Mathematical Model of Data Processing System for Information Support of Innovative Cluster Works

L. K. Bobrov(✉) and I. P. Medyankina

Novosibirsk State University of Economics and Management, Novosibirsk, Russia
{l.k.bobrov,i.p.medyankina}@edu.nsuem.ru

Abstract. Information systems of innovation management are an important component of cluster organization infrastructure. The creation of such systems involves working with a wide range of commercial information resources. These resources are often used to form thematic databases designed to meet the information needs of cluster organizations. It is highly relevant to find the topology of an information processing system which would minimize overall operational costs for information support of innovative cluster activities. The paper presents the mathematical statement of the problem concerning forming optimal topology of such system. It is shown that finding optimal topology of a distributed information processing system can be reduced to solving the problem of integer linear programming, which would allow minimizing the system operation costs.

Keywords: Mathematical modeling · Innovative clusters · Information support · Distributed data processing · Optimization models

1 Introduction

The expediency of using various forms of cooperation in organizing information activities was pointed out as early as in the 1990s [1–4]. The interest towards this topic was also provoked by the transition to the market economy in the former Soviet Union countries [5]. Currently, the relevance of this topic is determined by the need to develop innovative activities in conditions of limited financial resources.

Cluster policy is one of effective tools to develop territories and regions [6–9]. Cluster approach is most widely used in the European Union where it

This work was supported by a grant from the MES RK (project No. AP05134019 "Development of scientific and methodological foundations and applied aspects of building a distributed information support system for innovation activities, considering the specific features of each of the stages of the innovation life cycle".

Y. Shokin and Z. Shaimardanov (Eds.): CITech 2018, CCIS 998, pp. 64–70, 2019.
https://doi.org/10.1007/978-3-030-12203-4_7

was elevated to the rank of public policy. The Russian Federation has begun to actively support innovative territorial clusters since 2011.

In the Republic of Kazakhstan, the cluster approach is laid down in the State Program of Industrial Innovative Development for 2015-2019. The program provides active financial support for clusters with the highest development potential. They are selected on a competitive basis.

This support is realized in several directions, including the processes of forming supplier bases and creating information platforms. Information systems for innovation management are an important component of cluster organizations infrastructure. The creation of such systems involves working with a wide range of commercial information resources. These resources are often used to form thematic databases designed to meet the information needs of cluster organizations. It is highly relevant to find the topology of such information processing system which would minimize overall operational costs for information support of innovative cluster activities. There are many approaches to solving this class of problems, as well as relevant data processing models, beginning with the simplest set-theoretic models and ending with the most complex simulation models. This article shows that the optimal topology of the information support system for cluster organization can be obtained by means of solving an integer linear programming problem.

2 Mathematical Statement of the Problem

The mathematical model proposed in this article describes the process of creating an information product in terms of cooperation. It also allows solving the problem of optimal distribution of technological operations between members of an innovation cluster in order to achieve the minimum of overall costs for product creation. This model is constructed as follows. Let the analysis of the market and choice of the database subject show that the information flow for database formation is the association of R subjects (rubrics):

$$\Phi = \bigcup_{r=1}^{R} M_r \tag{1}$$

and the volume of each rubric can be estimated by the number of documents belonging to it:

$$|M_r| = V_r.$$

The level of interest of each of the partners $U_1, U_2, \ldots, U_N$ in the processing of each of the R rubrics can be estimated by the vector

$$\overline{p}_n = (p_{n1}, p_{n2}, \ldots, p_{nR}), \quad \left(n = \overline{1, N}\right). \tag{2}$$

In order not only to reflect the participant's interest in processing documents for each of the R rubrics, but also to compare the level of their interest towards any of the rubrics, the following condition must be met:

$$\sum_{r=1}^{R} p_{nr} = 1, \quad (0 \le p_{nr} \le 1)\,. \tag{3}$$

The values p_{nr} can be determined in several ways. In particular, the level of interest of each of the partners in the documents of the r^{th} rubric can be estimated based on the predicted number of queries to each of the M_r arrays from each participant. Then we denote by b_{nr} the number of queries from the partner U_n to the array M_r and get:

$$p_{nr} = b_{nr} / \sum_{r=1}^{R} b_{nr}. \tag{4}$$

More accurate estimates can be obtained using the notion of completeness of the answer in the database system. Here, as a criterion, the number of documents issued in response to each of the queries is used when searching all M_r arrays. Then, denoting by d_{nr} the total number of documents obtained by searching the array M_r for the entire set of queries B_n of the participant U_n, we have:

$$B_n = \sum_{r=1}^{R} b_{nr}, \quad p_{nr} = d_{nr} / \sum_{r=1}^{R} d_{nr}. \tag{5}$$

Introducing a system of weighting coefficients $(\beta_{n1}, \beta_{n2}, \ldots, \beta_{nR})$ to consider subjective factors which determine the interest of the participant U_n to the array M_r, we get:

$$p_{nr} = \beta_{nr} d_{nr} / \sum_{r=1}^{R} d_{nr}. \tag{6}$$

To simplify the time-consuming procedure of formulating a large number of queries, it is possible to use an approach based on the connection between the frequencies of terms in a database and the number of documents issued in response to a query. Then, using the results of a previously organized questionnaire of future customers (subscribers), and assuming that each of the queries includes only one term, we get a list of terms (normalized lexical units) for each of the U_n participants:

$$L_n = \left(l_n^1, l_n^2, \ldots, l_n^K\right).$$

Comparing each term with frequency dictionaries of each M_r array, we get:

$$\tilde{d}_{nr} = \sum_{r=1}^{R} F_r,$$

where

$$F_r = \sum_{k=1}^{K} f_{nr}^k,$$

and f_{nr}^{k} is the frequency of the k^{th} term from the list of partner U_n in the array M_r.

The technological process of information processing by the partners can be represented as an ordered sequence of simple or aggregated operations, the same for any M_r array:

$$O = \left(O^1, O^2, \ldots, O^Q\right). \tag{7}$$

Let us denote by t_n^{qi} the volume of unit costs for the i^{th} resource (taking into account the characteristics of software and hardware resources, as well as other factors affecting the real cost of data processing operations in the U_n center) for the operation of O^q by the partner U_n. Also, we consider the fact that the overall cost of i^{th} resource for each of the partners cannot exceed a certain limit value μ_n^i. Taking into account the volumes of the M_r arrays, let us introduce the indicator:

$$\tau_{nr}^{qi} = V_r t_n^{qi}, \tag{8}$$

which characterizes the cost of the i^{th} resource required by the partner U_n to perform the operation O^q on the array M_r. Then the overall cost can be described as:

$$H^1 = \sum_{i=1}^{I} \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{q=1}^{Q} \omega_{nr}^{q} \tau_{nr}^{qi}, \tag{9}$$

where

$$\omega_{nr}^{q} = \begin{cases} 1 - \text{if the } U_n \text{ participant performs the } O^q \text{ operation on the } M_r \text{ array;} \\ 0 - \text{otherwise;} \end{cases}$$

and the equality holds:

$$\sum_{n=1}^{N} \omega_{nr}^{q} = 1 \quad \left(q = \overline{1, Q};\ r = \overline{1, R}\right). \tag{10}$$

The latter means that each of the O^q operations on any M_r array is necessarily performed, and it is done only by one of the U_n partners, i.e. the principle of one-time processing of information is respected. Thus, there is a problem of minimizing the functional (9) with the equality (10) and limitations held:

$$\omega_{nr}^{q} = \{0, 1\}; \tag{11}$$

$$\sum_{r=1}^{R} \sum_{q=1}^{Q} \omega_{nr}^{q} \tau_{nr}^{qi} \leq \mu_i^n. \tag{12}$$

With the solution of this problem, it is possible to find such distribution of work between partners which allows achieving the minimum of the overall costs for creating an information product. However, this is true only if there are no subjective factors affecting the distribution of work between partners, and causing the need for some operations to be performed centrally whereas the solution of others (for example, information service operations themselves) is

decentralized. Based on the foregoing, the set of O operations can be divided into three disjoint subsets (O' means centralized operations; O'' means distributed operations; O''' means operations performed by each of the U_n partners), and further focus will be on distributed operations.

We impose a penalty on the execution of operations from O'' in case if the participant U_n (interested in the results of processing documents of the r^{th} rubric) does not perform operations on its processing. Logically, the amount of the penalty should be proportional to the level of interest of the participant and the cost of performing the operation O^q on the array M_r. In this case, the functional (9) will take the following form:

$$H^2 = H_1 + \sum_{i=1}^{I} \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{q:\, O^q \in O''} (1 - \omega_{nr}^q)\, p_{nr} \tau_{nr}^{qi}. \tag{13}$$

To simplify the functional, we introduce the notation:

$$\theta_{nr}^q = \sum_{i=1}^{I} \tau_{nr}^{qi}. \tag{14}$$

Then the minimized functional (13) can be rewritten in the form:

$$H^2 = \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{q:\, O^q \in O''} \omega_{nr}^q \theta_{nr}^q + \sum_{n=1}^{N} \sum_{r=1}^{R} \sum_{q:\, O^q \in O''} (1 - \omega_{nr}^q)\, p_{nr} \theta_{nr}^q. \tag{15}$$

Thus, the task of obtaining the optimal topology of the information management system for a cluster organization is described by means of the mathematical model (15), (10), (11), (12), which, as you can see, belongs to the class of integer linear programming models.

3 Results and Discussion

The results of numerical calculations for the proposed model allow us to determine the distribution of work between cluster members. This distribution allows to achieve the minimum of overall costs for operating the information support system of this cluster organization with given resource constraints. Figure 1 shows the graphical interpretation of the problem solution results (15) with constraints (10), (11), (12), and obtaining the network topology of distributed information processing in cluster organization conditions. In these conditions, any of the participants performs at least one operation on processing information of documents. There is at least one thematic rubric, and none of them can be duplicated.

Practical testing of the proposed model was carried out during the creation and implementation of an automated scientific and technical information system of the Siberian Branch of the Academy of Sciences. It covers research institutes of

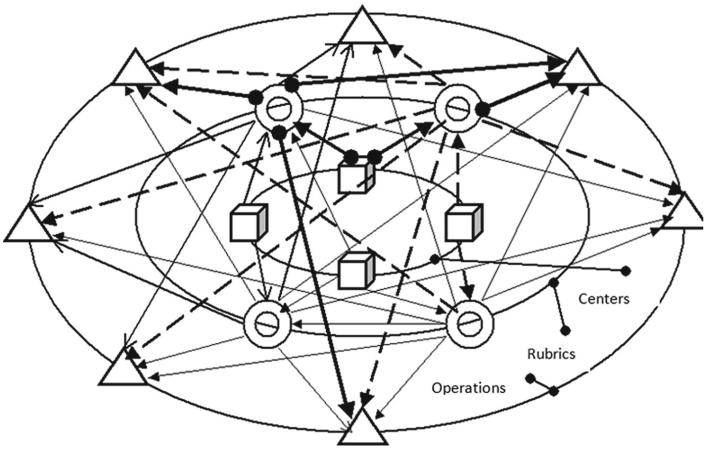

Fig. 1. Graphical interpretation of the results obtained

the Novosibirsk Scientific Center and information files in the fields of chemistry, biology, information sciences, environmental protection, etc. The results obtained allow to reduce financial costs for the creation and operation of the system almost twice [4].

4 Conclusion

The conditions of public-private partnership in solving the development problems of innovation activity cluster forms in the Republic of Kazakhstan involve new technologies of interaction between cluster organizations members. In some cases, this makes it possible to achieve a significant reduction in the cost of creating components of cluster organization infrastructure.

When forming the information infrastructure of a cluster, it is possible to organize a system of distributed information processing which would minimize the total cost of system operation.

References

1. Carpenter, K.H.: Competition? Collaboration? and cost in the new knowledge environment. Collect. Manag. **21**(2), 31–46 (1996)
2. Lesk, M.E.: The organization of digital libraries. Sci. Technol. Libr. **17**(3–4), 9–25 (1999)
3. Angelis, J.: A new look at community connections. III. Libr. **81**(1), 23–24 (1999)
4. Elepov, B.S.: Proektirovanie i e'kspluataciya regional'ny'x ASNTI. Nauka, Novosibirsk (1991)
5. Bobrov, L.K.: Strategicheskoe upravlenie informacionnoj deyatel'nost'yu bibliotek v usloviyax ry'nka. NGAE'iU, Novosibirsk (2003)

6. Anisova, N.A.: Gosudarstvennaya politika po podderzhke razvitiya klasterov: novacii, operezhayushhie teoriyu. E'konomika i upravlenie **3**(137), 75–86 (2017)
7. Novikova, I.V.: Klasterny'j podxod kak sposob povy'sheniya e'ffektivnosti investicionno-innovacionnoj politiki regiona. Regiony' v usloviyax globalizacii: problemy' innovacionno-investicionnogo razvitiya. SKFU, Stavropol (2014)
8. Plastinina, V.G.: Podxody' k analizu i ocenke e'ffektivnosti realizacii klasternoj politiki. E'konomika i predprinimatel'stvo **1**(78), 941–945 (2017)
9. Suroviczkaya, G.V.: Ocenka e'ffektivnosti realizacii klasternoj politiki na territorii regiona. Innovacii **3**(197), 58–60 (2015)