项目结题验收单

专家验收表（主持人所在单位组织 3-5 名专家对项目进行验收、自评。）

项目名称	基于数据挖掘的电子资源采购策略优化研究		
主持人	漆月	职务/职称	馆员
所在单位	西南大学图书馆		
专家意见	根据要求，我单位组织专家组对漆月同志承担的项目“基于数据挖掘的电子资源采购策略优化研究”有关成果进行了鉴定。在听取课题组汇报、阅读课题研究材料、质询答辩的基础上，形成如下意见： 一、研究课题具有一定的研究意义和价值。馆藏资源建设是图书馆的重要职能工作，如何通过合理的经费开销采购更符合读者需求的图书是图书馆一直关注的话题。该课题针对图书馆这一需求痛点，提出了一种面向图书采购决策的解决方案，能够应用于图书馆的服务建设之中，帮助图书馆更好地适应大数据时代的发展，提升图书馆的技术实力。 二、课题研究过程扎实，课题成果具有一定的实用价值。该课题采用文献研究法、案例分析法、调查法、总结法等方法进行研究，最终提出了一种基于动态规划技术的采购选择策略，以整体最优为目标推荐采购的图书组合，能够在一定程度上提高图书馆馆藏资源的综合质量及图书使用率。 三、课题研究的不足与建议 该课题虽然对采购图书的选择提出了有效的决策方案，但作为决策依据的图书评价指标的制定略显简单，对图书价值的表达不够充分。在下一步研究工作中，建议从该角度入手，构建更加多维度的图书评价指标体系，让图书采购的决策依据更加充分。 综上所述，专家组认为该项目已经达到了所申请的计划任务目标，同意结题。 （如需要可增加页数）		
专家签字		陀建波	王雪梅
职务/职称		研究馆员	研究馆员

项目编号：2021059

CALIS 全国农学文献信息中心研究项目
结题报告

项目名称：基于数据挖掘的电子资源采购策略优化研究

项目关键词：数据挖掘；动态规划；电子资源建设；高校图书馆

项目单位（盖章）：西南大学图书馆

通信地址：(详细地址含邮编)
重庆市北碚区西南大学中心图书馆

项目主持人：漆月

联系电话：15922836031

电子邮件：qqt.123@163.com

提交日期：2022 年 4 月 25 日

题目：基于数据挖掘的电子资源采购策略优化研究

关键词：数据挖掘；动态规划；电子资源建设；高校图书馆

1 研究背景、目的及意义

1.1 研究背景

尽管电子阅读已经兴起多年，纸质资源服务依然是图书馆最为基础和重要的职能之一，尤其在高校图书馆中，多数读者仍更倾向于利用纸本图书进行学术阅读。“十四五”规划纲要中对高等教育质量提出了明确要求，响应党和政府的号召，为高校师生提供完备的文献保障，为学校的教育事业发展与“双一流”建设提供优质的资源支持，是图书馆义不容辞的责任与义务。

然而，根据国家统计局发布的统计公报，近年来图书每年的出版量超过十万种，而高校图书馆每年用于纸质文献的购置经费以及图书馆从业人员数量却在逐年减少。面对种类众多、更新频繁的各学科文献，人数有限的采访人员越来越难以分辨资源的真实价值，导致采访存盲目性，也影响到图书馆馆藏质量。另一方面，读者对文献的需求并没有减少，反而随着时代和技术的发展不断叠加和更新，如何在经费有限的条件下，选择满足读者和学科发展需求的优质图书，不断优化馆藏资源结构，是当今图书馆采访人员普遍面临的压力和挑战。

1.2 研究目的

信息化时代逐渐改变着人们的学习和工作环境，为满足当代读者的数字阅读习惯，图书馆的服务重心也逐渐向电子化方向偏移。有研究表明，高校图书馆用于电子资源购置的经费投入正在逐年增多，并且已经超过了纸质资源的经费配置。但由于有限的预算以及电子资源的种类增多和价格上涨，图书馆不可能无限制的购买新资源，甚至需要在已订阅的电子资源中做出取舍。但有研究表明，在多数情况下图书馆的资源采购决策过程依赖于采访人员的主观判断，并存在不确定性。我们认为，馆藏建设应当基于有意义的数据，进行科学、合理的系统规划，而非图书馆员的个人行为。因此，本课题面向电子资源采购建设一套基于数据挖掘的决策支持模型，以读者行为数据为决策依据，利用大数据技术与统筹规划算法定量评定每一种资源数据库的价值，为图书馆的电子资源采购采购提供一种科学客观的评判手段，在预算范围内建设满足本馆读者阅读需求、符合学校战略发展的电子馆藏资源。

1.3 研究意义

（1）学术意义

本课题的研究将一定程度上弥补目前学术界对于电子资源采购策略方面探讨的不足，通过对采购策略的优化方案研究丰富和深化电子资源采访、使用与评价的理论体系。同时，本课题进一步研究了大数据技术在采购决策模型中的应用策略，为图书馆资源建设过程中的技术应用研究奠定思想基础。

（2）实践意义

经费紧张是高校图书馆普遍存在的情况，如何合理分配有限的采购经费、尽可能保障全校师生的学习与科研工作需求，成为了当前高校图书馆亟需解决的难题之一。本课题正是面向这一问题展开研究工作，为图书馆的电子资源采购提供一种可操作性实践方案。此外，本课题深入探讨了电子资源使用过程中的读者行为挖掘技术，为图书馆的未来决策提供可靠的数据支持。

2 研究内容及方法

课题设计了一种基于动态规划技术的决策支持模型，针对征订书单智能决策最优化采购方案，为文献采购工作提供一定参考。为解决图书数据量过载的问题，本课题对模型的数据处理及算法流程提出了优化方法，最后以西南大学图书馆为实例验证了该方法的可行性。具体内容如下。

2.1 基于动态规划的图书采购策略

本文主要解决的问题就是帮助采购人员从品目繁多的书单中给出最符合图书馆需求的选择建议。基本思路为，首先为一定数量的待选书目评估一个综合图书质量与本校需求度的得分，然后选择若干推荐采购的图书，确保图书价值之和不超过预算的同时，使其得分之和为最大。

由此建立一个面向图书采购选择的决策支持模型，输入为所有待选书目的价格和评分，输出基于动态规划算法得出的最优方案。首先设定图书种数为 N，图书馆对每种图书采购的复本量不超过 n 册，采购预算总经费为 C。若图书 i 的需求评分为 s_i，售价为 p_i，用函数 $f(x,y)$描述当预算为 y 时，前 x 种图书的最佳组合对应的得分之和。

首先讨论每种图书只买一册的情况。当完成前 i-1 种书的最优决策后，预算经费为 c，对于 i 将面临两种选择：①图书 i 的售价大于预算，则不能采购该书，此时经费仍为 c；②图书 i 的售价未超出预算，则是否采购该书取决于在 i 以外是否有更好的选择。如果决策选择 i，根据最优性原理，前 i-1 种书决策后的预算必为 c-p_i,而选择 i 后的最优得分应为 $f(i\text{-}1,c\text{-}p_i)+s_i$。由此可得出递推式（1）

$$\begin{cases} f(i,c)=f(i-1,c) & c<p_i \\ f(i,c)=\max(f(i-1,c),\ f(i-1,c-p_i)+s_i) & c\geq p_i \end{cases} \quad (1)$$

当 i=0 时，表示没有选择任何图书，最优得分为 0。当预算为 0 时，任何图书都不能购买，最优得分也为 0。由此可得出式（1）的边界条件：

$$f(0,c)=f(i,0)=0 \quad (2)$$

图书馆的入藏图书复本数并不是相同的，而是由采访人员决策需要的副本量。考虑每种图书需采购 0~n 册的情况，其实与式（2）的推导思想类似，只是需要增加每种图书购买几册的决策。假设图书 i 需要采购 k 册是一个最优决策，则此时最优解为 $f(i\text{-}1,c\text{-}p_i*k)+s_i*k$，可得到递推关系式（5）：

$$f(i,c)=\max\{f(i-1,c),f(i-1,c-p_i*k)+s_i*k\}\ \ 0\leq k\leq \min(n,[c/p_i]) \quad (3)$$

上式中，不选择 i 时，k=0，有 $f(i,c)=\max\{f(i\text{-}1,c),f(i\text{-}1,c\text{-}p_i*0)+s_i*0\}=f(i\text{-}1,c)$，由此公式（3）可简化为式（4）

$$\begin{aligned} f(i,c)&=\max\{f(i,c),f(i-1,c-p_i*k)+s_i*k\} \\ &=\max\{f(i-1,c-p_i*k)+s_i*k\} \quad 0\leq k\leq \min(n,[c/p_i]) \end{aligned} \quad (4)$$

这里 $f(N,C)$得出的最优决策的评价得分，并不能得出最优解是由哪些图书构成，故需设置一个数组 ch[N+1]来记录对图书 i 的决策结果，ch[i]的下标为图书 i，ch[i]的元素值为图书 i 的采购册数。决策过程部分实现代码为：

```
for(int i=1;i<=N; i++){
  pi=PRICE[i],si=SCORE[i]; //售价，评分
  if(C/pi<n) n=C/pi; // 取min(n, [c/pi])
  pi*=100 C*=100; //金额取整
  for(int c=1; c<=C; c++)//金额取整
    for(int k=1; k<=n&&c>=pi*k;k++){
      if(f[c]<f[c-pi*k]+si*k){
f[c]=f[c-pi*k]+si*k;
ch[i]=k;  //记录方案
      }
    }
}
```

图 1 决策代码

数组 ch 即为决策模型实际需输出的结果，通过枚举数组中每个元素，可以向采购人员直观展示征订书单中的每种图书是否建议采购，以及采购多少册复本。

2.2 算法分析

事实上，笔者曾进行过关于利用动态规划处理电子资源采购决策的相关研究[13]，于是想到将类似思想也应用与图书采购决策中。但在真实数据测试时发现，由于图书资源的数据量远大于电子资源，运行速度非常慢。由图 1 知算法复杂度为 O(NCn)，而图书采购流程涉及的 N 和 C 取值通常较大，循环次数相应较多，同时由于折扣等原因，图书售价通常精确到小数点后两位，导致循环次数还需增加 100 倍。因此必须对决策模型进行优化，以保障系统的运算效率。

2.3 模型优化

（1） 数据分割

在制定采购计划时，除确定预算经费总额外，图书馆通常还会根据学校的学科建设和发展需求规划经费的学科分配比例。因此进行最优解计算时，也可以将图书按学科分类，然后对每个学科的图书分别进行一次动态规划计算。假设需要采购 m 个学科的图书，每个学科的经费比例用 D(d_1,d_2,...,d_m)表示，在新书征订数据中，每个学科的征订图书种数为 N(N_1,N_2,...,N_m)，则有式（5）：

$$f(\mathrm{N},\mathrm{C}) = \sum_{j=1}^{m} f(N_j, \mathrm{C} * d_j) \quad \sum_{j=1}^{m} N_j = \mathrm{N}, \sum_{j=1}^{m} d_j = 1 \qquad (5)$$

此时，模型计算量由 100NC 变为$\sum_{j=1}^{m} N_j * (100\mathrm{C} * d_j) = 100\mathrm{C}\sum_{j=1}^{m} N_j * d_j$。按学科分割数据后得到的最优解与整体计算结果显然是不同的，但由于每年出版物的学科比例不一定符合学校的重点建设方向，因此以上做法将更加符合学科需求倾向。同时，因为$\sum_{j=1}^{m} d_j = 1$，显然$\sum_{j=1}^{m} b_j * d_j \le N$，因此数据分割后的计算量小于之前。当学科分类数量越多时，优化效果就越明显。

（2）算法优化

由式（4）可看出，$f(i,c)$只和 $f(i,c\text{-}p_i*k)$有关，c 的取值总是 p_i 的整数倍加一个余数，如果以 p_i 为单位将所有可能的 c 值划分为余数个组，可得到分组如表 1。

表 1 对经费 c 按 p_i 的余数进行分组

P_i 的余数	经费 c
0	c = 0, p_i, 2p_i, 3p_i, ...
1	c = 1, p_i+1, 2p_i+1, 3p_i+1, ...
...	...

p_i-1	$c = p_i$-1, 2p_i-1, 3p_i-1, ...

观察可知式（4）中每一次决策的状态转移都仅发生在同一分组内，而不同分组间不会发生状态转移。若令余数为 b，x=[c/p_i]，则 c=p_i*x+b。因此，可将决策函数改写为：

$$f(i,\ p_i * x + b) = \max\{f(i-1,\ p_i * (x-k) + b) + s_i * k\} \quad 0 \le k \le \min(n,x) \qquad (6)$$

令 x-k=k'，则有：

$$f(i, p_i * x + b) = \max\{f(i-1, p_i * k' + b) - s_i * k'\} + s_i * x \quad x - \min(n,x) \le k' \le x \qquad (7)$$

这里一对(x, b)唯一对应了一个 c 值，在 b 为静态值的情况下，求解过程变为从宽度为 k'的窗口中挑选最大值来更新当前值，因此可以利用一个单调队列来维护窗口最大值，方法如下：

① 枚举一个 i 和 b（0<=b<pi)

② 从小到大枚举 x，并使用单调队列维护 f(i-1,p_i*k'+b)-s_i*k'，关于 k'单调下降，记录 f(i,p_i*x+b)的更新。

改进算法的实现代码如图 2：

```
int Nj=N[j], Cj=C*d[j]; //对学科j进行规划
for(int i=1;i<=Nj; i++){
  pi=PRICE[i], si=SCORE[i]; //售价，评分
  if(Cj/pi<n) n=Cj/pi; // 取min(n, [c/pi])
  pi*=100, Cj*=100; //金额取整
  for(int re=0; re<pi; re++){//枚举余数
    head=tail=1; //队头与队尾指针
    for(int k=0; k<=(Cj-re)/pi; k++){
      int tmp=f[k*pi+re]-k*si;
      while(head<tail&&que[tail-1]<=tmp) tail--;
        que[tail]=tmp;
        num[tail++]=k;
        while(head<tail&&k-num[head]>n) head++;  //滑动区间长度不大于n
        if(f[pi*k+b]<que[head]+si*k){
          f[pi*k+b]=que[head]+si*k;
          ch[i]=k;
} } } } }
```

图 2 优化算法代码

3 结论与建议

3.1 结论

如果不进行数据分割，则每个余数分组中最多只有[C/p_i]个元素，对于一种图书 i，决策的总时间是 O([C/p_i]*p_i)=O(C)，所以 N 种图书的算法时间复杂度为 O(NC)。与第 2 节讨论算法的时间复杂度为 O(NCn)相比，有明显优化。在按学科分类后，优化算法循环次数为$\sum_{j=1}^{m} N_j * 100C_j$次，而优化前需要循环计算$100NC\sum_{k=1}^{n} k$次。假设共有 10 个学科，且重要程度和出版比例相同，采购的复本量最大为 4，优化算法的计算次数为(N/10)*(100C/10)*10=10NC，而优化前的计算次数为 100NC*10=1000NC，计算量减少了 100 倍。

本模型主要面向基于图书征订目录的线上采购工作提供决策支持，读者荐购及现场采选等采购方式仍需采访人员进行决策，但由于实际工作中线上采购的数据量远大于其它采购形式，因此本模型能够对图书采选工作起到较好的优化作用。作为完整系统架构时，允许采购人员首先在 UI 界面中人工挑选已经决定采购的图书及其复本数，系统将在运算前首先减去这批图书

的相关数据，再对剩余部分进行决策计算，并输出推荐的采购选择。

突如其来的新冠病毒给全社会都带来了巨大的影响，在各行各业逐渐恢复正常秩序的同时，中央政治局也做出了疫情防控常态化的工作部署。如此态势下，预算削减与经费不足将是图书馆短期内无法回避的压力。文献资源购置作为多数图书馆最大的一笔经费开销，采访工作必须更加精细化、精准化，用有限的经费为学校的“双一流”建设提供完备的保障。本课题研究了一种应用新技术优化图书馆工作的解决方案，以期为“十四五”时期图书馆事业发展和智慧化转型做出一点贡献。

4 项目成果（发表的文章、开发的软件、取得的实践效果等）

本课题已经开发了基于动态规划的图书采购决策支持原型系统，并在仿真实验中验证了方案的可行性，运行效果如下图所示。

5 参考文献

[1]张强.大学生数字阅读习惯调查及高校图书馆服务功能优化对策[J].中国管理信息化,2017,20(24):169-170.

[2]魏辅轶.中国图书馆学理论跨世纪的三次“重逢”与“莫比乌斯陷阱”[J].中国图书馆学报,2021,47(01):34-46.

[3]程蓓,吴帼帼.高校图书馆电子资源整合揭示及管理机制调查研究[J].图书馆工作与研究,2021(11):76-83.

[4]智研咨询.2021-2027 年中国图书出版产业运营现状及战略咨询研究报告[R].北京:北京智研科信咨询有限公司,2020,46-48.

[5]徐玲.以读者需求为导向的高校图书馆纸质图书服务实证研究[J].内蒙古科技与经济,2021(14):144-146.

[6]蔡迎春.智能选书：图书馆精准采购实现策略[J].数字图书馆论坛,2021(06):50-55.

[7]詹德潘. 基于图书征订目录建立图书荐购系统的技术研究[J]:现代情报，2004(8)：137-138.

[8]方向明, 沈玲. 图书馆读者荐购研究综述[J]. 图书情报工作, 2018, 62(9): 143-150.

[9]胡振宁.图书馆联盟电子书 PDA 模式及关键问题分析[J].中国图书馆学报,2016,42(04):75-87.

[10]符艺.公共图书馆读者决策采购应用研究与案例分析[J].四川图书馆学报,2021(02):29-32.