基于协同过滤的书籍推荐模型

摘要：文章根据题给数据，针对题目问题，基于协同过滤算法，共建立五个模型，并不断优化，程序运行结果显示模型较能反映实际问题。

问题一，考虑用户偏好书籍类型、关注好友评分两个因素对其评分的影响，建立**线性拟合**模型研究每个书籍标签在用户评分中的权重，并采用**梯度下降**法求解模型，在此基础上定义**偏好相似度**定量研究用户与关注好友之间的偏好相似程度，并作为另一个权重研究好友评分对用户评分的影响，并以**影响系数**反映用户好友评分对其评分的影响。MATLAB 程序运行结果显示，用户关注好友对其影响系数都较高，最终得出**用户阅读偏好以及关注好友评分是影响用户评分的两个关键因素**的结论。

问题二，考虑偏好相同的用户在对书籍评分上比较相似的特点，采用基于**协同过滤**算法的预测模型。首先采用皮尔逊相关系数描述两个用户之间的相似度，并建立基于用户的协同过滤预测模型；在此基础上，考虑用户之间公共阅读书籍较少的情况，采用杰卡德——皮尔逊相关系数更精确地描述用户之间的相似度，并建立改进的协同过滤预测模型。

随后，考虑相似书籍会得到同一用户相似评分的特点，采用皮尔逊相关系数描述书籍之间的相似度，建立**基于项目**的协同过滤预测模型。编写 MATLAB 程序求解模型，并从“user_book_score.txt”附件中选择 20%的评分数据作为测试集，结果显示三种模型的预测精度分别为 0.8347，0.9153，0.9319，最终使用模型三来预测目标用户评分。

问题三，从用户偏好书籍类型以及用户对书籍评分预测值两个角度建模。利用用户阅读历史数据，从用户已读书籍中某标签出现频数的角度出发，通过定义**偏好因子**的概念定量描述用户对书籍标签的偏好程度，并基于偏好因子选出用户最偏好的 Top10 标签，同时选出含有 Top10 标签的书籍组成候选推荐书籍集合。结合问题二的预测模型，计算出用户对于候选集合的评分预测值，最后将书籍的综合偏好因子以及评分预测值之积作为综合评估指标，选取评估指标值最大的三本书作为推荐书籍。

最后，文章对三个问题中的模型进行了客观评价，指出各个模型的优缺点，并针对模型的缺点指出了改进方向。

关键字：协同过滤　皮尔逊相关系数　线性拟合　偏好因子　梯度下降

目录

一、问题重述

互联网技术的发展将我们带进一个信息爆炸的时代，大量信息充斥着人们的生活。对于信息收集者来说，如何从大量的信息中寻找自己感兴趣的或是需要的信息成为极其困难的事情；而对于信息发布者来说，如何使自己的信息脱颖而出又成为其占领市场份额、盈利的必要条件。推荐机制成为解决信息冗余问题的一个重要工具，并在各大搜索关键词推荐、话题推荐、电子商务的产品推荐、社交网络交友推荐都有着广泛的应用。

题目给出了某著名网上书店的用户行为信息，包括对于书籍的评分数据、书籍的标签信息以及用户的社交关系，要求根据这些数据完成以下问题：

(1)分析影响用户对书籍评分的因素；

(2)通过建立模型，预测 predict.txt 附件中的用户对未看过书籍的评分；

(3)针对 predict.txt 附件中的用户，给每个用户推荐 3 本没看过的书籍。

二、问题分析

信息化时代将每个个人联系起来，计算机网络全面普及的背景使得信息的分享成为可能，网络推荐机制为人们获取所需的信息提供了方便。网站制作者可根据用户以往搜索关键词、关注话题、好友关注话题等项目设置推荐项目，而用户筛选有用信息的方式大多也是根据这些项目来选择的。

首先，针对题目给出的数据进行处理。题目所给可用于分析的数据为用户—书籍—得分数据、书籍—标签数据、用户—社交数据以及用户—所读书籍的数据，可先将这些数据制作成一系列矩阵后进行分析。针对问题一，我们认为用户对书籍的评分，一般受个人偏好的影响，附件中有部分书籍的标签可以反映书籍的类型，虽然全是用 ID 数字表示的，但同一个数字代表同一种书籍标签，总的书籍标签数是一定的，因此，可以从每种标签在用户评分中的权重中来考虑；其次，用户在选择书籍上也会参考自己所关注好友的选择。对于各种因素对其评分的影响，可采用主成分分析、多元线性拟合的方法进行分析。此处采用多元线性拟合的方法先对书籍标签的影响作分析。

针对问题二，需要对所给用户对书籍的评价进行预测，可采用协同过滤的方法挖掘“user_book_score.txt”附件中的统计信息。目前已有多种算法用于计算样本之间的相关系数，包括基于用户的协同过滤算法，改进后的协同过滤算法，以及基于项目的协同过滤算法，前两种算法都是基于一个假设：如果两个用户具有相似的偏好，当其中一个用户对某一商品有一定评价时，另一用户应当对着商品有相似的评价。当两个用户具有最高的相似度时，就可以用一个用户对某本书的评价近似估计另一个用户对相同书本的评价。而基于项目的皮尔逊则是基于另一个假设：当两个项目具有较高的相似度时，同一个用户给两个项目的评价应较为接近。因此，可以采用基于项目的协同过滤去估计相似项目的到的同一用户给的评分。在问题二中，采用三种算法分别进行预测，并选择部分已有评分数据作为测试集，最终评估算法的精确度，采用精确度最高的算法作为最终的预测算法。

针对问题三，给用户推荐书籍时，考虑用户阅读历史，挖掘出用户所青睐的书籍标签，并筛选出用户所喜欢书籍类型 top10，并结合问题二的模型，考虑书籍标签以及书籍预测评分两个因素，最终给出综合评分指标最高的三本书作为推荐书籍。

三、基本假设

（1）题目所给数据真实可靠；

（2）用户只对看过的书籍作评价；

（3）用户对所看书籍都理性评价；

（4）所选择数据为随机抽查获得的；

（5）用户选择书籍时会参考推荐信息与书籍评分。

四、参数说明

符号	说明
T	书籍标签矩阵
$t_{i,j}$	书籍 i 拥有第 j 个标签数目

$x_{i,j}$	第 j 个标签对用户 i 的权重
$r_{i,j}$	第 i 个用户对第 j 本书的评分；
$\hat{r}_{i,j}$	第 i 个用户对第 j 本书评分估计值
J_i	对用户 i 评分估计均方根值；
$\bar{r}_i$	第 i 个用户评分均值；
$r^*_{i,j}$	第 i 个用户对第 j 本书的预测评分；
I_{U_k}	用户 u_k 评价过的书籍集合；
U_i	第 i 个用户的邻居用户集合；
P_j	第 j 个项目的相邻项目集合；
$U_{i,j}$	对项目 i 和 j 都有评分的用户集合；
$sim(u,v)$	用户 u 与 v 之间的皮尔逊相关系数；
$\cos(u,v)$	用户 u 与 v 之间的余弦相似度；
u_i	任意一个用户集合 u_i；
u	目标集合；
$C_{i,j}$	用户 i 和 j 共同关注的书籍集合；
$J(u,u_k)$	用户 u 与 u_k 的杰卡德系数；
$\lambda_{i,j}$	用户 i 对标签 j 的偏好因子；
$F_{u,i}$	用户 u 对书籍 i 的综合评价指数；
I_{rec}	推荐书籍集合；

五、模型建立与求解

5.1 数据预处理

首先，对题目所给附件中的数据作预处理。将每个附件中的数据导入 JAVA 程序中，并将“book_tag.txt”、“user_read_history.txt”、“user_social.txt”文件中同一书籍的不同标签、同一用户历史所读的书籍以及同一用户对于不同单方向的好友整理成一系列向量。同时，为方便后期分析，找出“predict.txt”附件里面 6 个用户对应的好友、历史所读书籍分别构成一个向量，并通过编程找

出 6 个用户历史所读书籍里可以从“book_tag.txt”附件中找出对应标签的书籍，将这些书籍对应的标签同样构成一个向量。这样，就对题目所给的大量信息作了简单的统计整理，重要整理结果如下:“user_book_score”里共有用户 4071 个，已评论书籍 7840 本;“book_tag.txt”里共有书籍数目 8369 本，含标签总数 1129 个。其他相应向量与矩阵见附件。

5.2 问题一模型

5.2.1 数据筛选

用户对书籍的评分决定于多种因素，包括用户对所看书籍的类型的喜好程度，用户自身的评判尺度，用户所关注对象的评分，乃至于大众对于某本书籍的一致观点都会影响用户的评分。用户要么根据自己对于书籍的真实偏好，要么参考关注对象以及大众的意见。

根据题目所提供的数据信息，我们首先对这些数据作筛选。在众多的数据中，“book_tag.txt”给出了部分书籍的标签数据，而这些数据认为是反映书籍类型的数据，故对用户评分过程有一定影响；其次“user_social.txt”为反映用户关注对象的数据，在此也认为是有用的信息；而“user_read_history.txt”则只是反映用户阅读历史的数据，且并未直接给出用户对历史阅读书籍的评分，故认为此数据无明显用处;“user_book_score.txt”则可以反映用户的评分尺度，故认为是有用的数据。在此选择书籍类型、社交关系以及评分尺度作为影响用户评分的因素。

5.2.2 模型一、线性拟合模型

影响用户选择书籍并评分的一大因素是书籍类型，即书籍标签，为定量求取书籍类型对用户评分的影响，采用梯度下降的方法求取每个 tag 在用户评分中所的比重，并分析同一 tag 在用户所关注对象的评分中占的比重。此处采用线性拟合法建模，并用梯度下降法求取均方误差最小值，具体过程为:

(1)矩阵生成

将“user_book_score.txt”文档中所有被评价的书籍以及其对应的标签制成一个$m\times n$的书籍—标签 0—1 矩阵，

$$T=\begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,n} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{m,1} & t_{m,2} & \cdots & t_{m,n} \end{bmatrix}$$

其中，行代表被评价书籍总数，列代表标签总数，$t_{i,j}=0$或 1，当第i本书籍有第 j 个标签时，$t_{i,j}=1$，而没有第i个标签或是在“user_book_score.txt”文档中的书籍没有出现在“book_tag.txt”里面时，$t_{i,j}=0$。

同样的道理，生成一个$s\times m$的用户—评分矩阵

$$R=\begin{bmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m,1} & r_{m,2} & \cdots & r_{m,n} \end{bmatrix}$$

其中，$r_{i,j}$代表第i个用户对第 j 本书的评分，用户为所有参与评分的用户。两个矩阵都可以从题给数据中直接得出。

(2)系数设定

在求取每个书籍标签在用户心中地位时，只考虑书籍标签对评分引起的作用，以便进一步分析用户社交关系中其关注对象的书籍标签权重，通过探讨社交关系与用户倾向之间的关系来分析两者在用户评分中的作用。

设第 j 个标签对于第 i 个用户评分的权重为$x_{i,j}$，并形成一$s\times n$权重矩阵为

$$X=\begin{bmatrix} X_{11}\ X_{12}\ \ldots X_{1n} \\ X_{21}\ X_{22}\ \ldots X_{2n} \\ \ldots\ldots\ldots \\ X_{s1}\ X_{s2}\ldots X_{sn} \end{bmatrix}$$

其中，行代表着用户，列代表着标签，评分权重为待求变量。

(3)模型建立

每种标签对于用户来说具有不同的权重，当用户所阅读书目中存在自己所感兴趣的项目标签时，此项目标签对于提升用户评分具有一定的作用。通过线性拟合的方法可以求出各个项目对于用户评分的权重值。梯度下降法是一种利用函数在负梯度方向数值下降最快的原理设计的一种最优值迭代搜索算法。在

本模型中，通过梯度下降法求解残差最小值。建立的**线性拟合模型**

$$\hat{R} = XT^{T}$$

$$\begin{bmatrix} \hat{r}_{1,1} & \hat{r}_{1,2} & \cdots & \hat{r}_{1,m} \\ \hat{r}_{2,1} & \hat{r}_{2,2} & \cdots & \hat{r}_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{r}_{s,1} & \hat{r}_{s,2} & \cdots & \hat{r}_{s,m} \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{s,1} & x_{s,2} & \cdots & x_{s,n} \end{bmatrix} \begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,n} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{m,1} & t_{m,2} & \cdots & t_{m,n} \end{bmatrix}^{T}$$

即

而 $$\hat{r}_{i,j} = \sum_{k=1}^{n} x_{i,k} t_{i,k}$$

其中 $\hat{R}$ 为用户评分估计矩阵，$\hat{r}_{ij}$ 为用户 i 对书籍 j 的评分估计；这时，用户 i 关于 n 个权重值的均方差根为

$$J_i = \sqrt{\sum_{j=1}^{m} (r_{i,j} \hat{-} r_{i,j})^2} = \sqrt{\sum_{j=1}^{m} (\sum_{k=1}^{n} x_{i,k} t_{i,k} - r_{i,j})^2}$$

J_i 表示与用户 i 对应的评分估计均方根值，给定每个权重 $x_{i,j}$ 一个初值 $x^{(0)}{}_{i,j}$，进行迭代求取 J_i 的最小值，则有

$$\nabla J_i(X_i^{(k)}) = \left[\frac{\partial J_i(X_i^{(k)})}{\partial x_{i,1}}, \quad \frac{\partial J_i(X_i^{(k)})}{\partial x_{i,2}}, \quad \cdots \quad \frac{\partial J_i(X_i^{(k)})}{\partial x_{i,n}} \right]$$

式中 $X_i^{(k)}$ 表示第 k 次迭代后的用户 i 的权重向

$$X_i^{(k)} = \left[x_{i1}^{(k)}, x_{i2}^{(k)}, ..., x_{in}^{(k)} \right],$$

迭代过程为

$$X_i^{(k+1)} = X_i^{(k)} - a^{(k)} \frac{\nabla J_i(X_i^{(k)})}{\left\| \nabla J_i(X_i^{(k)}) \right\|},$$

当 $\left\| X_i^{(k)} - X_i^{(k+1)} \right\| \le \varepsilon$ 或 $\left\| \nabla J_i(X_i^{(k)}) \right\| \le \varepsilon$ 时，迭代过程结束，$a^{(k)}$ 代表第 k 次迭代的迭代步长，ε 代表迭代精度。

利用以上算法，编写 MATLAB 程序求解出各种标签对于各个用户的评分权重值，此处列举部分程序计算所得权重值，如下表所示。

书籍	书籍标签				
ID	8427	6391	5942	9230	7628
817168	0.1953	0.1024	0.0931	0.0834	0.0416
616799	0.0151	0.2102	0.0082	0.1021	0.0321
489646	0.1032	0.0092	0.2103	0.0023	0.1055
391403	0.2256	0.0036	0.1013	0.0032	0.0102

表 1 部分书籍标签对应书籍评分中的权重值

5.2.3 模型二、社交关系模型

（1）偏好相似度

用户社交关系对其评分过程也会有一定程度影响。举例来说，当用户关注对象普遍对某书籍评分较高或是较低时，有可能会直接影响用户对所读书籍的判断，这是源于人都有“跟风”的习惯；另一方面，目标用户关注另一用户的根本原因在于二者之间有着共同的喜好，即喜欢同一标签类型的书。因此，可以利用模型一的结果定性分析社交关系对用户评分的影响。

定义目标用户 u 与其关注用户 u_k 之间的偏好相似度为

$$H(u,u_k)=\frac{\sum_{j=1}^{n} X_{uj}\cdot X_{u_k j}}{\sqrt{\sum_{j=1}^{n} X_{uj}{}^2}\sqrt{\sum_{j=1}^{n} X_{u_k j}{}^2}}$$

其中，$x_{u,j}$ 为第 j 种标签在用户 u 评分中的权重。这个偏好相似度从用户之间对于书籍标签的偏好上来决定用户之间存在共同喜好的程度。通过 MATLAB 编程，结果显示用户 u 与其关注的对象 u_k 之间普遍都有较高的偏好相似度，说明社交关系确能对用户的评分造成一定的影响。下面建立模型说明。

（2）社交关系模型

由模型一知，$X_u=[x_{u,1},\ x_{u,2},\ \cdots\ x_{u,n}]$ 表示各个标签对于用户目标用户 u 的权重向量，而受关注用户权重向量为

$$X_{uk}=[x_{u_k,1},\ \ x_{u_k,2},\ \ \cdots\ \ x_{u_k,n}],$$

建立**模型**

$$Inf(u,u_k)=\frac{1}{5}\left|\frac{\sum_{i=1,i\neq p}^{n}x_{u,i}\overline{r_{u,i}}}{\sum_{i=1,i\neq p}^{n}x_{u_k,i}\overline{r_{u_k,i}}}\cdot\frac{\overline{r_{u_k,p}}}{\overline{r_{u,p}}}\right|$$

表示用户u_k对用户u的影响系数，式中，$\overline{r_{u,i}}$表示用户u对于标签为i的书籍的评分平均值，$\overline{r_{u,p}}$表示用户u与u_k权重矩阵中权重最高且最接近的标签所对应的书籍的评分平均值。此模型的意义在于，首先排除两个个体偏好最接近的标签对应的书籍的评分，考察二者评分尺度之比，代表两人对于其他书目的评判标准，而与右方分式相乘的意义在于先计算出目标用户本应给出的分数平均值，然后与其实际给分的平均值相除，最后乘五分之一是保证影响因子在[0,1]之间。

通过筛选合适的试验集，并运用 MATLAB 验证本模型正确性。发现目标用户所关注用户确实对其评分有一定程度影响。**因此，社交关系以及书籍标签是影响用户评分的关键因素。**

5.3 问题二模型

5.3.1 协同过滤求相关系数

结合实际情况，当两个用户u和v对于书籍集合C都有着较为接近的评分时，则当用户u对于另一个集合C_1或另一个书籍元素$i\notin C$时有一定的评分时，用户v对于集合C_1或是元素i应当也有着相似的评分。协同过滤就是基于这种假设的前提，提出了对两个个体之间相似度的衡量。协同过滤是一种用于商务推荐的很有效的算法，通过挖掘两个个体之间相似度来进行进一步的推荐、预测。常用的用于计算相似度协同过滤算法有余弦相似度算法和皮尔逊相关系数（PCC）法。

设用户的评分向量为$u\notin R^m$，另一个用户的评分向量为$v\notin R^n$，且$m\neq n$，为了在同一个向量空间中对两个向量作相似度分析，取$v,u\notin R^{\min(m,n)}$，即将两者映射到两个向量共同属性元素值的向量空间中

两个评分向量之间的余弦相似度为

$$\cos(u,v)=\frac{\sum_{i\in I_u\cap I_v} r_{u,i}\cdot r_{v,i}}{\sqrt{\sum_{i\in I_u\cap I_v} r^2{}_{u,i}}\cdot\sqrt{\sum_{i\in I_u\cap I_v} r^2{}_{v,i}}} \quad (1)$$

其中 $r_{u,i}$，$r_{v,i}$ 分别表示用户 u 和 v 对商品 i 的评分值。

而皮尔逊相关系数法考虑到了个人评分尺度的问题，因此较余弦相似度算法有更好的性能。两个用户 u 和 v 之间的皮尔逊相关系数即其相似度为：

$$sim(u,v)=\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r_u})\cdot(r_{v,i}-\overline{r_v})}{\sqrt{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I_u\cap I_v}(r_{v,i}-\overline{r_v})^2}} \quad (2)$$

其中，$sim(u,v)$ 表示 u 和 v 之间的相似度，I_u 和 I_v 分别便是被用户 u 和 v 评价过的书籍，$\overline{r_u}$ 和和 $\overline{r_v}$ 分别表示用户 u 和 v 对所有评价过的商品的集合，即 $I_u\cap I_v$ 中各个商品的平均值，$r_{u,i}$，$r_{v,i}$ 分别表示用户 u 和 v 对商品 i 的评分值。显然，用户 u 和 v 之间的相似度值 $sim(u,v)$ 处于区间 $[-1,\ 1]$ 之间。

从皮尔逊相关系数中我们看出，其考虑了不同个体由于评判尺度不同而对二者相似度的影响，因有的用户倾向于对每本书籍评分都较高，而有的用户则不然。故此算法更能消除评判尺度对两个用户之间相似度的影响。

5.3.2 模型一、PCC 预测模型

对于为读书籍评分预测问题，考虑协同过滤的基本假设，即具有相同阅读倾向的两个或多个用户，对于不同商品应相应都有较为近似的评价。基于用户的协同过滤（KNN）算法可以用于向目标用户提供推荐以及*对商品*进行评分预测。

基于皮尔逊相关系数的**预测模型**为：

（1）邻居用户形成

$$U=\{u_1,u_2,u_3\cdots\cdots u_k\} \quad (4)$$

其中，$$sim(u,u_1)=\max\{sim(u,u_1)\}\quad 0<i<N \quad (5)$$

$$sim(u,u_1)>sim(u,u_2)>\cdots>sim(u,u_k) \quad (6)$$

式中，u_i 表示第 i 个用户，$sim(u,u_i)$ 表示目标用户 u 与第 i 个用户之间的皮尔逊相关系数，U 为目标用户 u 的邻居用户，且 $u\notin U$，N 为所有用户数目。

将皮尔逊相似度最高的前 k 个用户作为目标用户 u 的邻居用户，并按照相似度大小依次排序，为后续预测作准备。

（2）目标评分预测

为了进行对目标项目的评分，采用邻居用户加权预测的方法实现，假设在目标用户 u 的邻居用户 U 中，对目标项目 i 作过评价的邻居用户构成的集合

$$U^*_i = \{u_1, u_2, u_3 \cdots u_k\} \tag{7}$$

目标用户对指定项目 i 的评分目标评分**预测模型**为

$$r_{u,j^*} = \overline{r_u} + \frac{\sum_{u_k \in U^*_i} sim(u,u_k) \cdot (r_{u_k,i} - \overline{r}_{u_k})}{\sum_{u_k \in U^*_i} (|sim(u,u_k)|)} \tag{8}$$

其中，$\overline{r_u}$ 表示用户 u 所有评价过的书的评分均值，$r_{u_k,i}$ 表示邻居用户集合 U^*_i 中用户 u_k 对第 i 本书籍的评分，$\overline{r_{u_k}}$ 表示用户 u_k 的与用户 u 共同评价过的书籍的评分均值，$|sim(u,u_k)|$ 表示两个用户之间相关系数的绝对值，因二者相关系数有正负之分。

而有

$$\overline{r_{u_k}} = \frac{1}{N_{I_{u_k}}} \sum_{I_u \cap I_{u_k} \neq \Phi} r_{u_k,j} \quad , \tag{9}$$

又有

$$\overline{r_u} = \frac{1}{N_{I_u}} \sum_{j \in I_u} r_{u,j} \tag{10}$$

式中，$r_{u,j}$ 表示目标用户对第 j 本书的评分，I_u 表示用户 u 的评价过书籍的集合，N_{I_u} 表示集合 I_u 的元素个数，式（9）中 $I_u \cap I_{u_k} \neq \Phi$ 表示只有用户 u 与 u_k 都对某本书评价过时，才对用户 u_k 所作过的书籍评分求平均值。

对这个**模型的解释**为：

1）首先，从目标用户 u 的邻居用户中筛选出对第 i 本书评价过的用户，因为只有当此邻居用户看过这本书并作过评价，其评价分支对于用户 u 对书本的评分才有参看意义；

2）对于公式（8），分母表示所有筛选出的邻居用户与目标用户之间皮尔逊相关系数的绝对值之和，分子则代表着先将所筛选出的邻居用户对第 i 本书籍的

评分与已有评分取差值，然后将这些差值作加权和作为分子。先取差值的目的在于需要考虑不同用户评论书籍好坏的尺度差异，而通过取差值更能反映某本书籍在用户自身评价尺度上的好坏；而取加权和的目的就在于其加权因子即 $sim(u,u_k)$ 反映的是二者阅读偏好的相似程度，与目标用户越相似的用户显然应具有更高的权重。

3）对于 $\overline{r_u}$，同样是基于目标用户 u 的评分尺度问题而设置的，$\overline{r_u}$ 通过公式（10）求取用户 u 评价过的所有分值的平均值得到，并作为用户 u 的评分尺度。

4）对于公式（9），在求取用户 u_k 的评分均值 $\overline{r_{u_k}}$ 时，只能求取用户 u_k 与用户 u 都看过的书籍的用户 u_k 的平均值，因为只有邻居用户与目标用户看过书籍的交集才能反映二者的共同偏好。

(3) 模型求解

1）求相关系数

将附件“user_book_socre.txt”导入 MATLAB 软件，并用u表示用户对所有读过书籍的评分向量，即

$$u=(r_{u_1},r_{u_2},r_{u_3}\cdots r_{u_i}\cdots r_{u_n})$$

用皮尔逊相关系数法求出每个用户与其他用户之间的皮尔逊相关系数，得出一个相关系数矩阵，部分相关系数表见附件 1。

将附件“predict_txt”中的用户作为目标用户，而将其中的书籍 ID 作为预测对象，通过 MATLAB 程序，过滤出目标用户的邻居用户。为避免数据的过于冗杂，选择系数 $k=200$，即筛选与目标用户相似度为前 50 位的用户作为目标用户的邻居用户。

2）评分预测

根据公式（10）编辑 MATLAB 程序，其具体计算流程如图 1-1 所示。

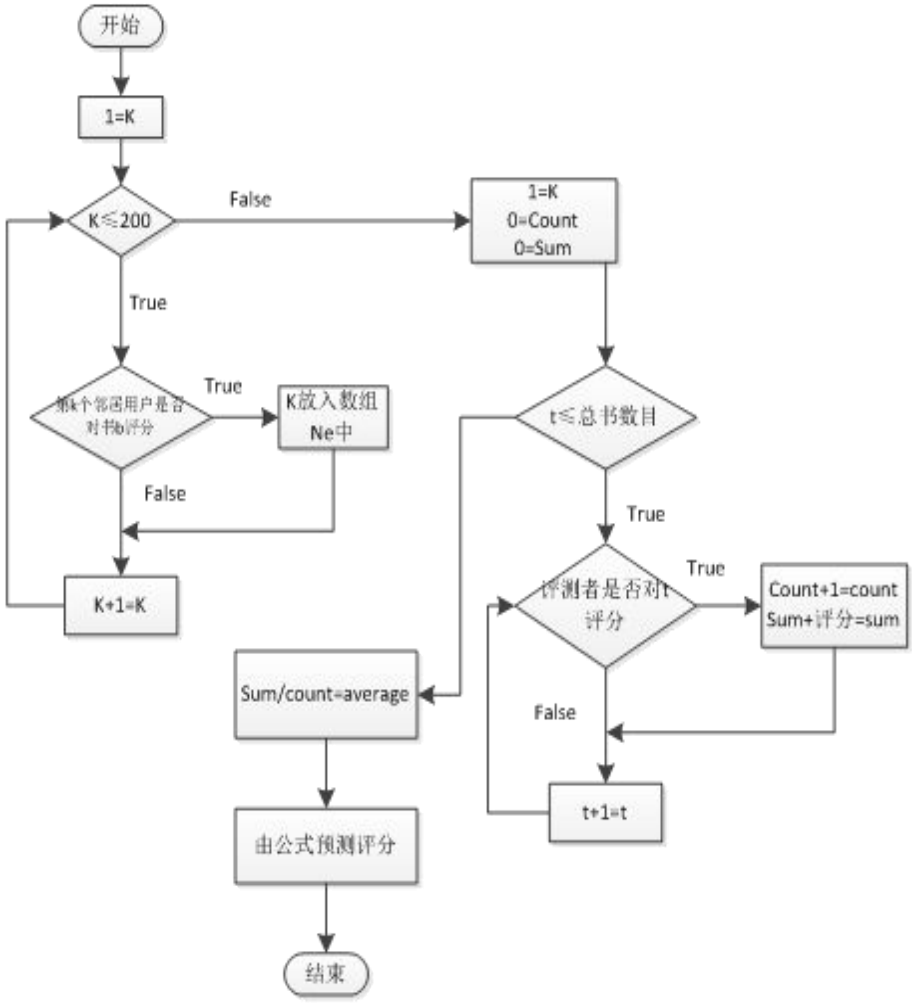

图 1. 模型一流程图

为检验模型正确性，先将附件里面的数据作为测试集，通过筛选部分用户与部分书籍，得出以下的误差曲线。

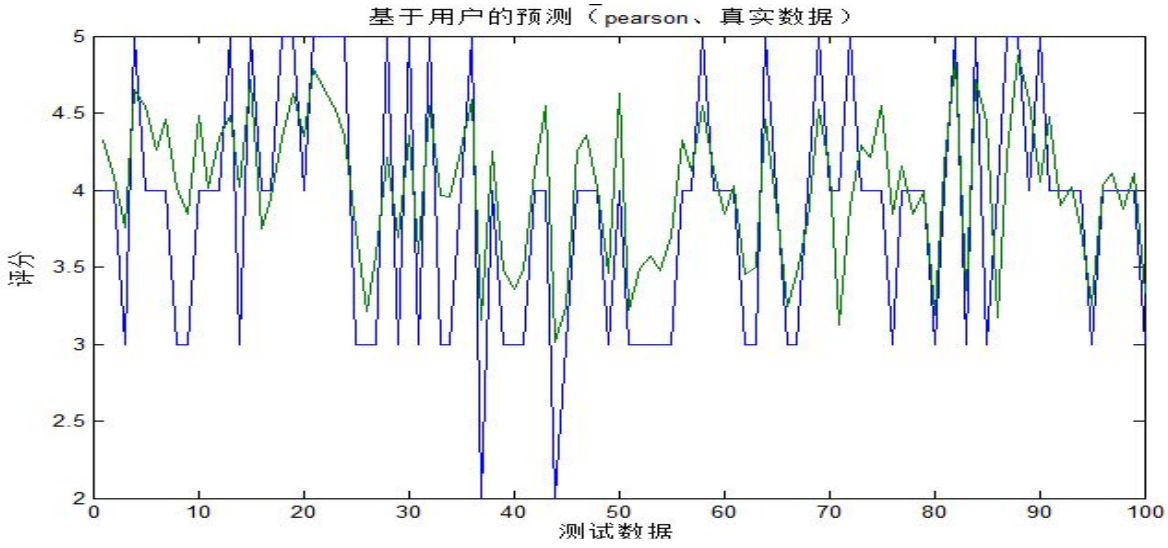

图 2. 基于用户的皮尔逊预测模型（绿线：PCC 预测值；蓝线：实际值）

从以上图像中可以直观看出，预测值与实际值之间有一定的误差。通过分析本预测模型得知，在计算两个用户之间相关系数的时候，只考虑了两个用户对于同一书籍的评分近似程度，而在某种程度上会导致较大的误差，因为可能出现两个人同时看过少数几本书而评分近似的情况，这种情况下利用本模型就可能导致两人的相关系数近似为 1 的情况，对整个预测就会产生一定的影响。为此，提出下面的基于杰卡德—皮尔逊相关系数（JacPCC）算法的预测模型。

5.3.3 模型二、JacPCC 预测模型

（1）杰卡德—皮尔逊 相关系数

杰卡德—皮尔逊相关系数(JacPCC)是基于杰卡德相关系数提出的一种改进的相关系数算法。皮尔逊相关系数算法是基于两个集合之间的共同对象所提出的，它仅考虑两个个体对具有同一属性的事物的评价，而忽略两个个体所处空间的维度数，即两个个体各自关注对象的全体。这种算法的显著缺点在于对某些特殊情况的误判，即当两个个体之间交集很小时，可能会出现两个个体本不具有强相关性时其皮尔逊相关系数却接近于一的情况。为此，改进的皮尔逊相关系数，也即杰卡德—皮尔逊相关系数则考虑了两个所评价的总体对两者相关系数的影响，通过杰卡德系数来反映整体对交集的影响。

杰卡德系数为 $$J(u,u_k)=\frac{|I_u\cap I_{u_k}|}{|I_u\cup I_{u_k}|}$$

式中，I_{u_k} 表示第 k 个用户评分集合，$|I_u\cap I_{u_k}|$、$|I_u\cup I_{u_k}|$ 分别代表集合 I_u 与 I_{u_k} 交集和并集中的元素个数，式子代表着目标集合与用户 k 集合之间交集在两者评价过书籍集合中的权重。

将杰卡德系数与皮尔逊系数相乘，就得到了杰卡德—皮尔逊系数为

$$sim(u,u_k)_J=\frac{|I_u\cap I_{u_k}|}{|I_u\cup I_{u_k}|}\frac{\sum_{i\in I_u\cap I_v}(r_{u,i}-\overline{r_u})\cdot(r_{u_k,i}-\overline{r_{u_k}})}{\sqrt{\sum_{i\in I_u\cap I_{u_k}}(r_{u,i}-\overline{r_u})^2}\sqrt{\sum_{i\in I_u\cap I_{u_k}}(r_{u_k,i}-\overline{r_{u_k}})^2}}$$

这样，当两个个体之间的交集较小时，用户集合 k 对目标集合之间的相关系数就相对较小，对目标用户的预测的影响就相对较小。

（2）改进的预测模型

在 JacPCC 的基础上，建立改进的预测模型，目标用户对书籍 i 的**预测值**为

$$r_{u,j^*} = \overline{r_u} + \frac{\sum_{u_k \in U^*_i} sim(u,u_k)_J \cdot (r_{u_k,i} - \overline{r}_{u_k})}{\sum_{u_k \in U_i^*} (|sim(u,u_k)_J|)}$$

其中 $sim(u,u_k)_J$ 为目标用户 u 与 u_k 之间的杰卡德—皮尔逊相关系数。

（3）模型的求解

编写 MATLAB 程序，再次对相同的测试集进行测试。得到如下所示的对比图。

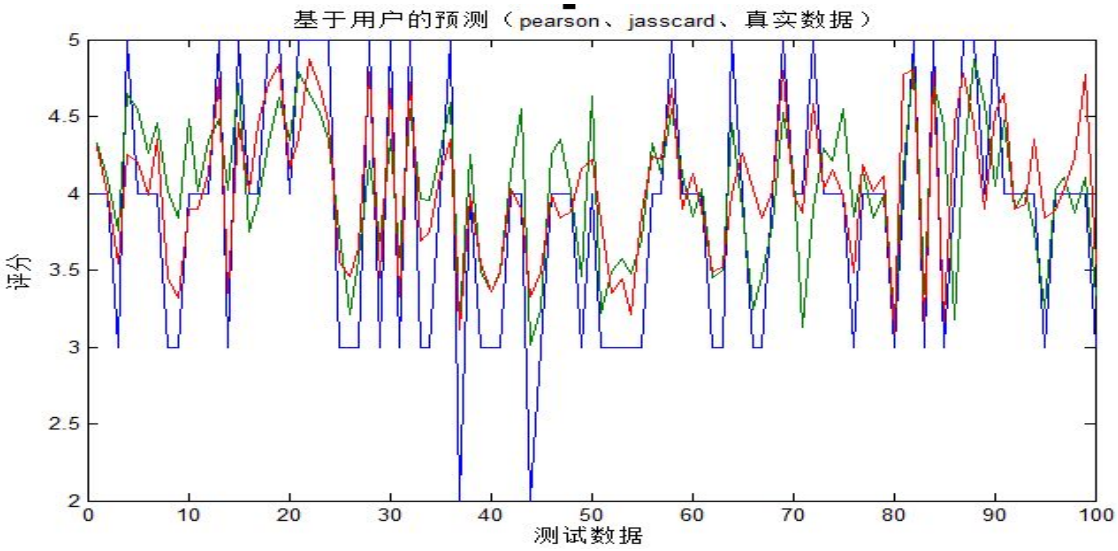

图 3. 基于用户的 JacPCC 模型与 PCC 模型对比图（红线：JacPCC）

对比两组图形，可以看出利用改进后的模型对于预测测试集数据具有更高的精确性。因此，利用改进后的模型对“predict.txt”附件中的用户关于书籍的评价进行预测。得出预测数据如下所示。

用户	书籍序号（与附件中的一样					
ID	1	2	3	4	5	6
7245481	4.1078	4.1918	4.3630	4.1985	4.0263	4.1938
7625225	3.5010	3.8133	3.7479	3.9250	3.7723	3.9030
4156658	4.1901	4.1763	3.9922	4.2538	4.0076	4.1294
5997834	4.1540	4.5587	4.2154	4.3694	4.3966	4.0877
9214078	4.3938	4.3729	4.2099	4.2088	4.2094	4.2362
2515537	3.8842	3.7179	3.9764	3.9427	3.7618	3.5544

表 2. 基于用户的 JacPCC 预测表

5.3.4 模型三、基于项目的评分预测模型

以上两种模型都是基于用户相似度的模型，其通过计算两个用户在公共项目集合上的相关系数进而研究两者之间偏好的相似程度。事实上，两个用户之间存在交集的集合可能只有很少的元素，这就会导致评分矩阵具有高度的稀疏性。

定义一个数据集的稀疏性为 $S_c = 1 - \frac{N_r}{N_u \cdot N_I}$，其中 N_r 代表所有评分总数，N_u 表示所有用户总数，N_I 表示所有书籍总数。显然，当每个用户都对所有书籍作过评价时，$S_C = 0$。通过计算机程序，计算出“predict.txt”的稀疏高达 0.9232，说明此数据内容具有高度的稀疏性。

基于项目的相关性分析问题与基于用户的相关性分析原理类似，只不过是由研究两个用户之间偏好的相似程度变为研究某两个项目之间的相似度。下面建立基于项目的预测模型。

（1）生成邻居项目集合

任意两本书籍 i 和 j 之间的皮尔逊相关系数为

$$sim(i,j) = \frac{\sum_{u \in U_{i,j}} (r_{u,i} - \overline{r_i}) \cdot (r_{u,i} - \overline{r_j})}{\sqrt{\sum_{u \in U_{i,j}} (r_{u,i} - \overline{r_i})^2} \sqrt{\sum_{u \in U_{i,j}} (r_{u,i} - \overline{r_j})^2}}$$

其中 $U_{i,j}$ 为对项目 i 和 j 都有评分的项目集合，$\overline{r_i}$ $\overline{r_j}$ 分别为项目 i 和 j 被评分的平均值。

同样的道理，生成项目 i 的邻居项目集合 $P_i = [I_1, I_2, \cdots, I_k]$，其中 $I_1, I_2, \cdots, I_k$ 分别代表与目标项目 I 最接近的前 k 个邻居项目。

（2）基于项目的评分预测模型

生成以上的邻居项目集合以后，就可以对项目 i 所能获得的评分进行预了。

预测模型为

$$r_{u,i}^* = \overline{r_i} + \frac{\sum_{j \in P_i} sim(i,j) \cdot (r_{u,j} - \overline{r_j})}{\sum_{j \in P_i} |sim(i,j)|}$$

式中，$r_{u,i}^*$ 表示用户 u 对项目 i 的预测评分，P_i 为项目 i 的邻居用户集合。式子代表着用项目邻居用户集的加权平均值去近似估计目标项目 i 所能够获得的评分。

（3）模型求解

仿照模型一、二的方法，编写 MATLAB 程序求解模型。最终得出各个用户对待评书籍的预测值，与基于 JacPCC 的模型预测值进行比较，如图所示。

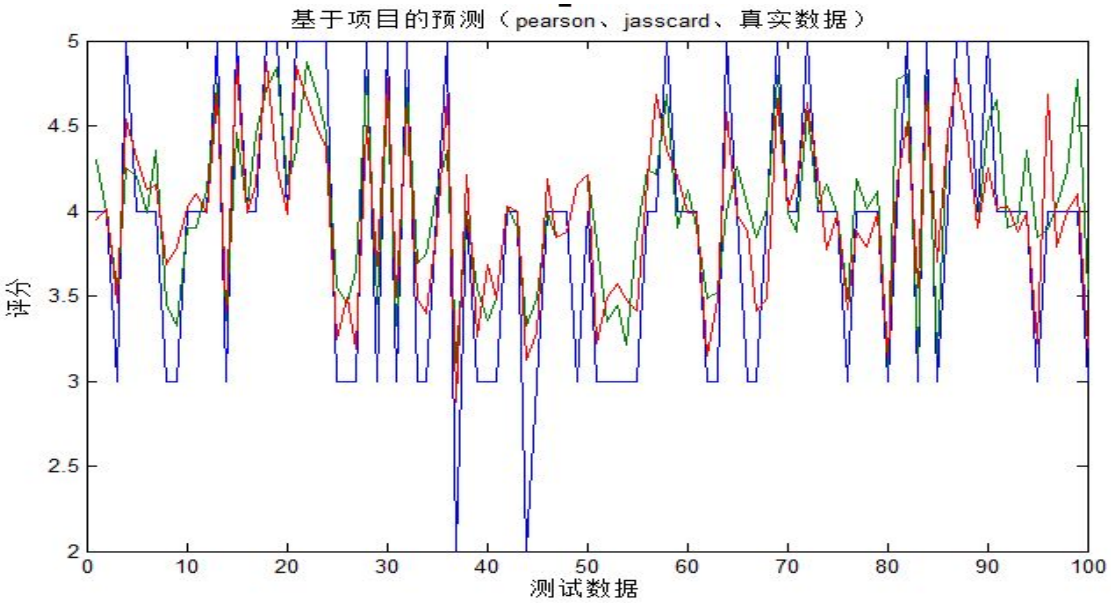

图 4. 基于项目的 PCC 与基于用户的 JacPCC 误差对比(绿线：基于项目的 PCC)

对比两种模型的预测情况，发现基于项目的模型预测精确度稍优于基于 JacPCC 的预测模型，而 JacPCC 模型的预测精确度又优于其他基于用户的预测精确度。显然，基于项目的模型具有较为明显的优势。

借助这个模型，对待评项目进行评分估计，得到其预测值，见附件。

5.4 问题三、推荐模型

根据问题二的讨论，对于问题三的推荐环节，主要考虑用户所青睐的书籍类型以及用户对书籍的评分预测值。在用户偏好不会随时间有大的变化的情况下，推荐用户所喜欢类型的书籍往往是比较合适的。同时，还需根据用户对某本书籍的评分预测值来综合考虑，用户评分的预测值越高，表明用户对某本书越可能产生兴趣。

（1）*Top*5 标签

题目给出了用户的阅读历史资料，包括需要做推荐的 6 个用户的历史阅读资

料，通过对这些资料的筛选，发现 6 名用户历史所读书籍大多可以在“book_tag.txt”文档里面找到其对应的书籍标签，因此，从用户历史数据中挖掘用户偏好信息成为可能。

设用户 i 对应的历史书籍集合为 H_i，其元素为用户 i 所读过的书籍，显然 $i \le 6$。将用户 i 的所有历史书籍与所有存在的标签构成一个标签矩阵

$$T=\begin{bmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,n} \\ t_{2,1} & t_{2,2} & \cdots & t_{2,n} \\ \cdots & \cdots & \cdots & \cdots \\ t_{m,1} & t_{m,2} & \cdots & t_{m,\mathrm{n}} \end{bmatrix}$$

其中，m 代表第 i 个用户的历史书籍总数，n 代表所有标签数目，

其中，$t_{i,j}=\begin{cases} 0 & tag_j \in Tag_{i,m} \\ 1 & tag_j \notin Tag_{i,m} \end{cases}$，$tag_j$ 表示第 j 个标签，$Tag_{i,m}$ 表示第 i 个用户历史读过的第 m 本书的标签集合。由此给出**偏好因子**的定义：

第 i 个用户对与标签 tag_j 的偏好因子为 $\lambda_{i,j}=\dfrac{\sum_{r=1}^{m_i} t_{r,j}}{\sum_{j=1}^{m_i}\sum_{r=1}^{m_i} t_{r,j}}$

式子的**意义在于：通过研究某种标签在用户历史书籍中出现的总次数占所有标签出现总次数的比值，来定量描述用户对具有某种标签的喜好程度。**因而显然有 $\sum^{n} \lambda_{i,j}=1$。

通过编程计算出 6 个用户对于所有标签的偏好因子，并选取 $Top5$ 标签作为候选，为下一步推荐做准备。这六个用户偏好标签 Top5 如下表所示:

用户 ID	TOP5 标签 ID/权重				
	1	2	3	4	5
7245481	6067/0.0317	6391/0.0263	3924/0.0175	6449/0.0171	7336/0.0136
7625225	6391/0.0237	6067/0.0191	7515/0.0186	2099/0.0186	5380/0.0169
4156658	6067/0.0253	6391/0.0252	3924/0.0175	6449/0.0161	9230/0.0156
5997834	6067/0.0264	6391/0.0264	6449/0.0215	4528/0.0182	5380/0.0149
9214078	6391/0.0248	3924/0.0232	9230/0.0205	7336/0.0192	6067/0.0167
2515537	3924/0.0295	7736/0.0259	6067/0.0222	5896/0.0197	6391/0.0192

表 3　Top5 权重表

（2）基于历史的推荐模型

通过（1），找出了用户 i 对应的 $Top5$ 标签，但是有着这些标签的书籍尚有很多，而且即使是用户所喜欢类型的书籍，也不一定会让用户满意。因此，可以考虑结合问题二的分数预测模型考虑书籍的推荐问题。

通过问题二的分析得知基于杰卡德—皮尔逊相关系数的预测模型预测精度优于皮尔逊相关系数预测模型，而基于项目的预测模型的预测精度又优于基于用户的杰卡德—皮尔逊相关系数预测模型。因此，在对“predict.txt”中的用户进行推荐时，可以使用问题二中模型三提出的基于项目的预测模型得到的预测值作为推荐的参考依据。

由于需要考虑书籍标签对推荐的影响，故使用“book_tag.txt”中的书籍作为推荐集合。针对用户 i，首先从这些文档中筛选出有一个或多个标签为 $Top5$ 标签的书籍。设用户 i 的 $Top5$ 标签集合为 Tag_i^5，用户已看过书籍集合为 B_d，筛选出的书籍集合为 B_c，其中第 j 本书的标签集合为 Tag_j，全部书籍集合为 B，则有

$$B_c \subset B, B_c \cap B_d = \Phi, Tag_j \cap Tag_i^5 \neq \Phi$$

筛选出具有 $Top5$ 标签的书籍之后，下面建立**基于历史的推荐模型：**

1）基于项目的目标集合评分预测

按照问题二模型三中的步骤，对已经过筛选的书籍进行评分预测时，首先求解出其项目邻居集合，书籍 i 与书籍 j 之间的相关系数为

$$sim(i,j)=\frac{\sum_{u\in U_{i,j}}(r_{u,i}-\overline{r_i})\cdot(r_{u,i}-\overline{r_j})}{\sqrt{\sum_{u\in U_{i,j}}(r_{u,i}-\overline{r_i})^2}\sqrt{\sum_{u\in U_{i,j}}(r_{u,i}-\overline{r_j})^2}},$$

其邻居项目集合为 $P_i=\lceil I_1, I_2, \ldots I_k \rceil$，$I_1, I_2, \ldots I_k$ 表示与目标项目最接近的前 k 个邻居项目，而基于项目的目标项目评分预测值为

$$r_{u,i}^{*}=\overline{r_i}+\frac{\sum_{j\in P_i} sim(i,j)\cdot(r_{u,j}-\overline{r_j})}{\sum_{j\in P_i}|sim(i,j)|}$$

2)推荐模型

将历史阅读数据纳入考虑之后，定义下面的**综合评价指数**

$$F_{u,i} = \lambda_{\Sigma}^{i} r_{u,i}$$

$$\lambda_{\Sigma}^{i} = \sum_{j=1}^{k} \lambda_{u,j}$$

其中，λ_{Σ}^{i} 表示书籍 i 的 Top5 标签的对应用户 u 的偏好程度的总和，k 书籍 i 含有的总 Top5 标签数目，因为一本书可能有多个标签，而当其中有几个标签都是用户所感兴趣的 Top5 标签时，这本书对于用户 u 就相应有更强的吸引力，其获得用户综合好评的可能也越大；将书籍 i 的评分预测值也作为因子的原因在于需要考虑这本书本身的优劣程度，因为即使某书籍本身所属题材是用户感兴趣的，但也可能因为其作者水平低劣的原因导致无法让用户满意。

整个筛选出推荐书籍的过程如下所示：

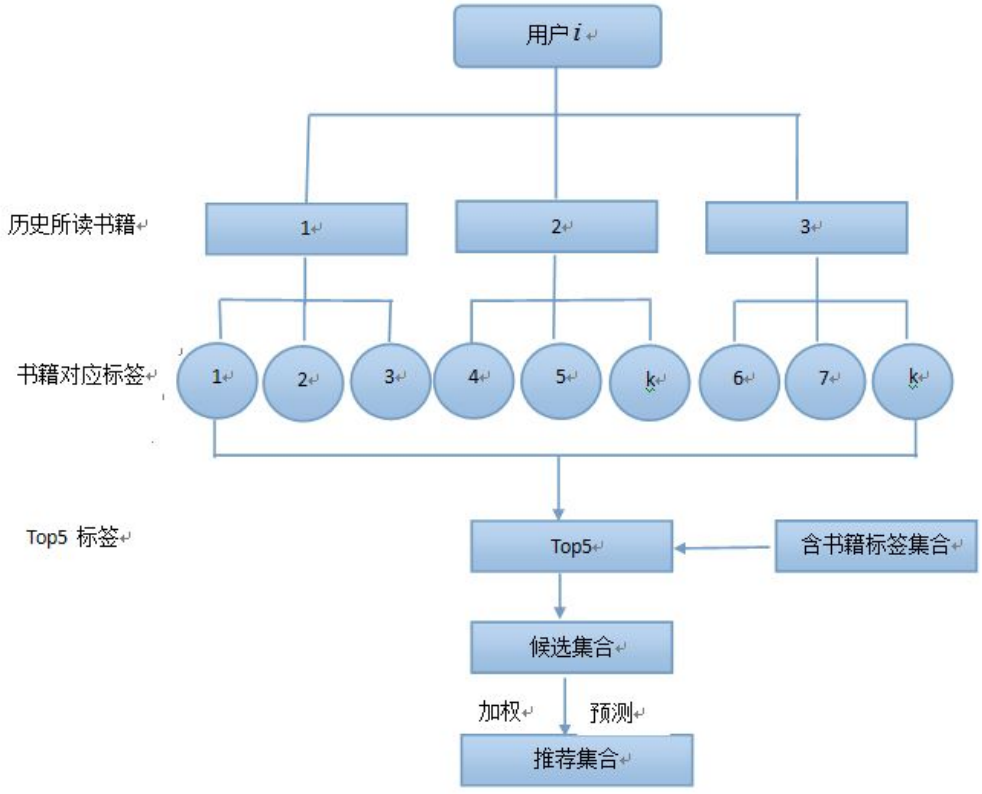

图 5. 推荐书籍生成过程

根据以上的定义，编写 MATLAB 程序求出 B_c 中各个元素的综合评价指数，并取其中最高的三个值所对应的书籍作为对应用户的推荐书籍。求出对 6 个用户的推荐书籍如下表所示。

六、模型评价

6.1 问题一

6.1.1 优点

针对问题一建立了两个模型。一个为线性拟合模型，研究书籍标签对用户评分的影响来，并采用梯度下降算法求得均方误差的最小值，最终求得各个标签在评分中占有的权重。该模型考虑合理，算法较优，有着较好的理论基础。另一个模型则为考虑社交关系影响的模型，通过定义用户偏好相似度的方法求取有着单方向社交关系的偏好相似度，并定量分析了单方向社交关系对用户评分的影响系数。该模型充分利用题目所给资源，考虑充分，并有量化数据支撑，具有较好的原创性与合理性。

6.1.2 缺点

模型一变量数目较多，且梯度下降算法时间复杂度较高，故模型求解时受计算机性能影响较大，而初值敏感性问题又可能导致迭代过程不收敛。

6.2 问题二

6.2.1 优点

模型采用推荐系统中广泛应用的协同过滤算法建立了三个模型。模型一采用皮尔逊相关系数求取两个用户之间的相关关系，并在此基础上采用基于用户的皮尔逊相关系数预测模型。模型二在模型一的基础上采用杰卡德—皮尔逊相关系数，弥补了模型一由于忽略两用户交集大小而可能会导致误判的缺陷。模型三在模型二的基础上，再次加以改进，将基于用户的预测模型进化为基于项目的预测模型，实验结果表明模型三的预测精度高于模型二。三种模型都有着较高的理论依据以及实践经验，且模型之间不断优化，使预测精度不断提高，具有较好的推广性。

6.2.2 缺点

模型一、二是基于用户的预测模型，由于只考虑两个用户之间所评论书籍的

公共部分，而忽略了用户之间公共部分很小的情况，因而会有误判的情况发生。而模型三则是基于项目的预测模型，虽在一定程度上提高了预测精度，但其没有从用户阅读历史的角度上来考虑，忽略了用户在阅读标签上的选择对评分的影响，预测评分仍具有一定的误差。

6.3 问题三

6.3.1 优点

通过考虑用户历史所读书籍数据直接分析用户对某种标签的喜好程度，并定义偏好因子来定量衡量，而后筛选出目标用户的 Top10 标签，继而生成具有 Top10 标签的书籍集合；结合问题二中基于用户的预测模型，对以上所述集合中的元素作评分预测，并用乘以对应书籍的综合偏好因子作为最后对书籍的综合评价。此模型充分考虑到用户阅读历史对其选择书籍上的影响，同时将用户偏好书籍的预测评分纳入考虑，具有一定的创新性。模型求解结果显示其预测精度较基于内容的预测更高，具有较好的推广性。

6.3.2 缺点

预测结果仍具有一定程度的误差，推荐效果尚不完美，还需将更多因素考虑进来，并需要大量的数据支撑。

参考文献:

[1]孙慧峰.基于协同过滤的个性化 web 推荐[D].北京：北京邮电大学，2012：42-46.

[2]程飞.基于用户相似性的协同过滤推荐算法研究[D].北京：北京邮电大学，2012:19-27.

[3]刘青文.基于协同过滤的推荐算法研究[D].北京：中国科学技术大学，2013:46-51.

[4]高惠璇.应用多元统计分析[M].北京：北京大学出版社,2007:249-258.

[5]邓爱林，朱扬勇，施伯乐.基于项目评分预测的协同过滤推荐算法[J].软件学报，2003,14(09):2-4.

[6]黄创光，印鉴，汪静，刘玉葆，王甲海.不确定近邻的协同过滤推荐算法[J].计算机学报，2010，33（8）：2-5.

[7]Goldberg D，Nich ol D，Oki B，Terry D.Using collaborative filtering to weave an information tapestry.Communications of the ACM，19921，35(12):54-67.

[8]Breese J，Hecherman D，Kadie C. Empirical analysis of the predictive algorithms for collaborative filtering. Processing of the 14th Conference on Uncertainty in Artifical Intelligence. 1998:29-47.

[9]Sarwar，Karypis G，Konstan J，Riedl J. Item- based collaborative filtering recommendation algorithms.Proceeding of the 10th International World Wide Web Conference.2001.279-284.

[10]吴湖，王永吉，王哲.两阶段联合聚类协同过滤算法[J].软件学报，2010,21（5）：139- 181.

[11]曾小波，魏祖宽，金在弘.协同过滤系统的矩阵稀疏性问题研究[J].计算机应用，2010，（004）：123- 134.

[12]李春，朱珍民，高晓芳.基于邻居决策的协同过滤推荐算法[J].计算机工程，2010,36（13）：27.

[13]Billsus D，Pazzani M J. Learning collaborative information filters. Proceedings of the fifteenth international conference on machine

learning，volume 54，1998. 54

[14]Pearson K.LIII. On lines and planes of the closest fit to systems of points in space. The London，Edinburgh，and Dublin Philosophical Magazine and Journal of Science，1901，2（11）：421- 435.

[15]Frank R.Giordano，William P.Fox，Steven B.Horton.数学建模[M].北京：机械工业出版社，203-245.

[16]卓金武，魏永生，秦健.MATLAB 在数学建模中的应用[M].北京：北京航空航天大学出版社，154-167.