

# 一种基于数据挖掘的图书荐购模型研究

李澎林, 郑 莉, 李 伟

(浙江工业大学 计算机科学与技术学院, 浙江 杭州 310023)

**摘要:**针对高校图书馆图书荐购无法满足读者需求的问题,设计了一种基于读者行为数据挖掘的高校图书荐购模型,该模型包括挖掘活跃书籍和匹配荐购书单两个过程。提出了一种基于改进的  $k$ -means 算法的活跃书籍挖掘方法,该方法通过设置可调节的阈值来确定样本集中的噪声点,然后再根据最大距离法选取样本的初始聚类中心,最后基于该算法得出活跃书籍。实验证明:相比传统  $k$ -means 算法,改进的  $k$ -means 算法在活跃书籍挖掘过程中稳定性好、准确率高,满足了高校图书馆图书荐购的需求。

**关键词:**数据挖掘;图书荐购模型;活跃书籍; $k$ -means 算法

**中图分类号:**TP311

**文献标志码:**A

**文章编号:**1006-4303(2019)01-0080-06

## Research on a book recommendation model based on data mining

LI Penglin, JIA Li, LI Wei

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

**Abstract:** In view of the problem that book recommendation method can not meet the needs of readers in university libraries, an university book recommendation model based on the data mining on readers' behavior is designed. The model includes two parts: mining active books and matching book recommendation list. An active book mining method based on the improved  $k$ -means algorithm is proposed. In this method, the noise points in the sample set are determined by setting an adjustable threshold, and the initial cluster center of the sample according to the maximum distance method is selected. Finally, active books based on the algorithm are obtained. Experiments show that compared with the traditional  $k$ -means algorithm, the improved  $k$ -means algorithm has good stability and high accuracy in the active book mining process, which meets the demand of books recommendation in university libraries.

**Keywords:** data mining; book recommendation model; active books;  $k$ -means algorithm

随着当今时代社会政治、文化和经济的高速发展,各学科交叉融合,书目信息在不断的增长。据《2016 年中国图书零售市场报告》显示,2016 年,全国图书零售市场动销品种数 1 725.09 万种,新书品种数约为 21.03 万种。新书品种从 2012 年到 2016 年始终在 20~21 万种。面对种类繁杂的图书资源,高校图书馆如何选购到既符合质量要求,又符合读

者需求的图书将面临着巨大的挑战。传统的文献采访<sup>[1]</sup>是当前图书馆收集图书荐购信息的主要方式,但该方式工作量大、效率低,很难及时把握读者多变的图书需求。通过查阅近 5 年国内外相关文献资料,发现基于计算机相关技术的面向读者的图书个性化推荐研究较多,面向图书馆的图书荐购研究相对较少。陈大莲<sup>[2]</sup>基于微薄平台对高校读者图书荐

收稿日期:2018-01-05

作者简介:李澎林(1968—),男,浙江上虞人,教授,研究方向为管理信息系统、移动互联网技术,E-mail:lp1@zjut.edu.cn。

购进行了探析;孔云等<sup>[3]</sup>提出了面向“互联网+”图书荐购模型;刘华<sup>[4]</sup>介绍了读者决策采购在美国大学的实践及其对我国的启示。随着大数据<sup>[5]</sup>时代的到来,一些学者开展了基于数据挖掘技术的图书荐购研究。唐小新等<sup>[6]</sup>利用数据挖掘技术聚类分析法和分布式异构技术对荐购系统进行了设计与实现;周伟等<sup>[7]</sup>探讨了数据挖掘技术、PDA模式和混合推荐算法在高校图书馆荐购系统中的应用。概括起来,传统文献采访已经无法适应当前高校图书馆荐购需求,现有基于计算机技术支持的图书荐购方法研究较少,且因缺乏对读者借阅等行为数据的分析,难以准确把握读者需求,并未被高校图书馆广泛采用。

为此,笔者设计了一种基于读者行为数据挖掘的图书荐购模型,该模型以读者图书馆行为数据为基础,首先利用改进的  $k$ -means 算法挖掘活跃书籍,然后对活跃书籍进行分词,提取其中的关键字,最后根据这些关键字与书商提供的书单或者荐购系统中的书单进行匹配,从而智能快速地给出荐购书单,大大提高了荐购的主动性、科学性、准确性和及时性。

## 1 基于读者行为数据挖掘的图书荐购模型

图书荐购模型主要包含挖掘活跃书籍、匹配荐购书单两个过程,具体如图1所示。

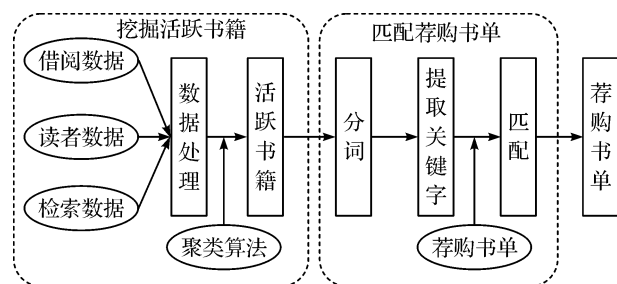


图1 图书荐购模型

Fig. 1 Book recommendation model

### 1.1 挖掘活跃书籍

挖掘活跃书籍是基于读者行为数据,通过聚类算法,得到既符合质量要求又受读者欢迎的图书的过程。高校图书馆业务系统和自动化系统中记录了大量的读者行为数据,包括读者进馆数据、借阅数据、资源预定数据和检索数据等。结合图书馆馆藏数据、读者个人数据(如学籍等),可建立读者完整的图书馆读者行为数据库。这些数据隐藏着读者个体在图书馆中对阅读的偏好,但荐购是面向全体读者,所以需要数据挖掘技术对数据进行分析。活跃书籍

挖掘的准确性决定着荐购书单的命中率,后续将着重讨论活跃书籍的挖掘过程。

### 1.2 匹配荐购书单

匹配荐购书单是活跃书籍与已有的荐购书单匹配的过程,是一个主动、自动化获得最终荐购书单的过程。目前高校图书馆荐购书单主要获得方式:1)荐购系统中读者推荐的书单;2)图书馆馆员以问卷调查等方式获得的书单;3)书商提供的书单。虽然上述3种方式提供的书单有一定的荐购合理性,但荐购的片面性也是显而易见的,需要对书单进行再次精细化选择。匹配荐购书单过程首先利用 Stanford CoreNLP 分词工具<sup>[8]</sup>,对活跃书籍书名进行分词处理,书名被分为名词、形容词、连接词、字符和数字等;然后去除分词中的数字、字符和连接词等无关词,获取书名关键字;最后,将提取的关键字与荐购书单进行字符串匹配,输出书名中含有关键字的书单,这些书单即为荐购书单。该过程相对较为简单,在此不作重点讨论。

## 2 活跃书籍的挖掘过程

挖掘活跃书籍是荐购模型的核心过程,具体挖掘过程如图2所示。

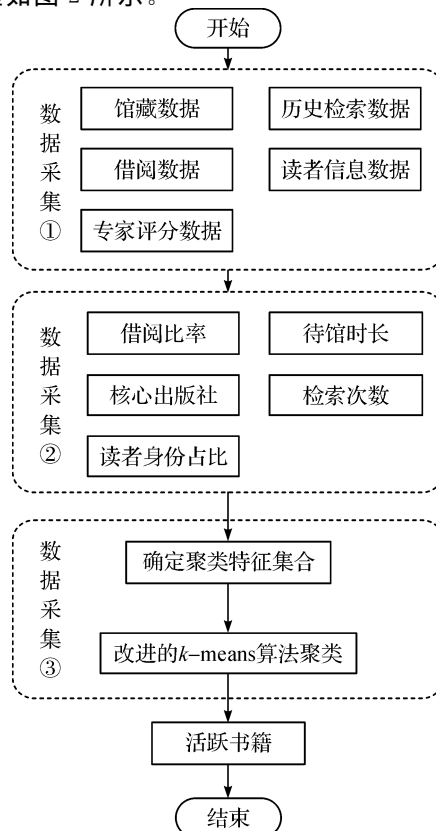


图2 活跃书籍挖掘过程图

Fig. 2 Active book mining process

## 2.1 原始数据的采集

研究的原始数据来自于高校图书馆业务系统和相关自动化系统,包括读者数据表(readerinfo)、馆藏数据表(colinfo)、专家数据表(expscore)、借阅数

据表(lendinfo)和检索数据表(searchinfo),其中,借阅数据表、检索数据表等记录着读者读书借阅行为,其他表构成了读者借阅行为的完整表达。其之间的关系如图3所示。

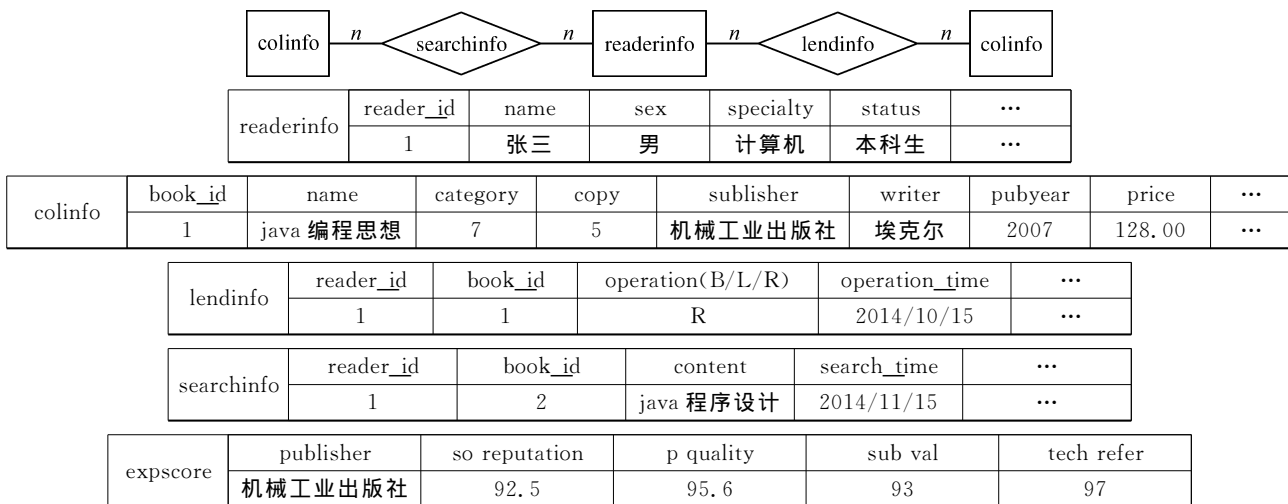


图3 数据存储表关系及结构

Fig. 3 The relationships and structures of data storage table

## 2.2 数据处理

读者行为决定了书籍的活跃度。将图书借阅比率、待馆时长、检索次数、活跃出版社和读者身份占比等5个特征作为评价活跃书籍的主要依据,也是后续  $k$ -means 算法重要的聚类特征,将聚类后特征值最大的簇作为活跃书籍。要获取上述5个特征,需要对原始数据进行分析计算。

### 2.2.1 借阅比率

借阅次数最直观的表现了读者对该书的需求。但图书馆馆藏复本数对借阅次数也有一定的影响。为了避免图书复本数不同对实验的影响,提出借阅比率计算式为

$$B_R = \frac{B_N}{C_N} \quad (1)$$

式中:  $B_R$  为借阅比率;  $B_N$  为借阅次数;  $C_N$  为复本数。

### 2.2.2 待馆时长

待馆时长的值越小表示该书被频繁借阅。主要根据书号(book\_id)、操作类型(operation)、操作时间(operation\_time)和复本数(copy)计算每本书的平均待馆时长。涉及借阅数据表(lendinfo)和馆藏数据表(colinfo)。

### 2.2.3 检索次数

检索次数对于发现读者需求十分重要,检索信息表中的检索记录是读者行为的真实记录,能客观、全面地反应读者的潜在需求。检索次数根据读者编号(reader\_id)、检索内容(content)和检索时间(search\_time)统计获得。

### 2.2.4 活跃出版社

将出版量大、质量高、利用率较高和读者影响力较大图书的出版社称为活跃出版社。针对要挖掘的活跃书籍,结合全校师生的需求和图书馆的馆藏结构,综合考虑影响文献采访的诸多因素,从读者借阅、专家推荐和历史采购等角度,构建活跃出版社数据挖掘模型为

$$P = P_A \times w_A + P_B \times w_B + P_C \times w_C \quad (2)$$

式中:  $P$  为某个出版社的综合评分;  $P_A, P_B, P_C$  分别为读者借阅、专家推荐和历史采购信息对出版社的评分;  $w_A, w_B, w_C$  分别为读者借阅、专家推荐和历史采购在活跃出版社评分挖掘中所设置的权重。根据  $P$  值的大小判断每本书的出版社是否为活跃出版社。

### 2.2.5 读者身份占比

读者身份占比表示借阅某本书的读者的身份不同,该书体现的价值也是不同的。教师、研究生和本科生代表着3个不同的身份,因此对不同身份的读者给予不同的考虑权重。读者身份占比计算式为

$$R = R_T \times w_T + R_G \times w_G + R_U \times w_U \quad (3)$$

式中:  $R$  为读者身份占比的值;  $R_T, R_G, R_U$  分别为借阅某本书的教师人数、研究生人数和本科生人数;  $w_T, w_G, w_U$  分别为教师、研究生和本科生3种身份所占的权重。

## 2.3 数据挖掘

经数据处理后获得聚类特征集合,应用数据挖掘中的聚类算法对其进行聚类,得出活跃书籍。采

用  $k$ -means<sup>[9-10]</sup> 聚类算法进行聚类,传统的  $k$ -means 算法准确率不高、稳定性不好,改进后的算法提高了准确率和稳定性。

### 3 改进的 $k$ -means 算法

#### 3.1 传统 $k$ -means 算法的不足

聚类算法是数据挖掘中的一类重要技术, $k$ -means 是聚类算法中最常用的算法之一,具有算法简单、收敛速度快以及能有效处理大数据集等多方面的优点<sup>[11]</sup>。但是, $k$ -means 算法也存在一定的局限性,聚类结果受初始聚类中心的影响较大<sup>[12]</sup>。针对  $k$ -means 算法的局限性,很多学者对  $k$ -means 算法进行了改进研究,任培花等<sup>[13]</sup>引入不确定域对数据对象进行描述并对数据预处理后采用累积距离的方法确定初始聚类中心,但引入数据对象的不确定

性因素会给算法带来复杂性问题;吕明磊等<sup>[14]</sup>提出了一种改进的  $k$ -means 算法,它依据“两个对象距离越近,相似度越大”这一依据来确定初始类心,但对异常数据并不敏感。

#### 3.2 改进的 $k$ -means 算法的思想

基于样本点距离越大,相似度就越小这一原则,为了避免选取的初始聚类中心是个噪声点,设置了一个可调节的阈值,用以确定样本集中的噪声点。如果初始聚类中心分类的数目小于特定的阈值,则将这个分类中的初始聚类中心标记为噪声点,在重新选取初始聚类中心时,去除该噪声点;阈值大小取决于样本集个数、分类个数和阈值参数。

#### 3.3 改进后 $k$ -means 算法的具体流程

在传统  $k$ -means 算法基础上,对初始聚类中心的选取进行了改进。改进后的  $k$ -means 算法流程如图 4 所示。

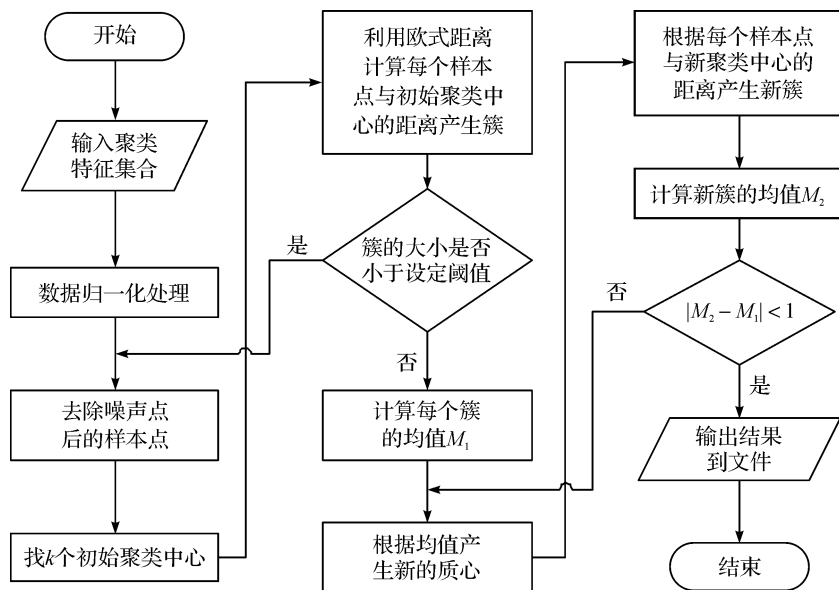


图 4 改进的  $k$ -means 算法流程图

Fig. 4 The flow of improved  $k$ -means algorithm

将数据处理后的聚类特征集合作为数据集输入该算法,并对数据进行归一化处理<sup>[15]</sup>,将取值映射在区间 $[0,1]$ 上。新建 1 个集合存放噪声点,在去除噪声点集合的样本集中选取  $k$  个初始聚类中心,利用欧式距离计算每个样本点与初始聚类中心的距离,将其划分至距离最小的初始聚类中心所在的簇。根据簇的大小判断其是否小于设定的阈值,若小于则将该初始聚类中心放入噪声点集合,重新选取聚类中心;否则计算每个簇的均值  $M_i$ ,根据均值产生新的质心,再根据欧式距离计算每个样本点与  $M_i$  的距离产生新簇,计算新簇的均值  $M_2$ ,若  $|M_2 - M_1| > 1$ ,继续调整簇,若  $|M_2 - M_1| < 1$  将得到的簇输出到

文件,算法结束。

初始聚类中心选取是本算法的关键环节,具体步骤为

步骤 1 假设聚类个数为  $K$ ,聚类中心用  $A$  表示,则有  $A = \{a_1, a_2, a_3, \dots, a_k\}$ ,样本点的个数为  $N$ ,噪声点个数为  $O$ ,属性为  $M$ ,则样本集合  $X = \{X_1, X_2, X_3, \dots, X_{N-O}\}$ ,计算任意 2 个样本点  $x_i, x_j$  之间的欧式距离  $D_{ij}$ ,找出最大的  $D_{ij}$ ,如果  $D_{ij} = \max_{i,j} |D_{ij}|$ ,则样本点  $x_i, x_j$  就作为初始聚类的前两个类心,即有  $a_1 = x_i, a_2 = x_j$ 。转步骤 2 计算剩下的  $K-2$  个聚类中心。

步骤 2 由步骤 1 计算出了聚类的前两个初始

聚类中心,假设此时已经确定了  $k$  个聚类中心( $2 \leq k \leq K-1$ ),则第  $k+1$  个聚类中心为  $a_{i+1}$ ,就是样本集中剩下的  $N-O-k$  个样本点与已经确定的前  $k$  个聚类中心的距离最小值  $D_i$ 。再找其中最大的  $D_i$ ,如果  $D_i = \max_i |D_i|$ ,则样本点  $x_i$  就是聚类的第  $k$  个初始聚类中心,这样重复步骤 2,直到  $K$  个初始聚类中心都找到为止。

步骤 3 噪声点的确定。设噪声点阈值参数为  $t$ ,根据初始聚类中心产生簇,判断是否每个簇中的个数大于阈值  $N/K'$ ,如果大于则继续算法下面的步骤,否则将这个簇中的初始聚类中心标记为噪声点,并在算法重新选取初始聚类中心时去除噪声点,

表 1 书籍活跃特征集

Table 1 Book active feature set

书号	借阅比率	待馆时长/d	是否为活跃出版社	检索数/次	读者身份占比
0000180913	1.60	298	0	72	1.0
0000545485	2.30	271	1	45	3.0
0000222808	4.30	296	0	24	1.8
⋮	⋮	⋮	⋮	⋮	⋮
0000355067	7.69	63	1	105	13.5
0000568205	1.50	339	0	36	1.1

利用改进后的  $k$ -means 算法对 2014,2015 年的 31 215,29 136 个样本集进行聚类,分别获得 4 735,3 490 本活跃书籍,最后利用分词工具对其书名进行分词,获得真正需要荐购的书单。例如:书号 0000355067 的书名为“学习 MySQL:[英文本]”,根据改进  $k$ -means 算法挖掘为活跃书籍,通过与荐购书单匹配,得到“高性能 MySQL(第 3 版)”为最终荐购图书,荐购过程结束。

#### 4.1 $k$ -means 算法改进的实验结果比较

采用传统  $k$ -means 算法与改进的  $k$ -means 算法对上述实验样本集进行准确率比较,聚类准确率指被正确分配到指定类的样本点个数与总样本点个数的比值。阈值的取值对改进算法的准确率有影响。在 2014,2015 年样本集和聚类个数一定的情况下,阈值参数  $t$  的取值对聚类准确率的影响如图 5 所示,横坐标代表  $t$  的取值,纵坐标代表准确率,经过大量的实验发现,当  $t$  取值为 1.0 时,每个类的个数都无法大于阈值  $N/K$ ,所以  $t$  从 1.5 开始取值验证。当  $t$  取值为 2.0 时准确率最高,而随着  $t$  的增大,阈值变小,对分类个数的限制变小,准确率受  $t$  值的影响变小甚至不变。而  $t$  的最优值与输入数据和聚类个数  $K$  有关系,当输入数据一定,聚类个数

直到每个簇的个数满足设定的阈值,才继续进行算法下面的步骤。

## 4 实验结果与分析

实验数据是某大学图书馆 2014,2015 年的借阅数据、馆藏数据、历史检索数据和读者信息数据等。其中读者借阅数据 513 148 条,馆藏 1 552 591 本,检索 9 005 486 条,涉及读者 26 267 人。基于前述所荐购模型相关算法,获得借阅比率、待馆时长、检索次数和读者身份占比;基于计算评分后,取排在 前 20% 的出版社为活跃出版社,最后获得影响书籍活跃度的 5 个特征项,如表 1 所示。

变化时,最优值也会随之改变。

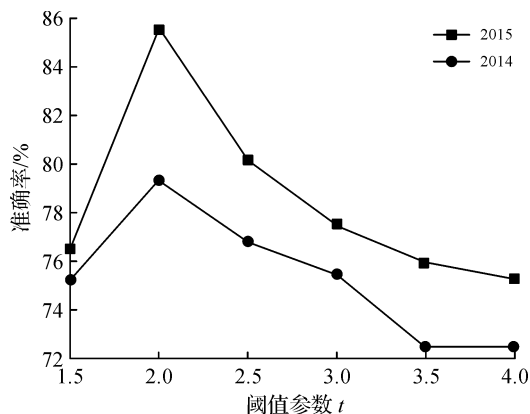


图 5 阈值参数不同准确率的变化情况

Fig. 5 The change of accuracy with different threshold parameters

将阈值参数  $t$  为 2.0 时的改进的  $k$ -means 算法与传统算法进行比较,实验结果如表 2 所示,传统算法由于初心每次都是随机选取的,所以每次的聚类结果都不同,导致准确率也不稳定;改进后的算法在相同的样本集中选取的初始质心是相同的,所以多次实验的聚类结果是相同的,准确率也稳定不变,而且每次实验的准确率都要比传统算法高,说明改进后的算法准确率和稳定性都要比传统算法好。

表 2 两种算法对 2015 年样本集的测试结果

Table 2 Two algorithms for the test results of the 2015 sample set

序号	传统算法		改进算法	
	随机质心	准确率/%	初始质心	准确率/%
1	541,1 084,21 456	80.67	205,1 849,28 285	85.64
2	234,6 785,11 576	73.41	205,1 849,28 285	85.64
3	17 435,1 169,27 345	77.49	205,1 849,28 285	85.64
4	5 456,14 563,25 635	77.32	205,1 849,28 285	85.64
5	3 253,3, 28 754	79.53	205,1 849,28 285	85.64
平均		77.68		85.64

## 4.2 图书荐购模型验证与分析

最后,通过分析读者对图书需求量和活跃度的变化情况,对图书荐购模型的合理性进行验证。首先挖掘出 2014 年的活跃书籍,提取书名的关键字,将其与 2015 年的借阅书籍进行匹配,其中 63.05% 的图书的借阅次数高于 2014 年的图书借阅次数,这说明读者对含有该关键字的图书的需求量在不断增加;将从 2014 年的活跃书籍提取的关键字与 2015 年的活跃书籍匹配,匹配到的图书占 2015 年活跃书籍的 71.43%,证明含有该关键字的图书在图书馆中非常活跃。两个实验结果证明了图书荐购模型方案科学可行。

## 5 结 论

设计了一种基于读者行为数据挖掘的高校图书荐购模型,包括活跃包括挖掘活跃书籍和匹配荐购书单两个过程;提出了一种基于改进的  $k$ -means 算法的活跃书籍挖掘方法,同时给出了影响书籍活跃度的 5 个特征项的计算方法;然后将挖掘的活跃书籍处理后与荐购书单进行匹配,最终获得图书馆真正需要购买的图书。实验证明:图书荐购模型方案科学可行。由于荐购书单过程相对简单,而且考虑到篇幅限制,只重点讨论了挖掘活跃书籍。后续工作可以针对图书购买的复本数进行分析研究,并给出建议,使得图书荐购模型更具适用性。

### 参考文献:

- [1] 蔡时连. 高校图书馆文献采访工作刍议[J]. 图书情报工作, 2016, 60(增刊 1): 108-112.
- [2] 陈大莲. 高校读者微博荐购图书的应用探析[J]. 图书馆工作与研究, 2014(2): 58-61.

- [3] 孔云, 田春燕, 资芸, 等. 面向“互联网+”的图书荐购模型研究与实践[J]. 现代情报, 2017, 37(3): 90-95.
- [4] 刘华. “读者决策采购”在美国大学图书馆的实践及其对我国的启示[J]. 大学图书馆学报, 2012, 30(1): 45-50.
- [5] 张引, 陈敏, 廖小飞. 大数据应用的现状与展望[J]. 计算机研究与发展, 2013, 50(增刊 2): 216-233.
- [6] 唐小新, 李高虎, 唐秋鸿, 等. 高校图书馆个性化电子图书荐购系统的设计和实现[J]. 现代图书情报技术, 2012(3): 83-88.
- [7] 周伟, 汪少华, 杨云. 基于数据挖掘和读者行为分析的图书馆荐书系统的研究与设计[J]. 图书情报研究, 2014, 7(4): 38-44.
- [8] MANNING C D, SURDEANU M, BAUER J, et al. The stanford corenlp natural language processing toolkit[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Baltimore, Maryland: Maryland Press, 2014: 55-60.
- [9] HARTIGAN J A. A  $k$ -means clustering algorithm[J]. Applied statistics, 1979, 28(1): 100-108.
- [10] MACQUEEN J. Some methods for classification and analysis of multivariate observations [C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California: California Press, 1967: 281-297.
- [11] 陆亿红, 翁纯佳. 基于三角模糊数的不确定性数据聚类算法[J]. 浙江工业大学学报, 2016, 44(4): 405-409.
- [12] 龙胜春, 傅佳琪, 尧丽君. 改进型  $k$ -means 算法在肠癌病理图像分割中的应用[J]. 浙江工业大学学报, 2014, 42(5): 581-585.
- [13] 任培花, 王丽珍. 不确定域环境下基于 DKC 值改进的  $k$ -means 聚类算法[J]. 计算机科学, 2013, 40(4): 181-184.
- [14] 吕明磊, 刘冬梅, 曾智勇. 基于改进  $k$ -means 算法的图像检索方法[J]. 计算机应用, 2013, 33(增刊 1): 195-198.
- [15] 陈鲤江, 景程, 吴姚鑫, 等. 数学表达式的归一化方法研究[J]. 浙江工业大学学报, 2012, 40(2): 229-232.

(责任编辑:陈石平)