Cost Minimization of Library Electronic Subscriptions

Laura Bigram, Patrick Hosein and Jonathan Earle

Abstract Many libraries, particularly those at Universities in developing countries, are facing challenging financial times. This has led to the need for budget cuts and more efficient management of limited resources. One of the major costs of an academic library are the fees paid for subscriptions to electronic journals, databases, conference proceedings and for costs associated with downloads of individual papers if there is no subscription to the corresponding resource. Typically the decision as to whether or not a particular subscription is acquired is done based on faculty member requests, information about the resource (such as cost) and policies of the library. However, with the availability of a wide range of collected statistics, (number of downloads, impact factors, etc.) one can make better informed decisions. In this paper we provide a decision support system in which we define a metric for the value obtained per access to a resource and then determine the minimum budget required to achieve a given total value of this metric.

1 Introduction

Due to the recent economic downturn in some countries, many academic institutions are having to slash their budgets resulting in a reduction in financial allocations to their libraries. Many academic libraries have been left with the challenge of providing the same value and experience to faculty and students despite the significant decrease in funding and spiraling increases in resource costs over the past few years [4]. Acquisitions of resources can account for several millions of dollars in a budget and so even small percentage reductions can result in significant savings. For example, in one particular

Laura Bigram, Patrick Hosein, Jonathan Earle
The University of the West Indies, St. Augustine, Trinidad
e-mail: {laura, patrick, jonathan}@lab.tt

library, approximately eighty two percent of the library's budget was spent on e-resource acquisitions for the year 2015. Decision support systems have been successfully used in the past to manage human resources in libraries (e.g., see [1]) with considerable savings. We focus on library cost savings for electronic subscriptions.

In this paper we provide a model for optimizing the selection of resources to subscribe to at the beginning of the academic year. We formulate the associated mathematical model and develop an algorithm for obtaining the solution. We then use the approach to determine the optimal subscriptions using real data and compare this solution with the actual performance in prior years.

We first develop a metric that measures the value of electronic resources to an academic community. For example, a simple metric is the number of downloads; more downloads implies more value to the community because of the increased knowledge provided. However, we go further and add other factors that affect value, such as impact factors and the importance of the field to the general goals of the university or country. Given this metric we can now measure the total value achieved for the given subscriptions made. What we then do is minimize the total cost while keeping the total value fixed. To our knowledge, this particular model is new to the area. The optimization approach is also one of our contributions. Note that, the work in [1, 2, 3] assumes material acquisition is budgeted by categories (books, journals etc.) whereas we optimize over all categories. The reason being that e-resources are becoming the preferred form of acquisitions and hence categorization can contribute to underutilization.

Although the model can take into account any of the resources acquired by the library, we only had access to a subset of the data, that of Science Direct e-journals. Since this formed a significant percentage (approximately 18%) of the resources, we believe that the results reflect what can be obtained if all resources are taken into account. We obtained statistics on e-journals in the Science Direct database for the years 2011 to 2015. This dataset consists of 1739 individual journals and for each of these it includes the title, a unique identifier, year, number of downloads for the year, annual subscription cost, cost per download and subject areas (keywords). One of our objectives is to check how closely the subscribed journals relate to the research goals of a university. We therefore also produced a list of the keywords that reflect such goals. Journals that have more common subject areas with the set of keywords in the research goals should be given more value.

2 Mathematical Model

In this section we provide the optimization objective and a mathematical model that will be used to achieve this objective. Note that the library pays for a service (access to resources) and the associated community receives some value from this service. The service cost is easily obtained but the associated value of a downloaded resource depends on various factors. In this paper, we take into account two of the factors that contribute to the value but others can be added in the formulation in the future. Note that in this paper we are focused only on subscription or pay-per-download services of electronic journals and not acquisition of books.

Let $\mathcal{J}$ denote the set of resources (e.g., Journals, Conference Proceedings, etc.) accessible through the library, either through subscriptions or pay-per-download. For any resource $j \in \mathcal{J}$ we use the following notation:

$$\begin{aligned}
S_j &= \text{the annual subscription rate for the resource}\\
P_j &= \text{the price per publication download for the resource}\\
I_j &= \text{the Impact Factor of the resource}\\
D_j &= \text{the number of completed downloads for the year}\\
R_j &= \text{the number of requested downloads for the year}
\end{aligned}$$

Note that some requests for a paper may not be satisfied because there is no journal subscription and no pay-per-download option or the cost for the download is considered too much or the budget for downloads has been exhausted and hence $R_j \geq D_j$.

We next take into account the relevance of the journal/conference resource to the goals of the university. In many developing states the Government has indicated specific areas that need to be developed as high priority as part of an overall vision for the country. For example, in a small island state there may be less interest in Quantum Physics research than in ICT since the latter will help in development of the country. If this is the case, we take this factor into account as follows. For any resource j, let K_j denote the set of keywords for the resource. Let O denote the set of keywords that represent the development objectives of the country. We assume that members of O are contained within the union of the sets K_j. We will weight each resource by the number of its keywords that are included in O.

2.1 Resource Value and Cost

The cost of a subscribed resource j is simply the annual cost of the subscription, S_j. For other resources the cost depends on the number of downloads and is given by P_jD_j. Naturally if $S_j < P_jR_j$ then one should subscribe to the resource otherwise it is better to pay for each download. Therefore we define the cost of resource j as a function of the number of satisfied requests x as:

$$C_j(x) = \min\{xP_j, S_j\} \tag{1}$$

Since we do not know the number of downloads in advance we will use historical data (with linear regression) to estimate the number of requests in the upcoming year.

For the value V_j of a resource we may take into account the Impact Factor as well as the goals of the university. This is given by,

$$V_j = 1 + \alpha I_j + \beta |K_j \cap O| \tag{2}$$

where α and β determine the importance of each factor. The third component is the number of common keywords between the resource and goals. For our numerical results we do not have Impact Factors and so we set $\alpha = 0$. For Governmental research goals we use $\beta = 0.1$.

2.2 Problem Formulation

The objective of this problem is to minimize the total cost for a given total value. However, as we mentioned before, this optimization requires knowledge of the number of requests per resource which is not known in advance. This will be obtained by using historical data to predict the expected number of requests for each resource. Let x_j represent the number of requests satisfied for resource j. This is the decision variable over which we must optimize. Note that the number of requests is integer but we relax this constraint and assume that x_j is continuous. We will find that, in the optimal solution of the relaxed problem, all but one of the optimal variables x_j will in fact be integral and hence we may only need to round one decision variable with negligible loss in optimality. We use $F(\mathbf{x})$ to denote the objective function value for a given decision vector $\mathbf{x}$. The optimization problem can be stated as follows:

$$\min_{\mathbf{x}} F(\mathbf{x}) \equiv \sum_{j \in \mathcal{J}} C_j(x_j) \tag{3}$$

$$\text{s.t.} \quad \sum_{j \in \mathcal{J}} V_j x_j = T$$

$$\text{and} \quad 0 \leq x_j \leq R_j \quad \forall\, j \in \mathcal{J}$$

where T denotes the total value to be achieved. We first provide an important property of the optimal solution.

Lemma 1. *The optimal solution has the property that, for all except maybe one resource, either all requests are satisfied or none is satisfied.*

Proof. This is a well known property of the solution obtained when taking the minimization of a sum of concave functions over a polytope. □

Although this property is helpful it is still difficult to check all possible solutions to find the optimal one. Next, we consider a modified version which can be solved and has a solution close to optimal. Consider the following lower bound to the objective function:

$$C'_j(x) = \min \left\{ P_j, \frac{S_j}{R_j} \right\} x \equiv G_j x \tag{4}$$

For convenience we use the representation $G_j x$ for some constant G_j for each j. Note that for all feasible x we have $C'(x) \leq C(x)$. Consider the problem 3 but with $C(x)$ replaced with $C'(x)$ or equivalently Gx since for a given resource the function is linear in x. Furthermore we convert the optimization problem to a maximization problem by considering the negative of the objective function value. This modified problem can be solved using Lagrange Multiplier methods by introducing λ, μ and γ. The Lagrangian is given by,

$$\mathcal{L}(\mathbf{x}, \lambda, \boldsymbol{\mu}, \boldsymbol{\gamma}) = -\lambda T + \sum_{j \in \mathcal{J}} -G_j x_j + \lambda V_j x_j + \mu_j x_j - \gamma_j (x_j - R_j) \tag{5}$$

$$\text{s.t } \mu_j x_j = 0, \;\; \gamma_j (x_j - R_j) = 0, \;\; \mu_j \geq 0, \;\; \gamma_j \geq 0 \quad \forall\, j \in \mathcal{J} \tag{6}$$

Taking partial derivatives and setting to zero we obtain:

$$\frac{\partial \mathcal{L}}{\partial x_j} = -G_j + \lambda V_j + \mu_j - \gamma_j = 0 \tag{7}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -T + \sum_{j \in \mathcal{J}} V_j x_j = 0 \tag{8}$$

Let us consider the various cases:

$$\begin{aligned}
\mu_j > 0, \gamma_j > 0 &\Rightarrow \quad x_j = 0, \quad x_j = R_j \quad \text{not possible} \\
\mu_j > 0, \gamma_j = 0 &\Rightarrow \quad x_j = 0, \quad \lambda < \frac{G_j}{V_j} \\
\mu_j = 0, \gamma_j > 0 &\Rightarrow \quad x_j = R_j, \quad \lambda > \frac{G_j}{V_j} \\
\mu_j = 0, \gamma_j = 0 &\Rightarrow 0 \leq x_j \leq R_j, \; \lambda = \frac{G_j}{V_j}
\end{aligned}$$

Therefore, we just need to find $\lambda = \lambda^*$ that satisfies these conditions. This can be accomplished by starting with the resource with the smallest ratio $\frac{G_j}{V_j}$ and satisfying all of the requests for that resource. We then go to the resource with the next smallest ratio and repeat. We keep track of the total value and stop when this total reaches T. In the case in which $V_j = 1$ we note that all optimal decision variables are integers and hence the optimal solution for the relaxed (approximate) problem is the same as that of the integer (approximate) problem. The solution of the modified problem will be used for our numerical results and so we next show that the error is bounded and in fact quite small. We start with the following Lemma.

Lemma 2. *Let* $\mathbf{x}^*$ *denote the optimal solution to the original problem 3 and let* $\mathbf{y}^*$ *denote the optimal solution to the modified problem then* $F(\mathbf{y}^*) - F(\mathbf{x}^*) \leq \max_{j \in \mathcal{J}} S_j$*.*

Proof. Note that the solution of the modified problem also has the property that for all but one resource, requests are totally satisfied or not satisfied. Suppose that some resource i is not at an extreme point. Since $C'(x) < C(x)$ then the optimal solution for the modified problem is less than that of the original problem so,

$$F(\mathbf{x}^*) \geq F'(\mathbf{y}^*) \tag{9}$$

where F' is used to represent the objective function value of the modified problem. Now note that for all resources j except i we have $C'(y_j) = C(y_j)$ since both functions are equal at extreme points. Hence, we can write $F(\mathbf{y}^*) = F'(\mathbf{y}^*) + C_i(y_i) - C'_i(y_i)$. Also, note that the maximum difference between $C_i(y_i)$ and $C'_i(y_i)$ is S_i and hence we can write $F(\mathbf{y}^*) \leq F'(\mathbf{y}^*) + S_i$. Inserting this into 9 we obtain $F(\mathbf{x}^*) \geq F(\mathbf{y}^*) - S_i$ and hence,

$$F(\mathbf{y}^*) - F(\mathbf{x}^*) \leq S_i \leq \max_{j \in \mathcal{J}} S_j \tag{10}$$

□

Note that, if in the optimal solution all resources are at extreme points then the optimal solution for the modified problem is also optimal for the original one. Since subscriptions for hundreds of resources are acquired and these contribute to F then the error is small when compared to the optimal function value. We will compute this error for our numerical results to illustrate this point.

3 Numerical Results

In this section, we provide numerical results to illustrate the performance of the proposed model as well as present our prediction model. We focus on the year 2014. The subscription decisions for 2014 must be made in 2013 and so to make a fair comparison we assume that only information available in 2013 can be used for the proposed algorithm. Using the publication download statistics for years prior to 2014 we predict download values for 2014. We use simple linear regression for this prediction. In addition to the journals that are active in 2014 (i.e., had at least one download), we also include in our model journals that are potential candidates for downloads. To do this, we look at historical data and if a journal previously had a subscription we predict what the number of downloads would have been had it been kept as an option. We use the Symmetric Mean Absolute Percent Error (sMAPE) to ensure that our predictions were sufficiently accurate.

Note that our model also requires the 2014 target value which is also not known in 2013. We again use historical data with linear regression to estimate this value for 2014. Given the predicted downloads and predicted target value we can run the optimization algorithm to determine the set of journals that should be subscribed, $\mathcal{J}_s^*$ and the set of journals for which the library should pay per download $\mathcal{J}_d^*$.

In order to compare the proposed approach with the actual 2014 subscriptions, we have to evaluate the resulting cost for both approaches given the actual downloads in 2014. Let $\mathcal{J}_s$ and $\mathcal{J}_d$ denote the subscription and non-subscription journal decisions made by the library for 2014 and $\hat{x}_j$ denote the number of pay-per-downloads for journal j. The total cost to the library would therefore be:

$$C_{lib} = \sum_{j \in \mathcal{J}_s} S_j + \sum_{j \in \mathcal{J}_d} \hat{x}_j P_j \tag{11}$$

Let us now consider the cost that would have resulted if the proposed approach was used. Different subscription decisions made in 2013 would have affected the resulting downloads. There may have been journals that the optimal solution decided not to purchase (either by subscription or pay-per-download) but the library solution had allowed it. In such cases the cost would be included for the library solution but not for the optimal solution. There may also be journals that the optimal solution decided to include (i.e. they were included in the past and the optimal solution includes them) but there are no downloads for these in 2014. For such cases we include the cost for the predicted values instead. For each such journal j, let x_j^* denote the optimal number of requests for the journal. The 2014 cost for the optimal solution can therefore be written as:

$$C_{opt} = \sum_{j \in \mathcal{J}_s^*} S_j + \sum_{j \in \mathcal{J}_d^*} x_j^* P_j \tag{12}$$

So C_{lib} provides the cost incurred based on the library's decisions while C_{opt} denotes the cost incurred if the optimal algorithm is used.

Let us now consider the target values. The proposed algorithm uses a predicted target value but this can be different to the value actually obtained in 2014. Let T_{opt} be used to denote the value achieved in 2014 if the optimal solution was used and let T_{lib} denote the actual value achieved in 2014. Since we have costs for different total values then we need to normalize them and so we compute the cost ratio as,

$$Q = \frac{C_{opt}/T_{opt}}{C_{lib}/T_{lib}} \tag{13}$$

This ratio determines the cost gain obtained with the optimization approach.

We also repeated this computation with perfect prediction. In other words, we ran the algorithm using the downloads actually experienced in 2014. This allows us to see the gain that is possible without including prediction errors. We performed these evaluations for two use cases $V = 1$ (i.e. just using downloads) and also for the value function for the keyword component using a value of $\beta = 0.1$. These are all included in Figure 1. Note the significant cost savings that are achievable by simply making better subscription choices. For the case of imperfect predictions and including research goals in our objective the cost for the optimized approach is approximately 10% of that of the present approach. The source of these savings can easily be verified by inspection. For example, for the journal *Surgery (Oxford)*, the library paid per download while the optimal solution decision was to pay for a subscription which resulted in a savings of $157,423. On the other hand in the case of *Journal of Sound and Vibration* the library subscribed to the journal but there were few downloads and so the optimal solution decision was to pay for each download rather than to subscribe and this resulted in a savings of $7,122. We also find that the error bounds are relatively small and will become even tighter when the complete dataset is considered.

Table 1 Performance Results

Year	**Prediction**	**Value Function**	**Cost Ratio Q**	**Error Bound**
2014	Perfect	Downloads	0.036	0.012
2014	Perfect	Downloads+Goals	0.038	0.092
2014	Regression	Downloads	0.051	0.077
2014	Regression	Downloads+Goals	0.091	0.075

References

[1] Chorba RW, Bommer MRW (1983) Developing academic library decision support systems. Journal of the American Society for Information Science 34(1):40–50

[2] Enger KB (2009) Using citation analysis to develop core book collections in academic libraries. Library and Information Science Research 31(2):107 – 112

[3] Ho TF, Shyu SJ, Wu YL (2008) Material acquisitions in academic libraries. In: Asia-Pacific Services Computing Conference, 2008. APSCC '08. IEEE, pp 1465–1470

[4] Jotwani D, et al (2014) Trends in acquisition and usage of electronic resources at indian institutes oftechnology libraries. Annals of Library and Information Studies (ALIS) 61(1):33–40