Analysis of the Demand for Bicycle Use in a Smart City Based on Machine Learning

Diana Koshtura[1], Myroslava Bublyk[2[0000-0003-2403-0784]], Yurii Matseliukh[3[0000-0002-1721-7703]], Dmytro Dosyn[4[0000-0003-4040-4467]], Liliya Chyrun[5[0000-0003-4040-7588]], Olga Lozynska[6[0000-0002-5079-0544]], Ihor Karpov[7[0000-0003-4885-5078]], Ivan Peleshchak[8[0000-0002-7481-8628]], Mariya Maslak[9[0000-0002-3322-740X]], and Oleg Sachenko[10[0000-0001-9337-8341]]

[1-7]Lviv Polytechnic National University, Lviv, Ukraine,
[9]National Technical University «Kharkiv Polytechnic Institute», Kharkiv, Ukraine
[10]Ternopil National Economic University, Ternopil, Ukraine

diana.koshtura.sa.2017@lpnu.ua[1], my.bublyk@gmail.com[2], indeed.post@gmail.com[3], mariya.maslak2016@gmail.com[8]

Abstract. The need to study the demand for bicycle sharing based on regression models of data analysis and prediction of results was investigated. To substantiate possibility of work of the main processes of the information systems, UML diagrams were created. To determine the peaks in demand for bicycles in a certain period of time, it was proposed to use regression models of data analysis. The proposed decision trees were recommended for modeling new datasets in Smart Cities, especially Lviv.

Keywords: Content Analysis, Data Set, Machine Learning, Predicting Demand, Smart City.

1 Introduction

The problem of forecasting the demand for bicycle rental has become more acute in recent months with the spread of atypical respiratory illnesses caused by the new coronavirus called SARS, MERS, and COVID-19. Implementation of the principles of a smart city allows us to solve this problem easily by using machine learning. The availability of relevant data sets is the cornerstone of the rapid implementation of the principles of smart cities in large metropolitan areas and in cities with a population of close to a million. Unfortunately, there is no dataset in Lviv Smart City related to the bike-sharing program. This led to the use of the Capital Bike Sharing dataset, which collects information about the famous bike-sharing program implemented in such a Smart City as megalopolis Washington DC [1]. In particular, the Bike Sharing data set from the UCI machine learning repository was used to forecast the need for the bicycles number [1-4].

The aim of the research was to substantiate the use of machine learning capabilities to predict the demand for bicycles, including regression models of data analysis. The following tasks have been considered and solved:

- To design the structure of an information system using UML diagrams;
- To justify regression models using the data analysis to determine the peak demand for bicycles in a certain period of time;
- To develop a solution tree for successful modeling of new datasets in smart cities such as Lviv.

2 Basic Principles of Research

The problem of increasing the quality of life of the urban population is closely related to the growing requirements for environmental protection, the protection level. This increases the relevance of travel within the city on bicycles, promotes the rapid spread of bicycle rental services, including their joint use [2-3]. Various bicycle sharing programs for a certain period of time are actively implemented in different cities (megacities or small towns). The population is provided with both manual and automated bicycle rental.

Bicycle sharing platforms allow you to collect a variety of data about the duration of the trip, its time, location, etc., about cyclists, their demographic characteristics, etc., as well as factors that directly or indirectly affect the implementation of these trips - weather, traffic, landscape [1-3]. These data sets are useful not only for researchers, but also have practical value for entrepreneurs, government officials. The collected data can be used in forecasting the demand for such services, making important management decisions to build strategic development plans for cities and towns, regions and even the development of economic activities such as tourism.

Unfortunately, there is no such data from Lviv in the UCI Machine Learning Repository data archive, but there is a valuable set of data related to one such bicycle-sharing program in Washington [1]. In view of the given set of bicycle sharing data, it was used as a sample to drafting the relevant data set to a dynamic information system for forecasting bicycle rental demand.

Among the stages of this problem solving, the key step is to study and understand the data. At this stage, it is important to analyze the study data. It is worth noting that the analysis of the research data is one of the most important phases in the whole work process and can help not only to understand the data set, but also to present some clear points that can be useful in the next steps [1-7]:

- Research, description and data visualization;
- Selection of data subsets and attributes for identification and analysis of the problem;
- Indication of missing data items (if any).

3 Analysis of Known Means of Solving the Problem

Moving on bike in Lviv Smart City is becoming increasingly popular - cycling infrastructure is developing, the number of cyclists is increasing. The bicycle has many advantages - environmental friendliness, maneuverability, health benefits. However, not everyone can buy a bike, so a good alternative is to rent a bike.

The first thing to consider is Nextbike, which is an automated bicycle rental system in Lviv Smart City, and it positions itself as an alternative to public transport, so it has the status of a municipal one. Its main advantage is that customers do not need to return the bike to the exact place where it is taken, because there are several rental stations. Their total number in Lviv Smart City is 22, where only 7 of them are functioning.

An analysis of cyclists, their route, number of hours spent and other characteristics such as weather or traffic has not been conducted, but the company has created a mobile application that will soon begin analyzing users data. To rent a bike, customers need a registration procedure, which can be done on the company's website through the mobile application. The registered account will be valid not only in Ukraine, but also in 23 countries and territories where the company operates. There are also several bicycle rentals in Lviv Smart City, but there are no special applications for data collection and analysis, forecasting the demand for bicycles. Some of these are Velobayk, Rental Centre, Veliki.ua. Due to traffic restrictions and heavy road traffic in Lviv Smart City, bicycle rental is gaining popularity. It is an environmentally friendly alternative to transport, and cycling is good for your health. That is why such a system is gaining popularity in Lviv Smart City. However, today there are no tools to analyze and collect data, it limits the ability to determine the demand for bicycles by its prediction.

4 System Analysis

The development of the system begins with a plan of work and construction of the system, which will help to understand the essence of the main task. If there are certain independent tasks at the development stages, they can be easily seen and divided into work, which will speed up the implementation of the system. First of all, at the beginning the system performs the task of reviewing and reading the data set. According to the recommendations given in [8-13] during the review of the data to be processed, the data set is checked for errors. Data analysis according to [14-22] is accompanied by the construction of certain diagrams (Fig.1). The last stage is the analysis of data on the basis of the main features, therefore results of the analysis of demand for use of bicycles are deduced, visualization of these data is carried out.

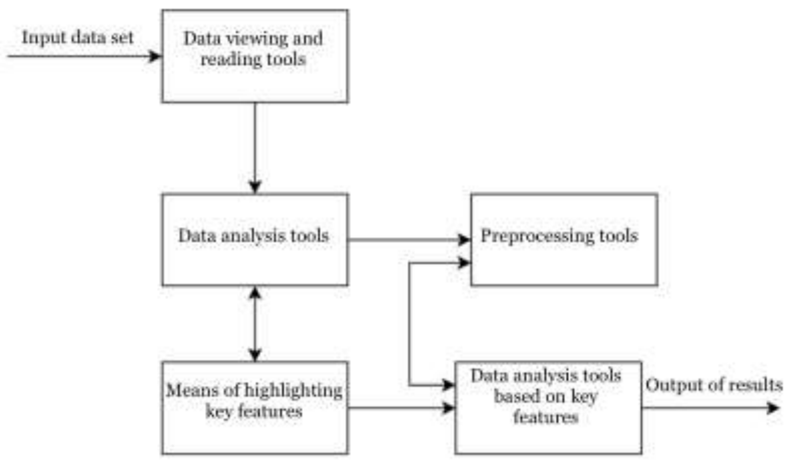

Fig. 1. The structure of the construction of algorithmic and software tools for the analysis of bicycle demand

The general structure of the analysis implementation stages of the bicycle demand is presented in Fig.2.

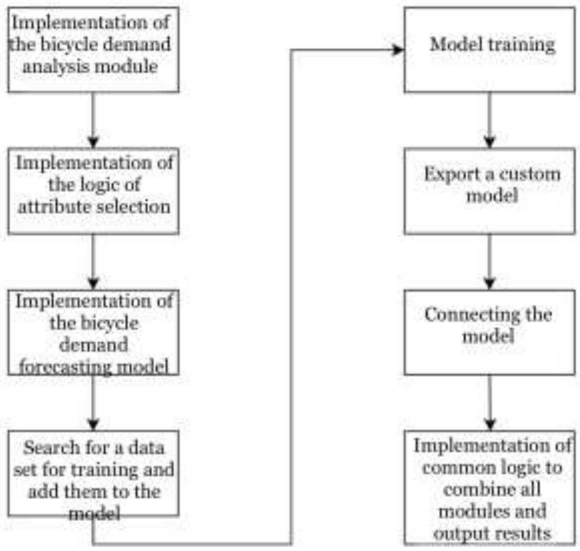

Fig. 2. Block diagram of the implementation stages

To describe the capabilities of the simulated system at the conceptual level, UML diagrams were created. Figure 3 shows the Use Case diagram to specify options for action sequences in the simulated system.

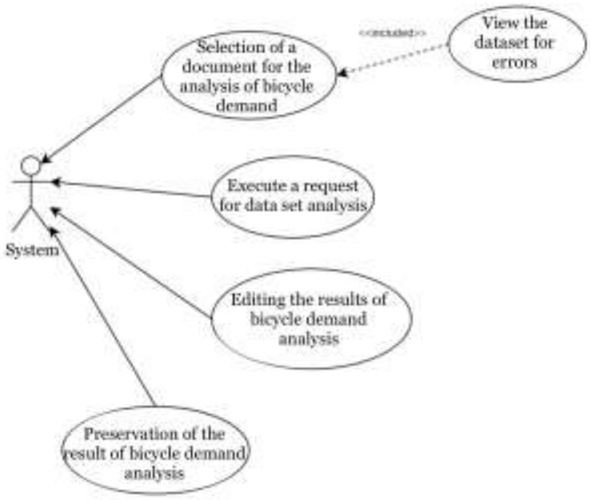

Fig. 3. Use Case Diagram

The system must first get a document with a dataset for further processing of this data [23-37]. The system then queries the analyses of this data set to create a bicycle demand forecast [38-47]. At these stages, it may be necessary to edit the dataset if errors in the dataset itself are detected in the previous stages of processing. Once you have finished editing, processing, and analyzing the datasets, you need to save the results.

The sequence diagram can used to understand the interaction of objects arranged in time. A lot of detailed steps, which the system will perform, can see in the following fig. 4 - in the sequence diagram.

The activity diagram is designed to describe the dynamic aspects of the system. In the general case, it is an opportunity to imagine the transition from one action to another. The main purpose of the activity diagram is to obtain the dynamics of system behaviour. Activity is part of the functioning of the system. The only limitation of the activity chart is that it does not display messages that are created and received from one part of a functioning system to another.

Therefore, the main functionality of activity diagrams can be defined as follows:

- Show the flow of activities in the system;
- Describe the sequence of transition from one activity to another;
- Describe the parallel, dividing course of events in the system.

Figures 5-6 present the Activity diagram for graphical representation of activities workflows and actions in the modeled system.

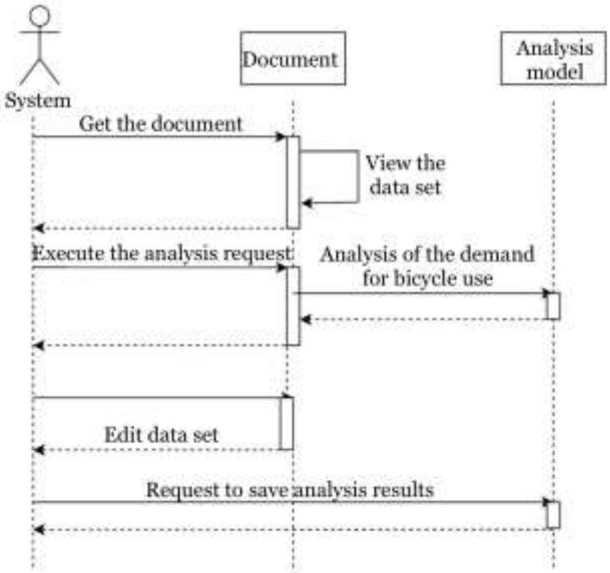

Fig. 4. Sequence diagram

Fig. 5. Activity diagram

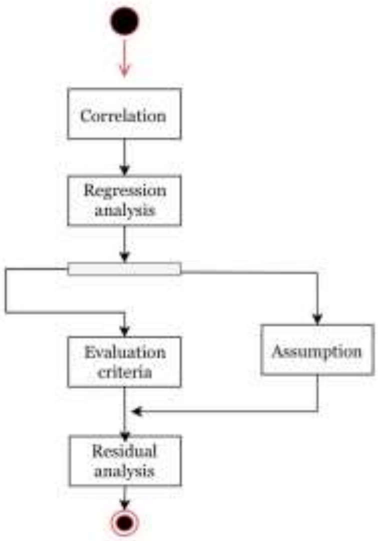

Fig. 6. Activity diagram

5 The Sample of Data Analysis

At the data analysis stage, the data is loaded into the analysis environment and explored in terms of matching the number of records to the number of attributes. The documentation for the dataset states that there is sharing of both bicycles and weather attributes [1-2]. For example, the attribute dteday would require the conversion of its type from the object (or string type) to the timestamp. Attributes such as weekday, holiday, season, etc., which are displayed as integers, would require conversion into categoricals, and so on. [1-3].

Visualization of the hourly cyclist distribution by seasons is presented in Fig. 7. As we can see, the smallest number of cyclists falls in the spring season, and the largest number of cyclists falls in the autumn season, where the number of cyclists is high within all 24 hours (Fig. 7).

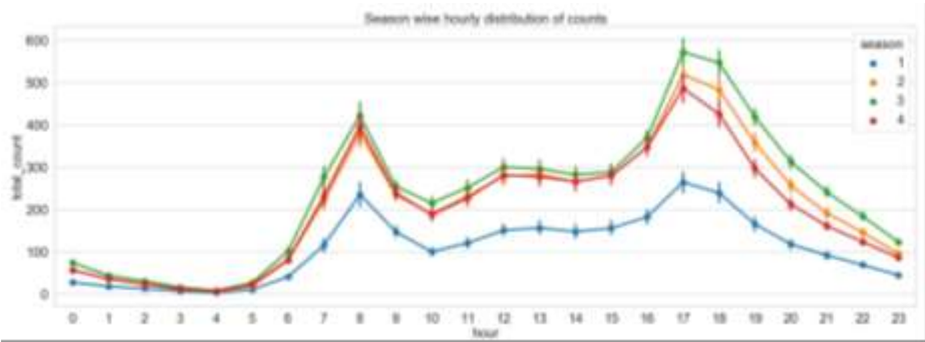

Fig. 7. Visualizing the hourly cycling distribution by seasons [1-3].

The dataset contains the value of cycling by days, weeks, and years. Visualization of the cycling distribution also shows specific trends in the increase of afternoon use on weekends, the increase of the next year comparing to the previous year, and so on. The coefficient of expansion is also much higher than in the previous year, although the maximum density for both is between 100 - 200 cyclists.

Thus, we built diagrams to understand how the system would work, developed UML diagrams to demonstrate how model training and analysis would be conducted, and visualized data from the dataset to better understand and correctly predict them. Now it is clear in which seasons there is the greatest demand for the use of bicycles - the summer-autumn period. Also, based on the knowledge of which hours bicycles are most often rented, and these are the morning and afternoon hours, you can make predictions about the appropriate use of bicycles.

6 Normality Test

To model a data set on the use of bicycles and solving problems in forecasting the demand for bicycles for a certain period, we use the concepts of regression analysis

[1-3]. We divided our data set into 67% and 33% accordingly for training and testing dataset, respectively. Normality Test is visualizing check for normality of the data. It helps us identify outliers, skewness, and so on. In Fig 8 it is presented the data plotting based on theoretical quartiles. To confirm normality on the sample plots showcasing data confirming the normality test are represented (Fig.8).

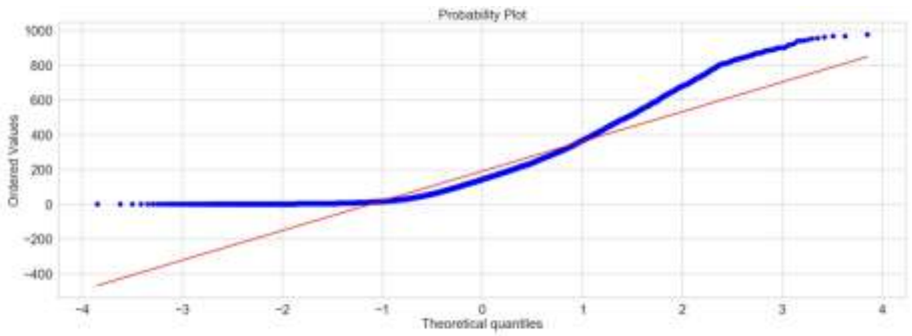

Fig. 8. Visualizing the normality test

7 Regression Based on the Decision Tree

We will explain the concepts and terminology associated with decision trees, for example. We have a set of models for renting bicycles from different owners. Suppose each data item has characteristics such as wheel size, number of gears, price, year of purchase, mileage, peak hours. The visualization shown in Fig. 9.

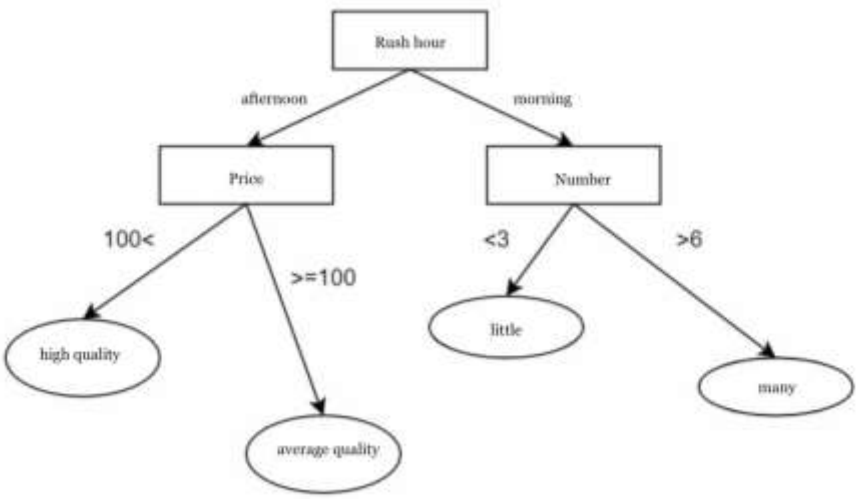

Fig. 9. Decision tree diagram

It demonstrates an exemplary solution tree with leaf nodes that tare focused on the target values [1-3]. The tree begins by breaking the data set at the root based on the time of day, which represents the bicycle rental in the morning by the left child and the right child by renting it in the afternoon, and similarly for other nodes.

The cost may depend on effort, time, and appropriate achievements. We now turn to the best model that the GridSearchCV object has helped to identify. [2-3]. As can be seen from the value of the R-square, productivity is quite comparable to learning (see Fig.10-11). It can be concluded that the Regressor Decision Tree better predicts the demand for bicycles compared to linear regression [2-3].

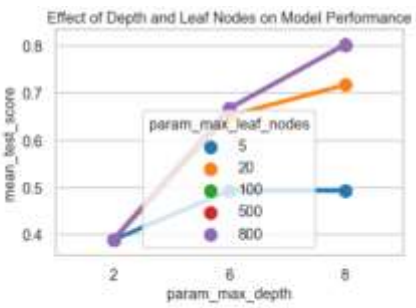

Fig. 10. Visualizing the average test score, the effect of tree depth and the number of leaf nodes

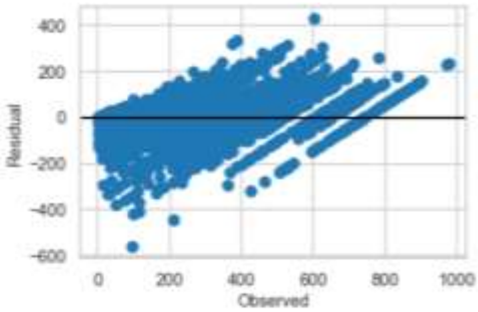

Fig. 11. The result of the inspection

8 Conclusions

In order to forecast the demand for bicycle sharing in Smart City, the following tasks have been considered and solved in this paper. As an example, the well-known Bike Sharing dataset was selected. This dataset is available from the UCI Machine Learning Repository. On the example of this dataset, a set of problems could be solved.

To forecast the demand for bicycle sharing in Lviv Smart City it was recommended to use regression models of data analysis. To demonstrate possible cases and the work of the main processes of information systems, UML diagrams were created. To determine the peaks in demand for bicycles in a certain period of time, it was proposed to use regression models of data analysis. The dataset transformations have been made, such as attributes renaming, types changing, and categories definition. The proposed decision trees were recommended for modeling new datasets in such a Smart City like Lviv. The Regressor Decision Tree better predicts the demand for bicycles compared to linear regression.

References

1. Fanaee-T, H., Gama, J.: Event labeling combining ensemble detectors and background knowledge, Progress in Artificial Intelligence, Springer Berlin Heidelberg, pp. 1-15 (2013):.
2. Sarkar, D., Bali, R., Sharma, T.: Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Apress, Berkeley. (2018) https://doi.org/10.1007/978-1-4842-3207-1
3. GitHub: Bike Sharing Dataset using Decision Tree Regressor (2018) https://github.com/dipanjanS/practical-machine-learning-with-python/blob/master/notebooks/Ch06_ Analyzing_Bike_Sharing_Trends/ decision_tree_regression.py
4. Wei Zhang, Xiaowei Dong, Huaibao Li, Jin Xu, Dan Wang: Unsupervised Detection of Abnormal Electricity Consumption Behavior Based on Feature Engineering, IEEE (2020)
5. Park Y.: A comparison of neural net classifiers and linear tree classifiers: Their similarities and differences. Pattern Recognition, 27(11):1493-1503. (2004)
6. Quinlan, J. R.: C4.5: Programs for Machine Learning. — San Mateo: Morgan Kaufmann Publishers Inc., P.302. (2003)
7. Wasserman, L.: All of statistics: a concise course in statistical inference. Springer. (2003)
8. Bublyk, M.I., Rybytska, O. M.: The model of fuzzy expert system for establishing the pollution impact on the mortality rate in Ukraine. In: Computer sciences and information technologies: Proceedings of the 2017 12th International Scientific and Technical Conference (CSIT 2017), 1, 253–256. (2017) DOI: https://doi.org/10.1109/STC-CSIT.2017.8098781.
9. Matseliukh, Y., Vysotska, V., Bublyk, M.: Intelligent system of visual simulation of passenger flows. In: CEUR Workshop Proceedings, 2604, 906. (2020) http://ceur-ws.org/Vol-2604/paper60.pdf
10. Bublyk, M., Matseliukh, Y., Motorniuk, U., Terebukh, M.: Intelligent system of passenger transportation by autopiloted electric buses in Smart City. In: CEUR Workshop Proceedings, 2604, 1280. (2020) http://ceur-ws.org/Vol-2604/paper81.pdf
11. Krislata, I., Katrenko, A., Lytvyn, V., Vysotska, V., Burov, Y.: Traffic flows system development for smart city. In: CEUR Workshop Proceedings, Vol-2565, 280–294. (2020)
12. Katrenko, A., Krislata, I., Veres, O., Oborska, O., Basyuk, T., Vasyliuk, A., Rishnyak, I., Demyanovskyi, N., Meh, O.: Development of Traffic Flows and Smart Parking System for Smart City. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 730-745. (2020).
13. Peleshko D., Rak T., Lytvyn V., Vysotska V., Noennig J.: Drone monitoring system DROMOS of urban environmental dynamics. In: CEUR Workshop Proceedings, Vol-2565, 178–19. (2020)
14. Lytvyn, V., Kowalska-Styczen, A., Peleshko, D., Rak, T., Voloshyn, V., Noennig, J. R., Vysotska, V., Nykolyshyn, L., Pryshchepa, H.: Aviation Aircraft Planning System Project Development. In: Advances in Intelligent Systems and Computing IV, Springer, Cham, 1080, 315-348. (2020)
15. Lytvyn, V., Dmytriv, A., Berko, A., Alieksieiev, V., Basyuk, T., Noennig, J., Peleshko, D., Rak, T., Voloshyn, V.: Conceptual Model of Information System for Drone Monitoring of Trees' Condition. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 695-714. (2020).
16. Alieksieiev, V., Markovych, B.: Implementation of UAV for environment monitoring of a Smart City with an airspace regulation by AIXM-format data streaming. In: Industry 4.0"

– Scientific-Technical Union of Mechanical Engineering “Industry 4.0”, Sofia, Bulgaria, 5(2/2020), 90–93. (2020), https://stumejournals.com/journals/i4/2020/2/90
17. Kunanets, N., Matsiuk, H.: Use of the Smart City Ontology for Relevant Information Retrieval. In: CEUR Workshop Proceedings, Vol-2362, 322-333. (2019)
18. Batiuk, T., Vysotska, V., Lytvyn, V.: Intelligent System for Socialization by Personal Interests on the Basis of SEO-Technologies and Methods of Machine Learning. In: CEUR workshop proceedings, Vol-2604, 1237-1250. (2020).
19. Kravets, P., Lytvyn, V., Vysotska, V., Burov, Y.: Promoting training of multi-agent systems. In: CEUR Workshop Proceedings, Vol-2608, 364-378. (2020)
20. Lytvyn, V., Vysotska, V., Osypov, M., Slyusarchuk, O., Slyusarchuk, Y.: Development of intellectual system for data de-duplication and distribution in cloud storage. In: Webology, 16(2), pp. 1-42. (2019)
21. Vysotsky, A., Lytvyn, V., Vysotska, V., Dosyn, D., Lyudkevych, I., Antonyuk, N., Naum, O., Vysotskyi, A., Chyrun, L., Slyusarchuk, O.: Online Tourism System for Proposals Formation to User Based on Data Integration from Various Sources. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 92-97. (2019)
22. Lytvyn, V., Vysotska, V., Veres, O., Rishnyak, I., Rishnyak, H.: The Risk Management Modelling in Multi Project Environment.. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 32-35. (2017)
23. Su, J., Sachenko, A., Lytvyn, V., Vysotska, V., Dosyn, D.: Model of Touristic Information Resources Integration According to User Needs. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 113-116. (2018)
24. Lytvyn, V., Vysotska, V., Burov, Y., Demchuk, A.: Architectural ontology designed for intellectual analysis of e-tourism resources. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 335-338. (2018)
25. Antonyuk, N., Vysotsky, A., Vysotska, V., Lytvyn, V., Burov, Y., Demchuk, A., Lyudkevych, I., Chyrun, L., Chyrun, S., Bobyk, I.: Consolidated Information Web Resource for Online Tourism Based on Data Integration and Geolocation. In: Proceedings of the International Conference on Computer Sciences and Information Technologies, CSIT, 15-20. (2019)
26. Artemenko, O., Pasichnyk, V., Kunanets, N., Shunevych, K.: Using sentiment text analysis of user reviews in social media for e-tourism mobile recommender systems. In: Computational Linguistics and Intelligent Systems, COLINS, CEUR workshop proceedings, Vol-2604, 259-271. (2020).
27. Shakhovska, N., Shakhovska, K., Fedushko, S.: Some Aspects of the Method for Tourist Route Creation. In: Advances in Artificial Systems for Medicine and Education II, 902, 527-537. (2019)
28. Antonyuk, N., Medykovskyy, M., Chyrun, L., Dverii, M., Oborska, O., Krylyshyn, M., Vysotsky, A., Tsiura, N., Naum, O.: Online Tourism System Development for Searching and Planning Trips with User’s Requirements. In: Advances in Intelligent Systems and Computing IV, Springer Nature Switzerland AG 2020, 1080, 831-863. (2020)
29. Lozynska, O., Savchuk, V., Pasichnyk, V.: Individual Sign Translator Component of Tourist Information System. In: Advances in Intelligent Systems and Computing IV, Springer Nature Switzerland AG 2020, Springer, Cham, 1080, 593-601. (2020)
30. Savchuk, V., Lozynska, O., Pasichnyk, V.: Architecture of the Subsystem of the Tourist Profile Formation. In: Advances in Intelligent Systems and Computing, 871, 561-570. (2019)

31. Zhezhnych, P., Markiv, O.: Recognition of tourism documentation fragments from web-page posts. In: 14th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET, 948-951. (2018)
32. Zhezhnych, P., Markiv, O.: Linguistic comparison quality evaluation of web-site content with tourism documentation objects. In: Advances in Intelligent Systems and Computing 689, 656-667. (2018)
33. Zhezhnych, P., Markiv, O.: A linguistic method of web-site content comparison with tourism documentation objects. In: International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT, 340-343. (2017)
34. Stoyanova-Doycheva, A., Ivanova, V., Glushkova, T., Stoyanov, S., & Radeva, I.: dynamic generation of cultural routes in a tourist guide. International Journal of Computing, 19(1), 39-48. (2020). http://computingonline.net/computing/article/view/1691
35. Berko, A., Alieksieiev, V.: A Method to Solve Uncertainty Problem for Big Data Sources. In: International Conference on Data Stream Mining and Processing, DSMP, 32-37. (2018)
36. Berko, A.Y., Aliekseyeva, K.A.: Quality evaluation of information resources in web-projects. In: Actual Problems of Economics, 136(10), 226-234. (2012)
37. Berko, A.Y.: Models of data integration in open information systems. In: Actual Problems of Economics, (10), 147-152. (2010)
38. Berko, A.Y.: Methods and models of data integration in E-business systems. In: Actual Problems of Economics (10), 17-24. (2008)
39. Berko, A.: Consolidated data models for electronic business systems. In: The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM, 341-342. (2007)
40. Sachenko, S., Beley O.: The information System of control risks. In: IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Aplications IDAACS, 270-274. (2001)
41. Sachenko, S., Pushkar, M., Rippa, S.: Intellectualization of Accounting System. In: International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 536 – 538. (2007)
42. Sachenko, S., Rippa, S., Krupka, Ya. Pre-Conditions of Ontological Approaches Application for Knowledge Management in Accounting. In: International Workshop on Antelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 605-608. (2009)
43. Sachenko, S., Rippa, S., Golyash, I.: Improving the Information Security Audit of Enterprise Using XML Technologies. In: Inretnational Conference on Intelligent Data Acquisition and Advanced Computing Systems:Technology and Applications, P. 933-937. (2011)
44. Basyuk T.: The Popularization Problem of Websites and Analysis of Competitors. In: Advances in Intelligent Systems and Computing, 689, Springer, Cham, 54-65. (2018)
45. Basyuk, T.: Innerlinking website pages and weight of links. In: International Scientific and Technical Conference on Computer science and information technologies (CSIT), 12-15. (2017)
46. Vasilevskis, E., Dubyak, I., Basyuk, T., Pasichnyk, V., Rzheuskyi, A.: Mobile application for preliminary diagnosis of diseases. In: CEUR Workshop Proceedings, Vol-2255, 275-286. (2018)
47. Basyuk, T.: Popularization of website and without anchor promotion. In: International Scientific and Technical Conference on Computer science and information technologies (CSIT), 193-195. (2016)