Alignment of Cluster Complexity at Network Systems

A.K. Enaleev
Senior researcher,
V.A. Trapeznikov Institute of Control Sciences of Russian Academy Sciences
Russia

V.V. Tsyganov
Chief researcher,
V.A. Trapeznikov Institute of Control Sciences of Russian Academy Sciences
Russia

This paper considers data management structures and cluster technologies in large-scale networks. Suboptimal network partitioning problems are formulated on the base of complexity index alignment. We propose methods for these problems solving, in particular the data clusters number and its boundaries determining. We describe a multi-stage iterative scheme for the semantic data mining from a large document with interdependent sections as well. At the first stage, a priori data mining complexity from these sections is estimated. Then we refine this complexity taking into account the revealed data mining from the adjacent sections. Based on this, the final partitioning of the data set of a big document into clusters is formed under circumstances of deadline and restrictions on financial resources. The proposed methods have been applied in some large-scale transport projects.

Keywords: *Big data, data mining, alignment, cluster, network partitioning, NP-hardness, heuristics.*

1. INTRODUCTION

In the Industry 4.0 paradigm, one of the main problems is the need to create horizontal and vertically integrated management structures, big data processing, fragmentation of networked production structures and value chains, digitalization, and the use of artificial intelligence [1-4]. In connection with these sections, in this article we adapt and summarize the study and experience of applying our developments in the field of managing large-scale transport networks in particular Russian railways [5, 6]. Management problems in such complex and large-scale production network structures are associated with the network division into clusters. At the same time, it is important to fragment the network into equally complex clusters in order to control, use artificial intelligence and process big data [2, 3].

An important feature of data retrieval is it volume and variety [7,8]. This is characteristic, in particular, of many areas of business intelligence [9]. For example, when developing large projects, it is necessary to deal with a large number of various documents and texts. These documents and texts often have significant volumes and complex structure, cross-references, contain data with a variety of semantics. To facilitate the work with such objects, it is necessary to solve the problem of splitting the entire set of data into cluster systems. When posing this problem, we must first formulate the criteria for splitting the data into clusters. In the composition of these criteria we will refer the volume (hereinafter "complexity") of the cluster and the semantics of the data contained in it. Accordingly, for this it is necessary to be able to measure both the complexity of the cluster and the data fragments included in the clusters, and to determine the clusters semantics. The definition and measurement of these characteristics are some difficult problems. Sometimes these problems are easily solved from the content and properties of practical applications. To begin with, we will assume that we are able to solve these problems, and then we will present some approaches to the definition of these characteristics.

Multiple components and procedures must be coordinated to ensure a high level of data quality and accessibility for the application layers, e.g., data analytics and reporting [7]. Also it is necessary to increase the velocity with which the data is retrieved [8].

In this connection, there arises the problem of developing technologies and algorithms that minimize the time of data mining. In addition, specialists of different profiles are needed to extract a variety of data. In practice, this leads to the creation of special teams and even hierarchical organizations to extract data (for example, when developing and examining large-scale projects). At the same time, the organizational structure (in particular, the composition and the number of experts) must meet the requirements of the minimum cost of data extraction. We consider methods, technologies and algorithms for data extraction in a two-tier "leader-expert" system that minimize not only the time but also the cost retrieval of data. This uses semantic analysis, in conjunction with the idea of relational data mining [10]. The source of relational tables can be either a real network (for example, transport or information), or a network formed by the researcher on the basis of an array analysis of big data. Such an array, for example, can be given in a document containing big data (a shortly - big document). In addition, the origin of the data is important for making a decision (especially in the context of expertise and evaluation of a complex project, in comparison with known analogs, using

Received: March 2019, Accepted: May 2019
Correspondence to: Prof Vladimir Tsyganov
Institute of Control Sciences, Profsoyuznaya 65,
117997, Moscow, Russia
E-mail: bbc@ipu.ru
doi:10.5937/fmet1904711E

benchmarking) [11]. In this case, we need to process not only a large array of data about the object but also a large array of data about similar objects.

Generally speaking, there is a large number of different variants of the problem of partitioning data into clusters in practice. Here, we consider two areas of research that differ widely in the problems statement, but they are similar in approach to their solution on the basis of the *Cluster Complexities Alignment* principle proposed below.

The first area relates to the partitioning into clusters of spatially separated agents having certain volumes of data and having communication channels with each other. Agents are busy with some operations with the information they have, as well as receiving and transmitting this information to other agents through the available communication channels. Here we consider the problem of partitioning clusters with given semantics of data from agents, as well as the problem of matching different types of partitions with different semantics of agents.

The second area is related to selective processing of big data. We propose a mechanism for allocating basic information clusters and their distribution among a limited number of agents (experts). This takes into account the alignment of the complexity of clusters and the distribution of semantics among them.

Another type of problem arising in the decomposition of social network data management into clusters and the formation of the corresponding hierarchical organizational structure, due to the appearance of elements in the structure of their own purposes (because the network is present people with their own interests). As a result, when the elements of the system interact in the context of conflicting goals, problems arise with deliberate distortion of the information circulating in the system. When considering mathematical models and productions of the problem, the methodology of synthesis of optimal hierarchical structures [12-14] and organizational management [12,15,16] is used.

The methods described in this article were used to solve the problem of clustering management in a large-scale transport network like Russian Railways. The proposed approaches apparently can be also distributed to solve transport problems in [17,18].

These methods are based on graph partitioning algorithms. The problem of partitioning graphs was considered in a large number of publications. It suffices to refer to a detailed review and classification of these methods in [19]. In contrast to the methods [19], this article proposes heuristic methods based on the specifics of the railway network partitioning problem [6]. These methods take into account additional requirements for cluster geometry. This allowed for the polynomial hardness of the algorithms.

2. MODELS OF DATA NETWORK PARTITIONING

We suppose that each element of large-scale network (its nodes and edges) include some semantic data. Let us consider a hierarchical system of data management at a social large-scale network including a Direction and its subordinate functional data management centers. We can think that this centers deal with different semantic data. For simplicity, we shell consider the case when there are only two centers. Each center manages within the framework of its functional responsibility (its regional subnets partition, in other words semantic clusters partition). We call regional subnets at any partition as the network data management clusters. These partitions can differ from each other. We suppose each of the regional subnets (cluster) has its own data cluster management body (briefly – manager). Note that the network can be broken up in different ways for each of the center. We call partition related to the first center the partition of the first type, and to the second center, respectively, of the second type (figure1).

Let us consider the network S consisting of n nodes. For each of the center which carries out its own kind of activity it is necessary to break up the network into sub networks of clusters. As already noted, network breakdowns into networks clusters may vary for each center. Let us denote $g^1 = \{g_i^1\}$ the network S partition for the first center on N clusters of the first type, and denote $g^2 = \{g_i^2\}$ the network S partition for the second center on N clusters of the second type.

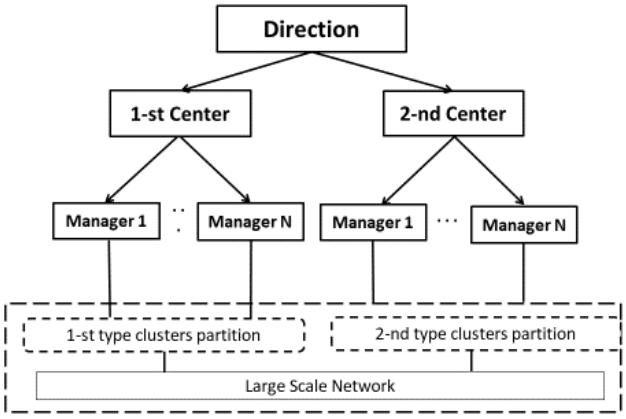

Figure 1. The structure of Network Partition model

Let us suppose that the partitioning into network clusters satisfies the conditions:

$$\bigcup_{i=1}^{N} g_i^m = S \text{ and } g_i^m \cap g_j^m = \varnothing$$

where m is the number of the partition type (in the case under consideration $m = 1$ or $m = 2$). The boundaries of each of the partitions pass through the nodes of the network. We supplement the network at each node with an edge that is a loop. For each type of partition, the loop at the node through which the boundary passes can only refer to one cluster of the corresponding type. We have N managers of the first type, and N managers of the second type.

We denote by G^1 and G^2 the sets of admissible partitions of the first and second types, respectively. Let a generalized indicator characterizing the complexity of data management (CM) for the considered center (for partition of the m-th type) is given:

$$K(g^m) = \bar{K}(K_0^{g^m}, K_1^{g^m}, ..., K_i^{g^m}, ..., K_N^{g^m})$$

where $K_0^{g^m} = K_0^{g^m}(N)$ is the indicator characterizing the CM by the social network for the *m*-th center, $K_i^{g^m} = K_i^{g^m}(\bar{l}_i^m)$ is an indicator characterizing the CM for the *i*-th manager in the partition g^m, $i = 1, ..., N$.

Here $\bar{l}_i^m$ is the set of parameters of the CM for the elements of the cluster *i* in the partition g^m (the meaning of these parameters will be clarified below).

We do not describe the methods for calculating indicators of complexity here, since they are specific to each applied problem. An example of complexity assessment is given in [20].

We assume that the function $\bar{K}(...)$ is non-decreasing, i.e. the value of $K(g^N)$ does not decrease in magnitude $K_i^{g^N}$. We assume $K_0^{g^m} = K_0^{g^m}(N)$, $K_i^{g^m} = K_i^{g^m}(\bar{l}_i^m)$ do not decrease in their arguments. A particular case of such a generalized indicator is additive,

$$K(g^m) = K_0^{g^m}(N) + \sum_{i=1}^{N} K_i^{g^m}(\bar{l}_i^m).$$

In general, the problem of optimizing the CM is posed as minimizing by choosing the number of clusters *N* in the partition and the partition g^m itself on the set G^N

$$\max_{N_{\min} \le N \le N_{\max}} \max_{g^N \in G^N} \bar{K}(K_0^{g^m}, K_1^{g^m}, ..., K_i^{g^m}, ..., K_N^{g^m})$$

where N_{min} and N_{max} define low and upper restrictions on the clusters number.

In general, such a problem is difficult to solve due to NP-hardness. Therefore, it is proposed to replace it (decompose) possibly with loss of the solution accuracy for two problems: estimates of the polygons number *N* in partitions, and of the partition itself. In this case, we propose to replace the search for an optimal partition on a set G^N by a search for cluster complexity alignment at a partition (the difference in CM should be minimal).

The principle of cluster complexity alignment at a partition: the difference in clusters complexity should be minimal:

$$\Delta^{g^{*N}} = \min_{g \in G^N} [\max_{1 \le i \le N} K_i^g(\bar{l}_i^{g_i^N}) - \min_{1 \le i \le N} K_i^g(\bar{l}_i^{g_i^N})]$$

where g^{*N} is the alignment partition.

The value $\max_{1 \le i \le N} K_i^g(\bar{l}_i^{g_i^N})$ determines the maximum cluster CM, and the value $\min_{1 \le i \le N} K_i^g(\bar{l}_i^{g_i^N})$ determines the minimum cluster CM in the given partition. We say that a partition is *alignment* if, with such a partitioning, all data management clusters have a CM as close to a value as possible. In other words, the difference between the CM of social network clusters in an alignment partition (briefly by the A-partition) is negligible for the conditions of the problems under consideration.

The above principle reflects a partitioning rule at which the clusters complexities are very close. The problem of cluster alignment is related to the problem of a graph partitioning and is an NP-hardness [19].

Problem 1. Formation of A-partitions for each type of partitioning separately.

Suppose that for each *m* there are restrictions on the number of clusters of the network, $N_{\min}^m \le N^m \le N_{\max}^m$. The problem of determining N^m and the suboptimal A-partition has the form:

$$\min_{N_{\min}^m \le N^m \le N_{\max}^m} [K_0(\bar{g}^m) + N^m R^{\bar{g}^m}] \qquad (1)$$

where $R^{\bar{g}^m} = \min_{g \in G^N} \max_{1 \le i \le N} K_i^g(\bar{l}_i^{g_i^N})$. In essence, it is necessary to define a partition ensuring the minimum value of the CM for the A-partitions. We note that the optimal number N^m of managers is determined by (1).

Thus, the problem of determining the optimal solution consists in calculating the optimal number of manager, and optimizing the boundaries (distribution to *alignment* subnets) for separate partition type (semantic). The structure of the problem (2), (3) allows us to decompose it into 2 subproblems in order to find a solution approximating to the optimum:

- the number of manager determination,
- defining the boundaries of *alignment* clusters at network partition.

Their consistent solution allows us to approach the optimum (the method of such decomposition we consider below).

The following statements of the problem are also possible. The search is not for an exact optimal solution, but for an approximate one, with a weakening of the requirements for the equilibrium of the partition. The concept of an absolute alignment indicator for a given partition g^m is introduced:

$$\Delta^{g^m} = \max_{1 \le i \le N^m} K_i^{g^m}(\bar{l}_i^m) - \min_{1 \le i \le N^m} K_i^{g^m}(\bar{l}_i^m).$$

The smaller the values of these indicators the more balanced is the partition. Setting an acceptable indicator of alignment Δ^*, it is possible to limit the number of options to be considered $\Delta^* \ge \Delta^{g^m}$.

2.1 Extension 1 of the Data Partitioning model

Suppose that for each node and edge of the network *S* there are given the indexes of complexity corresponding to each manager. This means that the indexes of complexity are corresponding to the type of network partition. We denote l_{ij}^m the complexity indices of the edge (*i, j*) in the given network for the *m*-th type of partitions. Note $l_{ij}^m = l_{ji}^m$. In the case node *i* is not connected to node *j*, we supplement the network with an edge (*i, j*) of zero complexity. It means $l_{ij}^m = 0$. The complexity of node *i* is defined as $w_i^m = l_{ii}^m$ i.e. the

complexity of the node is given by the complexity of the edge (loop) (*i, i*). Here $m = 1$ or $m = 2$.

We denote $Q^1_{k^1}$ the set of edges (*i, j*) and nodes *i* of the cluster with the number k^1 for the first type of partition and the cluster $Q^2_{k^2}$ with the number k^2 for the second type of partition. The numbers k^1 and k^2 correspond to the manager numbers of the clusters of the social network of partitions. We define the complexity of informational management of clusters of a network $L^{k^1}(Q^1_{k^1}) = \sum_{(i,j)\in Q^1_{k^1}} l^1_{ij}$, and $L^{k^2}(Q^2_{k^2}) = \sum_{(i,j)\in Q^2_{k^2}} l^2_{ij}$, as well as the complexity of coordinating work with a first type cluster k^1 with a second type of cluster k^2

$$Z^1_{k^1}(Q^1_{k^1}, Q^2_{k^2}) = (\sum_{(i,j)\in Q^1_{k^1}\cap Q^2_{k^2}} z^{k^1k^2}_{ij})(\sum_{(i,j)\in Q^2_{k^2}\setminus Q^1_{k^1}} l^2_{ij}),$$

where $z^{k^1k^2}_{ij}$ the additional complexity of coordinating actions on the edge (*i, j*) by the cluster's governing body of the first type with the cluster's governing body of the second type. Thus, the CM $K^{g^1}_{k^1}$ by the cluster k^1 is equal to $W^{k^1}(Q^1_{k^1}, Q^2_{k^2}) = L(Q^1_{k^1}) + Z^{k^1}(Q^1_{k^1}, Q^2_{k^2})$.

Problem 2. Clusters systems formation of equal complexity on the data management network with matching costs: finding the g^{1*} and g^{2*} partitions that are

$$W^{*k^1}(g^{1^*}, g^{2^*}) = \min_{g^1,g^2\in G} \max_{1\le k^1,k^2\le N} W^{k^1}(Q^1_{k^1}, Q^2_{k^2}).$$

Here $W^{*k^1}(g^{1^*}, g^{2^*})$ determines the equal complexity of the clusters in the first type partitioning, taking into account the costs of reconciling with the clusters of the partitioning of the second type. Here *G* denotes a given set of admissible partitions of the first and second types which determines the variety of the sets $Q^1_{k^1}, Q^2_{k^2}$.

Similarly, the other problem is to determine the clusters of equal complexity of the second type.

Problem 3. *The clusters boundaries harmonization of different complexity alignment partition types*. Determine the conditions under which the boundaries of the clusters of equal complexity of the first and second types coincide. This means $Q^1_k = Q^2_k$ for $k = 1, ..., N$.

The Maximum Coordination Condition of the Clusters Boundaries.

Statement. If for any pair of partitions of different types $g^1 = \{g^1_i\}$, $g^2 = \{g^2_i\}$ the conditions $z^{k^1k^2}_{ij} \ge \max(l^1_{ij}, l^2_{ij})$ for all $i, j = 1, ..., n$ are fulfilled then the partitioning into clusters of different types is the same. This means that $Q^1_k = Q^2_k$ for $k = 1, ..., N$. The coincidence of the partitions corresponds to their maximum consistency.

This statement corresponds to the condition of "strong fines" for deviating the state from the plan which is considered in the works on the theory of active systems [12]. In [21] this problem was investigated in the case when it is not necessary to fulfill the condition $z^{k^1k^2}_{ij} \ge \max(l^1_{ij}, l^2_{ij})$ but the 1-st manager should agree on its choice and take into account the interests of the 2-nd manager.

Theorem [21]. A sufficient condition for maximum consistency is the fulfillment of the "triangle inequality" for the complexity functions

$$Z^2(\hat{Q}^2, \hat{Q}^1) \le Z^2(\hat{Q}^2, \hat{Q}) + Z^2(\hat{Q}, \hat{Q}^1).$$

2.2 Extension 2 of the Data Partitioning model

In the model described above, we assume that the definition of the complexity of the edges and nodes given by the matrix $L^m = \left\| l^m_{ij} \right\|_n$ is unique. In practice, one can face the fact that for different manager the representations of the complexities matrix $L^m_{k^m} = \left\| l^m_{ijk^m} \right\|_n$ are different. This means that for each manager its own matrix is defined.

Suppose that the network is divided into clusters by the Direction on the basis of its information about the CM of the nodes and edges. Let us consider a simpler case where the costs of coordinating the interaction of clusters of different partitions types are absent or so large (see the assertion stated above) that divisions of different types obviously coincide. This means that problem 1 of partitioning into clusters of equal complexity for each of the partitioning types is solved independently. In this case, for both center, the partition problem is identical. Therefore, we can consider the task of splitting network into clusters for each center independently.

The main difference between the models is that the Direction does not know exactly the magnitude of the CM of nodes and edges for each type of partition. Suppose that each manager knows its CM. This means that, in fact, the elements of the matrix are $L^m_{k^m} = \left\| l^m_{ijk^m} \right\|_n$. Therefore, Direction requests from each manager the relevant information. Denote $V^m_{k^m} = \left\| v^m_{ijk^m} \right\|_n$ the information reported by the manager k^m. There are two possible cases. In the first case, all manager as well as Direction are interested in a partition of clusters of equal complexity and, therefore, are interested in reporting reliable information about their matrix $L^m_{k^m} = \left\| l^m_{ijk^m} \right\|_n$. Then there arises

Problem 4. To construct clusters of equal complexity for the case of different manager views of the CM expressed by matrices $L^m_{k^m} = \left\| l^m_{ijk^m} \right\|_n$.

In the second case, we will assume that each manager has its own interests in including certain edges and nodes of the network in its cluster. We will describe these interests for each manager k[m] using the matrix of benefits $F^m_{k^m} = \left\| f^m_{ijk^m} \right\|_n$. Let us suppose that the benefit matrix is known to the Direction. Let us suppose also that the Direction has a compensation fund of *B*, which he can use to encourage those managers for which "unprofitable" clusters of the social network are formed. The benefit of a cluster is determined by the sum of the benefits of the edges and nodes that are included in the cluster determined by the matrix $F^m_{k^m}$. Let's designate

$$P^{k^m}(Q^m_{k^m}) = \sum_{(i,j)\in Q^m_{k^m}} f^m_{ijk^m} + B^{k^m}$$

the benefit of the manager k^m where B^{k^m} is the amount of compensation for the manager k^m. Then there arises

Problem 5. To determine the minimum compensation fund B and its distribution among all the manager in which they are profitable to report reliable information, $V^m_{k^m} = \left\| v^m_{ijk^m} \right\|_n = L^m_{k^m} = \left\| l^m_{ijk^m} \right\|_n$.

This problem is closely connected with the problems of optimal control mechanisms synthesis in organizational systems [10-12]. In it, the procedure for forming the boundaries of clusters of a network and the distribution of a compensation fund can be linked to the construction of coordinated planning and incentive mechanisms. Let's consider some approaches to solving the presented problems with the exception of problem 5.

3. THE NUMBER OF NETWORK CLUSTERS ESTIMATION

As mentioned above, problem 1 can be decomposed into two subproblems. One of them is the estimation of the number of clusters of the network. First, assume that the CM of the entire network is equal to the sum of the CM of its subnets. Suppose also that we have A-partition, and the CM of the different organizational structures are approximately equal:

$$\min_{g^m \in G^m} \max_{1\le i\le N^m} K_i^{g^m}(\bar{l}_i^m) = R^{\bar{g}^m}.$$

Then we assume that the CM of all manager is the same and equal to the total CM of the entire network $\tilde{L} = NR^{\bar{g}^m}$.

The center spends time and money on monitoring the work of each manager. Therefore, one of the components of the CM center - $a_1 N$ is proportional to the number of subordinate organizational structures - clusters of the network. In addition, the center coordinates the interaction of the manager pairs. Therefore, another component of the its center can be estimated by a quadratic function $a_2 N^2$. Coefficients a_1 and a_2 characterize for example the time spent managing each manager and coordinating their interactions. Thus, the CM center is estimated by the sum $K_0(\bar{g}^N) = a_1 N + a_2 N^2$. Then problem (6) can be represented as a minimization with respect to N of the expression $K_0(\bar{g}^N) + NR^{\bar{g}} = a_1 N + a_2 N^2 + \tilde{L}$.

Let us consider the problem of minimizing costs (4). Suppose that the costs of the equisyllabic clusters of the network are described using an incremental cost function $\tilde{Z}(\tilde{L}/N)$. Then the problem of determining the conditionally optimal number of clusters of the network for the m-th center can be represented as minimization of $K_0(\bar{g}^N) + NR^{\bar{g}} = a_1 N + a_2 N^2 + N\tilde{Z}(\tilde{L}/N)$.

Let us approximate the cost function in the form of a quadratic function, namely:

$$\tilde{Z}(\tilde{L}/N) = b_1 \tilde{L}/N + b_2 \tilde{L}^2/N^2.$$

Then from the condition it is possible to determine the estimate of the optimal number of clusters N^*. From the necessary conditions of the extremum, we obtain the equation for estimating N*:$a_1 + 2a_2 N = b_2 L^2/N^2$ under the condition $N_{min} \le N \le N_{max}$.

4. NETWORK COMPRESSION

By compression (reduction) of a network, we call the transformation of the initial network to a simpler one, with a smaller number of edges and nodes, due to

- the union of some edges and nodes;
- a priory binding of individual edges and nodes to certain centers;
- restrictions on the ability to bind individual edges and nodes to certain centers. In this case, the centers (and, respectively, the clusters of the network) are defined, only to which one or another node or edge can be assigned in the process of forming the boundaries.

Let's single out two modes of network compression. The first mode determines the initial compression (reduction) of the network in order to reduce the size of the task and to take into account non-formalized factors that impose restrictions on the formation of clusters of the network. This compression is based on the analysis of a network specifics taking into account its technological features of information transfer, points of data "origin and repayment", of information impacts (origination and receipt of messages), types of security (data format), taking into account the technological interdependence of individual sections of the network.

The second compression mode used in the typical step in the algorithms of data analyzing on the network offered below, and the successive formation of the network clusters. Compression and the typical step of these algorithms are described by the following procedure. For ease of writing, this section omits an index characterizing the type of partition. This means that network transformations considered for the case when there is a single center. In the case of two center, we can add an index corresponding to the number of the center.

After carrying out the first compression mode, renumber the nodes of the received network so that the first N numbers receive the selected nodes (manager), *i*

= 1,...,*N*, ...,*n*. Suppose that for the complexity l_{ij} of the edge (*i, j*) is true $l_{ij} = l_{ji}$. The complexity of the *i*-th node is defined as $w_i = l_{ii}$. We denote $L^0 = \left\| l_{ij}^0 \right\|_n$ the initial matrix of edges and nodes of the network under consideration. Here, the superscript denotes the step number in consecutive network compression. Note that this matrix is symmetric, has dimension *n*, and its elements take non-negative values. In the case, the node *i* is not connected to the node *j* by an edge in the network under consideration, we supplement the network with an edge (*i, j*) of zero complexity. It means $l_{ij} = l_j = 0$. Suppose that the *N* selected nodes are not joined by edges of nonzero length.

We represent the formation of social network clusters as a consecutive assignment of edges and nodes to one or another selected node, which is the manager of the cluster, and the formation of a new network with a smaller number of nodes per unit (network compression). This transforms the matrix $L^{n-N} = \left\| l_{ij}^{n-N} \right\|_N$ into *n*–1 matrix $L^1 = \left\| l_{ij}^1 \right\|_{n-1}$ at the first step, and at the second step $L^2 = \left\| l_{ij}^2 \right\|_{n-2}$ in dimension n–2 and so on, until we obtain a matrix $L^{n-N} = \left\| l_{ij}^{n-N} \right\|_N$ of dimension *N* at the step (*n–N*). At the compression step, it is allowed to attach only one node and possibly several edges incident to the attached node.

Let us consider the first step of compression. Let the unselected node with the number *j* (*j* > *N*) connected to the edge (*i, j*) be attached to the selected node with the number *i* (*i* ≤ *N*), and $l_{ij} > 0$. Then the transformation of the complexities of nodes and edges of the network will be determined by the following relations $w_i^1 = w_i + w_j + l_{ij} + l_{jk}$, $l_{jk}^1 = 0$ where *k* is the number of the unseparated node such that $l_{ik} > 0$. Similar to the first step, the following compression steps are performed. The complexity recalculation formulas at the step τ have the form $w_i^{\tau+1} = w_i^\tau + w_j^\tau + l_{ij}^\tau + l_{jk}^\tau$ where τ = 1, ..., *n–N*.

Note that the compression formulas reflect the linear transformation the matrix $L^\tau = \left\| l_{ij}^\tau \right\|_{n-\tau}$ into the matrix $L^{\tau+1} = \left\| l_{ij}^{\tau+1} \right\|_{n-\tau-1}$. Thus, the compression step τ can be represented as $L^{\tau+1} = B^\tau L^\tau B^{\tau T}$ where B^τ is the transformation matrix at the τ-th step of dimension *n–*τ on n–τ–1, $B^{\tau T}$ is its transposed matrix. The transformation under consideration translates a symmetric matrix $L^\tau = \left\| l_{ij}^\tau \right\|_{n-\tau}$ into a symmetric matrix $L^{\tau+1} = \left\| l_{ij}^{\tau+1} \right\|_{n-\tau-1}$ in which there is no *j*-th row and column in the original numbering of rows and columns.

Thus, as a result of the entire sequence of steps described, we can write the final reduction of the original matrix L^0 to L^{n-N} in the form $L^{n-N} = BL^0B^T$ where $B = B^{n-N}B^{n-N-1}...B^1$ and $B^T = (B^{n-N})^T(B^{n-N-1})^T(B^1)^T$. As a result, by construction we obtain a diagonal matrix L^{n-N}, and the quantities on the diagonal set the values of the information management complexity indicators of the constructed network clusters $K_i^{g^N} = w_i^* = w_i^{n-N}$ where *i* = 1, ..., *N*. The described reduction procedure can be used in the basic step in algorithms for analyzing data on a network and locally optimal partitioning based on the directional construction of clusters of a network.

We can apply the described procedure for sequential network reduction to problem 1. It corresponds to obtaining the smallest difference in the values of the diagonal elements $K_i^{g^N} = w_i^* = w_i^{n-N}$ of the matrix L^{n-N} when choosing the partition.

This conclusion has the following geometric interpretation. As is known, symmetric matrices L^0,…, L^{n-N} correspond to quadratic forms. Since the elements of the matrices are nonnegative, these forms define ellipsoids in spaces of appropriate dimension. The reduction transformation at each step is the projection of this ellipsoid into a space having a dimension less than one. Eventually, a sequence of such projections forms an ellipsoid in a space of dimension *N*. An ellipsoid in a space of dimension *N* has a canonical form. With this interpretation, the problem of equal complexity consists in choosing such transformations that form this ellipsoid as close to the ball as possible.

Note. Generally speaking, formulas for the complexity of vertices recalculating can be non-linear, i.e. instead of the linear formula, can be:

$$w_i^{m+1} = Z_i(w_i^m + w_j^m + l_{ij}^m + l_{jk}^m)$$

where $Z_i(.)$ is a convex, nondecreasing function that determines, for example, management costs. In this case obviously, the geometric interpretation of the reduction is substantially more complicated.

5. HEURISTICS ALGORITHMS OF ALIGNMENT DATA MANAGEMENT CLUSTERS DETERMINING

Construction of heuristic algorithms for analyzing social network data is a local optimization. Some initial partitioning is determined, and then procedures for its sequential improvement are directed search of options and sequential expansion of subnets, until we get a complete network partition. The process of such an expansion is directed to improve at each step the indicator of the equilibrium of social network clusters.

Partitioning Algorithm "Nearest Centers of the Network". Consider the method of forming network clusters for the case of different matrices $L_{k^m}^m = \left\| l_{ijk^m}^m \right\|_n$. Let the net already reduced at some step τ be given. We get the CM of the reduced distinguished nodes $w_\theta = w_{k^\theta}^\tau$, where θ = 1, ..., *N* (we shall omit the reduction (reduction) order in the notation, since it is not important in the description of this algorithm). In the calculation of the CM, we use a representation of the matrix of complexities $L_{k^\theta}^\theta = \left\| l_{ijk^\theta}^\theta \right\|_n$ related to the θ manager.

Step of algorithm. We define the shortest distances between all manager (distinguished nodes) between the *s*-th and *t*-th manager of the reduced network in two variants, using matrices $L_s = \left\| l_{ijs}^s \right\|_n$ and $L_t = \left\| l_{ijt}^t \right\|_n$ respectively *s*, *t* =1, ...,*N*. If the shortest distance between centers is 0 it means that at the appropriate point the clusters are neighboring. This fact is fixed, and the edge of zero length is excluded from further consideration in the algorithm. We denote the shortest distances $\tilde{\lambda}_{st} > 0$ and $\tilde{\lambda}_{ts} > 0$ between the *s*-th and *t*-th, as well as the *t*-th and *s*-th the centers of the reduced network. Note that, generally speaking, $\tilde{\lambda}_{st} \neq \tilde{\lambda}_{ts}$ by force $L_s \neq L_t$. Let us determine the minimum distance between all pairs of centers $\tilde{\lambda}_{j^* i^*} = \min_{j \neq i} \tilde{\lambda}_{ji}$. Let it be a couple with numbers j^*, i^*. We will compare CM by clusters of a network corresponding to these reduced centers w_{j^*}, and w_{i^*}. Let $w_{j^*} > w_{i^*}$. Then in the reduction of the node *i**, we add an edge incident to the node *i** along the considered shortest path, and also a node connected by this edge to the center *i**. After this, we recalculate the CM of the corresponding clusters of the network. Then we again compare the CM w_{j^*} and w_{i^*} then add the edge and the node to the reduction of the center where the CM was smaller. In the case of equality of CM, we arbitrarily choose one of the centers. As a result of the described reduction along the shortest path, we obtain the distance between the centers *j**,*i** equal to zero. This zero edge is excluded from consideration.

The data analysis algorithm is completed when, after the next reduction, there are no shortest distances of non-zero length. The final reduction determines the partitioning into social network clusters.

Partitioning Algorithm "Nearest Network Boundary".

Step of algorithm. We define *t* such that $w_t = \min_{1 \le j \le N} w_j$. Let us consider the reduction of the allocated manager. This reduction is a subnet that is reduced ("compressed") to the reduced center t. We use the representation $L_t = \left\| l_{ijt}^t \right\|_n$ of the manager *t* for the network subnet, and we define the minimum "radius" which is the shortest path from manager to the "periphery". The boundary of the network is determined by the nodes with which the edges that are not part of the reduction in question are incident. To the node of the boundary corresponding to the minimal radius, we add an edge incident to this node. We add this edge and the associated node to the reduction of the *t*-th manager. We carry out this addition only from the number of edges not included in the reduction of other nodes. If there are several such edges then the selection rule from these edges establishes a modification of the considered algorithm. This completes the algorithm step. We pass again to the beginning of the described step. If we do not succeed in adding an edge (since the neighboring edge is in the reduction of another manager) we believe that a point of contact between neighboring clusters of the network has been found. This point is excluded from the boundary points to which the radius is calculated. The algorithm ends when all edges that are not included in any reductions are exhausted.

Partitioning Algorithm "Reducing Networks Order". At each step of this algorithm, reduction (compression) of the subnet with minimal $w = w_j$ is considered. In this reduction, we add an edge and the corresponding node which does not change the order of reduction of the distinguished manager considered, after which we recalculate $w = w_j$. If in the reduction under consideration it is not possible to find an edge that does not change the order of reduction then we go on to consider the reduction of another manager with the next largest increase *w*. If such manager was not found, then we increase the reduction order of the manager with minimum *w*. And so on until the entire network is broken down into clusters.

6. SELECTIVE EXAMINATION OF DESIGN IN NETWORK PROJECTS

During the examination of large and complex transport projects in Russia such as the development of the Moscow-Kazan high-speed railway project and the project of reconstruction the eastern section of the Baikal-Amur Railway, we faced to deal with big data received from various executors of project sections. The task was complicated by the fact that too short deadlines and limited funding were set. This made it impossible to attract the required number of qualified experts to the task. Under these conditions, there were problems of selective express analysis of big data, distribution of jobs, and payment for experts. It is necessary to take into account the interests of experts in the jobs. We propose a model and a procedure for these problems here.

These problems are closely related to the concept of the new industrial revolution considering the inclusion of operations with big data, the distribution of information between agents connected by a network structure in the production [2].

Similar problems were partially set and discussed in [22]. Close problems were considered also in a number of articles devoted to the automatic identification of valuable knowledge among scientific research for example [23,24] These papers also provide an overview of research on the issue and relevant links to more comprehensive reviews. We find similar aspects of the problem in [25].

Unlike some of these studies, we consider the problem of finding flaws, collisions and errors in big data. In addition, our target is to build a procedure for finding a solution in the context of lack of time for a detailed semantic data analysis, and taking into account the interests of the involved experts. We use some approaches from organizational management [12] as well.

In many applications, data distribution can be described by means of a network structure. For processing data in network structures, it is often useful to use the concept of complexity alignment [5,6].

Our approach is on selective processing of big data by means of data clusters complexity alignment. We

propose a mechanism for forming basic information clusters and allocating them among a limited number of experts. This takes into account the alignment of clusters complexity and the distribution of semantics between them.

Let $I=\{1,\ldots,n\}$ is the number set of all files concerning a complex project design, p_i – the volume of the i-th file (taking into account its complexity), l_{ij} – the data communication amount (the complexity) of the i-th file with the j-th file, $p_i \geq 0$, $l_{ij} \geq 0$, $l_{ii}=0$.

Consider m-th data cluster $I_m=\{i_j\}_m$ as some collection of n_m files i_j from the set $I=\{1,\ldots,n\}$ where $j=1,\ldots,n_m$. Let us $A_k=\{I_m\}_k$ is a collection of N clusters, $m=1,\ldots,N$. Here N is a fixed clusters number in the collection k. We denote by $A=\{A_m\}$ the set of admissible sets $A_k=\{I_m\}_k$ of N clusters collections where clusters non-intersect with each other, $N<n$,

$$I_m \bigcap I_j = \varnothing \text{ when } I_m \in A_k,\ I_j \in A_k,\ m \neq j\ . \quad (2)$$

We define the volume (complexity) of the m-th cluster as the sum of the volumes of the files included in it together with the volumes of their interrelationships and connections with external files

$$P_m = P_m(I_m) = \sum_{i \in I_m} p_i + \sum_{i \in I_m} \sum_{j \in I} l_{ij}\ . \quad (3)$$

Denote C_0 the initial cost (price) of the project, and c_i^0 – the initial cost of implementing the project share relating to the file i,

$$c_i^0 \geq 0,\ i \in I,\quad C^0 = \sum_{i \in I} c_i^0 \quad (4)$$

Let us assume:

- the amount of funds H allocated for the examination of the integrated project be provided;
- N experts are involved in the project's design examination, and the number of experts equals the number of clusters in $\{I_m\}$;
- d_i is the examination price for each i-th file.

Then the cost the i-th file processing is $d_i p_i$. Denote d_{ij} the price of the cross-reference analysis. Then the cost of working with the information cluster I_m is

$$h_m(I_m) = \sum_{i \in I_m} d_i p_i + \sum_{i \in I_m} \sum_{j \in I} d_{ij} l_{ij} \quad (5)$$

The following restriction holds is

$$\sum_{m \in A_k} h_m(I_m) \leq H \quad (6)$$

for any $A_k \in A$.

Note that this restriction implies that each expert is working on one data cluster.

Suppose that when working with a file i the expert j can reduce the cost of the project part associated with the file i by the amount $(1-\gamma ij)c_i^0$ where γ_{ij} the specified value, $0 \leq \gamma_{ij} \leq 1$. The value $(1-\gamma_{ij})c_i^0$ characterizes the "economy" as a result of the examination of the i-th file by expert j, and the value γijci0 corresponds to the resulting value of the project share associated with the file i after the examination by j-th expert. The parameter γij will be presented further in the form $\gamma_{ij}=1-(1-\gamma_i^0)v_j$ where γ_i^0 characterizes the level to which it is possible to reduce the cost of implementing the share of the project determined by the file i , v_j – a parameter characterizing the qualification of the expert, $0 \leq \gamma_i 0 \leq 1$, $0 \leq v_j \leq 1$.

Thus, we can reduce the cost of the project after the cluster collection $A_k=\{I_m\}_k$ examination on a value

$$\Delta(\{A_k\}) = \sum_{I_m \in A_k} \sum_{i \in I_m} v_m (1-\gamma_i^0) c_i^0\ .$$

The problem of forming an optimal system A_k consisting of N data clusters I_m selected for the examination we formulate as follows

$$\Delta(\{A_k^*\}) = \max_{A_k \in A} \Delta(\{A_k\}) \quad (7)$$

subject to the restriction (6).

We will supplement the model under consideration with the condition of taking into account the interests of the experts involved. Suppose that when determining the scope of job for each expert an appropriate salary must be established, and his job is accompanied by cost.

We present this fact in the form of the expert's objective function $w_m=h_m(P_m)-z_m(P_m,v_m)$. Here $h_m(P_m)$ is the amount of payment depending on the amount P_m of the data cluster I_m examination work defined by equation (5), z_m (P_m,v_m) is the expert cost function depending on the job volume P_m and the parameter v_m characterizing the expert's qualification. We will assume that the amount of work performed by an expert must satisfy the profitability condition for an expert

$$h_m(P_m) - z_m(P_m, v_m) \to \max_{P_m} \quad (8)$$

subject to the constraint (6).

The problem (7) under the conditions of taking into account the experts' goals (5) and restriction (8) were studied in game theory with the right of the first move of a selected player (here of the expert organizer) [12,15,16].

Clusters selection of equal complexity (Cluster Complexity Alignment). In the particular case $d_i=d_{ij}=d$ and the cost function has the form $z_m(P_m,v_m) = uP^2/(2v_m)$ where u is the normalization coefficient, we obtain the optimal volume of job for the cluster with the number m as a result of solving the problem defined by expressions (6) and (8) as follows

$$P_m^0 = Hv_m / (Vud)\ . \quad (9)$$

where $V = \sum_{j=1}^{N} v_j$.

This means that the optimal scope of the project clusters (the job complexity) is the same for all clusters of work in relation to the same value of the expert's skill level.

On this property we will base the following "principle of cluster complexity alignment", which is formulated as follows:

- clusters that combine files that are subject to analysis by specialists of the same qualification should have equal complexity;

• if the specialists have different qualifications, then the complexity of the clusters analyzed by them should be proportional to the productivity of these specialists.

This principle reflects condition (8) of the clusters utility for experts with a certain degree of approximation.

Problem statement: Determine $A_k^* = \{I_m^*\}_k$, $P_m^* = P_m(I_m^*)$, $h_m(I_m^*)$ such as (2), (6), (7) take place subject to the condition of "equal complexity"

$$P_m^* = P_m\left(I_m^*\right) \le P_m^0 \tag{10}$$

This problem comes down to the essentially well-known Multiple Knapsack Problem which is the NP-complete combinatorial optimization. They consider algorithms for this problem in [26].

If we do not adhere to strict formalism in applied problems, namely, to allow the variability of the right-hand parts of the constraints since parameters v_m are usually determined by expert means then it is reasonable to solve the problem by approximate methods, for example "greedy algorithm".

Note that for cases where the cost of processing files in the cluster does not exceed the cost of analyzing cross-references or the amount l_{ij} of cross-reference between documents significantly increases the volume (complexity) of the clusters which means

$$\sum_{i\in I_m} d_i p_i \le \sum_{i\in I_m}\sum_{j\in \mathrm{I}} d_{ij} l_{ij} \,. \tag{11}$$

It is advisable to create clusters containing as few of these references as possible at this cases.

In practice in a number of examples of project examinations, in particular, when inequality (11) takes place, and it is not necessary to obtain an exact solution of the problem formulated by conditions (7), (10). You can use the heuristic algorithm described below if it is sufficient to obtain an acceptable approximate solution.

To simplify the description of this algorithm, we will assume that the payment of each expert is sufficient, for processing, at least one most expensive file. Refusal of this assumption does not change the essence of the description of the algorithm, but leads to the inclusion of additional conditions.

The Heuristic Algorithm. We describe the heuristic algorithm for solving the problem formulated by conditions (7), (10).

We will use the greedy algorithm approaches. We number clusters (at the first stage they are empty) in order of decreasing skill indicators $v_j P_m^0$ of experts, j=1,...,N. Let's designate this order N^0. We calculate values

$$q_i = \frac{p_i + \sum\limits_{k\in I,\ i\ne k} l_{ik}}{h_i(p_i)}$$

for all files where

$$h_i(p_i) = d_i p_i + \sum_{k\in I,\ i\ne k} d_{ik} l_{ik}\ ,\ i=1,...,n.$$

Sort the quantities q_i in the order of their decrease. We form the order of file numbering n^0 in accordance with the resulting order of decreasing values q_i. Sequentially we include in N empty clusters ordered by one file of order. Thus, one file is included in the generated clusters. There are $(n - N)$ unallocated files. We denote I^1 the set of their numbers. Calculate for each cluster the remaining free volumes file

$$P_m^1 = P_m^0 - (p_m + \sum_{k\in I,\ m\ne k} l_{mk}) =$$

$$= \frac{Hv_m}{(Vud)} - (p_m + \sum_{k\in I,\ m\ne k} l_{mk}).$$

We proceed to the next (first) step of the algorithm. We number the set of clusters in order of quantities $v_m P_m^1$ decreasing. Thus, we obtain a new order N^1 of the clusters. In sequence of clusters N^1 we add the next files of order I^1 if

$$P_m^1 \ge g_m + \sum_{k\in I,\ m\ne k} l_{mk} \tag{12}$$

Otherwise, go to the next cluster and perform the above procedure to add the file under inequality (12). As a result, of this step we add a certain number k^1 of files into the clusters. We determine the set I^2 of unallocated files remaining in the number $n - N - k^1$. We repeat in the same way at the next iteration as at the first one until inequality (12) is satisfied for any cluster. This completes the algorithm.

7. APPLICATIONS

Developed in section 6 an approach based on cluster technologies and algorithms for leveling the complexity of the analysis of the semantic data of a large document with cross references was used in the examination (technological and price audit) of the construction projects for the high-speed Moscow-Kazan railway, 803 km long, reconstruction of the Baikal-Amur Mainline 4287 km long and other large-scale projects for the development of rail transport. Consider the application of this approach to the example of technological audit of the high-speed railway project "Moscow-Kazan". We adopted the following regulations during the audit.

At the first stage, the management of the audit work carried out a semantic analysis of the big data about the project under consideration. As a result, the necessity of examination of 453 documents, distributed according to the following sections of the project, was identified:

• the main railway tracks (the results of geological, hydrological, engineering and environmental studies, linking the route to the terrain and earthworks, the structure of the lower and upper layers of the railway, etc.) – 116 documents;
• bridge and other engineering structures – 45 documents;
• rolling stock (locomotives, wagons, trains, their maintenance and repair, requirements for wheel-rail interaction, etc.) – 97 documents;
• energy supply (contact networks, electric power substations, transformers, etc.) – 43 documents;
• organization and management of traffic (telecommunications, automation and telemechanics, etc.) – 48 documents;

• multimodal transport and transfer units – 39 documents;
• security (technical, environmental, etc.) – 24 documents;
• maintenance of infrastructure (management of failures, maintenance, etc.) – 41 documents.

To audit these sections of the project, 21 experts were involved, among which 453 documents were distributed. Basically, each expert had to analyze from 17 to 23 documents.

At the second stage, each expert conducted an analysis of the documents assigned to him, in order to identify 4-7 priority documents to be audited. The results of this analysis were agreed with the head of audit, and a list of 127 priority documents to be examined was formed. This list was brought to all experts. At the same time, the a priori difficulties in the analysis of each of the 127 priority documents were identified.

In the third stage, each expert conducted a semantic analysis of the links of each priority document assigned to it with priority documents considered by other 20 experts. The need to carry out this work was due, for example, to the fact that the work on the examination of the project of the main railways is connected with the work on the examination of bridge and other engineering structures, rolling stock, multimodal transport and transfer units, etc. The work on the expertise of the power supply and safety subsystems, to a greater or lesser extent, with work on the examination of all other sections, etc.

Based on the results of the analysis of semantic links, experts identified references to priority documents analyzed by their colleagues. The number of links usually did not exceed 4, and in some cases such references were absent altogether. Based on these references, the experts formed or "two-dimensional" relational tables reflecting the relationship of their work with audit work conducted by other experts. These "flat" tables corresponded to the experts' views on the dependence of their conclusions on data from other priority documents. At the intersection of the row and column of the relational table, each expert indicated his assessment of the complexity of the work on the analysis of the relevant part of the priority document referred to. This complexity was linked to the volume and variety of information contained in this part of the document.

After agreeing the relational tables with the head of the audit, a list of 246 cross-references between 127 priority documents was formed, and the complexities of the work on their analysis was determined. Taking into account the a priori complexities of analyzing priority documents, this made it possible to determine the a posteriori complexities of the expert jobs on the analysis of priority documents, taking into account their interdependence.

At the fourth stage, based on a posteriori works complexities, using the described in Section 2 algorithm, distributed between experts the funds allocated for the audit work (2.9 million rubles), for a limited time (2 months). The optimal solution was the analysis of 86 priority documents with 159 cross-references.

We present the final results of calculations by the proposed method skipping intermediate detailed calculations due to their bulkiness and obviousness.

Based on the results of the audit of these documents with cross-references, the cost of the Moscow-Kazan high-speed railway project was reduced by 2.2 billion rubles.

8. CONCLUSION

At the heart of the semantic technology proposed lies the indicator of the complexity of data mining. This indicator depends not only on the volume and variety of data but also on the interdependence of their fragments characterized by the quantity and quality of the links between these fragments. The complexity of data mining for each problem has its own, specific form.

The algorithms of the proposed technology are based on the principle of equalizing the complexity of data mining, and are heuristic. In this case, the entire data array in network is divided into fragments - clusters, the complexity of extracting data from which are approximately the same. We can't completely align these difficulties but we aim for maximum alignment.

To reduce the amount of search of vertices and edges of the graph reflecting the interdependence of data fragments in the network these algorithms form the most compact clusters. In this case, the compactness of the cluster is equivalent to its simplicity achieved by simplifying the links between data fragments (in particular, by reducing the number of consecutive references). This is the difference between our algorithms and the known partitioning methods [19].

Thus, the novelty of the developed cluster technologies and algorithms based on the equalization of the complexity of the analysis of semantic data is related to the interdependence between different sources of large data. Such sources can be:

- really existing structures, such as social or information networks;
- virtual network structures created by the researcher himself on the basis of an analysis of the entire set of information sources, for example when analyzing a document containing large data (a large document).

Systems, structures and technologies of information management of social networks have their own peculiarities, caused by a large number of related and geographically distributed processes. In this regard, there is a need for a multi-level organizational system of data management.

The necessity of its optimization generates the problem of partitioning social networks into subnets (clusters) - areas of responsibility of regional data management bodies. The paper describes the model and presents the problems of optimizing the number of centers and boundaries of organizational structures of information management on the network. To resolve this problem we introduce the notion of data management complexity. The problem of optimal network partitioning into information management clusters is set as the task of minimizing the complexity of information management due to its equalization. The ways of solution are determined on the basis of its decomposition into subproblems of determining the number of clusters and

their boundaries on the principle of equal complexity separately. Locally optimal partition algorithms for data mining are developed. They are based on a directional search of options including algorithms for the formation of cluster boundaries given the number and location of management structures on the network. The latter include heuristic data analysis algorithms: the nearest-center data analysis algorithm; algorithm for analyzing the data of the nearest boundary; algorithm for data analysis of the order of reduction.

We also propose cluster technologies and optimization algorithms of data mining from such virtual network structure as a large document with interdependent sections (for example, from the description of a large-scale project). The chief expert (for example, the head of audit) does not know all the relationships in advance. It has only a priori information about the block structure of a large document described, for example, in the form of a table of contents. In addition, the document has a large number of sections with low-value irrelevant for audit information. Note the location of these sections is unknown. Thus, there is a problem of dropout - excluding these sections from further consideration. And on the contrary, it is necessary to identify the essential interconnections of the most important sections. After that, the problem arises of distributing sections, by their complexity, between experts. However, the capabilities of each expert in data processing are limited. In this paper, we propose algorithms for solving this problem which have been repeatedly applied in the audit of large-scale projects for the development of rail transport in Russia. In general, in our opinion, the developed semantic technology of data mining is quite universal, and can be used for processing a variety of large data arrays.

REFERENCES

[1] Kagermann, H. et al.: Recommendations for implementing the strategic initiative INDUSTRY 4.0, in: Abschlussbericht des Arbeitskreises Industrie 4.0, DAT, Frankfurt/Main, pp. 5-105, 2013.

[2] Blanchet, M., Rinn, T., Thaden, G. and Thieulloy, G.: *INDUSTRY 4.0 - the new industrial revolution. How Europe will succeed*, Roland Berger Strategy Consultants GMBH, München, 2014.

[3] Bauernhansl, T., Hompel, M., and Vogel-Heuser, B.*: INDUSTRIE 4.0 in produktion, automatisierung und logistik - anwendung, technologie, migration*, Springer, Wiesbaden, 2014.

[4] Mueller, E., Chen, X.-L., Riedel, R.: Challenges and requirements for the application of Industry 4.0: A special insight with the usage of cyber-physical system, Chinese J. of Mechanical Engineering, Vol. 30, No. 5, pp.1050-1057, 2017.

[5] Enaleev, A., Tsyganov, V.: Structures and cluster technologies of data analysis and information management in social networks, Communications in Computer and Information Science, Vol. 754, pp. 683-696, 2017.

[6] Enaleev A., Tsyganov V.: Service support structure optimization of a large-scale rail company, CEUR Workshop Proceedings, Vol. 2098, pp.396-406, 2018.

[7] Ceravolo, P., et al.: Big data semantics, J. of Data Semantics, Vol. 7, pp. 65-75, 2018.

[8] Wu, X., Zhu, X., Wu, G.-Q., et al.: Data mining with big data, IEEE Transactions Knowledge Data Engineering, Vol. 26, No.1, pp. 97–107, 2014.

[9] Lim, E., Chen, H. and Chen, Q.: Business intelligence and analytics: research directions, ACM Transactions on Management Information Systems, Vol.3, No. 4, pp.1-17, 2013.

[10] Appice, A., Ceci, M. and Malerba, D.: Relational data mining in the era of big data. in: Flesca, S., Greco, S., Masciari, E. and Saccà, D. (eds), A comprehensive guide through the Italian database research over the last 25 years, Springer, Berlin, Heidelberg, pp. 323–339, 2018.

[11] Glavic, B.: Big data provenance: challenges and implications for benchmarking, in: Rabl, T., Poess, M., Baru, C. and Jacobsen, H.-A. (Eds*.)*: *Specifying Big Data Benchmarks*, Springer, Berlin, Heidelberg, pp.72–80, 2014.

[12] Burkov, V. et al.: *Mechanism design and management. Mathematical methods for smart organizations*, NOVA Publishers, New York, 2013.

[13] Voronin, A.A., Gubko, M.V., Mishin, S.P. Novikov, D.A.: *Mathematical models of organizations*, Lenand, Moscow, 2008 (in Russian).

[14] Gubko, M.V.: *Mathematical models of optimization of hierarchical structures*, Lenand, Moscow, 2006 (in Russian).

[15] Enaleev, A.: Optimal incentive-compatible mechanisms in active systems, J. Automation and Remote Control, Vol. 74, pp. 491-505, 2013.

[16] Enaleev, A.: Optimal incentive compatible mechanism in a system with several active elements, J Automation and Remote Control, Vol.78, pp. 146-158, 2017.

[17] Rajković, R.Z., Zrnić, N.D., Kirin S.D., Dragović B.M.: A review of multi-objective optimization of container flow using sea and land legs together, FME Transactions, Vol. 44, No. 2, pp. 204-211, 2016.

[18] Prah, K. Štrubelj, G.: Comparison of using different kinds of traffic data in best route analysis based on GIS, FME Transactions, Vol. 46, No. 4, pp. 668-673, 2018.

[19] Buluc¸ A., Meyerhenke, H., Safro, I., Sanders, P. and Schulz, C.: Recent advances in graph partitioning. Preprint, arXiv:1311.3144, 2013.

[20] Modrak, V., Krus, P. and Bednar, S.: Approaches to product variety management assuming configuration conflict problem, FME Transactions, Vol. 43, No. 4, pp. 271-278, 2015.

[21] Enaleev, A.K.: Coordinated partitions in organizational network structures, J. Automation and Remote Control, Vol. 79, No. 2, pp. 337-349, 2018.

[22] Putnik, G.D., Cruz-Cunha, M.M.: *Knowledge and technology management in virtual organizations:*

issues, trends, opportunities and solutions, IGI Global, Hershey, 2007.

[23] Riel, A.: Automatic knowledge extraction from manufacturing research publications, in: CIRP Annals - Manufacturing Technology, Vol. 60, No. 1, pp.477-480, 2011.

[24] Riel, A, Boonyasopon, P.: A. knowledge mining approach to document classification, 2009. https://www.researchgate.net/publication/47526651

[25] Schuh, G., König, C.: Determination of information demand for efficient technology monitoring, in: Proceedings of the 26th Intern association for management of technology conf., ASMET, Wien, pp. 851-865, 2017.

[26] Kellerer, H., Pferschy, U., Pisinger, D.: *Knapsack problems*, Springer, Berlin, Heidelberg, 2004.

УСКЛАЂИВАЊЕ СЛОЖЕНОСТИ КЛАСТЕРА У МРЕЖНИМ СИСТЕМИМА

А. К. Еналеев, Владимир В. Циганов

Овај рад разматра структуре за управљање подацима и кластер технологије у мрежама великих скала. Субоптимални проблеми партиције мреже формулисани су на основу усклађивања индекса сложености. Предлажен је метод за решавање ових проблема, посебно одређивање броја кластера података и њихових граница. Описана је вишестепена итеративна шема за семантичко претраживање података из великих докумената са међузависним секцијама. У првој фази, процењује се „*а-приори*“ комплексност претраживања података из ових секција. Онда, рафинише сеовасложеност узимајући у обзир откривених података из претраживања података из суседних секција. На основу тога, формира се коначна партиција скупа података великог документа на кластере, у околностима рокова и ограничења финансијских средстава. Предложене методе примењене су у неким транспортним пројектима великих скала.