Selection of Effective Methods of Big Data Analytical Processing in Information Systems of Smart Cities

Oleksii Duda[1][0000-0003-2007-1271], Nataliia Kunanets[2][0000-0003-3007-2462], Oleksandr Matsiuk[1][0000-0003-0204-3971], Volodymyr Pasichnyk[2][0000-0002-5231-6395], Nataliia Veretennikova[2][0000-0001-9564-4084], Anatoliy Fedonuyk[3][0000-0003-0942-227X], and Valentyna Yunchyk[3][0000-0003-3500-1508]

[1] Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine
[2]Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine
[3]Department of Higher Mathematics and Information Science, Lesya Ukrainka Eastern European National University, Lutsk, Ukraine
oleksij.duda@gmail.com, nek.lviv@gmail.com, oleksandr.matsiuk@gmail.com, vpasichnyk@gmail.com, nataver19@gmail.com, fedonyukanatan@gmail.com, uynchik@gmail.com

Abstract. The practice of using Big Data models is widespread implementing the procedures of information and technological support of processes occurring in urban resource social and communication networks. For Big Data, a list of characteristics is presented in the format 10V including volume, velocity, variety, validity, value, veracity, visibility, virtual, variability and valence. At the same time, the urgent task is to select the category of Big Data analytical processing tools for smart city needs. Further development of the Big Data concept leads to an expansion of the list of properties and a variety of tools for their analytical processing. It is suggested to use the expert estimation and hierarchy analysis methods proposed by T. Saati to select an effective Big Data analytical processing method for needs in a smart city. The procedures for this method are described below. The analysis was carried out among the following alternatives as managed machine learning, unmanaged machine learning, data mining, statistical analysis, data visualization. The obtained results show the highest efficiency of the method of managed machine learning. Furthermore, it should be noted that this method might be implemented in appropriate procedures considering the possible need to extend the sets of characteristics and their parameters. The proposed method allows to evaluate the advantages and disadvantages of the tools available on the market to work with big data in a situation when the city authorities decide on the need for appropriate investment.

Keywords: Big Data, analytical processing, expert estimation, hierarchy analysis.

1 Introduction

The current tendency of reorganization of cities, their transformation into comfortable ones for residents and guests, implies the formation of a complex system of modern technologies and tools that facilitate interaction with state authorities, obtaining quality services in the field of health, energy, gas and water supply, as well as transport. One of the smart city projects' implementation areas is the implementation of IoT technologies and the formation of information platforms to process the accumulated data.

1.1 Problem Statement

The use of IoT technologies in smart city projects is being explored by many scientists. It is noted in the paper [1] that due to the evolution of embedded Internet of Things (IoT) devices, networks are being formed to collect information from infrastructures and buildings. IoT technologies provide consumers, manufacturers, and utility providers with new ways to manage devices with smart meters to solve the problems of creating a smart network.

The authors [2] proposed the construction of a smart network that ensures information exchange between energy consumers and energy service providers. The formation of an extensive information flow management structure ensures the processing and analysis of real-time data by energy companies in Norway.

It is suggested [3] to develop a system using Raspberry pi for image processing, VB, sensors to display a message to civil and governmental authorities.

However, the method of hierarchy analysis has found its use in projects of such type.

Some authors analyze why big data technologies with elements of machine learning are effective to use in smart city projects [4]. It is proposed a procedure for selecting the best tools available on the market. The procedure is based on the knowledge of experts. And in our case, a specific group of experts preferred tools based on machine learning methods.

2 Choice of Big Data Analytical Processing Methods

Analyzing a wide variety of urban data sets, it can be stated that about 70 percent of all data on resource social and communication networks [5] are well-structured and could be described using a formal apparatus of functional dependencies and it is convenient to normalize and put it in an appropriate normal form. In case when a relational data model cannot be effectively applied to the processes of storing urban data collections, data warehouses and appropriate procedures for processing this data, such as cutting and aggregation, are used. In addition to the database and data warehouse, approximately 10-15% of smart city datasets are poorly structured and can be submitted using Big Data models [6]. In the context of a wide range of tasks generated during the implementation of information technology support procedures for projects occurring in urban resource social and communication networks, the use of Big Data models is an urgent task [7]. At the same time, it is necessary to give a clear and unambiguous answer to the question if the processed data corresponds to the full spectrum of characteristics and parameters inherent to the Big Data concept. A correct answer to this question allows you to form an expert opinion on the methods and tools needed to process such data. It is naturally arisen a question about what kind of tools is expedient to use in the procedures of preparing and making such decisions.

It is suggested to formulate answers to this kind of questions using the hierarchy analysis method. A description of the procedures for using this method in the context of solving the problem of identifying the full set of Big Data characteristics and parameters in information systems of information and technological support of projects implementing in urban resource social and communication networks in smart cities is given below.

Simultaneously, it should be noted that this method has to be implemented in procedures considering the need to extend the sets of characteristics and their parameters. Experts on the use of Big Data models from European Council for Nuclear Research (CERN, Geneva) suggest that they will grow substantially to 20, 30 and even 100v in the near future.

The choice of Big data analytical processing methods for the needs in a smart city [8] is proposed to be performed using one of the expert evaluation methods, which is based on the procedure of pairwise comparisons of variables for the hierarchy analysis proposed by T. Saati [9, 10]. The following algorithm is carried out to implement the hierarchy analysis method:

Step 1. A hierarchy with a three-tier structure is built. The top level of the hierarchy is the goal to be decided (see Fig. 1).

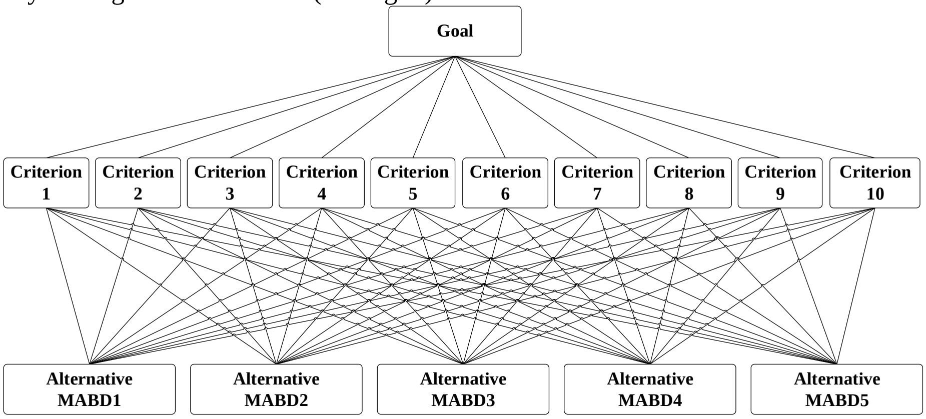

Fig. 1. The general scheme of the hierarchy analysis method of choosing a category for Big Data analytical processing in a smart city

Step 2. It is formed a criterion set at the second level, according to which alternative methods of analytical processing (MABD) of Big data urban collections are chosen.

According to the results of the study, a list of Big Data characteristics has been formed, which is presented in the format “10V”, containing the attributes as volume, velocity, variety, validity, value, veracity, visibility, virtual, variability, and valence. It is denoted $V^{BigData}\left(Attr^{BigData}\right)$, the set of Big Data description attributes $Attr^{BigData} = \left\{Attr_i^{BigData}, i = \overline{1,10}\right\}$, that provides the fixed characteristics.

$$Attr^{BigData} = \left\{ \begin{matrix} volume, velocity, \text{var}\,iety, validity, value, veracity, \\ visibility, virtual, \text{var}\,iability, valence \end{matrix} \right\} \quad (1)$$

For each of these characteristics, it is created the lists of parameters that can uniquely identify the presence of the i-attribute:

$$P_i^{BigData} = \left\{P_{J_i}^{BigData}, J_i = \overline{1, N_i}\right\} \quad (2)$$

where N_i is the number of parameters that describe the i-characteristic $Attr_i^{BigData}$ of Big Data.

Further development of the Big Data concept leads to an extension of the presented list of characteristics [11].

Step 3. It is set up a list of alternatives containing a few Big Data analytical processing methods, in particular:

Managed machine learning [12].

Unmanaged machine learning [13].

Data mining [14].

Statistical analysis [15].

Data visualization [16].

Available alternatives to Big Data analytical processing methods form the lower level of the hierarchy. The structure of the decision-making problem by the hierarchy analysis method on the choice of the method for analytical processing of Big Data urban collection to support the processes in the resource social and communication networks is presented in Figure 2.

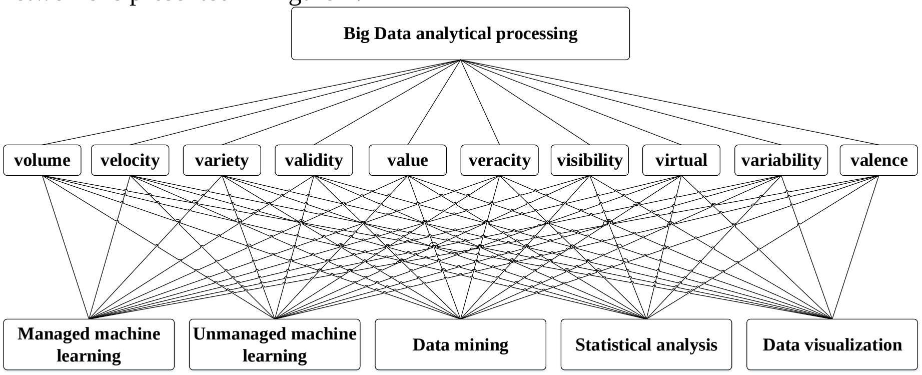

Fig. 2. The structure of the decision-making problem regarding the choice of methods for analytical processing of Big Data urban collections

In our case, the decision-making process is to choose one of the possible alternative categories based on the constructed priority vector. The priority will be a real number that corresponds to each alternative. The highest priority alternative would be considered a decision. The priority of the alternative will be called its weight.

Step 4. It is used the determination of the expert estimation scale for the hierarchy analysis method. In this case, an expert estimation scale or the importance of pairwise comparisons was used to evaluate the advantages of one object over another with values from 1 to 9. The overall content of the estimates is presented in Table 1.

Table 1. Determining the importance degrees for pairwise comparison matrices.

The importance degree	Definition	Comment
1	Equally important	The contribution of both objects is the same
3	Weak significance	Expert experience and judgment give the first object a slight advantage over the second one
5	Essential or great significance	Expert experience and judgment give the first object a great advantage over the second one
7	Great and obvious significance	The advantage of the first object over the second one is very large, almost obvious
9	Absolute significance	The evidence to the benefit of the first object is more than convincing
2, 4, 6, 8	Intermediate values between adjacent scale values	Used for situations where compromise solutions are required
Reciprocal of given values	If one of the above values is obtained for comparing the first object with the second one, then the result of comparing the second object with the first one is reciprocal	

Step 5. It is to construct matrices of pairwise comparisons for each of the above evaluation criteria and to calculate the numerical characteristics of these matrices, including the consistency index, the largest eigenvalue, and the index of ratio sequence. Each of these matrices contains the value of expert estimation on pairs of alternatives, which are the methods of Big Data analytical processing from the given list. The volume characteristic is the value of Big Data urban collections [17]. A pairwise comparison matrix for selecting the category of Big Data analytical processing methods by the volume attribute is presented in Table 2.

Table 2. A pairwise comparison matrix for choosing volume-based alternatives

Alternatives	Managed Machine Learning	Unmanaged Machine Learning	Data Mining	Statistical Analysis	Data Visualization
Managed Machine Learning	1	2	3	2	3
Unmanaged Machine Learning	0,5	1	4	3	4
Data Mining	0,33	0,25	1	2	1

Statistical Analysis	0,5	0,33	0,50	1	2
Data Visualization	0,33	0,25	1	0,5	1
Total	2,67	3,83	9,50	8,5	11

The results of calculating the weights of alternatives by the volume criterion are presented in Table 3.

Table 3. The weights of alternatives by volume criterion

Alternatives	Managed Machine Learning	Unmanaged Machine Learning	Data Mining	Statistical Analysis	Data Visualization	Total
Managed Machine Learning	0,3750	0,5217	0,3158	0,2353	0,2727	1,7205
Unmanaged Machine Learning	0,1875	0,2609	0,4211	0,3529	0,3636	1,586
Data Mining	0,1250	0,0652	0,1053	0,2353	0,0909	0,6217
Statistical Analysis	0,1875	0,087	0,0526	0,1176	0,1818	0,6266
Data Visualization	0,1250	0,0652	0,1053	0,0588	0,0909	0,4452
Total	1	1	1	1	1	5

The column elements of the Table 3 will be obtained thanks to the normalization procedure applied to the corresponding column elements of the Table 2. To estimate the principal eigenvector of a pairwise comparison matrix, it is calculated a vector of priorities whose elements are the weights of alternatives calculated in the form of algebraic sums of the elements of the corresponding rows in the Table 3, divided by the total number of alternatives, namely the number of line items in the Table 2.

Therefore, the best alternative is the Managed machine learning category by the volume criterion because it has the highest value of weight – 0,3441. For the pairwise comparison matrix constructed according to the volume criterion, the following parameters are calculated:

the estimate of the largest eigenvalue, calculated by the formula:

$$\lambda_{\max} = \sum_{i=1}^{n} w_i s_i \quad (3)$$

where w_i is the weight of the alternative with i number,

S_i is the sum of the column elements of the pairwise comparison matrix with i-number,

n is a number of alternatives;

the consistency index:

$$C_I = \frac{\lambda_{\max} - n}{n-1} \quad (4)$$

the index of ratio sequence:

$$C_R = \frac{C_I}{R_I} \quad (5)$$

For all further estimations of the weights of the alternatives and $n=5$, it will be used the same random index value $R_I = 1,25$.

After performing the calculations of the pairwise comparison matrix built on the volume criterion, these parameters will take the following values:

the estimate of the largest eigenvalue:

$$\begin{aligned}\lambda_{\max} &= 0,3441 \cdot 2,67 + 0,3172 \cdot 3,83 + 0,1243 \cdot 9,50 + \\ &+0,1253 \cdot 8,5 + 0,089 \cdot 11 = 5,5394\end{aligned} \quad (6)$$

the consistency index:

$$C_I = \frac{\lambda_{\max} - n}{n-1} = \frac{5,5394 - 5}{5-1} = 0,0898 \quad (7)$$

the index of ratio sequence:

$$C_R = \frac{C_I}{R_I} = \frac{0,0898}{1,25} = 0,0725 \quad (8)$$

Since $C_R = 7,25\% \le 10\%$, the matrix of pairwise comparisons according to the volume criterion is considered to be consistent one. Pairwise comparison matrices for selecting the category of Big Data analytical tools by criteria such as velocity, variety, validity, value, veracity, visibility, virtual, variability and valence are formed similarly to Table 2. The results of the weight estimation of the alternatives for these criteria are also formed in the same way as in Table 3. The best alternatives to these criteria and the corresponding calculated values are given in Table 4.

Table 4. The best alternatives and their corresponding weights

Criterion	The Best Alternative	Weight
Velocity	Unmanaged Machine Learning	0,3251
Variety	Managed Machine Learning	0,4955
Validity	Unmanaged Machine Learning	0,3149
Value	Managed Machine Learning	0,486
Veracity	Managed Machine Learning	0,414
Visibility	Managed Machine Learning	0,4143
Virtual	Data Mining	0,3632
Variability	Unmanaged Machine Learning	0,3719
Valence	Data Visualisation	0,4987

Similarly, to the volume criterion, it is estimated the largest eigenvalues – $\lambda_{\max}$, the consistency indices – C_I and the sequence of ratios – C_R, presented in Table 5.

Table 5. Variables of a pairwise comparison matrix

Criterion	$\lambda_{\max}$	C_I	C_R
Velocity	5,1297	0,0324	0,0262
Variety	5,2747	0,0687	0,0554
Validity	5,25	0,0625	0,0504
Value	5,3923	0,0981	0,0791
Veracity	5,3445	0,0861	0,0695
Visibility	5,1535	0,0384	0,0309
Virtual	5,3042	0,0761	0,0613
Variability	5,3897	0,0974	0,0786
Valence	5,3578	0,0895	0,0721

Since the inequality $C_R \leq 10\%$ is satisfied for all the criteria of the consistency ratio, then all pairwise comparison matrices are consistent.

Step 6. It is an evaluation of the importance of the criteria to determine the weight of the alternatives. The criteria are equally important to simplify the calculations. In the pairwise comparison matrix, all squares will be filled with the same values equal to one.

In the column of total value in Table 6, the alternatives of analytical processing methods are calculated in the form of arithmetic mean of weights.

Table 6. The estimation results of the weights of alternatives to the methods of Big data analytical processing

Category	Volume	Velocity	Variety	Validity	Value	Veracity
Managed Machine Learning	0,3441	0,3072	0,4955	0,2809	0,4860	0,4140
Unmanaged Machine Learning	0,3172	0,3251	0,1037	0,3149	0,2715	0,2691
Data Mining	0,1243	0,2072	0,0897	0,2222	0,1108	0,0922
Statistical Analysis	0,1253	0,0670	0,1097	0,0986	0,0595	0,1598
Data Visualization	0,0890	0,0936	0,2015	0,0834	0,0723	0,1016
Consistency index	0,0898	0,0324	0,0687	0,0625	0,0981	0,0861
Total	1	1	1	1	1	1
Category	Visibility	Virtual	Variability	Valence	Total	Total cost
Managed Machine Learning	0,4143	0,2825	0,3713	0,0463	3,4421	0,3442
Unmanaged Machine Learning	0,1961	0,0749	0,3719	0,0948	2,3390	0,2339
Data Mining	0,1852	0,3632	0,0941	0,1179	1,6067	0,1607

Statistical Analysis	0,0812	0,2131	0,0492	0,2423	1,2057	0,1206
Data Visualization	0,1232	0,0662	0,0882	0,4987	1,4178	0,1418
Consistency index	0,0384	0,0761	0,0974	0,0895	0,7390	0,0739
Total	1	1	1	1	10	1

Step 7. It is to select the methods with the highest weight. According to expert estimates obtained by the method of hierarchy analysis, it is the Managed machine learning category of analytical processing tools that is considered the best option for use in the analytical procedures of Big data urban collections. Let us estimate the average consistency ratio of the hierarchy according to the weighted index of the ratio sequence:

$$C_R = \frac{C_I}{R_I} = \frac{0,0739}{1,25} = 0,0596 < 0,1.$$

The result shows that the entire hierarchy is consistent. The results of the calculations are presented in Table 7 and Figure 3.

Table 7. Alternatives and weights of categories of Big Data analytical processing tools

Alternative	Priority
Managed machine learning	0,3442
Unmanaged machine learning	0,2339
Data mining	0,1607
Statistical analysis	0,1206
Data visualization	0,1418

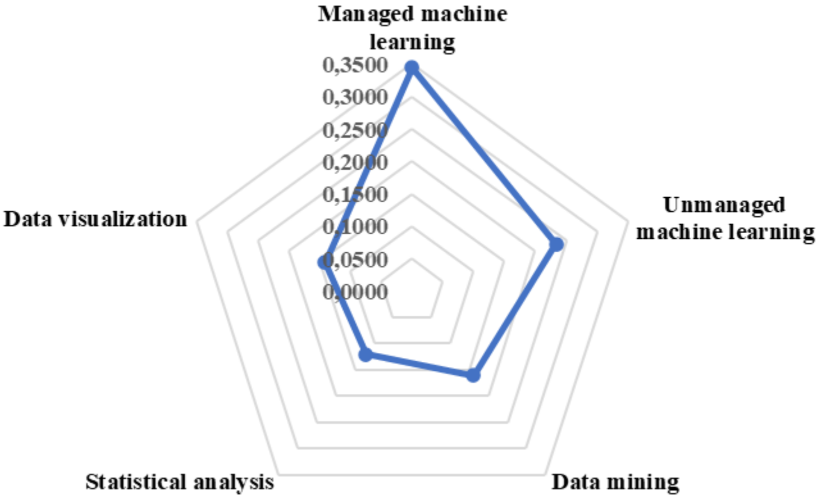

Fig. 3. Chart of the weight coefficient distribution of alternative Big Data analytical methods

3 Conclusion

A detailed analysis of Big Data processing methods was conducted, including managed machine learning, unmanaged machine learning, data mining, statistical analysis, and data visualization. A comparative analysis of their functionality was performed using the method of hierarchy analysis based on expert estimation. Furthermore, the evaluation was performed on 10 criteria such as volume, velocity, variety, validity, value, veracity, visibility, virtual, variability and valence. According to the results of the study, the managed machine learning is determined as the most effective method.

References

1. Devagkumar, U. Shah, Chiragkumar, B. Patel: IoT Enabled Smart Grid. International Journal for Scientific Research and Development, Vol.11, 40-42 (2016).
2. Bokolo, Anthony Jr, Sobah, Abbas Petersen, Dirk, Ahlers, John, Krogstie, Klaus Livik: Big data-oriented energy prosumptionservice in smart community districts: amulti-case study perspective. Energy Informatics, Vol. 2 (1), 1-26 (2019).
3. Gaikwad, C.M., Jadhav, A.A., Suryawanshi, N.N., Farakate, P.P.: Smart City Based on IoT. International Research Journal of Engineering and Technology, Volume 04, Issue 03, pp. 1265-1270 (2017).
4. Mohammadi, M., Al-Fuqaha, A.: Enabling cognitive smart cities using big data and machine learning: Approaches and challenges, IEEE Communications Magazine, Vol. 56, pp. 94-101 (2018).
5. Malik, K. R., et al.: A methodology for real-time data sustainability in smart city: Towards inferencing and analytics for big-data. Sustainable Cities and Society, 39, pp. 548-556 (2018).
6. Silva, B. N., et al.: Urban planning and smart city decision management empowered by real-time data processing using big data analytics, Sensors, 18.9, p. 2994 (2018).
7. Duda, O., Kunanets, N., Matsiuk, O., Pasichnyk, V.: Cloud-based IT Infrastructure for “Smart City” Projects. Dependable IoT for Human and Industry: Modeling, Architecting, Implementation, pp. 389-410 (2018). ISBN: 978-87-7022-013-2.
8. Duda, O., Matsiuk, O., Pasichnyk, V.: Information-Communication Technologies of IoT in the "Smart Cities" Projects. CEUR Workshop Proceedings, Vol.2105, T. I, pp. 317–330 (2018).
9. Saati, T.: Decision-making. Method of the analysis of hierarchies. Radio and communication (1993).
10. Kut, V., Pasichnik, V., Tomashevskyi, V., Kunanets, N.: The Procedures for the Selection of Knowledge Representation Methods in the “Virtual University” Distance Learning System. Advances in Intelligent Systems and Computing (AISC), Vol. 754, pp. 713–723 (2018).
11. Duda, O., Matsiuk, O., Karpinski, M., Veretennikova, N., Kunanets, N., Pasichnyk, V.: Information Technologies of Internet Devices and BigData in the “Smart Cities” Projects, in Proc. 13 Intern Scientific and Techn. Conf. on Computer Science and Information Technologies (CSIT), Vol. 2, Lviv, pp. 72-75 (2018). ISBN: 978-1-5386-6465-0.
12. Mohapatra, B. N., Panda, P. P.: Machine learning applications to smart city (2019).

13. Al-Turjman, F.: Smart Cities Performability, Cognition, & Security. Springer International Publishing (2019).
14. Honarvar, A. R., Sami, A.: Towards sustainable smart city by particulate matter prediction using urban big data, excluding expensive air pollution infrastructures. Big data research, 17, pp. 56-65 (2019).
15. Soomro, K. et al.: Smart city big data analytics: An advanced review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9.5 (2019).
16. Bouloukakis, M. et al.: Virtual Reality for Smart City Visualization and Monitoring. Mediterranean Cities and Island Communities, Springer, Cham, pp. 1-18 (2019).
17. Osman, A. M. S.: A novel big data analytics framework for smart cities. Future Generation Computer Systems, 91, pp. 620-633 (2019).