ANNOTATION

of dissertation work of Kassymova Dinara on the topic: "Research and development of methods for identifying and eliminating contradictions in big data", presented for the degree of Doctor of Philosophy (PhD) on the specialty 6D070300 - Information Systems

Relevance of research topics. At present, the processing of the data is one of the most urgent problems in many areas of science and engineering. Every two years, the volume of digitization in the world doubles. The volume of information at the moment is 33 zettabytes, and it is expected, that by 2025 the volume of data in the world will increase more than for five times up to 175 zettabytes. The rapid growth of data reflects the need in the development of automatic systems of analysis and data processing.

In the address of the President of the country Kassym-Jomart Tokayev dated September 1, 2020, "Digitalization is the basic element of all reforms", it was noted that "Digitalization is not a fashion trend, but the main tool for achieving national competitiveness. Working with "data" should be taken to a new level. One of the main tasks of the Government is to provide a unified database system and their further development". Laws have been adopted that allow Kazakhstan to become one of the international data processing and storage centers. Last year alone, more than 80 billion tenge was invested in digital mining. But we can not stop at reaching, we need to attract digital giants from all over the world. In the opposite case, in his words, this is what other countries do.

Revealing the contradictions relates to the problem of searching for data samples, that do not correspond to the given concept of normal behavior.

And the finding out the novelty in the data, related to the detection of contradictions, is usually initially included into the normal model after their discovery or rejected. In the works in different areas the foreign scientists also call them anomalies, ejections, contradictory observations, exceptions: Lishuai Lee, R. John Hansman (USA), Xiang Lu (USA), J. Albertengo, V. Hassan (Turin, Italy), Shiguang Wang, Dexin Yu, Xiaogang Ma and Xue Xin (China). Among the Russian researchers, who also have contributed to the theory and practice of analysis and the elimination of contradictions on the latent-semantic basis upon revealing the contradictions in the semantic close information are: Khomonenko A,D., Ageyev M.S., Dashonok V.L., Dobrov B.V., Krasnov S.A., Kuraleynok I.Ye., Logashev S.V., Nekrestyanov I.S., Lukashevich N.V. and others.

As well, among domestic scholars, obtained the scientific results in the processing of data large arrays in the field of recognition of images, processing of graphic and linguistic signals, classifying and clustering data big volumes by means of different methods are: M.N. Kalimoldayev, Ye.N. Amirgaliyev, R.I. Mukhamediyev, R.R. Musabayev, O.Zh. Mamyrbayev and a number of other authors.

The importance of contradictions' further disclosure is in the fact, that they are converted into meaningful information for using in the management over different processes and objects.

Recently, a great attention has been paid to the research of transport problems in large cities, especially, public transport, including Kazakhstan, with the use of intellectual analysis and big data processing.

Sources of big data are structured and unstructured data, obtained from sensors, video cameras and installed trackers, etc., information on the movement of people (using GPS and Wi-Fi signals), databases of registered vehicles, schedule of public transport routes, etc.

Basic tasks, solved with the help of intellectual analysis and processing of big data: forecasting of road traffic, using modern methods and algorithms: statistical methods, k Nearest Neighbor (kNN), space time correlation, methods of clustering, etc. Here are some authors who are actively involved into traffic issues: Jeffrey Dean (USA), Sanjay Gemawat (India), A. Alexandrov, Stefan Even, Max Heimel, Fabian Hueske (Germany), Odei Kao, Volker Markle, Eric Naikamp, Daniel Varneke, Daniel Kelleher, R. Massobrio, S. Nesmachnov (Uruguay), A. Chernykh, A. Avetisyan, G. Radchenko (Russia).

Despite the large number of articles and scientific papers of researchers in this field, it follows from the analysis of the current state of research that methods for detecting and eliminating contradictions in big data to improve the efficiency and speed of systems are not widespread enough. Taking into account the acute need of transport companies in Kazakhstan in research and development of methods for identifying and eliminating contradictions in the processing of big data, it was decided to choose the field of public transport as the object of research. Thus, the topic of the thesis "Research and development of methods for identifying and eliminating contradictions in big data" is relevant.

Objective of dissertation. Research and development of methods for automatic detection and elimination of contradictions in big data to improve the efficiency and effectiveness of decision-making based on statistical processing and machine learning.

Research tasks:

1. Analysis of existing methods for detecting and eliminating contradictions in big data and identification of key problems.

2. Research and development of a comprehensive method for automatic detection and elimination of contradictions in big data, based on statistical methods and machine learning algorithms, taking into account the specifics of the subject area.

3. Research and justification of the methodology for the formation of effective statistical methods and machine learning algorithms for the detection and elimination of contradictions, taking into account the specifics of the research area.

4. Creating an intelligent information system for the public transport sector using the developed integrated method for identifying and eliminating contradictions in big data and exploring the possibility of providing flexible and rapid changes in the bus route schedule based on it.

Research object: big data in public transport sphere.

Research subject: methods and tools for identifying and eliminating contradictions in big data.

Research methods: mathematical and simulation modeling methods, methods of clustering and classification, statistical methods, computer-aided learning, decisions taking theory, technology of software development.

The following is submitted for protection: Theoretical and comparative analysis of the justification of a complex method of automatic detection and elimination of contradictions based on statistical methods and machine learning algorithms, the effectiveness and accuracy of which is reflected in the results of calculations within the framework of the created intelligent information system of public transport.

Scientific novelty.

1. A complex method of automatic detection and elimination of contradictions in big data is proposed based on the methods of statistical analysis and machine learning and the developed methodology for selecting its components, taking into account the specifics of the subject area.
2. For the first time, an architecture, algorithm and program of an intelligent information system were created to identify and eliminate contradictions in the big data of urban public transport, which increases the efficiency and efficiency of decision-making in the management of public transport in the city.
3. An algorithm and a program for generating templates and a management solution for drawing up a preliminary schedule of a city bus route in an intelligent information system based on purified big data of the city's public transport have been developed.

Practical significance of the research: The information system based on analytical and statistical studies and machine learning algorithms increases the efficiency of public transport (bus fleet) by rationalizing route intervals as a result of automatic detection and elimination of contradictions in the formation and processing of big data. Research has also shown that it is possible to improve the efficiency of large-scale data processing in the field of public transport.

The main provisions for dissertation defense:

Theoretically and by comparative analysis, there is substantiated the comprehensive method for automatic detection and elimination of inconsistencies, based on statistical methods and machine learning algorithms, the effectiveness and accuracy of which is demonstrated by the results of calculations within the framework of the developed information system.

Personal contribution of the researcher. A comprehensive method has been developed for identifying and eliminating inconsistencies in big data. A numerical study and experimental evaluation of the proposed models and algorithms have been carried out. The architecture and software of an intelligent information system for the public transport sector has been developed using the developed integrated method for identifying and eliminating data contradictions.

Relationship of the dissertation topic with the plans of research work. The dissertation work was carried out within the framework of the project for grant funding of the Ministry of Education and Science of the Republic of Kazakhstan AR05133699 "Research and development of innovative information and telecommunication technologies, using modern cyber-technical means for the smart transport system of the city", project leader, Doctor of Technical Sciences, Professor, Chief Researcher of the RSE at the REM "Institute of Information and Computational Technologies", SC MES RK, Yavorskiy V.V.

Volume and structure of work. The dissertation consists of an introduction, three chapters, conclusion and references**.** The total volume of the dissertation: 111 pages of written text, including 49 figures, 15 tables, 134 references, 2 appendices.

In the introduction, the relevance of the work was identified and the problems associated with the topic were shown. The idea of work, the purpose and objectives of the research, scientific novelty and practical value of the research, research methods are disclosed.

The first chapter a large-scale analytical study of literary and Internet sources is given, which shows the trend of increasing published scientific works in the world due to the growth of global data that requires the development of theory and adequate means of intellectual analysis to identify specific ideas and useful models for making effective management decisions. The concepts of contradictions in big data and the types of contradictory data are considered, and the existing methods are analyzed and the main problems are identified. Despite the fact that data preprocessing is the main and important stage, it is shown that the problems of identifying and eliminating contradictions often arise at later stages of big data analytics. The analysis of existing methods and tools for identifying contradictions, their classification, categories and characteristics, showed the presence of many problems in the analysis and integration of data. Effective examples of using machine learning algorithms and statistical methods to identify contradictions are given. At the same time, the main problem is the need to choose methods and technologies that correspond to the specific goals and objectives of the analysis and the subject area, and therefore a specific subject area of big data was chosen for the study - urban passenger transport.

In the second chapter the functional scheme of big data analytics with a subsystem of a complex method for identifying contradictions and their automatic elimination is justified and proposed. The peculiarity of the proposed functional scheme is that the choice of the method used in each component is justified by the features and purpose of providing the specified parameters for identifying models based on them for the problem being solved in the subject area-big data analytics of urban passenger transport. The unified statistical methods that are justified and selected for the subject area are used at the following stages: data acquisition and preprocessing, the formation of a big data database, as well as the detection and elimination of contradictions in big data databases. The main areas of data preprocessing are grouped by type, taking into account the current problems in each of them. Based on their analysis, methods were identified that can be used to process data containing contradictions. This includes methods for removing data from the source data from a given dataset, so they have been analyzed and studied for use for a given subject area. An experimental comparative analysis of statistical criteria for selection for creating an effective algorithm for inferring the laws of change in the public transport variable with respect to the predictor is carried out. The Grubbs test was chosen because it detects inconsistent data with high accuracy and can work with large samples, and k-means algorithms, SVM algorithms, and DBSCAN algorithms were considered in detail, and experimental results were obtained. According to the results obtained by comparing clustering algorithms, the detection of inconsistent data using the k-means algorithm showed a higher percentage, and the calculation time was faster. Also, the selection of machine learning components of a multi - layer neural network-the construction of gradient images-was carried out and the corresponding learning algorithm was developed. The accuracy of the algorithms was compared and evaluated for different values of the contradiction parameters. The results obtained in different percentages confirm the effectiveness of the proposed complex method.

In the third chapter, the structure of the subsystem of a complex method for detecting and eliminating contradictions in big data in the field of public transport is developed, the components of which are justified in the second section. A distinctive feature of the proposed complex method is the implementation of: 1) a two-level contradiction detection system: using the Grubbs statistical test and testing the k-means algorithm. 2) training in a neural network to detect conflicting data and remove data from information obtained over a certain period of time. For the first time, the architecture of an intelligent public transport information system with a subsystem for identifying and eliminating contradictions in big data is proposed. A distinctive feature of this architecture is the presence of a three-component non-relational MongoDB database for storing large amounts of data, as well as the use of processing methods that allow you to process information in a distributed manner on multiple physical servers. Streaming data is collected and pre-processed in a real-time, non-relational MongoDB database. The module for deep learning of a neural network is designed to form templates and management decisions in the form of a preliminary schedule of a city bus route for each time period. Numerical experiments were conducted on the actual data for five bus routes, the results of which showed the correctness of the choice of the Grubbs criterion for solving data processing problems for bus transport management. The results of research on the identification of contradictions in the data of urban bus routes showed the effectiveness of the created information technology and the prospects for its further use and development. Based on the results, it was proposed to make changes to the public transport schedule. The experimental verification of the accuracy of the obtained models in the process of their processing and analysis was carried out, which showed the absence of conflicting values in the cleared data set. Thus, it was proved that the data set obtained after excluding contradictory data has guaranteed accuracy, and can be used in the field of public transport.

In the conclusion, the main results and conclusions of the thesis are described.

Confidence level and validation results.

The results of the dissertation were discussed and reported at the following scientific and methodological conferences: 20th IEEE International Conference on Soft Computing and Measurements, (SCM 2017, July 6, 2017, St. Petersburg; Russian Federation), XLVII international scientific and practical conference "Science yesterday, today, tomorrow "(Novosibirsk, 2017), XIV International Asian School-Seminar "Problems of Optimization of Complex Systems" (Cholpon-Ata, Kyrgyzstan, 2018, 2019), International Scientific Conference of the Institute of Computer Science and Technology of the Ministry of Education and Science of the Republic of Kazakhstan" Modern Problems of Informatics and Computational Technologies "(Almaty , 2018, 2019), III International Scientific Conference "Informatics and Applied Mathematics" (September 26-29, Almaty, 2018), International Scientific and Practical Conference "Innovative IT and Smart Technologies", dedicated to the 70th anniversary of Professor Utepbergenov I.T. (Almaty 2019), IV International Scientific and Practical Conference "Informatics and Applied Mathematics" dedicated to the 70th anniversary of T.N. Biyarov and W. Wujcik, 60th anniversary of E.N. Amirgaliev (Almaty, September 25-29, 2019), scientific seminars of the Institute of Information and Computer Technologies of the MES RK.

Publications. The main provisions of the dissertation were published in 19 scientific works, including: 8 - articles were published in editions, recommended by the Committee for Control in Education and Science of the Ministry of Education and Science of the Republic of Kazakhstan; 1 - articles have been published in editions, indexed in the Scopus database; 10 - publications in materials of international conferences, 2 - which are included in the Scopus database and 1 copyright certificate has been received**.**

1. Tashev A., Kuandykova J., Kassymova D., Akhmediyarova A. Detection and elimination of discrepancies in big data at transport applying statistical methods // Journal of Theoretical and Applied Information Technology. – 2020. - Vol.98. №9. – P.1435-1445. (Scopus, процентиль - 37)

2. Utepbergenov I.T., Kassymova D.T., Akhmediyarova A.T., Yeskendirova D.M. Approach to detecting and eliminating semantic contradictions in «big data // Bulletin of KazATC named after M.Tynyshpayev. - Almaty, 2017, - №2 (101). –

P.200-206.

3. Kassymova D.T., Yeskendirova D.M., Akhmediyarova A.T. Eliminating and identifying inconsistencies in big data // Bulletin of KazATCU, named after Tynyshpayev M. – Almaty, 2017. - № 3 (102). - P.76-80.

4. Utepbergenov I.T., Kassymova D.T., Musabekov N.R., Utegenova A.U., Muslimova A.K. Integrated Approach to Implementing the Virtual Information Infrastructure of the automated process control system // KazNU bulletin. – Almaty, 2015. - № 3(86). – P. 152-156

5. Utepbergenov I.T., Akhmediyarova A.T., Kassymova D.T. Concerning the task of modeling a regular city by means of Petri net // Bulletin of KazATCU, named after Tynyshpayev M. - Almaty, 2016. - №1 (96). – P.77-81.

6. Kassymova D.T., Akhmediyarova A.T., Shayakhmetova C., Turdalyuly M. Analysis of methods, used to identify and resolve inconsistencies in large data // KazNTRU bulletin. – Almaty, 2020. - № 2 (138). – P. 487-495.

7. Kassymova D.T., Akhmediyarova A.T., Bizhanova A.S. On the problem of controlling traffic lights at intersections // Bulletin of KazATK them. M. Tynyshpaeva. – Almaty, 2016. - №1(96). - P.74-77.

8. Khomonenko A.D., Kasymova D.T., Kuandykova D.R., Akhmediyarova A.T. The problems of resolving inconsistencies in big data // KazNRTU Bulletin. - Almaty, 2019. -№2(132). - P.418-424

9. Kassymova D.T., Buranbaeva A.I., Nurkamanova M.A. Comparative analysis of modern urban transport modeling systems // KazNRTU Bulletin. - Almaty, 2018. - №4(128). - P.170-176

10. Khomonenko A.D, Dashonok V.L., Kassymova D.T., Ivanova K.A. Approach to processing data from social networks for detecting public opinion on quality of educational services // 20th IEEE International Conference on Soft Computing and Measurements. SCM 2017. - St. Petersburg, 2017. - № 7970707. - P. 737-739. (Scopus)

11. Khomonenko A.D., Khalil M.M., Kassymova D.T. Probabilistic models for evaluating the performance of cloud computing systems with web interface // SPIIRAS Proceedings, 2016. - Issue 6(49). - P.49-64.

12. Dashonok V.L., Kassymova D. T. Problems of eliminating inconsistencies in big data. (RISC) // Proceedings of XLVII International scientific-practical conference «Science yesterday, today, tomorrow». – Novosibirsk: SibAc, 2017. - № 6(40). – C. 22-26.

13. Akhmediyarova A.T., Kassymova D.T. Analysis of variability and the use of big data of the metropolis // Mater. III Int. scientific. conf. "Computer Science and Applied Mathematics", dedicated. 80th anniversary of prof. Biyasheva R.G. and the 70th anniversary of prof. Aidarkhanova M.B. - Almaty, 2018. - Vol. 1. - S. 207-212.

14. Akhmediyarova A.T., Kassymova D.T., Nurkamanova M.A. Using genetic algorithm in traffic lights control. Proceedings of XIV International. Asian school-seminar «Issues of complex systems optimization». – Almaty, 2018. – Vol. 2. – P. 388-392.

15. Utepbergenov I.T., Akhmediyarova A.T., Kassymova D.T. Chvanova A.O. Implementation of intelligent systems in transport. // Mater. V int. scientific and practical conf. "Computer Science and Applied Mathematics". - Almaty: IIVT MES RK, 2020. - P. 332-340.

16. Kassymova D.T., Utepbergenov I.T., Yeskendirova D.M. Modern approaches to processing large amounts of data // Proceedings of the Intern. Satpayev readings "The role and place of young scientists in the implementation of the new economic policy of Kazakhstan". - Almaty, 2015. - Vol. IV. - P. 208-212.

17. Kassymova D.T., Utepbergenov I.T., Akhmediyarova A.T. Big data preprocessing: methods and perspectives // Mater. scientific. conf. IIVT KN MES RK "Innovative IT and Smart-technologies", dedicated. To the 70th anniversary of professor Utepbergenov I.T. - Almaty, 2019. - P. 256-264.

18. Akhmediyarova A.T., Kassymova D.T. Review of data collection methods for intelligent transport systems // Mater. IV international scientific practical. conf. "Computer Science and Applied Mathematics", dedicated. 70th anniversary of professors Biyarov T.N., Waldemar Vuytsik and 60th anniversary of professor Amirgaliev E.N. - Almaty, 2019. - P. 425-432.

19. Kalimoldaev M.N., Yavorskiy V.V., Sonkin M.A., Wojcik W., Utepbergenov I.T., Akhmediyarova A.T., Kassymova D.T., Klyueva E.G., Baidikova N.V., Esmagambetova M.M. Formation of storage and analysis of big data of movements in the city // Collection of mater. of the VI Internat. scientific-pract. conf. "Big Data and high-level analysis" - Minsk: Bestprint, 2020. - P. 59-70.

Certificate of entering data into the State Register of Rights to objects of copyright:

1. C.c. 12585. "Methods for identifying and eliminating contradictions"/ D.T. Kassymova, A.A. Tashev, I.T. Utepbergenov, A.T. Akhmediyarova; publ.15.10.2020, - 1p.