UDC 519.7; 519.8

MODELING OF THE OPTIMAL DATABASES ALLOCATION OF INFORMATION SYSTEMS BY THE AVAILABILITY OF INTERMEDIATE SERVERS

R. P. Krasniuk, G. G. Tsegelyk

Ivan Franko National University of Lviv

E-mail: krasniuk@ukr.net, kafmmsep@lnu.edu.ua

The problems of optimization of database placement in the modeling of distributed information systems is considered in the presence of intermediate storage servers as servers of data catching that play the role of data accumulation centers when using the batch mechanism data transfer placement. The mathematical formulated problems of optimal replication bases in the distributed information system is in the condition of minimizing the time of synchronization and minimizing the average time needed to find information. The precise solutions of the formulated problems using dynamic programming methods are obtained (in Bellman recursive equations).
The adaptation of the ant colony algorithm is carried out to construct solutions to the problems of optimal placement of replication databases. A series of numerical experiments are conducted and a conclusion is made as to the computational efficiency of the application of this approximate method in the case of increasing the number of nodes of the distributed information system.

Keywords: *optimization, mathematical modeling, distributed information systems, dynamic programming, Bellman recursive equations, ant colony algorithm.*

МОДЕЛЮВАННЯ ОПТИМАЛЬНОГО РОЗТАШУВАННЯ БАЗ ДАНИХ ІНФОРМАЦІЙНИХ СИСТЕМ ЗА НАЯВНОСТІ ПРОМІЖНИХ СЕРВЕРІВ

Р. П. Краснюк, Г. Г. Цегелик

Львівський національний університет імені Івана Франка

Розглянуто задачу моделювання оптимального розташування реплікаційних баз даних інформаційних систем за наявності серверів проміжного зберігання даних, що відіграють роль серверів їх кешування. Математично задачі оптимального розташування баз реплікації у розподіленій інформаційній системі сформульовано в умовах мінімізації часу синхронізації та середнього часу, необхідного для пошуку інформації. Отримано як точні розв’язки сформульованих задач з використанням методів динамічного програмування, так і наближені, що ґрунтуються на адаптації алгоритму мурашиної колонії.

Ключові слова: *оптимізація, математичне програмування, розподілені інформаційні системи, динамічне програмування, рекурентні рівняння Беллмана, алгоритм мурашиної колонії.*

Introduction. The design of databases in distributed information systems is a complicated problem that requires solving a number of tasks, among which the following can be specified:

– optimization of database placement by nodes of the distributed information system;

– synchronization of access to data and parallel processing of requests to ensure the required productivity of requests;

– support of duplicate data in multiple system nodes to reduce data transfer operations when queries are executed.

In accordance with the existing tasks of designing databases in distributed information systems, new approaches have emerged including the need for data replication (DR) – an asynchronous process of transferring changes of the source database to the database which belong to different nodes of the distributed system. The functions of DR are performed by special module of the database management system – the data replication server, which is called the replicator. Its task is to maintain the identity of the data in the target databases the data in the source database.

However, the use of this technology requires the solution of a number of optimization tasks, including the optimal placement of replication databases in a distributed information system in the conditions of limitations on the computing resources of nodes to minimize the time of data synchronization or to minimize the average time needed to search information provided there is no need for synchronization of the replication databases. In the technological aspect, these tasks can be solved by the presence of intermediate data storage servers which play the role of data accumulation centers for using the packet transmission mechanism. As a consequence, the study of optimal database distribution in information systems is relevant to the formulation of the corresponding mathematical models. The solution of the corresponding optimization problems is the subject of research of this paper.

Because of the importance of the problem of optimal distribution of databases in information systems the systematization of mathematical approaches and models for the modeling of distributed database systems was carried out in [1]. The investigation of mathematical models of multicriteria synthesis of physical structures of distributed databases was considered in [2] in which mathematical models were formulated and developed to implement in computer networks different topologies by criteria of minimum of reduced costs, access time to data and network traffic.

The concepts of construction and selection of distributed databases of information retrieval systems are devoted in [3]. This concept is based on the analysis and evaluation of qualitative and quantitative features determined by different technologies of database development to identify priority technologies and set its dominant features that will be involved in the automated design tools. Algorithmic provision of distributed databases is considered [4]. The life cycle of designing a distributed database is described as well as its stages. The problem of designing a distributed database is formulated and its complexity is estimated. The choice of genetic algorithms for the solution of the given problem is substantiated and an approach is proposed that allows us to consider the interdependence of the design stages.

The difference between this work and the results of other authors is the formation of new mathematical models for optimizing the distribution of replication databases, constructing both accurate and approximate solutions with the formation of efficient numerical algorithms. Exact solutions are constructed using dynamic programming methods. Bellman's recursive equations are obtained that are the independent results and can be used to analyze the accuracy of calculations of the corresponding optimization problems using numerical methods. Besides the ant colony algorithm was adapted to the mathematical models. According to the results of numerical experiments a conclusion is made regarding the computational efficiency of the application of this approximate method with the increase of the number of nodes of the distributed information system.

The formulation of mathematical models. *The problem of optimal distribution of replication bases in a distributed information system with servers of intermediate storage of data in terms of the synchronization time minimization is mathematically formulated.*

Let m databases need to be placed in n network nodes ($m < n$) in the presence of l ($l < m$) intermediate storage servers. Denote the bandwidth of the communication channel per unit time from the node i to the node j for the use of the intermediate storage server k through ρ_{ijk}. The average amount of data that needs to be transferred from node i to node j using the intermediate storage server k to synchronize replication databases is α_{ijk}. Then, if we introduce the coefficients x_{ijk} of using the communication channel in the optimization task, the mathematical model has formulation:

$$F = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{l} \frac{\alpha_{ijk}}{\rho_{ijk}} x_{ijk} \to \min , \tag{1}$$

$$x_{ijk} = \{0, 1\} , \quad \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{l} x_{ijk} \le 2m , \quad \sum_{i=1}^{n} \sum_{k=1}^{l} x_{ijk} \le m-1 , \quad i, j = 1,...,n; k = 1,...,l. \tag{2}$$

Mathematical formulation of the problem of optimal distribution of replication databases in the nodes of the information system with servers of intermediate storage of data provided that the average time needed to find information is minimized.

As in the previous case, m databases need to be distributed to n of the information system ($m < n$) in the presence of l ($l < m$) intermediate storage servers; λ_{ijk} – the intensity of requests from the node i to the database j for the use of the intermediate storage server k; β_{ijk} – the value of the data to be moved in response to the request between the node i of the information system and the base j; ρ_{ijk} – the channel capacity per unit time from the node i to the database j using the intermediate storage server k. Then if you enter coefficients x_{ijk} that determine whether the database j is located in the node i, the mathematical model of the problem in this case is such:

$$F = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} \frac{\lambda_{ijk}\beta_{ijk}}{\rho_{ijk}} (1 - x_{ijk})}{\sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{l} \lambda_{ijk}} \to \min , \tag{3}$$

$$x_{ijk} = \{0, 1\} , \quad \sum_{i=1}^{n} \sum_{k=1}^{l} x_{ijk} \ge 1; \; j = 1,...,m, \quad \sum_{j=1}^{m} \sum_{k=1}^{l} (1 - x_{ijk}) = 1; \; j = 1,...,n . \tag{4}$$

In the absence of intermediate storage servers the mathematical formulation of optimization problems (1), (2) and (3), (4) are somewhat simpler:

$$F = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\alpha_{ij}}{\rho_{ij}} x_{ij} \to \min , \tag{5}$$

$$x_{ij} = \{0, 1\} , \quad \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \le 2m , \quad \sum_{i=1}^{n} x_{ij} \le m-1 , \quad i, j = 1,...,n. \tag{6}$$

$$F = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\lambda_{ij}\beta_{ij}}{\rho_{ij}} (1 - x_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{m} \lambda_{ij}} \to \min , \tag{7}$$

$$x_{ij} = \{0, 1\} , \quad \sum_{i=1}^{n} x_{ij} \ge 1; \; j = 1,...,m, \quad \sum_{j=1}^{m} (1 - x_{ij}) = 1; \; j = 1,...,n . \tag{8}$$

An effective approach to constructing the analytic solutions for the formulated optimization problems (1)–(8) is the use of dynamic programming methods [5]. Without suggesting the intermediate calculations, the final results can be given – Bellman’s recursive equations for problems (5), (6) and (7), (8):

$$F_1\left(\{x_{ij}\}\right)=q_1\left(\{x_{1j}\}\right),\quad F_p\left(\{x_{ij}\}\right)=\min\left\{q_p(x_{pj})+F_{p-1}\left(\{x_{ij}\}\setminus\{x_{pj}\}\right)\right\}, \tag{9}$$

$$q_p(x_{pj})=\min_j\{\Lambda_{ij}\},\ j=1,2,...,n;\ p=2,3,...,m,$$

where Λ_{ij} in formula (9) for problem (5), (6) is equal to α_{ij}/ρ_{ij} and for problem (7), (8) – to $(\lambda_{ij}\beta_{ij})/\rho_{ij}$.

The formulation of the computational algorithm for constructing a solution of optimization problems. Because of the complicated above mentioned mathematical models, an effective way to construct an optimization problem solution is to use the ant colony algorithm [6, 7]. We can prove that the problems of the distribution of databases (1), (2) and (3), (4) can be reduced to the following:

$$F=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l}\gamma_{ijk}x_{ijk}\to\min;\quad x_{ijk}=\{0,1\},\quad i=1,...,n,\ j=1,...,m,\ k=1,...,l\ , \tag{10}$$

and additional conditions for the desired parameters of the problem x_{ijk}, which are not critical to the illustration of a common scheme of the algorithm. In formula (10), the coefficients γ_{ijk} are determined by "total costs" for the placement of the *i* database in the *j* node of the information system.

By using the MMAS algorithm [7] whose modifications are considered in this work, we have three additional conditions to the general scheme of the ant colony algorithm:

– in the renewal of the pheromones it takes only the better of the current ants iteration;

– the value of the pheromone level is limited to the interval $[L_{\min}>0,\ L_{\max}]$;

– at the beginning of the algorithm, the level of pheromones for all nodes is set to a level $L_{\max}$ which provides a better study of the set of solutions to the problems at the initial stage.

The ant colony algorithm discussed hereafter is iterative where, at each iteration, the colony (set) of ants generates a set of solutions in accordance with the levels of pheromones. We can obtain these results, the best ones will be selected which will update these levels of pheromones.

If we enter the designation for the appropriate pheromone level $L_t(i,j,k)$, where *t* is the current iteration index; *i* and *j* define the database indexes and the information system node respectively, the probability of placing the *i*th database in the *j*th node of the information system, provid that the use of the *k*th intermediate storage server, is

$$p_{s,t}(i,j,k)=\begin{cases}\dfrac{\left[L_t(i,j,k)\right]^q[\gamma_{ijk}]^{q-1}}{\sum\limits_{(p,r)\in S_{i,s}}\left[L_t(i,p,r)\right]^q[\gamma_{ipr}]^{q-1}}, & j,k\in S_{i,s};\\ 0, & j,k\notin S_{i,s};\end{cases} \tag{11}$$

In formula (11) the $S_{i,s}$ is the set of not used nodes information system, which can be hosted database *i* for *s*th ants' colony, *q* is parameter that defines the "greed" of the algorithm under the condition *q* = 0 the choice of the node will be determined by the lowest "total cost" γ_{ijk} and condition *q* = 1 determines the choice of the node only for the level of pheromones, which determines the rapid degeneration of the result to one suboptimal solution. Note that the formation of a set $S_{i,s}$ at each step of the algorithm is carried out using additional conditions for the desired parameters of the task of the corresponding optimization problem.

Among the solutions formed by each ant the best one is selected that will be used to update pheromone levels:

$$L_{t+1}(i,j,k)=\begin{cases} L_{\min}, & L^{*}\le L_{\min}; \\ L^{*}, & L_{\min}<L^{*}<L_{\max}, \quad L^{*}=\rho\cdot L_t(i,j,k)+\Delta L_t(i,j,k); \\ L_{\max}, & L^{*}\ge L_{\max}; \end{cases} \tag{12}$$

where $\rho\in(0{,}1)$ is the velocity of evaporation of pheromones; $\Delta L_t(i,j,k)=x_{ijk,t}^{\text{best}}$ is the contribution of the best ant iteration to the general level of pheromones in the nodes of the information system, where $x_{ijk,t}^{\text{best}}$ is the best solution in the iteration t which provides the least value of the target function (10).

Using the above considerations we will provide the scheme of ant colony algorithm for replication database distribution tasks.

Step 0. The value of the error of calculations ε, the value of the maximum number of iterations $t_{\max}$, the interval of change in the level of pheromones $[L_{\min}, L_{\max}]$, the value of the parameters of the "greed" of the algorithm q and the rate of evaporation of pheromones ρ are determined. For each node of the information system, the level of pheromone is set at the level $L_{\max}$ and the iteration index t = 1 and also the initial value of the target function $F_0=n\cdot l\cdot\left(\max\limits_{i,j,k}\gamma_{ijk}\right)$ are set.

Step 1. The initialization of the ant colony by one of the possible ways:

– coverage of nodes – the number of ants coincides with the number of nodes n information system when each ant is initially placed in the corresponding one of the node;

– accidental coverage – initially ants in the nodes of the information system are placed randomly when the number of ants and system nodes may not coincide;

– placement of the colony in focus – the whole colony of ants in each iteration is in one node of the information system. Number of ants in the colony can be arbitrary;

– migrating colony – the whole colony of ants in each iteration moves into an arbitrary, randomly chosen node of the system. The number of ants in a colony may also be arbitrary.

Step 2. Calculation of the solution of problem (10) for each ant from the colony by the use of probabilities (11).

Step 3. The choice of the best solution $x_{ijk,t}^{\text{best}}$ for the current iteration from the condition of obtaining the smallest value of the objective function with (10): $F_t=F(x_{ijk,t}^{\text{best}})=\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{l}\gamma_{ijk}x_{ijk,t}^{\text{best}}$.

Step 4. Checking of the conditions for termination of iterations: the found values of the target function in the last two iterative steps of the algorithm do not exceed the error of the calculations $|F_t-F_{t-1}|<\varepsilon$ or exceed the maximum number of iterations $t>t_{\max}$. For fulfilling these conditions the values $x_{ijk,t}^{\text{best}}$ found in the current step from the solution of the corresponding optimization problem. Otherwise, proceed to *Step 5*.

Step 5. Increase the value of the index of iteration per unit, carry out the calculation of the pheromone level by formula (12) and return to *Step 1*.

Remark 1. As it is clear from the above considerations, *Step 2* of the proposed algorithm allows parallel calculation of the solution of the optimization problem (10) for each colony ant. It is clear that this provides a reduction in the total running time of the algorithm.

Remark 2. From the practical point of view, in the tasks of placing replication databases, it is not necessary to obtain a global minimum – it is sufficient to obtain a locally optimal solution, which is provided by choose of coefficient q close to unit and coefficient $\rho > 0.5$. Fulfillment of these conditions ensures convergence of the algorithm though, possibly, to the local minimum.

The results of numerical analysis of the optimization problem by the formulated algorithm. As numerical experiments have shown, the ant colony algorithm provides the optimal solution to the tasks of placing replication databases for 250…500 iterations with arbitrary precision. Data for numerical experiments are chosen randomly, the obtained results of the calculation by an approximate algorithm are compared with exact solutions found using the Bellman’s recurrence equations. It should be noted that the effectiveness of the application of the ant colony algorithm increases with the increase of the problem dimension – the number n of distributed information system nodes.

Fig. 1 presents the dependence of the mean time for solving optimization problems (5), (6) and (7), (8) on the ant colony algorithm in comparison with the results found using Bellman's recursive equations (9), where the left figure corresponds to the case of the distribution of six databases between the eight nodes of the information system, and the right – an option for placing eighteen databases among the twenty-four nodes of the information system. The expected efficiency of the approximate method of finding the optimal value of the target function is observed with the increase of the dimensions of the problem.

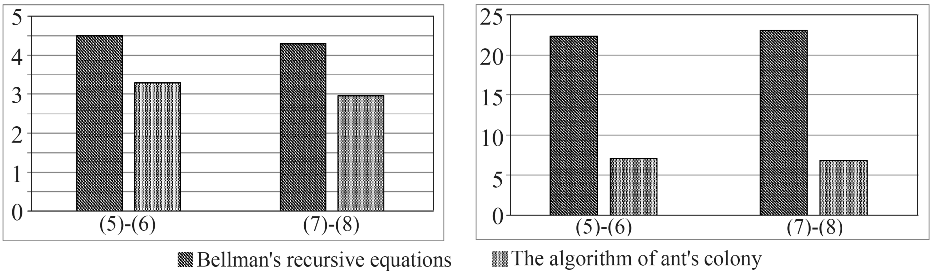

Fig. 1. Dependence of the mean time for solving optimization problems (5), (6) and (7), (8) (in seconds) by using of the ant colony algorithm and Bellman’s recursive equations.

The influence of the “greed” parameter of the algorithm q on the average relative error of the calculations of the solutions of optimization problems (5), (6) and (7), (8) is shown in Fig. 2. As it can be seen from the graphs of the dynamics of the relative error of calculations, there is a decrease in the error of calculations with an increase in the value of q which is conditioned by a more significant influence on the choice of node for placing the database value of the pheromones level than the magnitude of “greed” which, in the case of problems (5), (6) and (7), (8) is determined by the choice of nodes with the lowest “total cost” γij. Note that the graphs in Fig. 2 constructed for the value of parameter of the evaporation rate of pheromones $\rho = 0.7$.

Fig. 3 allows us to estimate the influence of the parameter ρ on the error of the calculations of the solutions of the optimization problems (5), (6) and (7), (8) for the choice of value $q = 0.7$. As you can see from the charts, we get the best results for $\rho = 0.9$. However, it should be noted that by increasing the parameter ρ the number of iterations of the algorithm to achieve the required accuracy increases. Therefore, we can conclude that for practical application, the results can be obtained for the values of the parameters ρ = 0.7 and q = [0.6; 0.8] with sufficient accuracy. The values of the remaining parameters describing the mathematical models (5), (6) and (7), (8) for computational experiments, a total of one hundred variants, were determined randomly.

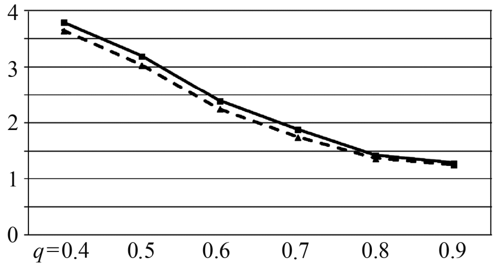

Fig. 2.

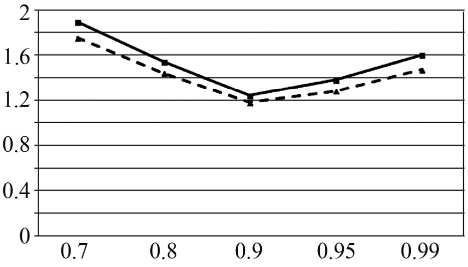

Fig. 3.

Fig. 2. Dependence of the average relative error of calculations of solutions of optimization problems (5), (6) and (7), (8) by the ant colony algorithm on the magnitude “greed” parameter q of the algorithm.

Fig. 3. Dependence of the average relative calculations error of solutions for optimization problems (5), (6) and (7), (8) by using the ant colony algorithm for the values of the evaporation rate of pheromones ρ parameter.

CONCLUSION

The paper presents the adapted ant colony method to construct solutions to the tasks of optimal placement of replication databases. According to the results of numerical experiments, a conclusion is made regarding the computational efficiency of the application of this approximate method with the increase in the number of the distributed information system nodes.

1. *Tsegelyk G. G.* The distributed database systems. – Lviv: Svit, 1990. – 168 p.
2. *Beskorovaynyi V. V., Ul’ianova O. S.* The mathematical models of multicriteria synthesis of physical structures of distributed databases // Eastern European J. of Adv. Techn. – 2010. – **Vol. 4**. – P. 44–48.
3. *Yakovlev Yu. S.* On the concept of construction and selection of distributed databases of information retrieval systems // Mathematical Machines and Systems. – 2013. – **Vol. 2**. – P. 35–53.
4. *Chumachenko Ye. I., Zakharov S. S.* The Algorithmic support of distributed databases // Artificial Intelligence. – 2013. – **Vol. 1**. – P. 49–54.
5. *Tsegelyk G. G.* The mathematical programming. – Lviv: Ivan Franko National University of Lviv, 2011. – 168 p.
6. *Dorigo M., Gambardella L. M.* Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem // IEEE Transact. on Evolutionary Computation. – 1997. – **Vol. 1**, Issue 1. – P. 53–66.
7. *Stützle T., Hoos H. H.* Max-Min Ant System // Future Generation Computer Systems. – 2000. – **Vol. 16**, Issue 8. – P. 889–914.

Received 11.09.2018