一种基于协作过滤的电子图书推荐系统*)

曾庆辉　邱玉辉
(西南师范大学计算机与信息科学学院　重庆400715)

摘　要　推荐系统中最常见信息过滤技术是基于内容的过滤和协作过滤，协作过滤由于有其自身的优点得到迅速发展，并得到广泛应用，但传统的协作过滤算法存在着稀疏性、扩展性和同义性等问题。本文提出一种基于评价矩阵列向量的图书协作过滤算法，并把这个算法应用到了一个数字图书馆的电子图书推荐系统中。此图书协作过滤算法主要计算图书之间的相似度而不是用户之间的相似度，可以大大降低计算量。实验也表明，这个算法比传统的基于用户的协作过滤算法有优势。

关键词　推荐系统，协作过滤，图书推荐，评价矩阵，数字图书馆

An E-Book Recommender System with Collaborative Filtering

ZENG Qing-Hui　QIU Yu-Hui
(School of Computer and Information Science, Southwest China Normal University, Chongqing 400715)

Abstract　Collaborative filtering and content-based filtering are the most common information filtering technology in recommender system. Collaborative filtering is becoming the popular one and has been used widely because of its good quality. But traditional collaborative filtering algorithm has the shortcomings of sparsity, scalability and synonymy. In this paper, we present a new collaborative filtering algorithm base on the column-vector of the evaluations matrix for an e-book recommender system in the digital library. The algorithm computes the similarity of books instead of the similarity of users, which can remarkably alleviate the workload. Our experiments suggest that the algorithm provides better performance than user-based algorithm.

Keywords　Recommender system, Collaborative filtering, Book recommending, Evaluations matrix, Digital library

1　引言

随着信息社会的到来，我们已经越来越依赖于因特网获得我们所需要的信息。但是，网上信息的迅速增长已经日益超出我们能够准确辨识的能力范围。如果我们想在网上找一本自己感兴趣的图书来阅读的话，那么通过传统的检索系统或搜索引擎所获得的大量相关信息往往并不是我们所需要的，也就是出现了所谓的“信息过载”和“信息迷向”现象。为了克服这种信息获取困难，推荐系统也就应运而生，它帮助我们从“信息迷向”中走出来，从众多相关信息中找出对我们来说最有价值的那部分。

推荐系统中最常见的两种信息过滤技术就是基于内容的过滤和协作过滤。前者主要考虑信息项本身的内容是否和用户兴趣相关，并计算这两者的相似度作为是否推荐的依据；协作过滤则是考虑和用户兴趣相类似的其他用户对某个信息项的喜好，并由此判断用户对此信息项是否感兴趣以决定是否推荐。

虽然这两种过滤技术无论在理论上还是在实践上都很成功，在信息检索和电子商务领域也都得到有效应用，但还是有各自的不足。对于基于内容的过滤来说，内容特征提取能力有限、无法推荐更多更新的信息资源和需要过多的用户反馈是其主要的缺点；而对于协作过滤来说，稀疏性、可信度和随着规模增长而带来计算复杂度的几何增长是其面临的主要问题。

本文中，我们将讨论如何在数字图书馆中对读者进行图书的推荐。由于基于内容的过滤的特征提取的能力有限：通常只能对资源进行比较简单的特征提取，在一些领域，目前还没有有效的特征提取方法，如：图像、视频、音乐等；即使文本资源，其特征提取方法也只能反映资源的一部分内容，例如，难以提取关于信息质量的信息，而这些特征可能影响用户的满意度。因此，我们采用协作过滤的技术，同时针对协作过滤技术存在的主要问题做了相对应的改进，在一定程度上克服了稀疏性带来的不利影响，提高了推荐的可信度，而且采用了预先计算的方法，降低了实时计算的难度。

本文第2节对现有的协作过滤技术和算法进行分析总结；然后在第3节对图书推荐系统做了整体描述，其中着重讨论了应用在系统中的协作过滤算法；第4节做了实验验证；最后一部分对整个系统进行总结，说明今后要做的工作。

2　协作过滤技术

协作过滤技术已经成功地广泛应用于各种推荐系统中，例如GroupLens，应用协作过滤技术，可以自动地为用户寻找与其兴趣相似的其他用户，用户对系统中的文章做出自己的兴趣评价，系统也为用户推荐其他用户感兴趣的文章。与众多需要用户显式评价的推荐系统不同，Phoaks等则应用隐式用户评价为用户推荐信息。在商业领域的应用，著名的电子商务网站Amazon.com根据客户对产品的评价或者客户的购买历史等信息应用协作过滤策略为客户推荐购买产品。总而言之，应用协作过滤技术的目标就是针对某一特定的用户，根据他先前对事物的评价和其他相似用户的对事物的评价，计算出他对其未评价过的新事物的感兴趣程度以便系统向他进行推荐。

*)本文研究得到重庆市自然科学基金资助，项目编号：CSTC，2004BB2086。曾庆辉　硕士研究生，主要研究方向：信息过滤，推荐系统；邱玉辉　教授，博士生导师，主要研究方向：人工智能。

一个典型的协作过滤可以形式化描述为:参与协作过滤的有 m 个用户 $U=\{u_1,u_2,\cdots,u_m\}$ 和代表 n 个事物的信息项 $I=\{i_1,i_2,\cdots,i_n\}$,每一个用户 u_i 有一个集合 I_{u_i},表示用户 u_i 已经做过兴趣评价的事物。这里有一个特殊用户 $u_a\in U$ 称作当前用户(active user),协作过滤算法就是为该用户预测可能喜欢的事物并把它推荐给用户。一般有两种形式来表示结果:

·预测(Prediction) 是指计算一个数值 $P_{a,j}$,它表示用户 u_a 对信息项 i_j 的感兴趣程度,$i_j\notin I_{u_a}$。这里值 $P_{a,j}$ 是系统预测的,其取值范围和用户的评价级别范围一样。

·推荐(Recommendation) 是指系统给出 N 个信息项 I_r,它们代表系统推荐给当前用户的可能最感兴趣的 N 个新事物,$I_r\in I$ 且 $I_r\cap I_{u_a}=\Phi$。

协作过滤的一般过程有三步,如图1所示。

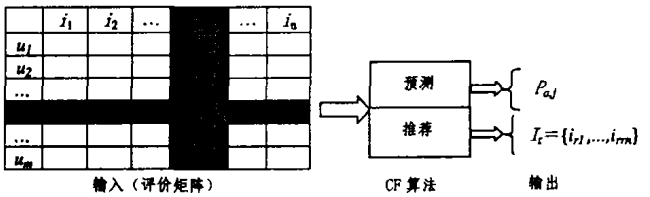

图1 协作过滤过程

第一步,协作过滤算法把 m 个用户对 n 个信息项的兴趣评价当成一个 $m\times n$ 的评价矩阵 A,并作为算法的输入。矩阵中的每一项 $a_{i,j}$ 表示第 i 个用户对第 j 个信息项的兴趣评价,用户对某一事物的兴趣评价一般通过用户显式提交而获得,也可以通过分析用户的浏览记录,浏览时间,网页超链接等信息来隐式获取。通常用某一范围之内的数值大小表示兴趣评价级别的高低,数值0也通常用来表示用户尚未对此信息项做出兴趣评价。

第二步,也是整个协作过滤最重要的一步,要根据用户对信息的评价矩阵,计算出用户之间的相似度,建立相似用户组。进行用户相似度计算的方法有以下几种:向量夹角余弦(Cosine-based Similarity),用户相关相似度(Correlation-based Similarity),预设选票(Default Vote),倒转使用频率(Inverse User Frequency),实例放大(Case Amplification)等。这其中向量夹角余弦方法最为简单,但为了使相似度计算更加准确,需要单独考虑两个用户都评价过的信息项,因此通常使用得更多的是用户相关相似度,计算公式如下:

$$S_{i,j}=\frac{\sum_k (r_{i,k}-\bar{r}_i)(r_{j,k}-\bar{r}_j)}{\sqrt{\sum_k (r_{i,k}-\bar{r}_i)^2}\sqrt{\sum_k (r_{j,k}-\bar{r}_j)^2}}$$

其中,$S_{i,j}$ 是用户 i 与用户 j 的相似度,$r_{i,k}$ 是用户 i 对信息 k 的喜好评价值,$r_{j,k}$ 是用户 j 对信息 k 的喜好评价值,$\bar{r}_i$ 是用户 i 对所评价信息的评价均值,$\bar{r}_j$ 是用户 i 对所评价信息的评价均值。

当相似用户组建立之后,根据和当前用户在同组中每个成员对某信息的评价信息,预测当前用户对该信息的偏好程度;或者筛选出最受该组用户欢迎的信息。

第三步,输出。根据第二步的计算结果判断当前用户对某个信息是否感兴趣或者将筛选出的结果推荐给当前用户。

有学者根据协作过滤中采用的方法,将其分为 Memory-Based 和 Model-Based 两大类。

1)Memory-Based 方法在进行推荐时要比较计算用户的历史记录,以找出与用户历史记录相似也即兴趣相似的其他用户。这其中最常见的方法就是最近邻居法(Nearest Neihbors)。

2)Model-Based 方法主要将用户的历史记录通过统计方法或者机器学习的方法要建构用户的兴趣模型,进而再利用这一兴趣模型来进行推荐。建模时所使用的方法有:潜在语义索引(Latent Semantic Indexing,LSI),关联规则(Association Rule)以及贝叶斯网络(Bayesian Network)等。

协作过滤技术也存在着一些潜在的不足,这些不足之处包括以下几点。

·稀疏性。在计算用户相似性 $s_{i,j}$ 时,我们可以从公式中发现,必须存在两个用户都做出评价的同一信息项 k,如果不同用户评价过的信息没有重叠的部分,即不存在信息项 k 的话,相似性的计算根本无法进行;而且如果要得到较为准确的结果,重叠的信息项还不能太少。因此,如果信息项得到的用户评价比较少,使得重叠的信息项更少,必然导致推荐质量下降。

·扩展性。协作过滤系统的用户和信息项的增长都有一定的限制,以最常见的最近邻居法为例,如果评价矩阵是 $m\times n$ 维的矩阵,那它的算法时间复杂度将达到 $O(m^2\times n)$。如果评价矩阵规模很大,那带来的计算复杂度将使协作过滤失去实际意义。

·同义性。由于协作过滤只对资源的重叠程度进行分析,使得形式不同而内容相同的资源将被看作不同的资源,导致相似兴趣的用户难以建立联系。

3 基于协作过滤的电子图书推荐系统

随着互联网的兴起和发展,有关数字图书馆的研究和建设也越来越引起大家的重视,但是就目前建设的数字图书馆来看,往往都是包括电子图书在内的文献信息资源的堆砌,导致读者往往需要花费很大的精力才能找到自己真正需要的文献信息。针对这一问题,我们设计了一个数字图书馆中的电子图书推荐系统(以下简称"本系统"),应用协作过滤技术,为每一个读者提供符合其阅读兴趣的个性化图书推荐服务。

本系统首先取得用户对图书的评价信息,然后综合专家对每一类图书的推荐意见,对每一类图书形成一个评价矩阵;协作过滤模块将分析每个评价矩阵,从中找出阅读这一类图书的相似兴趣用户组;然后根据同一组内各用户已有的对不同图书的评价,判断当前用户对此图书的喜好程度,并把可能是当前用户最感兴趣的那些图书作为推荐的结果返回给用户;用户对系统推荐图书进行阅览和评价,用户的评价将返回给系统更新评价矩阵。本系统整体结构如图2所示。

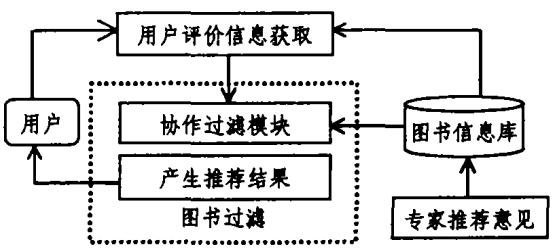

图2 应用协作过滤的电子图书推荐系统总体结构

本系统采用了协作过滤技术来对用户进行图书推荐,为了尽量避免典型协作过滤技术的不足带来的不利影响,我们根据应用的实际情况做了相应的改进。

为了保证在评价矩阵为空(如系统刚投入使用)以及评价

矩阵信息不足的情况下系统的正常运行，“专家推荐意见”被引入到系统中。“专家推荐意见”是指各学科专家或图书管理人员对图书的分类和相对应于每一类的推荐图书，相当于专家系统里“知识库”的概念。因此即使评价矩阵为空的情况下，系统也可以采纳“专家推荐意见”把图书推荐给用户。而且，本系统引入“专家推荐意见”还克服了典型协作过滤系统中信息推荐过度依赖于普通用户评价的不足。在典型的推荐系统中，如果同一兴趣组内的所有用户都没有阅读过的图书，那么这本图书是不可能被系统推荐给用户的，即使那是一本很有阅读价值的图书；但是在本系统中，有了专家的推荐意见的这本图书就会推荐给用户，实现“为书找人”，让真正有价值的图书不会找不到它的读者。为了便于系统协作过滤算法的运算，专家推荐意见将和普通用户对图书的评价信息一起记录在一个相同的评价矩阵中；同时为了体现专家推荐意见的特殊性，专家对图书的评价信息将赋予更高的权重，我们可以通过本系统的训练集数据(training set)来确定权重系数的大小。

本系统的核心部分是协作过滤模块，所有的推荐结果的产生都是先由协作过滤算法计算相似性后进行判断才能得到。典型的协作过滤系统往往都是从评价矩阵的行向量这个角度来计算相互之间的相似性，即计算用户之间的相似性。如常用的最近邻居法，就是从矩阵行向量中找出与当前行向量最相似的那几个，也就代表了当前用户的几个“最近邻居”。但是对于本系统来说，图书的信息相对来说是固定的，但是用户的信息和数量却是相对来说变化比较大的，如果采用基于行向量的方法，每次用户信息或数量有变化都要重新计算整个行向量之间的相似度，无疑会给系统带来很大的压力。我们考虑从评价矩阵的列向量，即从图书信息角度来考虑协作过滤算法，通过图书之间的相似性，来判断读者对其未做评价的图书的感兴趣程度。

当本系统针对某一个读者进行图书推荐时，协作过滤算法如下：

0)获取用户对不同图书的评价信息，建立图书评价矩阵；这一步应在进行协作过滤之前完成；

1)从评价矩阵中查找当前用户已经评价过的图书信息，构成已评价图书集 N；

2)选定目标图书 i；

3)从评价矩阵中查找这些用户：这些用户既评价过集合 N 中任一本图书又评价过目标图书 i；所有这些用户构成用户集 U；

4)计算集合 N 中图书 j 与图书 i 的相似度，为了避免单纯的夹角余弦相似度计算中因为不同用户采用不同的评价级别而带来的误差，算法采用调整的夹角余弦计算相似度，计算公式如下：

$$s_{i,j}=\frac{\sum_{u\in U}(r_{u,i}-\bar{r}_u)(r_{u,j}-\bar{r}_u)}{\sqrt{\sum_{u\in U}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum_{u\in U}(r_{u,j}-\bar{r}_u)^2}}$$

其中，$r_{u,i}$是用户 u 对信息 i 的喜好评价值，$\bar{r}_u$ 是用户 u 对所评价信息的评价均值；注意，用户 u 必须同时对图书 i 和图书 j 做过评价；

5)在集合 N 中根据每本图书与图书 i 的相似度的大小按降序排列选定 k 个与图书 i 最相似的图书，构成相似图书集 $I=\{i_1,i_2,\cdots,i_k\}$，它们与图书 i 的相似度记为 $\{s_{i1},s_{i2},\cdots,s_{ik}\}$；

6)从当前用户对相似图书集 I 中每本图书的已有评价计算当前用户对图书 i 的评价，采用计算公式如下：

$$P_{u,i}=\frac{\sum_{n\in I}(S_{i,n}\times r_{u,n})}{\sum_{n\in I}|S_{i,n}|}\times\alpha$$

其中，$s_{i,n}$是图书 i 与 n 之间的相似度，$r_{u,n}$是用户 u 对信息 n 的喜好评价，α 是专家推荐的权重系数，若图书 i 非专家推荐图书，则为 $\alpha=1$；

7)选定另一目标图书，重复3至6，直到无目标图书可选；

8)选取预测评价值最高的那几本目标图书推荐给读者。

上述算法的第1步和第3步其实就是分别对评价矩阵的行向量和列向量进行降维，再加上第5步只选择 k 本图书进行进一步计算，而在实际运行中往往 $k\ll n$，实际上也是进一步的降维过程；而且，由于图书的信息是相对固定的，因此图书之间的相似性还可以进行预先计算(即算法的2至5步)，并把计算结果——相似图书集 $I=\{i_1,i_2,\cdots,i_k\}$存放在一个快表里，当系统需要的时候，就直接从快表里查得相应的相似图书信息，然后直接进行算法的第6步，将很快得到最后的推荐结果。由以上可见，图书协作过滤算法与典型的协作过滤算法相比，通过降维大大降低了算法计算量，通过预先计算将大大提高算法的运行效率，最终可以缩短系统的响应时间。

4 实验结果

实验的目的是为了验证本系统是否能够有效运行，得到满意的推荐结果。实验主要是对系统的图书协作过滤算法进行测试。

实验的数据来源于西南师大数字图书馆中读者对超星电子图书库的访问数据。超星电子图书库是西南师大数字图书馆的一个重要数字资源库，有40万册各类电子图书，每周都有上千人次浏览和下载。在实验中，我们采用从读者的行为来判断读者对图书的感兴趣程度，避免了要求读者对图书进行评价这种显式反馈给读者带来的额外负担。读者阅读某本图书的页数多少往往可以反映读者对这本图书的感兴趣程度，在实验中我们分析了每种评价级别所对应的阅读页数，然后系统就可以据此得到读者对图书的评价信息，最后这些评价信息转换成读者-图书评价矩阵。读者对图书的评价级别从1到10等，1代表很不感兴趣，10代表很感兴趣，其他则处于中间状态，数字越大表示越感兴趣。

本文的实验只对计算机类部分图书进行推荐，从实际评价数据得到评价矩阵是一个20×25的矩阵，非零元素有72个，矩阵稀疏度按照公式 $Sparsity=1-\frac{\text{矩阵中非零元素个数}}{\text{矩阵中所有元素个数}}$ 来计算，为0.856。

为了进行对比验证，我们同时也实现了一个采用最近邻居法的协作过滤算法与本系统所采用的协作过滤算法进行实验对比。

本文的实验是在一台安装 Windows XP Professional 的 PC 机上进行，其 CPU 为 P4 2.0GHz，内存为512MB。算法用标准 C 程序实现。

第一个实验是评价准确性。平均绝对误差 MAE(Mean Absolute Error)是一个广泛使用的衡量评价准确性的指标，主要是计算系统预测推荐值与用户真正评价值之间的偏离程度。针对每一个兴趣预测级别对$\langle P_i,q_i\rangle$，其中 P_i 为系统预测值，q_i 为用户评价值，MAE 首先对所有预测级别对$\langle P_i,q_i\rangle$计算出它们之间的绝对误差值$|P_i-q_i|$和所有这些预测级别对的绝对误差和，然后计算它们的平均值，相应的公式为：

$$MAE=\sum_{i=1}^{N}|P_i-q_i|/N$$

MAE 值越小，推荐系统给用户提供的预测级别越准确。

在实验中，我们将事先从评价矩阵中随机去掉5个已有的评价值，然后分别运行本系统采用的协作过滤算法和基于最近邻居法的协作过滤算法得到新的评价值，再与原有的评价值进行误差统计，得到各自的 MAE 值；同时，我们还将评价矩阵的用户数5到20进行变化，得到 MAE 值的变化结果如图3所示。

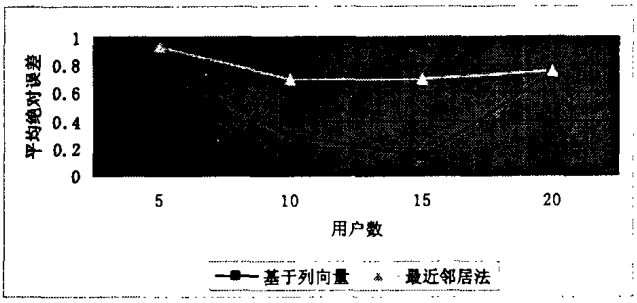

图3 两种算法在不同用户数情况下的 MAE 值

第二个实验是评价有效性。本实验所谈的推荐系统有效性主要指系统推荐给读者的图书顺序应该与读者的兴趣高低是一致的。这里我们借用排列的逆序数这个概念，我们规定系统推荐的图书的顺序应该按照读者的兴趣高低来排列，并计算系统实际推荐图书顺序的逆序数，用 t 来表示。t 越小，说明推荐系统越有效。

在实验中，我们将事先从评价矩阵中随机去掉同一用户5个不同级别评价值，然后分别运行两个算法得到新的评价值，把图书按新的评价值从大到小排列的顺序与图书按原有的评价值从大到小排列的顺序进行比较，得到逆序数 t，同时还计算了这5个评价值的 MAE。我们还将评价矩阵的用户数5到20进行变化，得到 t 值的变化结果和对应的 MAE 值如图4所示，柱状图表示逆序数，折线图表示平均绝对误差。

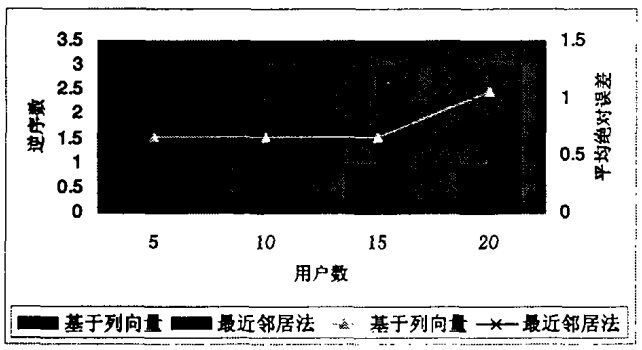

图4 两种算法在不同用户数情况下的 t 值和 MAE 值

从以上两个实验的结果我们可以看出，在大部分的情况下，本系统采用的协作过滤算法的 MAE 值要比采用最近邻居法的协作过滤算法的 MAE 值小，而且产生的图书推荐顺序的逆序数也更低，因此无论从准确性的角度还是从有效性的角度来看，本系统所使用的基于列向量的协作过滤算法比典型的基于用户的协作过滤算法（如对比实验采用的最近邻居法）更有优势。

本系统在实际应用中也体现了这个优点，作者本人做了个小测试，在系统现有的评价矩阵中加入了作者对某些图书的评价信息，其中特意对有关 ASP 编程的图书均给予比较高的评价级别，而对有些关于 XML 的图书给予比较低的评价级别。系统运行后给作者推荐了清华大学出版社出版的《ASP 应用大全》；而倘若系统使用最近邻居法来实现的话，那么推荐结果将是中国水利水电出版社出版的《XML 实用大全》，排在第二位的才是《ASP 应用大全》。

小结及展望 推荐系统已经成为现代信息社会中越来越重要的帮助人们获取信息的技术手段，它可以让用户获取到他们真正感兴趣的信息，为了满足数字图书馆读者的个性化服务的要求，本文就应用协作过滤技术设计了一个图书推荐系统。系统中应用的协作过滤算法从图书之间相似性的角度出发，有效地克服了稀疏性、扩展性等传统协作过滤系统的不足，提高了准确性和有效性。

本文提出的图书推荐系统也存在需要进一步研究的地方，如对用户兴趣的获取方式上，在相似图书的判断上加入基于内容的分析等，这些都将在以后的研究中进行深入探讨。

参 考 文 献

1 Sarwar B, Karypis G, Konstan J, et al. Item-based Collaborative Filtering Recommendation Algorithms. WWW10, Hong Kong, 2001

2 McNee S M, Albert I, Cosley D, et al. On the Recommending of Citations for Research Papers. CSCW'02, New Orleans, Louisiana, USA, 2002

3 Sarwar B, Karypis G, Konstan J, et al. Analysis of Recommendation Algorithms for E-Commerce. In: Proc. of the ACM E-Commerce 2000 Conf. 2000. 58～167

4 Breese J S, Heckerman D, Kadie C. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. In: Proc. of the 14th Annual Conf. on Uncertainty in Artificial Intelligence, 1998. 43～52

5 Zhang Tong, Iyengar V S. Recommender Systems Using Linear Classifiers. Journal of Machine Learning Research, 2002(2): 313～334

6 Adomavicius G, Tuzhilin A. Recommendation Technologies: Survey of Current Methods and Possible Extensions. Working paper, Stern School of Business, New York University, 2003

7 吴志宏. 以隐性回馈为基础的自动化推为机制:[硕士论文]. 朝阳科技大学，中国台湾