doi:10.16104/j.issn.1673-1891.2021.02.014

高校图书馆个性化图书推荐算法研究

张德青,程　锦

(安徽三联学院计算机工程学院,安徽 合肥 230601)

摘　要:构建图书推荐系统,不仅可以让用户快速有效地获取所需图书信息,减少信息过载,同时也可以较好地发挥图书馆馆藏图书资源的潜在价值。在综述了几种常用推荐算法的基础上,给出了基于协同过滤的推荐算法的实现过程,并针对其冷启动和数据稀疏性问题给出了优化方案及优化后的算法实现流程。结果表明:在算法中引入用户特征属性与用户聚类方法,有效降低了数据稀疏性问题,提升了算法的推荐效率,一定程度上解决了图书推荐系统中的推荐算法设计。可以将该优化后的算法运用于图书馆的图书推荐系统设计中。

关键词:推荐算法;相似度;数据稀疏

中图分类号:TP391.3;G258.6　　**文献标志码**:A　　**文章编号**:1673-1891(2021)02-0078-04

Research on Personalized Book Recommendation Algorithmfor University Libraries

ZHANG Deqing, CHENG Jin

(School of Computer Engineering Anhui Sanlian University, Hefei Anhui 230601, China)

Abstract: Constructing a book recommendation system can not only allow users to quickly and effectively obtain the required book information and reduce information overload, but also can make better use of the potential value of the library´s collection of book resources.On the basis of summarizing several common recommendation algorithms, the implementation process of recommendation algorithm based on collaborative filtering is given, and the optimization scheme and the optimized algorithm implementation process are given for its cold start and data sparsity. The results show that the user characteristic attribute and user clustering method were introduced into the algorithm, which effectively reduced the data sparsity problem, improved the recommendation efficiency of the algorithm, and solved the recommendation algorithm design in the book recommendation system to a certain extent. The optimized algorithm can be used in the design of library book recommendation system.

Keywords: recommendation algorithm; similarity; data sparsity

0 引言

伴随着互联网+、大数据时代的到来,各类信息数量激增直至信息过载。如何在海量数据中高效准确地获取所需信息、提升图书馆藏利用率成为近年来的研究热点,个性化推荐系统应运而生,并广泛应用于电商、搜索引擎、智慧服务推荐等多个领域当中。同时,推荐系统也在高校图书馆中掀起了研究热潮,以清华大学开发的 Open Bookmark[1] 和南京大学的 Book+图书推荐系统为代表,中国人民大学、浙江大学等均开发了自己的图书个性化推荐系统,大大提升了读者书目资源检索的效率。

个性化图书推荐系统通过获取读者的相关信息,采用推荐算法将计算出的用户可能感兴趣的图书信息资源主动推荐给读者,变被动检索为主动推荐,进一步发挥潜在价值书目资源的价值。因此,推荐算法的研究便成为整个个性化图书推荐系统设计的核心内容。本文从高校图书馆为出发点,结合目前图书馆的建设现状,通过分析选择协同过滤推荐算法作为主要的图书推荐算法,并重点介绍其实现过程,同时结合实际情况,考虑到了个人特征信息对推荐结果的影响因素而提出了算法的优化方案。

收稿日期:2020-10-07

基金项目:安徽三联学院校级自然科学基金项目(KJYB2019008);安徽省教育厅自然科学基金重点项目(KJ2019A0891)。

作者简介:张德青(1984—),女,安徽合肥人,副教授,硕士,研究方向:机器学习。

1 常用推荐算法

个性化推荐这一概念产生于 20 世纪末,随后美国明尼苏达大学的 GroupLens 小组着手对推荐算法展开深入研究并推出了首个推荐系统,提出了协同过滤技术。伴随着“互联网+”时代的发展,推荐算法被应用于更广泛的领域。相较于国外的推荐系统,国内的研究起步相对较晚,“当当网”于 2006 年率先实现了对用户提供个性化的推荐服务后,受到了用户的青睐。基于图书的个性化推荐系统的研究就此拉开帷幕,但仍有一些高校图书馆至今尚未研发或使用任何推荐系统。

1.1 主流算法介绍

目前,在个性化图书推荐领域中经常使用的算法有基于协同过滤的推荐算法、基于关联规则的推荐算法、基于内容的推荐算法和基于混合的推荐算法。

基于协同过滤的推荐算法最早被应用于邮件推荐系统中,该算法可以分为:基于用户的协同过滤推荐算法(User-Based CF)和基于项目的协同过滤推荐算法(Item-Based CF)。基于用户的协同过滤推荐算法是通过挖掘用户的历史行为数据,使用最近邻算法(KNN)计算相似用户集,进而在该用户集中进行物品推荐。而基于项目的协同过滤推荐算法则是通过计算不同用户对不同物品的评分获得物品间的关系,基于物品间的关系对用户进行相似物品的推荐。目前,该算法是推荐系统中使用频率高的算法之一。

基于内容的推荐算法是利用机器学习的方法从内容的特征描述中匹配到用户感兴趣的信息[2],更适合应用于文本推荐领域[3],它克服了协同过滤推荐算法中可能存在的冷启动的问题,只要获取用户的历史行为数据,即可进行推荐,且随着历史行为数据的增加,推荐的信息更加准确,其弊端在于无法挖掘用户潜在兴趣的信息。

基于关联规则的推荐算法就是从大量的数据中找到不同项目之间的内在联系,目的在于挖掘出强规则[4]。此算法最重要的是通过支持度 Support(X,Y)的计算从而找到最大频繁项集,进而通过计算置信度 Confidence(X,Y)找到强规则,Apriori 算法则是用来生成频繁集的常用的算法之一。虽然关联规则算法能计算出项目间的关联,但也容易产生无效的规则。

基于混合的推荐算法则是对各种推荐算法的优缺点进行扬长避短的一种组合推荐。针对具体应用背景的不同,算法的组合方式也不同,通常可以有加权、变换、特征组合与特征扩充等多种方式来达到准确度高的推荐效果。这里拟采用加权的方式来实现对算法推荐效果的优化。

1.2 图书推荐算法分析

以上几种推荐算法已经被广泛应用于各类推荐系统中,基于不同推荐系统的不同特点,所选择的推荐算法的策略差异性也较明显。高校图书馆的个性化推荐系统所面向的用户群体较为稳定,主要是全体在校师生;且同一院校的用户知识结构相近,可以按照专业对用户进行分类;主要以图书推荐为主,较少涉及其他领域。在系统设计之初,不仅要考虑到读者当前的兴趣需求,也要兼顾挖掘内在关联图书,提高馆藏图书的使用率。基于以上分析的用户群体的特点,本文将研究基于用户的协同过滤推荐算法来有效地解决高校图书馆图书推荐的有效性和准确性问题。

2 基于用户的协同过滤推荐算法

2.1 算法思路

基于用户的协同过滤推荐算法的算法思路可以分 3 步进行。

第 1 步:根据读者用户的历史行为数据建立对书籍的 $m\times n$ 的评分矩阵 $\boldsymbol{U}$,如式(1)所示。

$$\boldsymbol{U}=\begin{pmatrix} U_{11} & U_{12}\cdots U_{1j}\cdots U_{1n} \\ U_{21} & U_{22}\cdots U_{2j}\cdots U_{2n} \\ \vdots & \vdots \quad \vdots \\ U_{i1} & U_{i2}\cdots U_{ij}\cdots U_{in} \\ \vdots & \vdots \quad \vdots \\ U_{m1} & U_{m2}\cdots U_{mj}\cdots U_{mn} \end{pmatrix} \tag{1}$$

式中:m 表示参与计算的用户人数;n 表示图书数量,册;A_{ij}表示用户 i 对第 j 本图书的评分,分值越高说明用户对该图书的偏好程度越高。

第 2 步:在第 1 步的基础上继续计算出待推荐的目标用户的相似用户集合。通常,计算相似度的算法常用的有余弦函数法和 Pearson 相关度算法 2 种,分别如式(2)和(3)所示。

$$\mathrm{Sim}\ (x,y)=\frac{\sum_i U_{x,i}\cdot U_{y,i}}{\sqrt{\sum_i U_{x,i}^2}\sqrt{\sum_i U_{y,i}^2}} \tag{2}$$

式中:$\mathrm{Sim}(x,y)$表示用户 x 和 y 的相似度,$U_{x,i}$ 表示用户 x 对图书 i 的评分,$U_{y,i}$ 表示用户 y 对图书 i 的评分。

$$\mathrm{Sim}_1(x,y)=\frac{\sum_i (U_{x,i}-\overline{U}_x)(U_{y,i}-\overline{U}_y)}{\sqrt{(U_{x,i}-\overline{U}_x)^2}\sqrt{(U_{y,i}-\overline{U}_y)^2}} \tag{3}$$

式中：$\overline{U_x}$ 和 $\overline{U_y}$ 分别表示用户 x 对所有图书的评分均值和用户 y 对所有图书的评分均值。

虽然 Pearson 相关度算法较余弦函数法复杂一些，但对于评分数据不规范时，Pearson 相关度算法可以给出更好的结果，本文选择公式(3)作为计算用户相似度的公式。

第 3 步：进行图书推荐。根据第 2 步的计算结果，求出目标用户的 K 个最近邻用户，进一步计算出用户对每一待推荐图书的预测评分来确定最终推荐的图书结果，如式(4)所示。

$$P_{x,i} = \sum_{y \in K} \mathrm{Sim}(x,y) \cdot U_{y,i} \qquad (4)$$

式中：$P_{x,i}$ 表示用户 x 对于推荐图书 i 的评分。通过公式(4)可以计算出用户 x 对于所有推荐图书的预测评分，进而得到排名前 K 位的图书，并推荐给用户 x。

2.2 算法实现

表 1 模拟了安徽三联学院计算机工程学院 12 名学生(用户)的借阅数据。

表 1 借阅数据

用户	图书名								
	A	B	C	D	E	F	G	H	I
U1	1	0	1	0	1	0	1	0	0
U2	1	0	0	1	1	0	1	0	1
U3	0	1	1	0	0	1	0	1	0
U4	1	1	1	0	1	1	0	0	0
U5	1	1	1	0	1	0	0	0	1
U6	1	1	0	1	0	0	1	1	0
U7	1	0	1	1	0	0	0	0	1
U8	0	1	1	0	0	0	1	1	0
U9	1	0	0	0	1	0	1	1	1
U10	0	0	1	1	0	1	0	1	0
U11	1	1	1	1	0	1	0	0	0
U12	1	1	1	1	0	1	0	0	0

注："1"表示已借阅，"0"表示未曾借阅。

2.2.1 建立用户评分矩阵

根据以上借阅记录，建立用户评分矩阵，如表 2 所示。用户对图书的评分数据范围为 1~5 分。

表 2 用户评分矩阵

	A	B	C	D	E	F	G	H	I
U1	3	0	1	0	3	0	3	0	0
U2	4	0	0	1	2	0	1	0	3
U3	0	3	4	0	0	2	0	4	0
U4	2	3	3	0	3	1	0	0	0
U5	3	3	3	0	3	0	0	0	4
U6	4	2	0	4	0	0	1	3	0
U7	5	0	3	4	0	0	0	0	4
U8	0	3	5	0	0	0	2	5	0
U9	3	0	0	0	5	0	4	1	3
U10	0	0	3	3	0	4	0	2	0
U11	4	1	2	3	0	3	0	0	0
U12	4	1	2	4	0	4	0	0	0

使用公式(3)的 Pearson 相关度算法计算出用户间的相似度，结果(保留到小数点后 3 位)的散点图如图 1 和 2 所示。

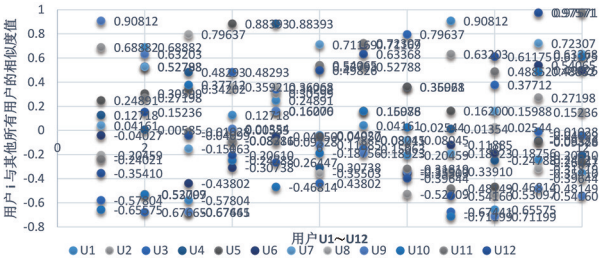

图 1 用户相似度散点图

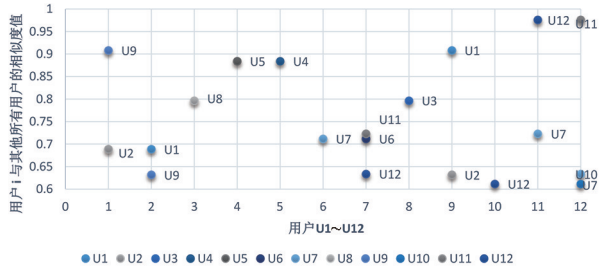

图 2 用户相似度>0.6 的相似度散点图

分析以上计算结果，可以将用户进行分类。使用公式(4)求解用户的预测评分，计算产生的用户未曾借阅过的图书的预测评分结果如图 3 所示，取评分值最高的前 N 项，从而得到 Top-N 推荐结果。

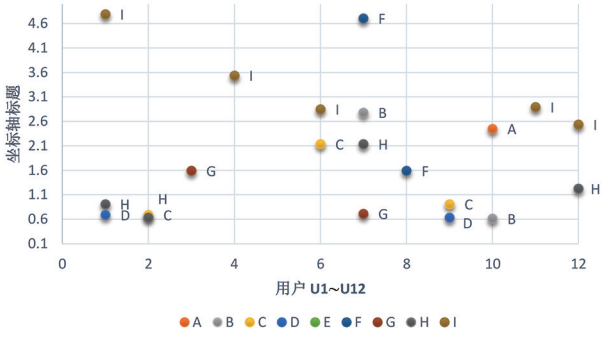

图 3 推荐预测评分结果

2.2.2 算法优化

虽然上述算法可以计算出用户相似度，并进行图书推荐，但还需要改进。首先是未考虑到用户自身信息的差异性，未将用户特征信息考虑至用户评分矩阵中，导致推荐的准确性不足；同时还存在大量用户对借阅记录的无评分操作，同时图书馆图书数量和种类庞大，都将会产生大量的数据稀疏问题，降低了推荐算法的效率。因此，现针对以上提出的 2 个问题提出以下的优化方案。

1)引入用户的特征的权重信息，提升算法的准确率。在现有的基础上，考虑到用户的性别、学院、专业、年级等信息，建立用户信息编码向量，然后求

出用户特征信息向量间的欧氏距离。

$$\mathrm{Sim}_2(x,y)=\sqrt{(R_{x,i}-R_{y,i})^2} \quad (5)$$

得到综合相似度计算公式:

$$\mathrm{Sim}(x,y)=\alpha\cdot Sim_1(x,y)+(1-\alpha)\cdot Sim_2(x,y) \quad (6)$$

式中:α($0<\alpha<1$)表示权重系数,通过多次反复实验,根据实验结果确定 α 的取值。该综合相似度计算公式兼顾了用户特征信息和用户的原始评分数据,使得相似度评价更加客观。

2)建立用户分层聚类,降低评分矩阵的稀疏性。根据用户所属学院进行第一层分类,用户感兴趣的图书作为第二层分类,将不属于同一范围的用户清洗、删除。对于新用户可以使用注册信息得到初始评分,大大减少 0 评分数据。实现对数据的降维处理,减轻计算量。这里采用 Slope-one 算法[5]进行计算,先计算出用户 u、v 对所有共同评分图书书目的均分差值:

$$R(u,v)=\frac{\sum_i(r_{u,i}-r_{v,i})}{|N_u\cap N_v|}\ (i\in(N_u\cap N_v)) \quad (7)$$

则用户 v 对图书 j 的缺失填充评分为:

$$P_{v,j}=r_{u,j}-R(u,v) \quad (8)$$

式中:$N_u\cap N_v$ 表示用户 u 和 v 共同评分的个数,$r_{u,i}$ 表示用户 u 对图书 i 的评分,$r_{u,j}$ 表示用户 u 对图书 j 的评分。

2.2.3 改进后的算法流程及结果

1)建立用户对图书的原始评分矩阵,定义二维数组 $U_{i,j}$(式 1)。

2)建立用户分层聚类,对数据进行降维处理,使用公式(8)去预测矩阵中的未评分项,即二位数组 $U_{i,j}$ 中的 0 值。

3)引入用户的特征信息,反复实验,确定参数 α 的值,建模得到用户的综合相似度计算公式(式(6))。

4)最后使用公式(4)计算出用户对各本待推荐图书的预测评分,依次从高到低,得到预期的 Top-N 推荐。

在算法中引入了用户特征属性与用户聚类方法,有效降低了数据稀疏性问题,提升算法的推荐效率,一定程度上解决了图书推荐系统中的推荐算法设计。

3 结语

推荐系统目前已成为各类网站设计的重要功能,本文兼顾了当前高校图书馆图书推荐的实际需求,从分析常用的推荐算法着手,详细给出了基于用户的协同过滤推荐算法的计算过程,并通过实验验证了推荐结果的有效性。同时针对基于协同过滤算法存在的数据稀疏性问题,提出优化方案,进而可以将该算法运用于高校图书馆的推荐系统设计中。大数据技术的发展使得各系统之间不再相互独立,借助于机器学习技术获取异构系统中的大量相关数据并为推荐系统所用,提高推荐准确性的同时,其数据处理与用户偏好数据提取将成为下一步研究的重点。

参考文献:

[1] 冯翱,刘斌,卢增祥,等.Open Bookmark——基于 Agent 的信息过滤系统[J].清华大学学报(自然科学版),2001,41(3):85-88.

[2] 顾立志.电子商务主要推荐技术研究[J].计算机光盘软件与应用,2014,17(8):41-42.

[3] 张宇航,姚文娟,姜姗.个性化推荐系统综述[J].价值工程,2020(2):287-292.

[4] 黄坤元.高校图书推荐系统算法与模型研究[D].呼和浩特:内蒙古大学,2019.

[5] 戎静月.改进的协同过滤推荐算法研究[D].保定:河北大学,2020.