doi:10.16104/j.issn.1673-1891.2020.01.017

医学专业图书推荐算法比较研究

邱煜炎,吴福生
(蚌埠医学院图书馆,安徽 蚌埠 233000)

摘要:[目的]基于医学读者行为,提供合理的医学专业图书推荐算法。[方法]利用蚌埠医学院图书管理系统借阅历史记录,构建用户-项目偏好指标评价体系,针对目前流行的各种推荐算法进行对比实验。[结果]基于矩阵分解的推荐算法均方根误差分值较低,表现较好。

关键词:推荐算法;医学资源;推荐系统
中图分类号:G252; TP391.3 **文献标志码:**A **文章编号:**1673-1891(2020)01-0079-05

A Comparative Study of Medical Books Recommendation Algorithms

QIU Yuyan, WU Fusheng
(Library of Bengbu Medical College, Bengbu, Anhui 233000, China)

Abstract: [Objective] Based on the behavior of medical readers, we provide a reasonable medical books recommendation algorithm. [Methods] Historical records of book borrowing in the library management system of Bengbu Medical College are used to develop a user-project preference index evaluation system, and comparative experiments were conducted on a variety of present popular recommendation algorithms. [Results] The recommendation algorithm based on matrix factorization has lower root-mean-square error and better performance.

Keywords: recommendation algorithm; medical science resource; recommendation system

0 引言

随着信息技术的高速发展,信息过载问题日益严重。推荐系统通过建立用户与信息产品之间的二元关系,利用已有的选择过程或相似性关系挖掘每个用户潜在感兴趣的对象,进而进行个性化推荐,其本质就是信息过滤。它是目前解决信息过载问题最有效的工具。

作为推荐系统的核心,推荐算法一直是学界研究的热门领域。目前比较流行的推荐算法有基于关联规则、基于内容、基于人口统计学以及协同过滤和基于矩阵分解的隐因子推荐算法。其中,协同过滤算法由于在准确度和多样性方面综合表现优异,受到业界追捧,落地应用比较成熟。而针对推荐系统数据集稀疏性特点设计的矩阵分解算法,目前也越来越受到学术界和工业界的重视。本文主要利用蚌埠医学院图书管理系统读者借阅历史记录,构建读者-图书偏好评分模型[1],并将清洗后的数据切分成训练集和测试集两部分,针对协同过滤技术下基于用户、基于物品以及矩阵分解下基于物品和基于用户和物品的推荐算法进行对比实验,为医学专业图书推荐系统的设计提供参考[2]。

1 推荐算法综述

推荐算法是推荐系统的核心,算法的质量直接影响推荐系统的性能。因此,学术界和工业届都非常重视设计和优化推荐算法模型上。

Chen 和 Sycara K 等人提出基于文献内容的推荐算法[3],其设计思想是利用自然语言处理等文本挖掘技术,充分提取物品的特征描述以及用户属性和用户对物品评价描述的信息,这建立了用户项目偏好相似性函数。然后学习训练分类器,该分类器根据用户和项目特征确定用户偏好。

Goldberg D等提出协同过滤算法[4],其思想是根据用户产生的历史行为记录,利用欧式距离、余弦相似度或者皮尔逊相关系数等相似度计算函数对相关特征进行对比,找到与其最相似的一个用户集,然后推荐系统利用这些用户集评分最高项作为

收稿日期:2019-07-29
基金项目:安徽省教育厅自然科学重点研究项目(KJ2019A0393);蚌埠医学院校自然科学基金项目(BYKY1762)。
作者简介:邱煜炎(1982—),男,安徽蚌埠人,讲师,硕士,研究方向:机器学习、推荐系统。

推荐备选。在工业届,协同过滤算法受到商业推荐系统的广泛欢迎。

基于人口统计学的推荐可以解决推荐系统冷启动的问题,仅凭借人口统计相关信息即可产生推荐结果[5]。该算法仅使用用户基本信息如性别、年龄、学历、职位等信息,衡量用户的相似性,将与当前相似用户偏好的物品推荐给当前用户。这类推荐系统可以有效地解决推荐系统中常见的冷启动问题。

针对评分数据稀疏的问题,基于知识的推荐系统应运而生,它以一种交互的方式通过特定的领域知识向用户推荐符合要求的物品[6]。其算法分为两种类型:基于限制的推荐和基于用例的推荐。第一种通过用户对物品属性需求限制的交互,系统将需求根据知识转化为规则,并根据这些规则返回满足的物品。第二种,在基于用例的推荐系统中,用户在键入属性之后,系统通过相似函数返回相关物品。

Netflix Prize比赛始于2006年,无数的专家学者因为其丰厚的奖金去深入研究并设计各种推荐算法。该比赛产生出一系列矩阵分解模型,而LFM(Latent Factor Matrix)隐语义模型是最著名的,之后成为工业届推荐系统的主流算法。LFM算法的思想是对用户物品评分矩阵进行奇异值分解,然后采用随机梯度下降的拟合方法进行模型参数优化,计算出的矩阵模型隐式包含用户兴趣和物品特征,基于此方法去预测用户对物品的偏好及评分。

2 基于真实数据集的评价指标构建

2.1 数据采集

本文数据来源于蚌埠医学院图书馆的金盘图书管理信息系统后台的Oracle数据库,选取时间段为2014年1月1日至6月30日,拟设计的数据表字段主要包括图书登录号、读者登录号、操作类型(借书、还书、续借、罚款等)、处理时间和图书分类号。读者覆盖全体教职工和在校学生。

2.2 数据表合并

为统一数据集,需要对多个数据表的字段进行合并提取。数据集采用多表联合查询进行提取,多表数据源包括借书信息表、还书信息表、续借信息表、R类图书信息表和借阅信息表。

2.3 数据清洗

数据处理在对读者操作类型(J、H、X、F、S、C、D、K、M)研究分析后,选取J、H和X三种操作类型,创建了三个信息表:借书信息表、还书信息表、续借信息表,见表1、表2、表3。

借书信息表Oracle语句:

Create table 借书 as select 条形码,读者条码,登录号,主键码,操作类型 as 借书,处理时间 as 借书时间 from 流通日志 t where 处理时间 between to_date ('2014-03-01', 'yyyy-mm-dd') and to_date ('2014-06-01', 'yyyy-mm-dd')and 操作类型 not IN ('X','M','S','H','K','C','D','F')order by 处理时间;

表1 借书信息表

条形码	读者条码	登陆号	主键码	借书	借书时间
*	*	*	*	J	*

还书信息表Oracle语句:

Create table 还书 as select 条形码,读者条码,登录号,主键码,操作类型 as 还书,处理时间 as 还书时间 from 流通日志 t where 处理时间 between to_date('2014-03-01', 'yyyy-mm-dd') and to_date ('2014-06-01', 'yyyy-mm-dd')and 操作类型 not IN ('X','M','S','J','K','C','D','F')order by 处理时间;

表2 还书信息表

条形码	读者条码	登陆号	主键码	还书	借书时间
*	*	*	*	H	*

续借信息表Oracle语句:

Create table 续借 as select 条形码,读者条码,登录号,主键码,操作类型 as 续借,处理时间 as 续借时间 from 流通日志 t where 处理时间 between to_date('2014-03-01','yyyy-mm-dd') and to_date ('2014-06-01', 'yyyy-mm-dd')and 操作类型 not IN ('H','M','S','J','K','C','D','F')order by 处理时间;

表3 续借信息表

条形码	读者条码	登陆号	主键码	续借	借书时间
*	*	*	*	X	*

从借书信息表和还书信息表中可以得到借书总表,操作是用还书时间减去借书时间得到借书天数,再通过两个信息表,联合建立新表。此外,需要把关于医学图书的借阅信息抽取出来,形成的新表就是读者对医学类图书的借阅信息(医学图书借阅信息表)。通过中国图书分类法对医学类图书分类信息中,寻找R类图书即创建R类图书信息表,见表4、表5。

R类图书信息表Oracle语句:

Create table 医学图书 as select t.条形码,s.索书号 from 流通日志 t,馆藏典藏库 s where t.条形

码=s.条形码 and 索书号 like 'R%';

表4　R类图书信息表

条形码	索书号
*	R…

R类图书借阅信息表Oracle语句:

Create table 医学图书 as select distinct t.条形码,t.读者条码,t.主键码,t.登录号,t.操作类型 as 借书,s.操作类型 as 还书,(s.处理时间-t.处理时间)as 借书天数 from 借书 t,还书 s,医学图书 m where t.操作类型 != s.操作类型 and m.条形码 = s.条形码 and t.条形码 = s.条形码 and t.登录号=s.登录号 and t.读者条码=s.读者条码 and t.主键码 = s.主键码;

表5　R类图书借阅信息表

条形码	读者条码	主键码	登录号	借书	还书	借书天数
*	*	*	*	J	H	*

2.4 评分数据集构建

本文利用读者借阅时间构建读者-图书兴趣度偏好模型[7-8],见图1。图中,α点表示盲目借阅时间点,β点为一般图书最大借书时间点。若读者借阅图书时间在0到α内,则表示盲目借阅;在α到β内表示正常借还;在β到2β内,蓝线表示短期超期,橙线表示续借。

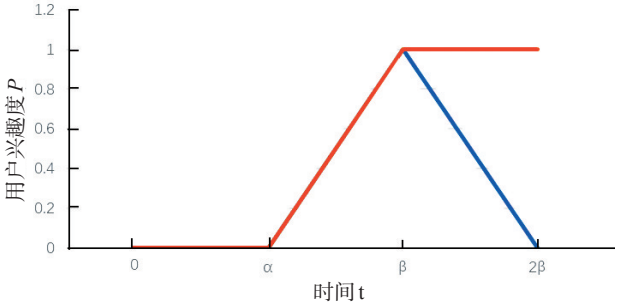

图1　图书借阅时间t与兴趣度p关系折线图

基于此思想,将医学图书借阅信息表按借书天数分成8类:分数一(盲目借阅,借阅天数<5 d)、分数二(5 d≤借阅天数<20 d)、分数三(20 d≤借阅天数<35 d)、分数四(35 d≤借阅天数<50 d)、分数五(50 d≤借阅天数<60 d),超期的按照分数六(60 d≤借阅天数<70 d)、分数七(70 d≤借阅天数<85)、分数八(85 d≤借阅天数<100 d),由于表中最高借阅天数没有超过100 d,故没有超期的其它评分类。

各评分创建所需的Oracle语句如下:

Create table 盲目借阅 as select 条形码,读者条码,主键码,借书,还书,借书天数 from 借阅 t where 借书天数≤5;

Alter table 盲目借阅 add(评分 int default '1' not null);

表6　盲目借阅信息表

条形码	读者条码	主键码	借书	还书	借书天数	评分
*	*	*	J	H	*	1

Create table 分数二 as select 条形码,读者条码,主键码,借书,还书,借书天数 from 借阅 t where 借书天数 between 5 and 20;

Alter table 分数二 add(评分 int default '2' not null);

表7　分数二的信息表

条形码	读者条码	主键码	借书	还书	借书天数	评分
*	*	*	J	H	*	2

Create table 分数三 as select 条形码,读者条码,主键码,借书,还书,借书天数 from 借阅 t where 借书天数 between 20 and 35(或between 85 and 100);

Alter table 分数三 add(评分 int default '3' not null);

表8　分数三的信息表

条形码	读者条码	主键码	借书	还书	借书天数	评分
*	*	*	J	H	*	3

Create table 分数四 as select 条形码,读者条码,主键码,借书,还书,借书天数 from 借阅 t where 借书天数 between 20 and 35(或between 70 and 85);

Alter table 分数四 add(评分 int default '4' not null);

表9　分数四的信息表

条形码	读者条码	主键码	借书	还书	借书天数	评分
*	*	*	J	H	*	4

Create table 分数五 as select 条形码,读者条码,主键码,借书,还书,借书天数 from 借阅 t where 借书天数 between 35 and 50(或between 50 and 60);

Alter table 分数五 add(评分 int default '5' not null);

表10　分数五的信息表

条形码	读者条码	主键码	借书	还书	借书天数	评分
*	*	*	J	H	*	5

3 对比实验

3.1 相似矩阵

本实验分别基于用户和物品构建余弦系数相

似矩阵,修正余弦系数相似矩阵以及皮尔逊系数相似矩阵,然后通过均方根误差值(Root Mean Squared Error,RMSE)进行实验对比。

余弦相似度(cosine):通过不同的读者对不同的图书的评分,形成读者-评分的N维向量空间,若读者未对某图书进行评分则设为0分[9]。以读者A与读者B为例,读者A、B之间的余弦相似矩阵见式(1)。

$$\mathrm{sim}(A,B)=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}}\quad。\qquad(1)$$

修正的余弦相似性(adjusted cosine):通过不同的读者对不同的图书的评分,形成读者-评分的N维向量空间,若读者未对某图书进行评分则设为0分。在此内容基础上,考虑对不同读者的评分尺度,主要方法是使用读者评分减去图书的平均分[20]。以读者A与读者B为例,读者A、B之间的修正余弦相似矩阵见式(2)。

$$\mathrm{sim}(A,B)=\frac{\sum_{i=1}^{n}(A_i-\bar{I})\times(B_i-\bar{I})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{I})^2}\times\sqrt{\sum_{i=1}^{n}(B_i-\bar{I})^2}}\quad。(2)$$

皮尔逊相关系数(Pearson correlation coefficient)是反映两种不同的变量线性相关程度的统计量,是数学意义上的一种线性相关系数。在使用统计量的过程中需要对两个变量(观测值X_i和均值$\bar{X}$)进行处理,其中N为样本总量,通过相关系数r表示两个变量间线性相关的程度,相关系数r的绝对值越大表明相关性越强,反之越弱[10]。以读者A与读者B为例,读者A、B之间的皮尔逊相关系数矩阵见式(3)。

$$\mathrm{sim}(A,B)=\frac{\sum_{i=1}^{n}(A_i-\bar{A})\times(B_i-\bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}\times\sqrt{\sum_{i=1}^{n}(B_i-B)^2}}\quad。(3)$$

3.2 RMSE指标

均方根误差是预测值与真实值偏差的平方,与观测次数N比值的平方根,见式(4)。在实际测量中,观测次数N总是有限的,真值只能用最可信赖的最优值代替,均方根误差是用来衡量观测值同真值之间的偏差。在研究中使用时,数值越小,表明误差越小。由于RMSE值较大,因此采用以10为底的RMSE对数值(以下简称RMSE值)进行对比。

$$\mathrm{RMSE}(X,h)=\sqrt{\frac{\sum_{i=1}^{n}(h(x_i)-y_i)^2}{n}}\quad。\qquad(4)$$

3.3 实验结果及对比分析

3.3.1 基于用户协同过滤

利用基于用户的协同过滤算法进行10组实验,实验结果形成关于余弦相似系数矩阵、修正余弦系数矩阵和皮尔逊系数矩阵的RMSE值的折线图,其中X轴表示组数,Y轴表示RMSE值,见图2。

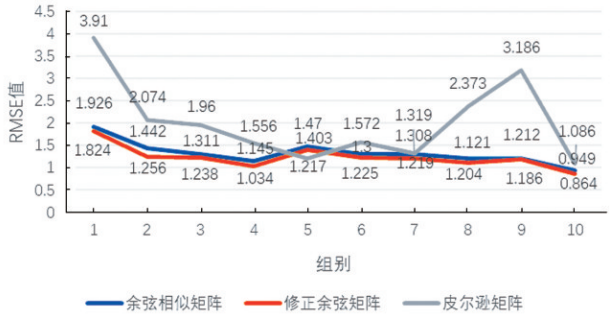

图2 基于用户协同过滤算法各相似矩阵的RMSE值

通过各系数矩阵的RMSE值折线图可以看出,皮尔逊系数矩阵的RMSE值的起伏程度大,余弦相似矩阵、修正余弦矩阵与皮尔逊系数矩阵折线趋势相近。皮尔逊系数矩阵的RMSE值的均值高于其它两个系数矩阵的RMSE值的均值。通过计算可知皮尔逊系数矩阵的RMSE值的均值在2.1以下,其它两个系数矩阵的RMSE值的均值在1.4以下,上下幅度差值在为3.1。

3.3.2 基于物品系统过滤

利用基于物品的协同过滤算法进行10组实验,实验结果形成关于余弦相似系数矩阵、修正余弦系数矩阵和皮尔逊系数矩阵的RMSE值的折线图,见图3。

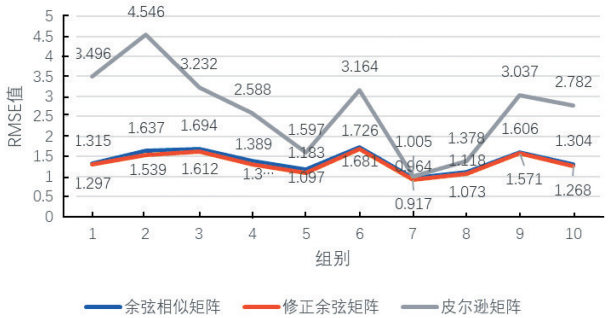

图3 基于物品协同过滤算法各相似矩阵的RMSE值

通过各系数矩阵的RMSE值折线图可以看出,皮尔逊系数矩阵的RMSE值的起伏程度大,余弦相似矩阵、修正余弦矩阵与皮尔逊系数矩阵折线趋势相近。皮尔逊系数矩阵的RMSE值的均值高于其它两个系数矩阵的RMSE值的均值。通过计算可知皮尔逊系数矩阵的RMSE值的均值在2.7以下,上下幅度差为3.6,其它两个系数矩阵的RMSE值的均值在1.4,上下幅度差值在1以内。

3.3.3 矩阵分解下基于用户和基于物品实验结果

对矩阵分解下基于用户和基于物品的推荐算

法分别进行实验,形成关于3组系数矩阵的RMSE值的2个折线图,分别见图4和图5。

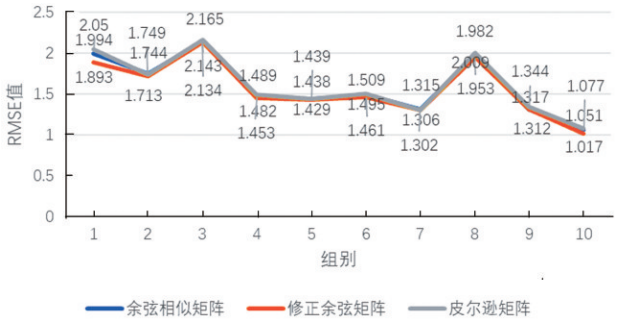

图4　各系数矩阵的RMSE值折线图

通过各系数矩阵的RMSE值折线图可以看出,余弦相似矩阵、修正余弦矩阵与皮尔逊系数矩阵折线趋势相近。余弦相似矩阵、修正余弦矩阵与皮尔逊系数矩阵RMSE均值相近。通过计算可知三个系数矩阵的RMSE值的均值在1.7以下,上下幅度差值在1.1以内。

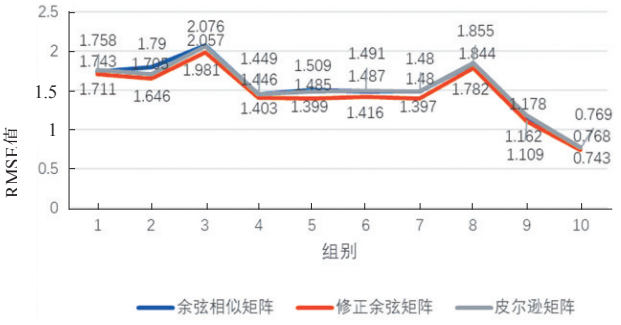

图5　各系数矩阵的RMSE值折线图

通过各系数矩阵的RMSE值折线图可以看出,余弦相似矩阵、修正余弦矩阵与皮尔逊系数矩阵折线趋势相近。余弦相似矩阵、修正余弦矩阵与皮尔逊系数矩阵RMSE均值相近。通过计算可知3个系数矩阵的RMSE值的均值在1.6以下,上下幅度差值在1.4以内。

在矩阵分解前后,基于用户和基于物品的均值相近。矩阵分解下基于用户和基于物品的各系数矩阵的RMSE值的均值在1.8以下,普遍低于未矩阵分解的基于用户和基于物品的各系数矩阵的RMSE值的均值。

基于用户和基于物品的RMSE值起伏程度相似,上下波动基本一致,矩阵分解下的基于用户和基于物品RMSE值起伏程度平缓,明显优于基于用户和基于物品的RMSE值起伏程度。

在矩阵分解前后,基于用户幅度差值基本保持一致,无明显变化。未矩阵分解的幅度差值均在3.0以上,矩阵分解后的幅度差值均在1.5以下。

对比后,基于用户和基于物品自身均值相近、起伏程度相仿和幅度差值相近,无明显差异。矩阵分解下基于用户和基于物品的均值、起伏程度和幅度差值均优于未进行矩阵分解下的推荐算法。

4　结论

本文利用蚌埠医学院图书馆图书管理信息系统中医学类图书的历史借阅记录,构建读者图书行为矩阵。研究通过对余弦相似矩阵、修正余弦系数矩阵和皮尔逊系数矩阵RMSE值的计算,进行了对协同过滤技术下基于用户、基于物品和矩阵分解下基于用户和基于物品推荐算法结果的对比实验。实验结果表明,在真实数据集环境下,基于用户和基于物品的协同过滤算法自身比较无明显变化,而矩阵分解下基于用户和基于物品要优于推荐算法。下一步研究可以添加新的推荐结果评价指标如F1、AUC、汉明距离评估多样性指标等,多重评价指标下尝试多种推荐算法如基于统计、基于知识、基于因子分解机以及基于深度学习的推荐算法进行对比实验,利用真实的医学图书借阅信息,完善推荐算法的比较研究。

参考文献:

[1] 林晓霞,刘敏,杨晓东,等.融合信任相似度的高校图书馆个性化推荐研究[J].数字图书馆论坛,2018(8):14-19.

[2] 何胜,熊太纯,柳益君,等. Spark的高校图书馆文献推荐方案及实证研究[J].图书情报工作,2017,61(23):129-137.

[3] CHEN L,SYCARA K.WebMate: A personal agent for browsing and searching[C].1998.

[4] GOLDBERG D,NICHOLS D,OKI B M,et al.using collaborative filtering to weave an information tapestry[J]. Communications of the ACM,1992,35(12):61-70.

[5] LINDEN G.Amazon.com recommendations: item-to-item collaborative filtering[J].IEEE Internet Computing,2003,7(1):76-80.

[6] FELFERNIG A,TEPPAN E,GULA B.Knowledge-Based recommender technologies for marketing and sales[J].International Journal of Pattern Recognition & Artificial Intelligence,2008,21(2):333-354.

[7] 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(9):1621-1628.

[8] 嵇晓声,刘宴兵,罗来明.协同过滤中基于用户兴趣度的相似性度量方法[J].计算机应用,2010,30(10):2618-2620.

(责任编辑:曲继鹏)