:fanjing@zjut.edu.cn(通信作者)。

的问题，即对书籍的评分数据较为稀少甚至缺失，因此传统协同过滤推荐算法在书籍推荐上的进一步应用受到了限制。

为解决上述用户评分缺失的问题，李建廷等人[3]通过用户的使用记录，将简单的用户行为转化为用户评分数据，即有用户使用记录的为1，没有记录的为NULL，构造0-1评分矩阵，这样的评分方式也能够用来进行电子书籍的推荐，但因过于简单导致准确度较低。不同用户阅读同一本书所花时间的不同，往往代表他们对该书的喜爱程度是不一样的。

为了解决用户评分数据不准确所带来的问题，本文工作的主要思想是将用户的各类阅读行为转化为用户的精细评分，进而填充整个用户项目评分矩阵来实现书籍的协同推荐，从而提高推荐性能。基于此，本文的主要贡献是对用户的阅读行为做进一步提取和建模，并将这些行为量化成为协同过滤需要的评分矩阵。通过分析用户在某款手机阅读软件（出于隐私考虑，此处略去软件名称）上的行为日志信息，对比几种用户阅读行为对用户兴趣所产生的影响，选择从用户阅读时间和阅读次数两方面来进一步刻画用户的行为，分别根据用户阅读书籍时所花费的平均时间建立了Time-based评分模型，根据阅读该本书籍的频次建立了Frequency-based评分模型。然后在此基础上，建立Time-Frequency的混合模型（简称T-F模型），用来综合评估用户对该本书籍的喜好程度，计算用户-书籍评分矩阵，再利用该矩阵进行基于用户的协同过滤推荐。最后，为了验证本文方法的有效性，通过与基于0-1评分矩阵的用户协同过滤推荐的实验进行对比，发现本文所提出的T-F模型在推荐准确度上提高了10%以上，证明了将用户阅读书籍的时间和频次转化成用户对项目的有效评分更加符合用户行为的真实情况，同时也解决了缺少用户评分所带来的问题。

2 相关工作

在推荐系统研究中，与本文相关的研究工作主要分为两个领域：(1)利用用户行为数据进行推荐；(2)提高推荐准确度的方法。

用户行为数据最普遍的存在形式是日志，因此它很容易被推荐系统用来作为分析用户偏好并做推荐的依据。用户行为在个性化的推荐系统中主要可以分为两种[4]，一种是显性反馈行为（explicit feedback）[5]，另一种则是隐性反馈行为（implicit feedback）。显性反馈行为包括用户明确表示对某一物品的喜好行为，如用户对该物品进行了评分，或者选择喜欢还是不喜欢。目前大多推荐系统都是利用这种基于显示评分的方式进行推荐。但实际中用户往往并不会频繁主动地对项目进行评分，通过对大量用户行为数据的分析研究表明，用户在访问了100个项目中，平均只会对2个项目进行评分[2]。在使用应用的行为过程中，如果没有遇到自己非常喜欢或者特别讨厌的内容或项目，用户往往会采用不进行任何评分的沉默式浏览行为方式。

隐性反馈行为则统一指不能明确反映用户喜好的所有行为。最具有代表性的就如通过在某项目上的页面停留时间来判断用户对该项目的喜好程度[6]，以及Yin等人从心理学角度出发研究逗留时间与用户阅读兴趣间的关系。本文工作属于这一类研究，但现有对隐性行为研究的工作所考虑到的行为往往较为单一，只考虑单个比较重要的行为来展开推荐。另外，这类工作只是在已知用户真实兴趣的情况下做进一步的研究来验证隐性行为对推荐结果有所影响，很少有研究将这一类隐性行为转换为用户的主要兴趣评分，进而直观地产生系统推荐。Yin等人虽然对阅读时长做了喜好程度的评分转化，但其只能通过标识喜欢与不喜欢两种程度来反映用户的两种状态。而本文则对不同程度的喜好进行了具体量化和评分转化，即不同的评分值可以表示不同的喜好程度，进而能够更好地区分用户，最后可以帮助提高推荐准确度。

评分预测算法的核心则是提高推荐准确度方法，因此当前有很多关于评分预测算法的研究。评分预测算法主要可以分为以下几类：平均值算法、基于邻域的方法、隐语义模型与矩阵分解模型。

平均值预测用户对物品的评分算法是最简单的一种评分预测算法，主要有计算全局平均值、计算用户评分平均值、物品评分平均值、用户分类对物品分类的平均值等几种算法。本文所采用的基于邻域的评分预测方法中，也有采用计算评分平均值的方法来获得相似用户，是在平均值预测的基础上进一步完成基于邻域的评分预测算法的。

基于邻域的方法则是本文所采用的评分预测方法。目前主要的两种算法就是基于用户的邻域算法和基于物品的邻域算法[7]。另外，通过研究用户项目内容排名[8]、相似性排名的测量[9]等方法，可以提高邻域算法的推荐准确度。

在获得推荐结果后，利用RMSE（均方根误差）作为评测指标，来比较算法的推荐度等。均方根误差RMSE公式如下：

$$RMSE=\sqrt{\sum_{(u,i)\in T}(r_{ui}-\hat{r}_{ui})^2}/|Test| \quad (1)$$

其中，$Test$表示测试集的数据，r_{ui}表示用户u对书籍i的真实评分，$\hat{r}_{ui}$表示推荐算法预测用户u对书籍i的评分。均方根误差值越小，表示推荐结果越准确。

隐语义模型与矩阵分解模型的主要思路是通过降维的方法将评分矩阵补全，进而解决评分稀疏的问题，可以分为传统的SVD分解（奇异值分解）[10]与Simon Funk的SVD分解[11]等。Simon Funk的SVD分解在之后被称为Latent Factor Model，简称LFM，即隐语义模型。它的思想其实就是可以直接通过训练集中的观察值利用最小化RMSE来学习P，Q分解矩阵，进而最小化训练集中的预测误差。LFM在传统SVD分解的基础上，通过求参数的偏导数找到最速下降方法，进而利用迭代法不断优化参数，解决了传统SVD分解方法复杂度高等问题。

虽然隐语义模型也能够完成向目标用户推荐书籍的任务，但由于隐语义模型需要用户评分数据的支撑且较邻域算法复杂，因此本文不采取此类方法，而采用传统的基于邻域的算法来完成本文的实验。

3 用户阅读行为数据分析与准备

我们收集了一款手机阅读软件中4万多用户连续30天的阅读日志信息，该数据集中包含有19000余本书籍。

该数据集中没有任何用户对书籍的评分数据，但包含了用来描述用户信息、时间、用户行为这3大类数据的共24个字段，如用户名、设备IMEI(International Mobile Equipment Identity)，用户阅读的具体行为操作，如打开书、翻页等。利用该数据集可以知道具体的某一用户在某一个时间段看了哪一本书，另外也能通过统计得到用户在一天中总共阅读了多少次，以及针对某本书总共阅读了多长时间、每次持续多长时间等具体信息数据。

通过整理分析上述用户阅读书籍的行为数据，统计得到用户主要的阅读书籍行为有18种，如表1所列。

表1 用户阅读书籍行为数据表

用户主要行为	具体eventId数据
阅读书籍(18种)	书籍封面点击、封面翻页、封面滑动点击、书籍目录点击、书籍正文翻页、新书籍目录点击、新书籍正文翻页、选择书籍正文主题、书籍正文进入时主题、添加本地书籍、书籍翻页、书籍翻页效果、推送消息点击放入书架、添加至购物车、添加模块、删除模块、推荐标签、推送消息打开

将这18种阅读书籍行为进行归类。其中，书籍封面点击、封面翻页、书籍目录点击、书籍正文翻页、新书籍目录点击、新书籍正文翻页、选择书籍正文主题、书籍正文进入时主题、书籍翻页等类似的共9种阅读行为可以归类到阅读书籍操作这一大类，通过这些字段能够得到用户的阅读时长和阅读次数等信息。添加本地书籍、推送消息点击放入书架、添加至购物车等其余9种用户行为可以归类到用户的书籍喜爱程度表现这一类。但由于在本数据集中，这类操作并不是系统强制用户必须做的，只有极少的用户会选择这类操作，因此无法用这些行为数据来有效标识用户对某本书的喜好程度，在此不做赘述。基于此，本文将考虑能够客观反映用户阅读书籍喜好的9种行为数据来进行评分推荐。

通过分析用户阅读天数和阅读书籍次数两类数据，发现一部分的用户阅读一本书籍的次数为1，阅读天数为1，时间也只是短短的几十分钟。这种现象的存在可能有着多方面的因素：(1)用户对这本书并不感兴趣，读了一次之后就发现并不是自己喜欢的书；(2)用户因为时间关系再也没有读过这本书；(3)因为书籍本身的原因，如内容的长度、版权、更新等问题，引起了用户的一次性阅读。以上这些原因都有可能导致这类数据的产生。这类数据对于真实反映用户喜好程度的意义并不大，因此我们需要考虑将这部分数据进行过滤，即本文工作中，当用户阅读某本书籍的天数和次数均为1时，将这类用户行为数据进行清洗，不把此类数据作为建模的数据进行处理。

为得到用户的阅读频次，对源数据进行了进一步的预处理。在源数据中，字段SessionStep是用来记录用户每次使用app时所产生的步数。例如表2所列，前4行和后3行的用户行为代表两次用户阅读行为。用户登录后点击了1次书籍封面，并进行了2次书籍正文翻页的行为，此时SessionStep对每一条步骤都进行了记录，分别从1到4。而下一条数据产生的时间和这次用户登录后所发生的一系列行为时间并不连贯，且SessionStep计数又从1开始，那么我们就认为这连续的4条行为是用户阅读一次书籍的行为，而当SessionStep重新从1开始计数时，用户的行为将是另外一次阅读行为。因此，我们能够通过统计SessionStep为1的个数，结合用户名称和书籍Id，来得到用户阅读书籍的次数。

表2 用户阅读书籍行为计数举例

UserName	eventId	SessionStep	Time
A	Login	1	07-01 12:00:00
A	书籍封面点击	2	07-01 12:00:11
A	书籍正文翻页	3	07-01 12:00:15
A	书籍正文翻页	4	07-01 12:05:24
A	Login	1	07-01 16:12:03
A	书籍封面点击	2	07-01 16:12:56
A	书籍正文翻页	3	07-01 16:13:09

另外还存在一种现象，即用户阅读书籍次数明显增加，但每次阅读时间只有短短的几秒，之后又会伴随一个登录行为，然后又继续之前的阅读。我们认为这是由网络不稳定或其他因素导致的，而这些行为所发生的时间实际上是具有连续性的，因此在这种情况下我们认为用户只阅读了一次。

通过计算用户阅读书籍的costTime和阅读同一本书的日期并记录用户阅读该书籍的行为次数(即统计SessionStep为1的次数)，就能获得用户对该本书籍的阅读时间和阅读次数。本文下面将主要从用户阅读时间和用户阅读书籍次数这两方面着手，来建立用户评分模型，产生推荐。

4 阅读时间与频次模型

4.1 时间模型(Time Model)

现实生活中我们通常可以认为，用户阅读时间越长，用户对书籍内容的兴趣度也就越高。目前也有一些研究表明，用户行为的驻留时间和用户的兴趣度有着密切关系[12,13]。孙光福等人在文献[12]中提出一种基于时序行为的SequentialMF推荐算法，其证明了用户驻留时间与兴趣度成正比关系。基于此，本文考虑将用户的阅读时间行为转化为用户对书籍的评分，我们认为，在普通情况下，用户平均每天的阅读时间越长，评分也应越高。当平均每天阅读时间趋于最大值时，就说明用户的兴趣有着很好的保持度，也表示用户会对该书籍给予高评分。同时，对阅读同一本书籍的所有用户来说，我们定义平均每天所花时间最长的用户对这本书的评分为1，其余用户所花平均时间越接近最长时间，评分也就越高，具体相关定义如下。

定义1 定义 $Interest(t)_{ub}$ 为用户 u 阅读书籍 b 所花时间对应的评分数据。t_i 表示用户 u 在 i 天阅读书籍 b 所花的时间，d 表示用户 u 看书籍 b 的天数，$i\in[1,d]$，$t_{\max}(b)$ 表示所有用户阅读书籍 b 所花平均时间中的最大值。构造以下公式来表示用户阅读某本书籍所花时间的评分值：

$$Interest(t)_{ub}=(\sum_{i=1}^{d}t_i)/(d*t_{\max}(b)) \tag{2}$$

例如，用户 A 在3天内都看了书籍 b，第一天共花了180min，第二天共花了300min，第三天共花了240min。而用户 B 是所有阅读书籍 b 的用户中所花平均时间最长的，平均每天阅读时间为300min，因此 $t_{\max}(b)$ 也为300。$Interest(t)_{Ab}=(180+300+240)/(3*300)=0.8$，由此得到用户 A 对书籍 b 的评分为0.8，用户 B 对书籍 b 的评分则为1。

由以上公式可知，由于 $(\sum_{i=1}^{d}t_i)/d$ 为平均每天阅读时间，区间为 $[0,t_{\max}(b)]$，因此 $Interest(t)_{ub}\in[0,1]$。

另外，用户在阅读过程中，还会经常碰到推送广告等导致阅读时间很短或者因临时有事离开而页面却一直处于打开状态导致后台系统记录的阅读时间变长的情况。因此需要设定一个时间阈值，将不正常的时间清洗掉。现有研究表明，一般阅读时长小于3～5s则无法反映用户的真实兴趣，因此我们将阅读一屏幕文字所花时间的最小阈值设置为5s，最大阈值则设置为[14]：

$$T_{\max}=(\sum_{i=1}^{m}T_i)/m \tag{3}$$

其中，$T_1,T_2,\cdots,T_m$ 表示用户每次阅读书籍的时长，m 表示用户阅读该书籍的总次数，T_i 表示用户阅读书籍所花的所有时间中前10%长的阅读时间，当用户阅读时长超过该阈值时，取 $t_{\max}(b)=T_{\max}$。将最大阈值设为时长在前10%时间的平均值，可以过滤掉时间数值过大的情况，去除用户阅读书籍时的异常现象，保证时间数据在一个正常的阈值范围中。

通过以上公式得到用户评分矩阵，再利用基于用户的协同过滤算法得到推荐结果，并将推荐结果同0-1评分矩阵的结果进行RMSE(均方根误差)计算和对比，误差值越小，代表推荐结果与真实结果越接近，因此表示推荐结果越准确。对比结果如图1所示。

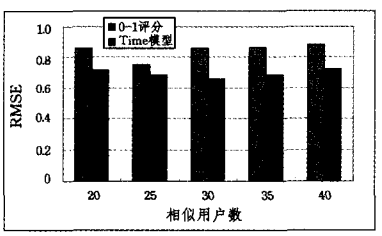

图1 Time模型与0-1评分方法结果对比

由图1可知，不论相似用户数的取值为多少，使用Time模型计算得到评分所获得的RMSE均比0-1评分矩阵所获得的RMSE要小。因此，相较于基于0-1矩阵，利用用户驻留时间进行对书籍的评分构建(定义1)并执行推荐，可产生较为准确的结果。

4.2 频次模型(Frequency Model)

用户行为次数也对用户评分有着一定的影响，Wu[15]通过计算基于用户浏览行为和访问次数的兴趣度，发现在一般情况下，用户访问次数越多，用户对项目的兴趣度也就越高。付关友等人[16]提出了用户兴趣度与阅读时间和访问次数的线性方程，证明了阅读时间和访问次数与用户兴趣度成正比关系。但他们的工作是通过已知的用户评分数据以及用户访问次数来进行推荐，进而证明用户的访问次数与用户的兴趣度成正比，并没有通过用户的访问次数对推荐结果产生影响。本文则希望能够将用户访问次数通过Frequency模型转换为用户对书籍的评分，进而对目标用户产生书籍推荐。因此在他们验证工作的基础上，我们通过将用户每天阅读书籍的行为数据转化为用户阅读书籍的次数，构成频次模型，来继续验证此类工作是否具有成正比的共同特征。

由于用户每天用于阅读的时间有长有短，因此取用户平均每天阅读的次数来作为用户阅读某一本书籍的评分依据。在大部分情况下，用户看的次数越多，可以在一定程度上表明用户对这本书的喜好程度越强，因此可以考虑频次系数是否对推荐有影响。

下面给出有关频次模型的定义：

定义2 定义 $Interest(f)_b$ 为用户 u 阅读书籍 b 所花次数对应的评分数据。f_i 表示用户 u 在 i 天阅读书籍 b 所花的次数，$i\in[1,d]$，$f_{\max}(b)$ 表示所有用户阅读书籍 b 时所花平均阅读次数最多的一个。构造以下公式来表示用户阅读某本书籍所花次数的评分值：

$$Interest(f)_{ub}=(\sum_{i=1}^{d}f_i)/(d*f_{\max}(b)) \tag{4}$$

例如，用户 A 阅读书籍 b 花了3天时间，第一天阅读了2次，第二天阅读了1次，第三天阅读了6次，平均阅读书籍 b 所花的次数为3。用户 B 为阅读书籍 b 所花平均次数最多的用户，为每天10次，那么 $f_{\max}(b)$ 取为10，$Interest(f)_{Ab}=(2+1+6)/(3*10)=0.3$，由此得到用户 A 对书籍 b 的评分为0.3，用户 B 对书籍 b 的评分为1。

同样地，为了使用户阅读书籍平均所花次数所对应的评分数据控制在[0,1]区间内，且方便计算处理，将用户阅读某本书籍所花的平均次数除以阅读该书籍的所有用户中所花次数最大的数值。与 $Interest(t)_{ub}$ 相似，由上述公式可得：$Interest(t)_{ub}\in[0,1]$。

通过以上评分公式得到用户书籍的评分矩阵，同样利用基于用户的协同过滤算法得到推荐结果，将产生的推荐结果与0-1评分矩阵进行对比，得到如图2所示的对比结果。

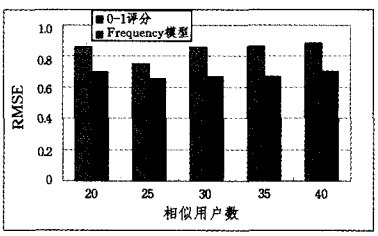

图2 Frequency模型与0-1评分方法结果对比

由图2可以看到，左侧一栏的数值均表示由0-1评分矩阵做推荐所获得的RMSE(均方根误差)，右侧一栏的数值则表示使用Frequency模型计算评分做推荐所获得的RMSE，不论相似用户数取多少，使用Frequency模型所产生推荐结果的RMSE值均比0-1评分矩阵所获得推荐结果的RMSE要低。由此得出，将用户阅读次数转化为用户评分，产生的推荐结果要比0-1评分所得到的推荐结果更加准确。

5 基于T-F模型评分的书籍协同过滤推荐

根据第4节的介绍，用户的阅读时间和阅读次数都在一定程度上与用户的兴趣度评分成正比。但如果单从驻留时间这一方面来评判用户对该书籍的偏好，会很容易受到其他一些因素的干扰。如：用户在阅读某本书籍的过程中临时有事离开，而界面仍然停留在所看的书籍上，停留时间却仍在持续增加。单一的时间模型只从时间停留的长度上来判断用户对书籍的评分不够准确。同样，若仅通过用户阅读次数来计算用户对书籍的评分，也存在一些不足。例如用户网络不稳定会使用户反复登录阅读软件，导致用户在这一段时间内的阅读次数增加，但用户每次阅读书籍只有几秒。虽然数据显示用户阅读书籍的次数增加，但这样的行为记录不能真实反映用户的阅读喜好。

为了提高用户对书籍喜好程度表示的准确度，本文将用

户阅读的持续时间长度和频繁次数两方面结合来计算用户对书籍的客观评分，提出了时间(Time)-频次(Frequency)模型，简称 T-F 模型。由于 Time 模型和 Frequency 模型中所得到的评分数据均为表示用户兴趣度的数值，没有单位限制且数值范围都在[0,1]中，因此考虑可以将 Time Model 和 Frequency Model 按照一定权重相加，从而得到将用户行为转化为用户 u 对书籍 b 评分的线性方程。$Interest_{ub}$ 表示用户 u 对书籍 b 的预测分数：

$$Interest_{ub}=[w_1(\sum_{i=1}^{d}t_i)/t_{\max}(i)+w_2(\sum_{i=1}^{d}f_i)/f_{\max}(i)]/d \tag{5}$$

由于不同用户阅读不同书籍所产生的 $t_{\max}(i)$ 与 $f_{\max}(i)$ 都不相同，时间 t 与频次 f 对总体评分的影响也会不同，因此需要定义 w_1、w_2 分别为 T_b、F_b 的参数权值，并通过改变 w_1、w_2 的大小来整体控制阅读时间、阅读次数对产生的评分的权重影响，$w_1+w_2=1$；$w_1,w_2>0$。

基于 T-F 模型的协同过滤书籍推荐系统流程如图 3 所示。其中，虚线框内所表示的用户阅读行为评分子系统即为本文的主要贡献。将项目日志文件中的阅读行为数据进行清洗、过滤、分类，在基于分析数据所得结论上再进行进一步的数据清洗，得到所需的用户名称、事件类型等数据。由于源数据中并没有模型中所提到的数据，因此通过将各数据进行重新整合计算，得到了模型中所需的平均阅读天数、平均阅读次数等最终数据，并将模型进行重组，得到了最终的 T-F 模型。通过 T-F 模型计算出用户对书籍的评分，得到用户-书籍评分矩阵，做到了将用户隐式行为转化为用户对书籍的评分，解决了原始数据集中不存在用户评分所带来的困难，进而展开书籍协同过滤推荐计算。

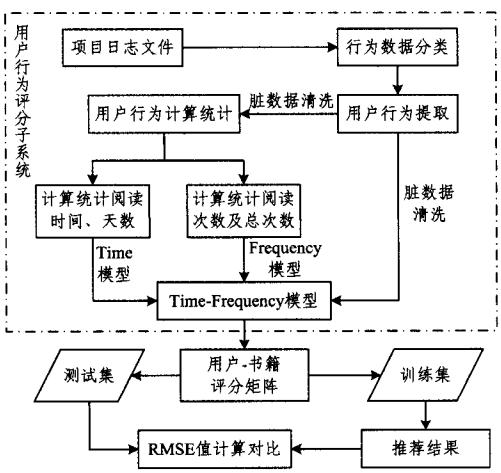

图 3　基于 T-F 模型的协同推荐架构

在获得评分数据后，将通过计算相似度来进一步实现推荐算法[17]。本文主要采用一种基于邻域的算法——基于用户的协同过滤算法(简称 UserCF)来进行相似度计算，另外还会采用基于项目的协同过滤算法来获得实验结果进而验证本文所提方法的合理性。首先利用余弦相似度公式得到用户兴趣相似度，公式如下：

$$w_{uv}=|N(u)\cap N(v)|/\sqrt{|N(u)|\times|N(v)|} \tag{6}$$

其中，$N(u)$，$N(v)$分别表示用户 u 和用户 v 有过评分的书籍集合。$|N(u)|$表示用户 u 评分过的书籍集合的对象个数。

基于用户的协同过滤算法与基于项目的协同过滤算法预测评分的方法相似，因此可以采用相同的公式不同的参数来预测评分。在得到用户兴趣相似度之后，利用该相似度，采用基于用户的邻域算法来预测一个用户对一本书籍的评分，公式如下：

$$\hat{r}_{ui}=\bar{r}_u+\frac{\sum_{v\in S(u,K)\cap N(i)}w_{uv}(r_{vi}-\bar{r}_v)}{\sum_{v\in S(u,K)\cap N(i)}|w_{uv}|} \tag{7}$$

其中，$S(u,K)$表示与用户 u 最相似的 K 个用户集合或与项目 u 最相似的项目集合，$N(i)$表示对书籍 i 评过分的用户集合或用户 i 评过分的项目集合，r_{vi} 是用户 v 对书籍 i 的评分，$\bar{r}_v$表示用户 v 对他评过分的所有书籍的分数平均值或项目 v 的平均分，w_{ij} 是用户或项目之间的相似度。

通过计算用户/项目之间的相似度，预测得到了用户对未阅读书籍的评分，获得了用户-书籍的评分矩阵。将每个用户所预测得到的评分进行递减排序，根据评分从高到低的顺序就能为用户推荐相对应评分的用户最感兴趣的 k 本书籍，其中 k 的值可自行选取。

6　实验评估

在实验过程中，需要计算用户对每本书籍的平均每天阅读时长和平均每天阅读次数。根据式(2)与式(4)，还将分别对阅读过同一本书籍的所有用户所花平均时长和次数进行统计，找出该情况下的最大时长和最大次数。当计算用户阅读该书籍的时长对应评分和频次对应评分时，就需要找到该本书籍对应的最大时长和最大频次，分别将其赋值给公式中的 $t_{\max}(b)$和 $f_{\max}(b)$。

通过 T-F 模型得到用户-书籍评分矩阵后，按约 4∶1 的比例将该评分矩阵的数据分为了训练集和测试集，测试集主要用来对预测所得到的结果进行推荐结果真实度的检验和比较。

首先应确定 Time 和 Frequency 的权重 w_1、w_2。本文通过设置 $w_1=0.9$、$w_2=0.1$，$w_1=0.8$、$w_2=0.2$，$w_1=0.7$、$w_2=0.3$，$w_1=0.6$、$w_2=0.4$，$w_1=0.5$、$w_2=0.5$ 这 5 组数据，设置相似用户数为 25 和 30 两种场景，得到相应的评分数据后分别进行推荐，并通过计算 RMSE 均方根误差来比较这几组数据的推荐准确度，如图 4 所示。

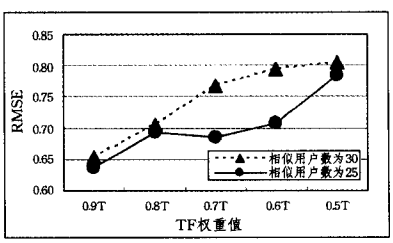

图 4　权重与结果对比

由图 4 可以看到，在相似用户数为 30 的情况下，随着 Time 权重的减小，RMSE 的值越来越大。当 Time 时间权重从 0.8 减小为 0.7 时，RMSE 呈现加速增长趋势。而当权重继续减小到 0.6 及以下后，增长率又趋于平稳，但总体趋势还是随着 Time 权重的减小和 Frequency 权重的增加，RMSE 值变大。由于用户阅读时间所获得的评分精度会比阅读次数所

对应的评分精度更大，因此用户阅读时间能够更精确地反映用户阅读书籍的行为和喜好。在相似用户数为 25 的条件下，总体的 RMSE 趋势也是随着 Time 权重的减小而增加的，但当 Time 的权重设置为 0.8 时，发现其 RMSE 要比 Time 权重为 0.7 时的略大，可能原因是当前相似用户集合中用户们阅读书籍所花平均次数比较接近，虽然 Time 权重减小，但 Frequency 也能够准确进行书籍推荐，因此导致了 Time 权重虽然减小，但误差没有随之增加的现象。

当相似用户数为 30 时，RMSE 整体值均要比相似用户数为 25 的大，这有可能是因为当相似用户数变大后，由于不同用户所阅读的书籍不尽相同，因此通过评分推荐的书籍也相应变多，推荐书籍的增加使得推荐效果有了一定的偏差。例如当有几本书籍评分相同时，系统就会根据评分排名推荐，而这个排名顺序则是按照数据的先后顺序得到的，因此在最后的推荐准确性上会产生一定的波动。

综上，随着 Time 权重增大，推荐效果趋于准确。由图 5 可看到，当 Time 权重占 0.9，0.8 时，推荐效果总体比较好。虽然当相似用户数在 25 的情况下，Time 权重为 0.8 时的推荐效果要略差于权重为 0.7 时的效果，但综合相似用户数为 25、30 这两种情况，还是认为 Time 的权重越大，推荐效果更准确。因此在接下来的实验中，将设置 Time 权重为 0.9 和 0.8 两种情况，与 0-1 评分矩阵做对比。

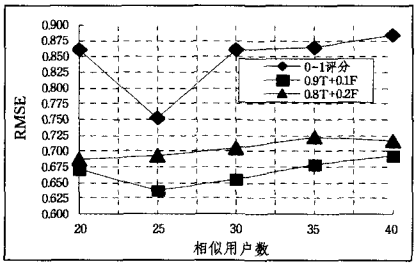

图 5　实验结果

本文通过等差改变相似用户的数量，每组递增 5 个相似用户数，产生了 5 组用户评分对比数据。由图 5 可以看到，在不同相似用户数的情况下，本文所提出的基于 T-F 模型的用户评分推荐算法均具有比传统 0-1 评分推荐算法要小的 RMSE 值。另外，在相似用户数为 25 的情况下，不论是采用 0-1 评分推荐算法还是取 Time 权重为 0.9 的算法，均具有最小的 RMSE 值。在采用 Time 权重为 0.8 的算法时，虽然在相似用户数为 25 的情况下 RMSE 值不是最小，但比相似用户数为 20 的 RMSE 值只大了 1%，因此我们可以认为相似用户数为 25 时，采用推荐算法的准确度较高。当相似用户数大于 25 时，随着相似用户数量的增加，推荐准确度降低。这一现象表明，不论采用何种方法，当取相似用户数为 25 时，RMSE 都能够保持一个较小值，推荐结果较为准确。

这一共同点不仅存在于对比时间和频次的权重实验所得到的推荐结果中，同时也在通过独立的 Time 模型和 Frequency 模型所得到的推荐结果中得到了验证。因此通过本数据集可以得到，取相似用户数为 25 所获得的推荐效果最好。

由此可见，与传统的在没有评分的情况下基于用户行为进行推荐的方法相比较，本文所提出的 T-F 模型的推荐效果有一定的优势。

另外，将 $w_1=0.9$、$w_2=0.1$ 的 T-F 模型与 T 模型和 F 模型的推荐结果做了比较，如图 6 所示。由 T-F 模型所得到的 RMSE 值均比 T 模型和 F 模型小，验证了将 T 模型和 F 模型融合到一起进行书籍推荐的可行性。

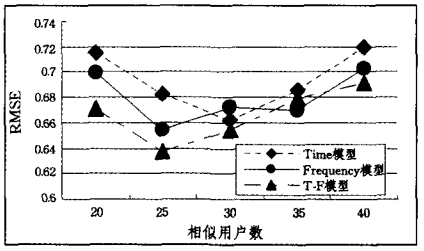

图 6　各类实验对比

继续讨论基于项目的协同过滤算法是否也存在相同的情况。图 7 给出了在基于项目协同过滤算法下得到的推荐结果。

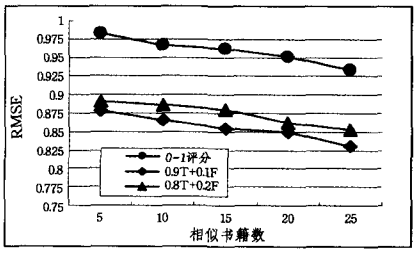

图 7　基于项目协同过滤算法的结果

选取了相似书籍数为 5、10、15、20、25 这 5 种场景下的推荐结果，由结果图能够发现，随着相似书籍数的增加，所有方法的推荐准确度增大。这是符合实际情况的，相似书籍数量越多，意味着强关联关系的书籍越多，一旦确立了更多书籍之间的关系，那么推荐的结果一定是更加准确的。另外，选取 Time 时长权重为 0.9 时的 RMSE 值要比其他两种情况的值小，而传统的 0-1 评分方法所得到的推荐结果最差，这与基于用户的协同过滤推荐算法得到的结果一致。但值得注意的是，本组实验所获得的 RMSE 值都要比基于用户的协同过滤算法所获得的值要大。也就是说，本文所采用的数据集采用基于用户的协同过滤算法会更加准确。由于本文数据集中并没有对书籍的详细描述，而且每个用户在一个月之内所阅读的书籍并不多，但数据的总数量是十分庞大的，这就导致了书籍之间的关联性并不强烈，因此所获得的推荐结果并不理想。

结束语　本文主要解决了在进行电子书籍协同过滤推荐过程中用户评分数据缺失导致无法进行准确推荐的问题。通过分析用户阅读书籍的行为数据，提出了一种 Time-Frequency 模型，即利用用户行为数据计算得到用户阅读书籍时长和用户阅读书籍次数两项数据，进而获得该两类的评分数据，建立用户评分矩阵，最后将该评分用于传统的基于用户的协同过滤算法得到推荐结果。通过实验对比，验证了本文的方法能够在没有用户评分数据的情况下提高推荐准确度，并客观反映用户对书籍的真实喜好程度。实验结果表明，本文基于用户行为的评分模型能够在一定程度上提高推荐准确度，具有实际的研究意义。

（下转第 54 页）

由表1与表2可以看出，对于语料的切分，本文中所提供的词典结构以及双向最大匹配法，与传统的单向最大匹配分词法和双向最大匹配分词法相比较，能够显著提升分词的精度与分词的速度；而与基于最大概率的双向最大匹配法相比，分词的准确率与召回率相差无几，却大大提升了分词的速度。

结束语 本文基于传统中文分词使用的正向最大匹配法与逆向最大匹配法，设计了一种既可用于正向最大匹配也可用于逆向最大匹配的词典结构，并设计与之相应的双向最大匹配分词法，提升了传统双向最大匹配分词的速度，并在最后加入了互信息歧义处理，进一步提升了分词的精度。相比较于基于最大概率的分词算法，本文提出的双向最大匹配分词算法在速度上有优势。因而本文提出的基于词典的双向最大匹配分词算法可以用于对词语分词精度要求更高的中文语言处理系统中，如车载人机交互的语音识别系统以及非结构化-结构化信息储存系统或非结构化大数据信息处理系统，以提高系统对信息的识别能力、预测分析能力，提升系统分析预测的准确度。

参 考 文 献

[1] 奉国和，郑伟. 国内中文自动分词技术研究综述[J]. 图书情报工作，2011，55(2)：41-45

[2] 罗智勇，宋柔. 现代汉语通用分词系统中歧义切分的实用技术[J]. 计算机研究与发展，2006，43(6)：1122-1128

[3] 吴育良. 百度中文分词技术浅析[J]. 河南图书馆学刊，2008(8)：115-117

[4] 莫建文，郑阳，首照宇，等. 改进的基于词典的中文分词方法[J]. 计算机工程与设计，2013，34(5)：1802-1807

[5] 吴旭东. 正向最大匹配分词算法的分析与改进[J]. 科技传播，2011(20)

[6] 王瑞雷，栾静，潘晓花，等. 一种改进的中文分词正向最大匹配算法[J]. 计算机应用与软件，2011，28(3)：195-197

[7] 张李义，李亚子. 基于反序词典的中文逆向最大匹配分词系统设计[J]. 现代图书情报技术，2006(8)：42-45

[8] 赵艳红，费洪晓. 一个基于改进的反序分词词典的中文分词算法[J]. 深圳职业技术学院学报，2004，3(4)：28-31

[9] 罗桂琼，费洪晓，戴弋. 基于反序词典的中文分词技术研究[J]. 计算机技术与发展，2008，18(1)：80-83

[10] 丁振国，张卓，黎靖. 基于 Hash 结构的逆向最大匹配分词算法的改进[J]. 计算机工程与设计，2008，29(12)：3208-3211

(上接第 41 页)

参 考 文 献

[1] 荣辉桂，火生旭，胡春华，等. 基于用户相似度的协同过滤推荐算法[J]. 通信学报，2014，35(2)：16-24

[2] Yin Pei-feng, Luo Ping, Wang C-L, et al. Silence is Also Evidence: Interpreting Dwell Time for Recommendation from Psychological Perspective[C]// the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2013. Chicago: ACM Press, 2013: 989-997

[3] 李建廷，郭晔，汤志军，等. 基于用户浏览行为分析的用户兴趣度计算[J]. 计算机工程与设计，2012，33(3)：968-972

[4] Ricci F, Rokach L, Shapira B. Recommender Systems Handbook [M]. New York: Springer, 2010

[5] Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback[C]// 25th Conference on Uncertainty in Artificial Intelligence, 2009. Virginia: AUAI Press, 2009: 452-461

[6] Yi Xing, Hong Liang-jie, Zhong Er-heng, et al. Beyond Clicks: Dwell Time for Personalization[C]// 8th ACM Recommender Systems Conference, 2014. Foster City: ACM Press, 2014: 113-120

[7] Gong Song-jie. A Collaborative Filtering Recommendation Algorithm Based on User Clustering and Item Clustering[J]. Journal of Software, 2010, 5(7): 745-752

[8] Gong Song-jie, Cheng Guang-hua. Mining User Interest Change for Improving Collaborative Filtering[C]// 2nd International Symposium on Intelligent Information Technology Application, 2008. Shanghai: IEEE, 2008: 24-27

[9] Yu Li, Liu Lu, Li Xue-feng. A Hybrid Collaborative Filtering Method for Multiple-interests and Multiple-content recommendation in E-Commerce[J]. Expert Systems with Applications, 2005, 28: 67-77

[10] Billsus D, Pazzani M J. Learning Collaborative Information Filters[C]// 15th International Conference on Machine Learning, 1998. Madison: ICML, 1998: 46-54

[11] Funk S. Netflix Update: Try This at Home[EB/OL]. http://sifter.org/~simon/journal/20061211.html. 2006 December

[12] 孙光福，吴乐，刘淇，等. 基于时序行为的协同过滤推荐算法[J]. 软件学报，2013，24(11)：2721-2733

[13] Koren Y. Collaborative filtering with temporal dynamics[C]// 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009. Paris: ACM Press, 2009: 89-97

[14] Claypool M, Le E, Waseda M, et al. Implicit interest indicators [C]// The 6th International Conference on Inteligent User Interfaces (IUI), 2001. New York: ACM Press, 2011: 33-40

[15] Wu Xiao-jun, Feng Ai-gang, Yin Qin-ye. Universal Discrete Model and Linear Algebra Representation for Variant OFDM-CDMA Systems[C]// IEEE International Symposium on Cireuits and Systems, 2001. Sydney: IEEE, 2001: 213-216

[16] 付关友，朱征宇. 个性化服务中基于行为分析的用户兴趣建模[J]. 计算机工程与科学，2005，27(12)：76-78

[17] Sarwar B, Karypis G, Konstan J et al. Item-Based collaborative filtering recommendation algorithms[C]// 10th International World Wide Web Conference, 2001. HongKong: ACM Press, 2001: 285-595