Mathematical model for optimizing distributed information systems

A N Mazalov, A V Yakovlev, V V Alekseev and I S Gaidashov
Department of Information System and Information Security, Tambov State Technical University, 106 Sovetskaya St, Tambov, 392000, Russia

E-mail: alexmazalov2293@yandex.ru

Abstract. An approach to optimizing distributed information systems is discussed in this paper. An argument explaining the demand for such distributed information systems nowadays is presented as well. The topic is crucial because it is impossible to find an optimal solution due to the curse of dimensionality of an NP-complete problem. The problem of data distribution by nodes of a distributed information system was defined. A mathematical model for optimizing distributed information systems was developed to solve the problem. In Stage 1 an analysis of criteria optimization of resource distribution among distributed information systems was carried out. The following definitions were established in this paper: information entity, attribute of the entity, distributed database. From the perspective of distributing entities among nodes a distributed information system was determined. As a result of describing elements of the mathematical model, a function of information flows was obtained. The function deals with data flows during distributed query execution. Including certain restrictions, such as the prohibition of hosting one entity on different nodes, the problem of distributing entities among nodes is presented. It is established that the problem of distributing resources is related to integer programming which, by itself, is concerned with optimizing distributed database structures, where the objective function indirectly depends on the variable of distribution and presence of restrictions.

1. Introduction
Currently, plenty of enterprises employ distributed information systems (DIS) for their business. The reason for such a trend, compared to local systems, is that a DIS raises productivity while lowering costs. Not only does that lead to an increase in the dimensions of a DIS but it also comes to the point where an amount of stored and transmitted data also grows. Consequently, there’s a challenge for developers to boost the productivity of a DIS even further. A solution to this problem can be achieved by optimizing resource distribution by nodes [1, 2, 3].

The topic is especially important considering that it is impossible to find an optimal solution because of the curse of dimensionality of an NP-complete problem [4].

2. Mathematical model
The problem of data distribution and hosting will be referred to as the resource distribution and hosting problem on DIS nodes.

Choosing an optimization criterion, which assesses how effective a DIS is, is the beginning of a solution to the problem.
An analysis of the subject area suggested that the authors use the following optimization criteria: lowering the amount of input and output data \([5]\), average data flows, overall data costs, query response time \([6]\), average feedback time of the system, query execution time and the coefficient of transaction availability in the database \([7]\), accumulated data flows between databases and users while taking into account database integrity violation risks.

As a result of the analysis it was established that previously adopted methods used NP-complete problems in the process of solving and they didn’t reduce their exponentially growing complexity.

A DIS topology will be represented as a network connecting nodes (a DIS node is a local computer network with several workstations and one or a few database servers) via data transmission channels:

\[ G = (V, E), \]  

where \( V = \{v_i, l=1, \ldots, L\} \) is a set of nodes on a network \( G \); \( E = \{e_{m}, m=0, \ldots, M\} \) is a set of arcs on the network \( G \).

Connections between the vertices are written in the form of a weighting coefficient matrix \( \Gamma = [\gamma_{ij}] \):

\[
\Gamma = \begin{pmatrix}
\gamma_{11} & \cdots & \gamma_{1L} \\
\vdots & \ddots & \vdots \\
\gamma_{L1} & \cdots & \gamma_{LL}
\end{pmatrix}
\]  

(2)

An Information Entity (s) is «a specific or abstract object and connections between the objects». An Attribute of the Information Entity (A) is «an observed property of the entity or observed property of a connection between entities of a problematic area».

The given DIS is designed to solve a set of problems \( B = \{b_p, p=1, \ldots, P\} \). It uses data on instances of an information entity \( S \), which is hosted on DIS nodes:

\[
S = \{s_{r_1, \ldots, r_w}, r_1 = 1, \ldots, R_1, \ldots, r_w = 1, \ldots, R_W, w = 1, \ldots, W\}
\]  

(3)

or

\[
S = \bigodot s_{r_1}, s_{r_1} = \bigodot s_{r_1, r_1}; \ldots; s_{r_1, \ldots, r_W} = \bigodot s_{r_1, \ldots, r_W},
\]

(4)

where \( S \) is a set of entities used for solving the problem; \( r_1=1, \ldots, R_1, \ldots, r_W = 1, \ldots, R_W, w = 1, \ldots, W \) is the entity index number, which determines the location of the entity within the dependency hierarchy, thus independent entities of level \( w = 1 \) have a single-digit index (e.g., \( s_{11}, s_{12}, \ldots, s_{1R_1} \)), whereas the subentities entering level \( w = 2 \) have a two-digit index (e.g., \( s_{11,1}, s_{11,2}, \ldots, s_{11,R_2} \)); \( \bigodot \) is a combining operation which enables obtaining of entities of the previous level.

The concept of creating indices of entities is shown in figure 1.

The description of entities is given in the form of a reference database directory.

The following part is devoted to describing the DIS from the perspective of distributing entities among network nodes.

The structural description of entities is provided in Table 1.

Let us suppose that on every \( l-th \) node a strictly defined set of problems is being solved:

\[
F_l = \{f_{l,n}, n = 1, \ldots, N_l\}, l = 1, \ldots, L, F_l \subset B
\]  

(5)

Let us introduce a new definition, Frequency Response of a Problem (FRP), which is a number of repetitions of every problem being solved on the \( l-th \) node per unit of time.

The FRP on the \( l-th \) node can be given as:
\[ Q = \| q_{l,n} \|. \quad (6) \]

where \( l \) is the node number where the problem is being solved; \( n \) is the problem number.

The following set of entities is applied to solve the \( n \)-th problem:

\[ D_{l,n} = \{ d_{l,n,k}, k = 1, \ldots, K_{l,n} \}, \quad n = 1, \ldots, N, l = 1, \ldots, L, D_{l,n} \subset S; \]

\[ D_{l,n}^- \bigcap D_{l,n}^- = \emptyset; D_{l,n}^- \bigcup D_{l,n}^- = D_{l,n}^-; K_{l,n}^- + K_{l,n}^- = K_{l,n}. \quad (7) \]

where \( D_{l,n} \) is a set of entities used for solving the \( n \)-th problem on the \( l \)-th node. The set \( D_{l,n} \) is an injection of the set \( S \) into the set \( F_{l,n} \). At the same time the set \( D_{l,n} \) consists of \( D_{l,n}^- \) and \( D_{l,n}^- \). The former one is a set of nonrelocatable entities used for solving the \( n \)-th problem on the \( l \)-th node, while the latter one is a set of relocatable entities used for solving \( n \)-th problem on the \( l \)-th node.

**Figure 1.** Hierarchical model of entities.

**Table 1.** Structural description of entities.

| Description | Mathematical notation |
|-------------|-----------------------|
| A set of entities used in the DIS | \( S = \{ s_{l,1,W} \cdot r_1 = 1, \ldots, R_1, r_w = 1, \ldots, R_W, w = 1, \ldots, W \}; \) |
| s_{r1,...,rw} = \{ < A_{j(r_1,...,rw)}, s_{r1,...,rw} >, j = 1,...,J_{(r_1,...,rw)} \} | |
| Description of entities from the level \( W \) | \( s_{l,1,...,rw} = \{ < A_{j(r_1,...,rw)}, s_{l,1,...,rw} >, j = 1,...,J_{(r_1,...,rw)} \} \) |
| A set of attributes describing set \( S \) objects | \( A = \{ A_{j(r_1,...,rw)}, j = 1,...,J_{(r_1,...,rw)} \} \) |

Let us introduce a term, Frequency Response of the Entity (FRE), which is a number of queries to the entity \( d_{l,n,k} \) in the process of solving the \( n \)-th problem on the \( l \)-th node per unit of time.

The FRE can be given as:

\[ U = \| u_{l,n,k} \|. \quad (8) \]

where \( l \) is the node number where the problem is being solved; \( n \) is the problem number where the entity is used during the solution; \( k \) is the entity number.
Ф is the injection of the set S into the set V. Ф is a set of entities hosted on nodes including Ф_i, which is a set of entities hosted on the l-th node. Ф_i also has Ф^*_i, a set of nonrelocatable entities, and Ф^~_i, a set of relocatable entities:

\[ \Phi_i \subseteq \Phi; \Phi^*_i \subseteq \Phi_i; \Phi^~_i \subseteq \Phi_i; \]
\[ \Phi^*_i \cap \Phi^~_i = \emptyset; \Phi^*_i \cup \Phi^~_i = \Phi_i; X^*_i + X^~_i = X_i. \] (9)

The formula of obtaining the cost of delivering an information entity from an origin node to a destination node, where the problem will be solved, is as follows:

\[ \Psi_{l,n,k} \left( I_n \left( \Phi \right) \right) = \max \left( \Psi_{l,n,k,x} \right), \kappa \in R_{l,n}, \] (10)

where R_{l,n} is the lowest cost path connecting the source and host nodes; κ is the ark belonging to the path.

Table 2 contains the description of the mathematical model elements for data distribution if a DIS is active.

| Elements | Mathematical notation |
|----------|------------------------|
| Network G | \( G = \left( V, E \right) \) |
| Set V of vertices of G | \( V = \left\{ v_i, l = 1, \ldots, L \right\} \) |
| Set E of arcs of G | \( E = \left\{ e_m, m = 0, \ldots, M \right\} \) |
| Weighting coefficient matrix | \( \Gamma = \left[ v_{i,k} \right] \) |
| Set B of system-solved problems | \( B = \left\{ b_p, p = 1, \ldots, P \right\} \) |
| Set \( F_i \) of problems to be solved on the l-th node | \( F_i = \left\{ f_{l,n}, n = 1, \ldots, N_i \right\}, l = 1, \ldots, L, F_i \subseteq B \) |
| Frequency Response of the Problem | \( Q = \left\| q_{l,n} \right\| \) |
| Set \( D_{l,n} \) of entities used for solving the n-th problem on the l-th node | \( D_{l,n} = \left\{ d_{l,n,k}, k = 1, \ldots, K_{l,n} \right\} = \left\{ D^*_l, D^~_l \right\}, n = 1, \ldots, N_i, l = 1, \ldots, L \) |
| Set \( D^*_l \) of nonrelocatable entities used for solving the n-th problem on the l-th node | \( D^*_l = \left\{ d^*_l,k^*, k^*_l = 1, \ldots, K^*_l \right\}, n = 1, \ldots, N_i, l = 1, \ldots, L \) |
| Set \( D^~_l \) of relocatable entities used for solving the n-th problem on the l-th node | \( D^~_l = \left\{ d^~_l,k^~, k^~_l = 1, \ldots, K^~_l \right\}, n = 1, \ldots, N_i, l = 1, \ldots, L \) |
| Frequency Response of the Entity | \( U = \left\| u_{l,n,k} \right\| \) |
| Set \( \Phi \) of entities hosted on the nodes | \( \Phi = \left\{ \Phi_i, l = 1, \ldots, L \right\}, \bigcup_{l=1}^{L} \Phi_i = S \) |
| Set \( \Phi_i \) of entities hosted on the l-th node | \( \Phi_i = \left\{ \phi_{l,x}, x = 1, \ldots, X_i \right\} = \left\{ \Phi^*_i, \Phi^~_i \right\}, l = 1, \ldots, L \) |
| Set \( \Phi^*_i \) of nonrelocatable entities hosted on the l-th node | \( \Phi^*_i = \left\{ \phi^*_l,x^*, x^*_l = 1, \ldots, X^*_i \right\}, l = 1, \ldots, L \) |
| Set \( \Phi^~_i \) of relocatable entities hosted on the l-th node | \( \Phi^~_i = \left\{ \phi^~_l,x^~, x^~_l = 1, \ldots, X^~_i \right\}, l = 1, \ldots, L \) |
Table 2. Description of mathematical model elements.

| Elements                                                                 | Mathematical notation |
|-------------------------------------------------------------------------|-----------------------|
| Set $H_{l,n,k}$ of attributes describing the $k$-th entity in the process of solving the $n$-th problem on the $l$-th node | $H_{l,n,k} = \{ h_{l,n,k,i} : i = 1, \ldots, L \}$, $k_{l,n} = 1, \ldots, K_{l,n}$, $n = 1, \ldots, N$, $l = 1, \ldots, L$, $H_{l,n,k} \subseteq \Lambda$ |
| The $i$-th attribute can be described by a tuple                        | $h_{l,n,k,i} = \{ < \Omega_{l,n,k,i}, \Theta_{l,n,k,i}, Y_{l,n,k,i} > \}$ |
| Size of the $i$-th attribute of the $k$-th entity in the process of solving the $n$-th problem on the $l$-th node | $Y_{l,n,k,i}$ |

Based on the description provided above a function of data flows during the execution of distributed queries was obtained:

$$Z(\Phi) = \sum_{i=1}^{L} \sum_{n=1}^{N} \sum_{k=1}^{K} \psi_{l,n,k} (l_n (\Phi)) \rho_{l,n,k} \sum_{i=1}^{l_n} Y_{l,n,k,i};$$

(11)

where $\psi_{l,n,k} (l_n (\Phi))$ is the cost of delivering data from an origin entity $l_n$ to a destination node $l$ ($l \neq l_n$).

The following restrictions must be defined in the process of solving the entity distribution problem on DIS nodes:

- one information entity mustn’t be located on different nodes (12):

$$\bigcap_{l=1}^{L} \bigcap_{x=1}^{X_l} \phi_{l,x} = \emptyset; \quad \bigcup_{l=1}^{L} \bigcup_{x=1}^{X_l} \phi_{l,x} = S$$

(12)

- nodes have limited storing capabilities (13):

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \rho_{l,n,k} \sum_{i=1}^{l_n} Y_{l,n,k,i} \leq \Lambda_l,$$

(13)

where $\Lambda_l$ is the size intended for data distribution and reserved by nonrelocatable entities; $\rho_{l,n,k}$ is the number of instances of an entity [5].

The first restriction exists due to the fact that optimal data distribution does not imply any duplication. In the meantime, the issue of data storage security requires additional measures. That is why there should be redundant nodes on the network to ensure data security since they will store current replicas of a DIS.

The second one is related to the hardware capabilities of a node. It allows controlling the value of input information on a node ($\Lambda_l$). For nodes, which lack the capability of sharing their resources with others, this value is equal to 0.

Considering the aforementioned information, the problem of entity distribution among nodes can be determined as follows: define such entity distribution ($\Phi$) among DIS nodes which would lower the objective function (query processing time in the process of solving problems on a given interval) (11) with restrictions (12) and (13) in place

$$\Phi \rightarrow \min_{\Phi} Z(\Phi).$$

(14)

The problem of distributing resources is related to integer programming which, by itself, is concerned with optimizing distributed database structures, where the objective function indirectly depends on the variable of distribution and presence of restrictions.
3. Conclusion
The developed mathematical model for optimizing a DIS is able to raise its productivity by reducing query processing time in the process of solving problems in limited time intervals.

References
[1] Date C J 2003 An Introduction to Database Systems (Boston : Addison Wesley)
[2] Tanenbaum A S 2003 Distributed Systems. Principles and Paradigms (Upper Saddle River (N J): Prentice Hall PTR)
[3] Schwarzkopf E A, Choporov O N, Razinkin K A, Yurasov V G and Mazalov A N 2020 Mathematical and algorithmic support for the early detection process automation of potentially dangerous content in Internet resources IOP Conf. Series: Materials Science and Engineering 862 052037 DOI:10.1088/1757-899X/862/5/052037
[4] Garey M and Johnson D 1979 Computers intractability: a guide to the theory of NP-completeness (San Francisco: W. H. Freeman) p 338
[5] Apers P M G 1988 Data allocation in distributed database systems ACM Trans. Database Syst 13(3) DOI:https://doi.org/10.1145/44498.45063
[6] Johansson J M, March S T and Naumann J D 2003 Modeling Network Latency and Parallel Processing in Distributed Database Design Decision Sciences Journal 34(4) DOI:10.1111/j.1540-5414.2003.02409.x
[7] Mukkamala R, Bruell S C and Shultz R K 1988 Design of partially replicated distributed database systems: an integrated methodology ACM SIGMETRICS conf. 16(1) DOI:https://doi.org/10.1145/1007771.55617