Model of optimum integration of diverse geodata for the benefit of management of forestry

E Istomin, S Stepanov, Y Petrov*, O Kolbina, A Sidorenko

Department of applied information science, Russian State Hydrometeorological University, 79 Voronezhskaya Street, St. Petersburg 192007, Russian Federation

*Corresponding email: yaroslav.petrov025@gmail.com

Abstract. The description of model of optimization of processing of the diverse space distributed data is provided in work. Approach to implementation of consolidation of information from different sources is considered. The control algorithm of diverse data taking into account filtering of fragments of all arrays of information is provided. Key information and technical aspects of application of model of processing of diverse geodata for the solution of tasks of management of forestry are described.

1. Introduction

In many tasks of steady management of the woods, it is necessary to process different geodata of monitoring of forestry quickly. Using the geospatial data collected by means of the remote sensing of Earth (RSE) and land measurements we have the possibility of forecasting of influence of adverse factors for holding actions which minimize damage. In turn receiving the qualitative analysis of forestry and landscape information is connected with need of continuous integration of diverse data among themselves, at the same time there are different problems in the course of integration of diverse data.

The solution of problems of consolidation of diverse geospatial information demands consideration of the problem of development of new model of data integration. Use of different mechanisms of access to data will allow to simplify access to diverse geospatial information. In problem definition of data processing it is possible to allocate the primal problem which is of great importance - development of model of access to the distributed diverse space data.

For reduction of labor input of connection of new data sources, the most optimal solution is the possibility of support of any of widely known interfaces of access to data, such approach will allow to process data not only from the structured sources supporting own language of requests and certain models of data presentation but also from unstructured sources.

Deep consideration of process of integration and interaction of diverse data among themselves, revealed the number of problems, are basic of which:

- Lack of the uniform applied software for processing of heterogeneous geospatial data, or its availability which purpose is the solution of the number of the typified tasks which is adjusted on specific scope of application.
- Lack of accurately formulated algorithms of processing and interrelation of heterogeneous data.
- Partial incompatibility of formats of geospatial data.
- Heterogeneity of geospatial data.
- Incorrect formulation or total absence of metadata - the description of sets of geospatial data.
Processing of diverse information in the uniform system will allow to use different data in total for the made calculations and operations that will promote obtaining results in the form of expanded statistical information for the objective. The model of integration of diverse data needs to be developed taking into account the revealed problems in the studied models and systems.

2. Methods and Materials
The general scheme of information transfer includes the sender of information, the channel of communication and the receiver of information. The main characteristic of channels of communication is their flow capacity. The volume of the transferred information of $V$ is calculated on formula (1):

$$V = qt$$

where $V$ – the volume of the transferred information, $q$ – flow capacity of the channel (in bits a second or similar units), the flow capacity of the channel is equal to amount of information which can be transferred on it in unit of time, and $t$ – information transfer time (formula of the general view 2),

$$t = \frac{V}{q}$$

In some cases availability, both several communication channels, and one is possible.

The first case when there are multiple channels of information processing (3),

$$t_{gen} = \max_{1 \leq i \leq m} \frac{V_i}{q_i}$$

where, $V_i$ – information sources (the specific format), $q_i$ – multiple communication channels, and $t_{gen}$ – the general time of passing of information.

The second case when there is one channel of information processing (4),

$$t_{gen} = \frac{\sum_{i=1}^{m} V_i}{q}$$

where, $V_i$ – information sources (the specific format), $q_i$ – communication channel, and $t_{gen}$ – the general time of passing of information.

The more it is processed data from diverse sources, the longer to become the turn of passing of information via the channel, along with it also data processing time increases.

All volume of diverse data (information array) can be calculated on the following formula (5):

$$V_i = \sum_{j=1}^{n} a_{ij}$$

where, $n$ – writing field length (matrix element), $\sum_{j=1}^{n} a_{ij}$ - the sum of all elements of data array, and $V_i$ — the volume of the transferred information, $n$-order of data source.

For the further research there is no need to use all volume of the actual information of $V_i$. Only required information is necessary for us for the solution of the applied task. In the investigation of what it is necessary to make data filtering.
It is necessary to make multiple operation on the exception of some sequences of data for removal of information. To be set originally the task of determination of sign by which sequences of data, unnecessary for the further research, will be removed.

Mathematical data processing is set by the following condition (6, 7):

\[
t^{*}_{gen} = \frac{\sum_{i=1}^{m}V^{*}_i}{q}
\]  

\[
V^{*}_j = \min \sum_{i=1}^{n}a_{ij}
\]

provided that \( \alpha_{ij} = \beta_{ij} \), where \( V^{*}_i \) – the minimum value on the site, \( \alpha_{ij} \) – signs required information, and \( \beta_{ij} \) – criterion of determination of signs of required information.

Data filtering stages:

1. The condition/criterion of determination of signs of required information - \( \beta_{ij} \). The criterion \( \beta_{ij} \) is set by the operator of the system in advance, proceeding from the established purposes of the applied task. Further, the criterion will be put in the knowledge base for substitution in similar options of processing of diverse data, at the solution of the number of applied tasks.

2. Data filtering is made. Data filtering is carried out on predefined conditions/sign \( \beta_{ij} \).

- if \( \alpha_{ij} \neq \beta_{ij} \), then a condition is considered not executed and the zero identifier is appropriated to elements of the table (as a result of comparison operation separate elements \( \alpha_{ij} = 0 \)), respectively these elements will be removed/are excluded from the table;
- if \( \alpha_{ij} = \beta_{ij} \), then a condition is considered executed and these elements will be defined as required values and will continue further data processing.

Provided that the volume of data will be reduced, depending on the chosen information for the further research, that time of information processing to be reduced. Calculation is made on formula (8):

\[
t = \frac{\sum_{i=1}^{n}V_i}{q}
\]

As a result of filtering, having selected only fields of data tables, necessary for further processing, there is the reduction of volume of the processed information that will allow to reduce considerably time of data processing.

3. Results and Discussion

As a result of the conducted research on reduction of volume of the processed information, the processing algorithm taking into account data filtering ‘Figure 1, 2’ was created.
Figure 1. Algorithm of processing of diverse data, part 1.
Figure 2. Algorithm of processing of diverse data, part 2.
After completion of the algorithm the integrated data become available to the user of the system for the further complex analysis of information.

Architecture of model of integration of diverse data

On the basis of the given algorithm of processing are diverse data, the model can be separated into three main components:

- The scheduler who receives the plan of performance of the request from external system components (the optimizer, or directly from language tools). On the basis of the received plan defines what components of the kernel of the performer are necessary for performance of this request, makes for each of them the task which the subquery for this specific components enters, allocated from the general plan of performance of the request and some meta information.

- Performers of operations of expanded relational algebra. Each performer can support different operations, algorithms of their performance, using different technologies. It should be noted that for addition of new operations, it is only enough to implement the special interface of the performer, thus, the scalability concerning support of new operations and different algorithms of their performance is reached.

- Mechanisms of access to data which, in turn provide the interface for extraction of data from different sources. Thus, the scalability concerning support of new data sources (for extraction of data from new base, or like documents, enough to implement for it the adapter) is reached.

Implementation of the concepts concerning submission of information and operations for their processing requires detailed creation of tools for processing of diverse information. Decomposition of the instrument of data processing on such functional components as the module of requests, the processing module, the module of integration and the kernel of execution of the system – the scheduler ‘Figure 3’ is supposed.

![Figure 3. Scheme of processing system of diverse data](image)

Execution of the request begins when to the scheduler input information arrives. The scheduler creates tasks for each new source and sends data on system modules. Further, there is processing of information according to the received task.

In new architecture the module of requests is interrelation of base of metadata which stores drivers of access to diverse data with the module of integration of diverse data. For ensuring access to databases different technologies, the set of libraries by means of which there is the interaction to the database are used. The procedure connection of the new source is performed by the expert. All further system operation (requests, downloading data in storage) is made taking into account the algorithm of data handling.
The feature of the module of integration consists in data of the known drivers of access to databases to the uniform automated intelligence system of integration of diverse data sources. The main idea of such approach allows to minimize participation of the person in the procedure of integration of diverse data. To each expansion there corresponds the interface of data processing.

The module of processing allows to choose from all available diverse data arrays only fields of the data table, necessary for further processing. Automation of processing of diverse data allows to receive the new database. Selection of data is carried out on the signs which are in advance set by the operator.

As a result of interactive interaction with the system, connection of new data is carried out, the module defines the procedure of processing of diverse data which best of all answers the nature of the specific objective of integration and types of the used data. As a result of integration information is transferred to the canal of data processing, services, necessary for its implementation, are caused and the procedure of integration is performed.

4. Conclusions
The solution of problems of integration of diverse geospatial information for the benefit of management of forestry will demand adaptation of the offered model, both from the point of view of used data, and in terms of functionality of services of integration. The model of processing of diverse geodata will provide to users the following advantages:

- The model describes the optimal solution which allows to support widely known and often used interfaces of access to data, such approach will allow to process data not only from the structured sources supporting own language of requests and certain models of data presentation but also from unstructured sources.
- The model of processing of diverse data, will allow to reduce time of processing of geodata, due to reduction of the processed information by the built-in procedures of search of criteria of required data.
- The model will allow to use effectively access to space geodata in different formats without direct participation of all data.

The offered model will allow to provide support of adoption of management decisions on management of forestry and can become the effective tool for governing bodies of forestry at the different administrative levels. The model is considered as the basis for implementation of the geographic information system supporting operational management of diverse geodata and their analytical processing for support of procedures of decision-making in the field of forestry. The model can be considered as "superstructure" for development of "road map" of steady management of the woods. The general structure of model of integration can be taken as the basis during creation of similar models in other scopes of space data and also will allow to simplify access to diverse geospatial information. It is necessary to develop the specialized software allowing to obtain expanded information and to carry out more careful analysis of the obtained statistical data for use of model of management of diverse data.

References
[1] Castano S, De Antonellis V and De Vimercati March 2001 S.C.D. Global Viewing Of Heterogeneous Data Sources, IEEE Transactions on Knowledge and Data Engineering Vol 13, Issue 2, pp 277-297
[2] Cheng S, Li Y, Tian Z and Cheng W 26 February 2019 A Model For Integrating Heterogeneous Sensory Data In Iot Systems, Computer Networks Vol 150, pp 1-14
[3] Fize J, Roche M and Teisseire M 7 February 2019 Matching Heterogeneous Textual Data Using Spatial Features, IEEE Int Conf. on Data Mining Workshops, ICDMW Volume
[4] Gaddam H K, Rao K R 2019 Correction To: Speed–Density Functional Relationship For Heterogeneous Traffic Data: A Statistical And Theoretical Investigation, Journal of Modern Transportation 27, 1, pp 61-74, 10.1007/s40534-018-0177-7)
[5] Hathaway R J, Bezdek J C, Pedrycz W 1996 A Parametric Model For Fusing Heterogeneous
**Fuzzy Data, IEEE Transactions on Fuzzy Systems** Vol 4 Issue 3 pp 270-281

[6] Istomin E P, Abramov V M 2017 *Knowledge Database In Geoinformation Management Of The Territory Development*, Int Multidisciplinary Scientific GeoConf. Surveying Geology and Mining Ecology Management, SGEM, 17 (21), pp 951-959.

[7] Li Q, Li Y and Gao J 2014 *Resolving Conflicts In Heterogeneous Data By Truth Discovery And Source Reliability Estimation*, Proceedings of the ACM SIGMOD Int Conf. on Management of Data, pp 1187-1198

[8] Samoilenko S and Osei-Bryson 1 March 2019 *K.-M. A Data Analytic Benchmarking Methodology For Discovering Common Causal Structures That Describe Context-Diverse Heterogeneous Groups*, Expert Systems with Applications Vol 117, pp 330-344

[9] Smith R N, Aleksic J and Butano D and other December 2012 *Intermine: A Flexible Data Warehouse System For The Integration And Analysis Of Heterogeneous Biological Data*, Bioinformatics Vol 28, Issue 23, pp 3163-3165

[10] Wang T, Li P and Wang X 15 February 2019 *Comprehensive Survey On Mobile Data Offloading In Heterogeneous Network*, Wireless Networks Vol 25, Issue 2, pp 573-584

[11] Zuech R, Khoshgoftaar T M and Wald R 1 December 2015 *Intrusion Detection And Big Heterogeneous Data: A Survey*, Journal of Big Data Vol 2, Issue 1, article number 3

[12] Kolbina O, Istomin E, and Stepanov S *Review Of Technological Methods Of Access To Distributed GIS Databases* Information technologies and systems: management, economics, transport, law: Coll. tr. Int practical conf. "Infogeo-2014", p 116