ISSN 2710-0766
DOI 10.31891/CSIT

THE INTERNATIONAL SCIENTIFIC JOURNAL

COMPUTER SYSTEMS AND INFORMATION TECHNOLOGIES

No 3-2023

МІЖНАРОДНИЙ НАУКОВИЙ ЖУРНАЛ

КОМП'ЮТЕРНІ СИСТЕМИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

2023

COMPUTER SYSTEMS AND INFORMATION TECHNOLOGIES

INTERNATIONAL SCIENTIFIC JOURNAL

Professional publication of category "B"

Published since 2020 year ***Four time a year***

Khmelnitskyi, 2023, № 3 (12)

Establishers: **Khmelnytskyi National University** (Ukraine)
Associated establisher: **Institute of Information Technologies** (Slovakia)

National Library og Ukraine named after V.I.Vernadsky http://nbuv.gov.ua/j-tit/csit

The journal is included in scientometric databases:

Index Copernicus https://journals.indexcopernicus.com/search/details?id=69998&lang=en
Google Scholar https://scholar.google.com.ua/citations?hl=uk&user=HW1XpMsAAAAJ
CrossRef http://doi.org/10.31891/CSIT

Recommended for publication by the decision of the Academic Council of Khmelnytskyi National University, protocol № 2 from 29.09.2023

Editorial board address: International scientific journal "Computer Systems and Information Technologies", Khmelnytskyi National University, Institutska str. 11, Khmelnytskyi, 29016, Ukraine

☎ **(0382) 67-51-08**

e-mail: csit.khnu@gmail.com

web: http://csitjournal.khmnu.edu.ua/
http://lib.khnu.km.ua/csit_khnu.htm

Registered by the Ministry of Justice of Ukraine
Certificate of state registration of the print media
Series KB № 24924-14864PR dated 12.07.2021

КОМП'ЮТЕРНІ СИСТЕМИ ТА ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

МІЖНАРОДНИЙ НАУКОВИЙ ЖУРНАЛ

Фахове видання категорії "Б". Наказ Міністерства освіти і науки України про затвердження рішень Атестаційної колегії №420 від 19.04.2021 року

Засновано в 2020 р. ***Виходить 4 рази на рік***

Хмельницький, 2023, № 3 (12)

Засновник і видавець: **Хмельницький національний університет** (Україна)
Асоційований співзасновник: **Інститут інформаційних технологій** (Словаччина)

Наукова бібліотека України ім. В.І. Вернадського http://nbuv.gov.ua/j-tit/csit

Журнал включено до наукометричних баз:

Index Copernicus https://journals.indexcopernicus.com/search/details?id=69998&lang=en
Google Scholar https://scholar.google.com.ua/citations?hl=uk&user=HW1XpMsAAAAJ
CrossRef http://doi.org/10.31891/CSIT

Рекомендовано до друку рішенням Вченої ради Хмельницького національного університету, протокол № 2 від 29.09.2023

Адреса редакції: **Україна, 29016,**
м. Хмельницький, вул. Інститутська, 11,
Хмельницький національний університет
редакція журналу "Комп'ютерні системи та інформаційні технології"

☎ **(0382) 67-51-08**

e-mail: csit.khnu@gmail.com

web: http://csitjournal.khmnu.edu.ua/
http://lib.khnu.km.ua/csit_khnu.htm

Зареєстровано Міністерством юстиції України
Свідоцтво про державну реєстрацію друкованого засобу масової інформації
Серія КВ № 24924-14864ПР від 12 липня 2021 року

CONTENTS

ЗМІСТ

https://doi.org/10.31891/csit-2023-3-1
UDC 004.9

Tetiana HOVORUSHCHENKO, Sergii ALEKSOV, Yurii POPOV, Vyacheslav BACHUK
Khmelnytskyi National University

DECISION-MAKING METHOD FOR TEMPERATURE CONTROL IN THE SMART HOME

The current challenge is to provide automatic decision support in a smart home. A study of the top solutions of well-known smart homes has shown that existing solutions usually do not provide for fully automatic control in a smart home, but are focused either on automatic control in conjunction with manual control or user-controlled control. Therefore, the goal of this study is to support decision-making for fully automatic temperature control in a smart home.

Human well-being and performance depend on the meteorological conditions of the environment in which a person is located. The most important condition for high performance, rest, and health is the creation and maintenance of an optimal home microclimate. One of the main parameters of the indoor microclimate is temperature. The room temperature control subsystem ensures the optimal temperature level and allows for individual adjustment for each family member.

The developed rules for determining the optimal room temperature allow you to evaluate the existing temperature parameters for further automatic operation of the smart home temperature control subsystem in residential premises of various types. The purpose of the temperature control subsystem is to provide comfortable conditions in residential premises of various types in terms of their temperature regime.

The developed decision-making method for temperature control in a smart home, which is the basis of the smart home temperature control subsystem, provides a comfortable and optimal (taking into account building and sanitary and hygienic standards) temperature in the corresponding living space.

The results of the functioning of the smart home temperature control decision-making method have shown that the developed method provides for the recognition of situations (optimal temperature, low temperature, high temperature) and support for decision-making on the temperature regime in a certain type of residential space (turning on heating devices, turning on cooling devices, no action, etc.).

Keywords: decision-making support, smart home, temperature control subsystem, air temperature.

Тетяна ГОВОРУЩЕНКО, Сергій АЛЕКСОВ, Юрій ПОПОВ, В’ячеслав БАЧУК
Хмельницький національний університет

МЕТОД ПРИЙНЯТТЯ РІШЕНЬ ЩОДО КЕРУВАННЯ ТЕМПЕРАТУРОЮ У РОЗУМНОМУ БУДИНКУ

Наразі актуальною задачею є забезпечення можливості автоматичної підтримки прийняття рішень у розумному будинку. Дослідження топових рішень відомих розумних будинків показало, що наявні рішення, як правило, не передбачають повністю автоматичного керування у «розумному будинку», а орієнтовані або на автоматичне керування спільно із ручним керуванням, або на кероване користувачем керування. Тому метою даного дослідження є підтримка прийняття рішень щодо повністю автоматичного керування температурою у розумному будинку.

Самопочуття і працездатність людини залежать від метеорологічних умов середовища, в якому вона знаходиться. Найважливішою умовою високої працездатності, відпочинку і здоров’я людини є створення і підтримка оптимального мікроклімату житла. Одним із основних параметрів мікроклімату приміщень виступає температура. Підсистема керування температурою приміщення забезпечує оптимальний рівень температури і дає можливість індивідуального налаштування для кожного члена сім’ї.

Розроблені правила для визначення оптимальності температурного режиму приміщення дозволяють оцінити наявні температурні параметри для подальшого автоматичного функціювання підсистеми керування температурою розумного будинку в житлових приміщеннях різного типу. Метою підсистеми керування температури є забезпечення комфортних умов у житлових приміщеннях різного типу з точки зору їх температурного режиму.

Розроблений метод прийняття рішень щодо керування температурою у розумному будинку, який є основою підсистеми керування температурою розумного будинку, забезпечує комфортну та оптимальну (з врахуванням будівельних та санітарно-гігієнічних норм) температуру у відповідному житловому приміщенні.

Результати функціювання методу прийняття рішень щодо керування температурою у розумному будинку довели, що розроблений метод передбачає розпізнавання ситуацій (оптимальна температура, низька температура, висока температура) та підтримку прийняття рішень щодо температурного режиму у житловому приміщенні певного типу (ввімкнення обігрівальних приладів, ввімкнення охолоджуючих приладів, жодних дій, тощо).

Ключові слова: підтримка прийняття рішень, розумний будинок, підсистема керування температурою, температура повітря.

Introduction

A cyber-physical system is a system that combines physical and digital elements. It can include various components such as sensors, microcontrollers, software, cloud services, and other technologies. Such systems can be applied in various industries, including industry, transportation, medicine, agriculture, and others. This is a clear example of the Internet of Things concept [1].

The cyber-physical system "Smart home" is a combination of smart objects that can facilitate the exchange of information between objects and residents to connect the smart home with the outside world of the Internet. A smart

home is a system that helps to make a home more comfortable and safe. It is a system that combines IoT and IT technologies to control home devices and systems, such as lighting, heating, air conditioning, security, etc. This makes it possible to improve the convenience and safety of living in a home [2-4].

A smart home is a set of solutions that automate everyday activities, relieving the owner of routine tasks. "A smart home is not a set of devices that are controlled remotely, but a single system for managing such devices (an ecosystem) that provides specific benefits to the user, such as visibility of control, convenience, and saving time and effort. Such an ecosystem should perform certain actions in response to specific situations without human intervention [5].

In general, a smart home is a system of devices that includes equipment, sensors, and other elements that can perform various actions both without human intervention (according to developed and programmed scenarios) and on human command [6, 7].

The cyber-physical system "Smart home" is designed to automate and control various devices in the house, such as lighting, heating, temperature, security, etc. The system consists of sensors, controllers, and management software that work together to collect data, analyze it, and make decisions based on the set parameters. As a result, residents of the house receive comfort and security, as well as save energy and reduce utility costs [8, 9].

A smart home should be able to recognize situations, perform certain actions and make the necessary decisions depending on the situation without human intervention [10].

So, *the current challenge* is to provide automatic decision support in a smart home.

Let's consider some top solutions of the known smart homes [11-16]:

1) Amazon Alexa - easy to set up using a mobile application; uses Wi-Fi and Bluetooth as a data transmission method; the most convenient is voice control; supports only English;

2) Google Home - an ecosystem with the well-known and advanced voice assistant Google Assistant; the Google Home mobile application allows you to control devices connected to the assistant from different manufacturers; supports English and Russian; control capabilities are somewhat limited compared to Alexa;

3) Apple HomeKit - to manage such an ecosystem, the Home mobile application is enough, which allows you to quickly integrate different devices and use the user-friendly interface of iOS devices to set up work and create various scenarios; the voice control service is the Siri oboe service; supports English and Russian; has a small number of compatible devices and a rather high cost;

4) Xiaomi Smart Home is an open-type ecosystem based on the ZigBee protocol, which allows you to supplement your smart home with standard-compliant gadgets from any manufacturer, including various switches, security sensors, relays, locks, etc; Wi-Fi enabled devices (smart sockets, video cameras, light bulbs, household appliances, etc.) can be easily added to the ecosystem; it is affordable; modification and gradual expansion are available; the Mi Home app is used for control; the voice assistant supports only Chinese;

5) Ajax - a closed wireless ecosystem of Ukrainian origin designed to ensure home security; "Security System of the Year" at the Security & Fire Excellence Awards; controlled by the Ajax Security System application, which is simple and intuitive; autonomous power supply for up to 15 hours without power supply; simultaneous connection of up to 100-150 devices; supports alarm management by multiple users (up to 50 accounts); detectors are protected against loss of communication and are noise immune; provides not only notification of the owner in case of a security breach, but also automatic notification of the central monitoring station of specialized companies;

6) Nero is a wide-ranging ecosystem; it works on Z-Wave and Intro III wireless protocols; it has a limited selection of compatible devices; it automates the adjustment of the room's microclimate (heating, air conditioning), lighting control, gate control, protection against intrusions, etc.; all control is carried out using the free NeroHome application; it is compatible with most video cameras manufactured by HiWatch and Hikvision; it is easy to set up; it supports the Russian language.

A study of the top solutions of well-known smart homes has shown that existing solutions usually do not provide for fully automatic control in a smart home, but are focused either on automatic control in conjunction with manual control or user-controlled control. Therefore, *the goal of this study* is to support decision-making for fully automatic temperature control in a smart home.

Decision-Making Method for Temperature Control in the Smart Home

Human well-being and performance depend on the meteorological conditions of the environment in which a person is located. The most important condition for high performance, rest, and health is the creation and maintenance of an optimal home microclimate. One of the main parameters of the indoor microclimate is temperature. The room temperature control subsystem ensures the optimal temperature level and allows for individual adjustment for each family member.

Violation of the limits of the thermal regime of the room provokes a deterioration in health and exacerbation of chronic diseases. At high air temperatures, much of the heat is lost through evaporation. Along with sweat, the body loses water, vitamins, and mineral salts, which disrupts metabolism. Over time, this causes an increase in body temperature, increased heart rate, weakening of the cardiovascular system, decreased gastrointestinal activity, etc. This is accompanied by headaches, malaise, decreased attention and coordination of movements, and decreased

performance. Thus, performance at 24 degrees decreases by 15%, and at 28 degrees – by 30%.

Low indoor temperatures create all the conditions for the emergence and exacerbation of respiratory diseases (rhinitis, bronchitis, pleurisy, pneumonia), musculoskeletal diseases and diseases of the peripheral nervous system (myositis, rheumatism, neuritis, radiculitis), as well as exacerbation of other chronic diseases. In low air temperature conditions, there is a risk of hypothermia due to increased heat transfer. Prolonged and even short-term exposure to cold causes a variety of reflex reactions of a general and local nature. They affect not only the areas directly affected by the cold, but also distant parts of the body. For example, cooling of the feet causes a decrease in the temperature of the nasal and throat mucosa, which leads to a decrease in local immunity and the appearance of a runny nose, cough, and sore throat.

The regulatory document in this regard is SSN 3.3.6.042-99 "Sanitary norms of microclimate of industrial premises" and SBN B.2.5-67:2013 "Heating, ventilation and air conditioning" specify temperature indicators depending on the type of premises.

According to the state building and sanitary & hygienic norms and standards, the temperature level for premises where people live and work should be within the permissible range of 16.5º to 26.5º, depending on the type of premises and seasons of the year – Table 1.

Table 1

Air temperature standards in residential premises of different types

Type of room	Optimal air temperature during the heating period (October 15 - April 14)	Optimal air temperature during the cooling period (April 15 - October 14)
Bedroom, living room, dining room, office room	20-24º	23-26º
Kitchen, dressing room	16.5-22.5º	16.5-22.5º
Bathroom	23.5-26.5º	23.5-26.5º

Let's develop *the rules for determining the optimal room temperature:*

1) if the current date (variable *d*) is in the range [15.10; 14.04] and the room type is "bedroom" and the room temperature (variable *t*) is in the range [20; 24], then the room temperature is optimal, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "bedroom" and $t < 20$, then the room temperature is low, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "bedroom" and $t > 24$, then the room temperature is high;

2) if *d* $\in$ [15.10; 14.04] and the room type is "living room" and *t* $\in$ [20; 24], then the room temperature is optimal, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "living room" and $t < 20$, then the room temperature is low, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "living room" and $t > 24$, then the room temperature is high;

3) if *d* $\in$ [15.10; 14.04] and the room type is "dining room" and *t* $\in$ [20; 24], then the room temperature is optimal, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "dining room" and $t < 20$, then the room temperature is low, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "dining room" and $t > 24$, then the room temperature is high;

4) if *d* $\in$ [15.10; 14.04] and the room type is "office room" and *t* $\in$ [20; 24], then the room temperature is optimal, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "office room" and $t < 20$, then the room temperature is low, otherwise if *d* $\in$ [15.10; 14.04] and the room type is "office room" and $t > 24$, then the room temperature is high;

5) if *d* $\in$ [15.04; 14.10] and the room type is "bedroom" and *t* $\in$ [23; 26], then the room temperature is optimal, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "bedroom" and $t < 23$, then the room temperature is low, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "bedroom" and $t > 26$, then the room temperature is high;

6) if *d* $\in$ [15.04; 14.10] and the room type is "living room" and *t* $\in$ [23; 26], then the room temperature is optimal, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "living room" and $t < 23$, then the room temperature is low, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "living room" and $t > 26$, then the room temperature is high;

7) if *d* $\in$ [15.04; 14.10] and the room type is "dining room" and *t* $\in$ [23; 26], then the room temperature is optimal, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "dining room" and $t < 23$, then the room temperature is low, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "dining room" and $t > 26$, then the room temperature is high;

8) if *d* $\in$ [15.04; 14.10] and the room type is "office room" and *t* $\in$ [23; 26], then the room temperature is optimal, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "office room" and $t < 23$, then the room temperature is low, otherwise if *d* $\in$ [15.04; 14.10] and the room type is "office room" and $t > 26$, then the room temperature is high;

9) if the room type is "kitchen" and *t* $\in$ [16.5; 22.5], then the room temperature is optimal, otherwise if the room type is "kitchen" and $t < 16.5$, then the room temperature is low, otherwise if the room type is "kitchen" and $t > 22.5$, then the room temperature is high;

10) if the room type is "dressing room" and *t* $\in$ [16.5; 22.5], then the room temperature is optimal, otherwise if the room type is "dressing room" and $t < 16.5$, then the room temperature is low, otherwise if the room type is "dressing room" and $t > 22.5$, then the room temperature is high;

11) if the room type is "bathroom" and *t* $\in$ [23.5; 26.5], then the room temperature is optimal, otherwise if the room type is "bathroom" and $t < 23.5$, then the room temperature is low, otherwise if the room type is "bathroom" and $t > 26.5$, then the room temperature is high.

The developed rules for determining the optimal room temperature allow you to evaluate the existing temperature parameters for further automatic operation of the smart home temperature control subsystem in residential

premises of various types. The purpose of the temperature control subsystem is to provide comfortable conditions in residential premises of various types in terms of their temperature regime.

The decision-making method for temperature control in the smart home, which is the basis of the smart home temperature control subsystem, consists of the following steps:

1) the user of the subsystem checks the correctness of the current date: if the date is incorrect, then sets the start date; then the current date is updated automatically in the smart home;

2) the user of the temperature control subsystem selects the type of room (bedroom, living room, dining room, office room, kitchen, dressing room, bathroom);

3) measuring the room temperature with a temperature sensor (for example, Sonoff DS18B20 or Connect Home-s01);

4) searching for a rule in a set of the rules for determining the optimal room temperature;

5) if, according to the found rule, the air temperature in the room is optimal, the temperature control subsystem does not perform any additional actions;

6) if, according to the found rule, the air temperature in the room is low, the temperature control subsystem should heat the air, for which it turns on the available heating devices (boiler, fireplace, air conditioner in heating mode, etc.);

7) if, according to the found rule, the air temperature in the room is high, the temperature control subsystem should cool the air, for which it turns on the available cooling devices (fan, air conditioner in cooling mode, etc.);

8) switching to step 3 after 30 minutes to ensure a constant response to changes in the room temperature.

The developed decision-making method for temperature control in a smart home, which is the basis of the smart home temperature control subsystem, provides a comfortable and optimal (taking into account building and sanitary and hygienic standards) temperature in the corresponding living space.

Example of Functioning the Decision-Making Method for Temperature Control in the Smart Home

According to the first two steps of the decision-making method for temperature control in the smart home, the user sets the current date (20.11) and selects the room type "office room".

The third step of the method involves measuring the air temperature in the room with a temperature sensor ($t=16^o$). Next, a rule is searched for in the rules set to determine the optimal room temperature. Since $d \in$ [15.10; 14.04] and the type of room is "office room" and $t < 20$, the air temperature in the room is low. Since, according to the found rule, the air temperature in the room is low, the temperature control subsystem should heat the air, for which it turns on the existing heating devices (boiler and fireplace).

After 30 minutes, step 3 of the decision-making method for temperature control in the smart home is repeated – the temperature of the air in the room is measured again with the temperature sensor ($t=19^o$). Next, a rule is searched for in the rules set to determine the optimal room temperature. Since $d \in$ [15.10; 14.04] and the type of room is "office room" and $t < 20$, the air temperature in the room is low. Since, according to the found rule, the air temperature in the room is low, the temperature control subsystem should heat the air, so the existing heating devices (boiler and fireplace) remain on.

After 30 minutes, step 3 of the decision-making method for temperature control in the smart home is repeated – the temperature of the air in the room is measured again with the temperature sensor ($t=22^o$). Next, a rule is searched for in the rules set to determine the optimal room temperature. Since $d \in$ [15.10; 14.04] and the type of room is "office room" and $t \in$ [20; 24], the air temperature in the room is optimal. Since, according to the found rule, the air temperature in the room is optimal, the temperature control subsystem should not perform any additional actions (heating or cooling) at this time, and the heating devices are turned off.

After 30 minutes, step 3 of the decision-making method for temperature control in the smart home is repeated – the temperature of the air in the room is measured again with the temperature sensor ($t=21^o$). Next, a rule is searched for in the rules set to determine the optimal room temperature. Since $d \in$ [15.10; 14.04] and the type of room is "office room" and $t \in$ [20; 24], the air temperature in the room is optimal. Since, according to the found rule, the air temperature in the room is optimal, the temperature control subsystem should not perform any additional actions (heating or cooling) at this time

After 30 minutes, step 3 of the decision-making method for temperature control in the smart home is repeated – the temperature of the air in the room is measured again with the temperature sensor ($t=20^o$). Next, a rule is searched for in the rules set to determine the optimal room temperature. Since $d \in$ [15.10; 14.04] and the type of room is "office room" and $t \in$ [20; 24], the air temperature in the room is optimal. Since, according to the found rule, the air temperature in the room is optimal, the temperature control subsystem should not perform any additional actions (heating or cooling) at this time

After 30 minutes, step 3 of the decision-making method for temperature control in the smart home is repeated – the temperature of the air in the room is measured again with the temperature sensor ($t=19^o$). Next, a rule is searched for in the rules set to determine the optimal room temperature. Since $d \in$ [15.10; 14.04] and the type of room is "office room" and $t < 20$, the air temperature in the room is low. Since, according to the found rule, the air temperature in the room is low, the temperature control subsystem should heat the air, for which it turns on the existing heating devices

(boiler and fireplace).

The results of the functioning of the smart home temperature control decision-making method have shown that the developed method provides for the recognition of situations (optimal temperature, low temperature, high temperature) and support for decision-making on the temperature regime in a certain type of residential space (turning on heating devices, turning on cooling devices, no action, etc.).

Conclusions

The current challenge is to provide automatic decision support in a smart home. A study of the top solutions of well-known smart homes has shown that existing solutions usually do not provide for fully automatic control in a smart home, but are focused either on automatic control in conjunction with manual control or user-controlled control. Therefore, the goal of this study is to support decision-making for fully automatic temperature control in a smart home.

Human well-being and performance depend on the meteorological conditions of the environment in which a person is located. The most important condition for high performance, rest, and health is the creation and maintenance of an optimal home microclimate. One of the main parameters of the indoor microclimate is temperature. The room temperature control subsystem ensures the optimal temperature level and allows for individual adjustment for each family member.

The developed rules for determining the optimal room temperature allow you to evaluate the existing temperature parameters for further automatic operation of the smart home temperature control subsystem in residential premises of various types. The purpose of the temperature control subsystem is to provide comfortable conditions in residential premises of various types in terms of their temperature regime.

The developed decision-making method for temperature control in a smart home, which is the basis of the smart home temperature control subsystem, provides a comfortable and optimal (taking into account building and sanitary and hygienic standards) temperature in the corresponding living space.

The results of the functioning of the smart home temperature control decision-making method have shown that the developed method provides for the recognition of situations (optimal temperature, low temperature, high temperature) and support for decision-making on the temperature regime in a certain type of residential space (turning on heating devices, turning on cooling devices, no action, etc.).

References

1. Fuada S., Hendriyana H. UPISmartHome V.2.0 – A Consumer Product of Smart Home System with an ESP8266 as the Basis. Journal of Communications. 2022. P. 541–552.
2. Optimal Energy Scheduling of Appliances in Smart Buildings Based on Economic and Technical Indices / I. Muda et al. Environmental and Climate Technologies. 2022. Vol. 26, no. 1. P. 561–573.
3. Smart Home Personal Assistants: Fueled by Natural Language Processor and Blockchain Technology / S. A. Ansar et al. 2022 Second International Conference on Interdisciplinary Cyber Physical Systems (ICPS), Chennai, India, 9–10 May 2022. 2022.
4. Sung W.-T., Hsiao S.-J. Creating Smart House via IoT and Intelligent Computation. Intelligent Automation & Soft Computing. 2023. Vol. 35, no. 1. P. 415–430.
5. Yaici W., Entchev E., Longo M. Internet of Things (IoT)-Based System for Smart Home Heating and Cooling Control. 2022 IEEE International Conference on Environment and Electrical Engineering and 2022 IEEE Industrial and Commercial Power Systems Europe (EEEIC / I&CPS Europe), Prague, Czech Republic, 28 June – 1 July 2022. 2022.
6. Development of smart application for house condition survey / A. Hyder Chohan et al. Ain Shams Engineering Journal. 2022. Vol. 13, no. 3. P. 101628.
7. Energy Savings in Buildings Based on Image Depth Sensors for Human Activity Recognition / O. Mata et al. Energies. 2023. Vol. 16, no. 3. P. 1078.
8. Tunable White Light for Elders (TWLITE): A Protocol Demonstrating Feasibility and Acceptability for Deployment, Remote Data Collection, and Analysis of a Home-Based Lighting Intervention in Older Adults / J. E. Elliott et al. Sensors. 2022. Vol. 22, no. 14. P. 5372.
9. Ayu Lestari R., Yusmaniar Oktiawati U. Full state feedback and feed forward control of servo smart window using MATLAB/Simulink. Indonesian Journal of Electrical Engineering and Computer Science. 2022. Vol. 28, no. 3. P. 1355.
10. Smith N. Smart Bee Houses: Designing to Support Urban Pollination. ACI'21: Eight International Conference on Animal-Computer Interaction, Bloomington IN USA. New York, NY, USA, 2021.
11. He Y., Tian J., Cao Y. Intelligent home temperature and light control system based on the cloud platform. 2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 15–17 April 2022. 2022.
12. Kumar T., Srinivasan R., Mani M. An Emergy-based Approach to Evaluate the Effectiveness of Integrating IoT-based Sensing Systems into Smart Buildings. Sustainable Energy Technologies and Assessments. 2022. Vol. 52. P. 102225.
13. Nchena L. Smart House Assistive Technologies for Senior Citizens. 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), Ruzomberok, Slovakia, 26–28 September 2022. 2022.
14. OTP-Based Smart Door Opening System / P. Srinivasan et al. Intelligent Communication Technologies and Virtual Mobile Networks. Singapore, 2022. P. 87–98.
15. Saha A., Das P. S., Banik B. C. Smart Green House for Controlling & Monitoring Temperature, Soil & Humidity Using IOT. 2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP), Vijayawada, India, 12–14 February 2022. 2022.
16. Smart sensors network for accurate indirect heat accounting in apartment buildings / Y. Stauffer et al. Journal of Building Engineering. 2022. Vol. 46. P. 103534.

Tetiana Hovorushchenko **Тетяна Говорущенко**	DrSc (Engineering), Professor, Head of Computer Engineering & Information Systems Department, Khmelnytskyi National University https://orcid.org/0000-0002-7942-1857 E-mail: govorushchenko@gmail.com	Доктор технічних наук, професор, завідувач кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Sergii Aleksov **Сергій Алексов**	PhD student of Computer Engineering & Information Systems Department, Khmenlnytskyi National University https://orcid.org/0000-0001-8764-675X E-mail: aleksov@gmail.com	Аспірант кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Yurii Popov **Юрій Попов**	PhD Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University E-mail: popov@gmail.com	Аспірант кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Vyacheslav Bachuk **В'ячеслав Бачук**	PhD Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University E-mail: v.bachuk@gmail.com	Аспірант кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет

https://doi.org/10.31891/csit-2023-3-2
UDC 004.032.26

Yevhen KHOLIAVKA, Yuliia PARFENENKO
Sumy State University

FORECASTING PEAK LOAD ON THE POWER GRID

In the modern world, precise forecasting of peak electricity consumption stands as a pivotal pillar in the efficient management of power grids. The paramount importance of this task necessitates a comprehensive examination of various forecasting methodologies, leveraging hourly electricity consumption data and a diverse array of predictive models.

This article is dedicated to a thorough analysis of distinct peak load forecasting methods, elucidating the research methodology encompassing data preprocessing, model selection, and parameter optimization. The models under scrutiny encompass a spectrum of techniques, including ARIMA, SARIMA, LSTM, GRU, and Random Forest. To gauge their performance, a suite of evaluation metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), R-squared, and Receiver Operating Characteristic Area Under the Curve (ROC AUC) were employed.

The findings of this investigation underscore the nuanced strengths and limitations inherent to each forecasting model when tasked with predicting peak electricity consumption. Notably, certain approaches exhibit superior accuracy in short-term forecasting scenarios, while others excel in long-term predictions. The selection of the optimal forecasting method becomes contingent upon the specific conditions, constraints, and objectives of the study at hand.

The LSTM and GRU models, representing deep learning neural networks, manifest their prowess in addressing the intricate dynamics of electricity consumption data. Their capacity to discern intricate patterns, nonlinearities, and long-term dependencies positions them as formidable contenders in the domain of long-term peak consumption forecasting.

The Random Forest model emerges as a versatile choice, adept at accommodating the multifaceted characteristics of electricity consumption data. Its ability to autonomously identify complex dependencies, nonlinear relationships, and seasonal patterns while considering external factors amplifies its utility across a broad spectrum of forecasting scenarios.

This comprehensive work is of great importance for the practical study of various methods of forecasting peak electricity consumption. The results obtained from this analysis have significant implications for improving power grid management strategies, ultimately contributing to microgrid stability and resilience.

Keywords: forecasting, peak consumption, electricity, clean energy, Random Forest, neural networks.

Євген ХОЛЯВКА, Юлія ПАРФЕНЕНКО
Сумський державний університет

ПРОГНОЗУВАННЯ ПІКОВОГО НАВАНТАЖЕННЯ НА ЕЛЕКТРИЧНІ МЕРЕЖІ

У сучасному світі точний прогноз пікового споживання електроенергії використовується як одне з ключових значень для ефективного управління електромережами. Вирішення цієї задачі вимагає ретельного розгляду різних методів прогнозування, використовуючи годинні дані споживання електроенергії та різні моделі прогнозування.

Ця стаття присвячена аналізу різних методів прогнозування пікового навантаження, використовуючи дослідницьку методологію, що включає обробку даних, вибір моделі та оптимізацію параметрів. Моделі, що розглядаються, охоплюють широкий спектр методів прогнозування, включаючи ARIMA, SARIMA, LSTM, GRU та Random Forest. Для оцінки їх ефективності було використано низку метрик оцінки, таких як середня абсолютна помилка (MAE), коренева середня квадратична помилка (RMSE), середня абсолютна відсоткова помилка (MAPE), R-квадрат та площа під кривою характеристики отримувача (ROC AUC).

Результати цього дослідження підкреслюють сильні та слабкі сторони кожної моделі прогнозування споживання електроенергії. Зокрема, деякі підходи проявляють себе з більшою точністю в короткострокових сценаріях прогнозування пікового споживання електроенергії, тоді як інші відзначаються в довгострокових прогнозах. Вибір оптимального методу прогнозування стає залежним від конкретних умов, обмежень та цілей конкретного дослідження.

Моделі LSTM та GRU, що представляють собою нейронні мережі глибокого навчання, проявляють свою ефективність в розгляді складних динамік даних щодо споживання електроенергії. Їх здатність розпізнавати патерни, нелінійності та довгострокові залежності робить їх потужними конкурентами в області довгострокового прогнозування піку споживання.

Модель Random Forest виявилась універсальною, здатною адаптуватися до багатогранних характеристик даних споживання електроенергії. Її здатність автономно визначати складні залежності, нелінійні відносини та сезонні патерни, враховуючи зовнішні фактори підвищує її корисність в широкому спектрі сценаріїв прогнозування.

Ця робота має значення для практичного вивчення різних методів прогнозування піку споживання електроенергії. Результати, отримані з цього аналізу, мають значні значення для вдосконалення стратегій управління електромережами, що в кінцевому рахунку сприяє стабільності та стійкості енергетичної мікромережі.

Ключові слова: прогнозування, пікове споживання, електроенергія, чиста енергетика, Random Forest, нейронні мережі.

Introduction

Forecasting peak electricity consumption is of great importance for effective planning and management of the energy microgrid. This allows early detection of periods of high consumption and adaptation of resources to ensure the best response to changes in demand.

Planning and management of the energy microgrid involves using resources in an efficient way, minimizing costs and ensuring the stability of energy supply. Overtime forecasting helps determine the need for backup resources during peak load periods, helping to maintain network resilience and avoid supply failures.

Peak electricity consumption is one of the key characteristics of the energy system, which reflects the maximum load on the electrical network during a certain period of time. This characteristic is critically important for proper planning and management of power supply, energy security, and efficiency of the power system [1].

With the introduction of smart meters and the development of collective data technologies, more objective and detailed data on electricity consumption became available [2]. These data can be used to develop predictive models that provide accurate forecasts of peak consumption [3]. Forecasting peak consumption is of great importance to energy companies, allowing them to effectively plan electricity production, avoid congestion and ensure reliable and stable electricity supply.

This article is devoted to a comprehensive exploration of diverse techniques and strategies employed in predicting peak electricity consumption. It encompasses both conventional statistical methods and contemporary advancements rooted in machine learning and deep learning. The primary aim of this study is to scrutinize and identify effective methodologies for forecasting peak consumption, further facilitating a comprehensive comparison of their respective performance and accuracy.

Furthermore, an ensemble of models was taken into account, encompassing Random Forest, a technique that amalgamates multiple models to enhance prediction accuracy [4]. Moreover, specific libraries tailored for time series analysis were employed, such as Neural Prophet, leveraging neural networks for predictive purposes.

Subsequent sections delve into an in-depth exploration of the employed methods, their performance assessment, and a comparative analysis. The objective is to advance the cause of efficient and stable electricity provision by means of predictive examination of peak electricity consumption.

The study delves into the prediction of peak electricity consumption by analyzing a time series dataset spanning four years, which includes hourly electricity consumption figures. The primary objective is to forecast upcoming peak consumption values based on historical data, employing a range of forecasting methods. This endeavor aims to enhance the efficiency and reliability of power system operations.

Related works

The forecasting of peak electricity consumption has gained importance in recent years due to its importance for the management of energy processes and the stability of the power grid. Time series forecasting, including the energy sector, uses a variety of methods. These methods can be classified into three categories: statistical methods, machine learning methods, and deep learning methods [4]. Among the statistical methods considered are ARIMA, SARIMA and ETS, which are based on the analysis of trends and seasonality. Machine learning techniques include Random Forest, Gradient Boosting, SVM, and k-NN, which are used to detect complex dependencies in data [5]. Deep learning techniques such as LSTM, GRU, and 1D CNN are able to interact with sequential data and capture long-term dependencies [3]. This section delves into a survey of research that has investigated different methodologies for peak load forecasting.

The article [6] considers a wide range of methods and approaches to load forecasting in energy systems using smart networks. The authors review the literature on load forecasting and highlight the main trends and challenges related to this area. The article examines various methods, including statistical approaches, machine learning methods, and artificial neural networks, their advantages and disadvantages. In addition, the authors of the paper analyze important factors affecting load forecasting, such as weather, seasonality, geographic and social aspects. They also consider the implementation of smart grid technologies in the load forecasting process and emphasize the importance of accurate forecasting to ensure grid efficiency and reliability.

In the work [7], a data-driven approach for load forecasting in smart grids is proposed. The approach combines statistical and machine learning methods to predict load demand. The authors employ techniques like autoregressive integrated moving average (ARIMA), exponential smoothing, support vector machines (SVM), and artificial neural networks (ANN) to enhance load prediction accuracy.

Another study is presented in [8], introduces a novel load forecasting method for smart grids. This method relies on deep learning using long-short-term memory (LSTM) to simulate dynamic load changes. The authors demonstrate significant improvements in prediction accuracy compared to traditional methods.

The research [9] presents a load forecasting technique for smart grids based on cloud computing and LSTM neural networks. This approach offers enhanced prediction accuracy, scalability for large grids, and adaptability to varying conditions.

Moreover, the article [4] delves into the application of Kalman and filtered Monte Carlo methods for load forecasting. By analyzing unlinked time series models, the authors forecast peak and total electricity demand. Utilizing data containing peak demand and electricity production information, they observe consumption trends, identify outliers, and establish inter-day relationships.

These articles collectively contribute to the field of load forecasting for smart grids, introducing advanced methods to improve accuracy, flexibility, and efficiency in predicting electricity demand.

The obtained results indicate the convergence of the Monte Carlo and Gibbs methods Sampling when estimating model parameters, in particular covariances. The authors analyze changes in covariances between different

components of the model and indicate correlations between different days of electricity consumption. They also emphasize the dynamics of autocorrelation, which indicates a relationship between days and electricity demand.

Methodology

To achieve the goal of forecasting peak electricity consumption, the following methodology is proposed:

- Data Preparation and Research:
  - Collect historical electricity consumption data, including hourly values, for four years.
  - Conduct data analysis to understand distribution, trends, seasonality, and possible anomalies in the data set.
  - Preprocess the data by addressing issues with missing values, outliers, and feature normalization.
- Selection and Engineering Features:
  - Identify relevant attributes that may affect peak power consumption, such as 'month', 'day_of_week', 'day_length' and 'night'.
  - Create additional features that can reflect patterns or variations in energy consumption.
- Selection of Models:
  - Use a variety of predictive models, both traditional and machine learning-based, to account for different aspects of the time series.
  - Selected models include ARIMA, SARIMA, LSTM, GRU, NARX and ensemble models such as Random Forest and Gradient Boosting.
- Model Training and Evaluation:
  - Divide the data set into training and test sets. For time series, it is important to apply chronological separation to simulate real-world conditions.
  - Train each selected model on the training set and tune the hyper parameters as needed.
  - Evaluate the performance of models on the test set using appropriate evaluation metrics such as MAE, RMSE, and MAPE.
- Detection and Treatment of Emissions:
  - Apply anomaly detection techniques to identify unusual patterns or outliers in data that may affect forecasting accuracy.
  - Resolve detected anomalies through data imputation or by considering their impact during the modeling process.
- Ensemble Approaches:
  - Explore ensemble methods to combine predictions from multiple models to improve the accuracy and reliability of peak energy demand forecasting.
  - Evaluate the performance of ensemble models using metrics such as F1-Score, Precision-Recall, and ROC-AUC.
- Visualization and Interpretation:
  - Use visualization libraries such as Matplotlib to create visual representations of forecasting results, comparing predicted values with actual consumption.
  - Analyze patterns and insights from visualizations to make informed decisions.

The proposed methodology aims to use a combination of traditional time series forecasting models, machine learning algorithms and ensemble methods to achieve accurate and reliable forecasting of peak energy consumption.

To effectively forecast peak electricity consumption, a dataset with hourly electricity consumption metrics and consistent weather-related features is essential. The dataset should ideally be free of gaps or missing values to ensure accurate predictions. These hourly measurements provide the necessary granularity to capture fluctuations in electricity demand, while the weather-related attributes contribute to understanding external factors that influence consumption patterns.

Choosing an appropriate model for forecasting peak electricity consumption is a key task in research, as the effectiveness and accuracy of predictions depends on its correctness. In this section, an in-depth examination of diverse methods and models utilized for peak performance prediction will be conducted. This analysis will encompass their merits, drawbacks, and domains of applicability.

ARIMA (Autoregressive Integrated Moving Average) and SARIMA (Seasonal ARIMA) [5] are popular time series forecasting methods. They are based on a combination of autoregressive (AR), moving average (MA) and integrated (I) models. SARIMA includes a seasonal component to ARIMA.

The ARIMA model uses three parameters [13]: *p, d,* q, where:

- *p* is the degree of autoregression (the number of previous observations to be included in the model).
- *d* is the order of differentiation (how many times it is necessary to take the difference between consecutive observations to make the series stationary).

– q is the degree of the moving average (the number of previous forecasting errors to be included in the model).

The ARIMA model can be represented by the formula:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (1)$$

where: $Y_t$is the value of the time series at time *t*,
c is a constant component,
ϕ_i– autoregression coefficients,
$\phi_{t-i}$is the value of the time series at previous time points,
θ_i– moving average coefficients,
$\varepsilon_t$is the prediction error at time *t*.

The SARIMA model includes an additional seasonal component [13], which allows simulating seasonal changes in the time series. For this, the SARIMA model has three more parameters: *P, D, Q*, and *s*, where:

– P is the degree of seasonal autoregression,
– D is the order of seasonal differentiation,
– Q is the degree of the seasonal moving average,
– s is the period of seasonality (the number of observations per seasonal cycle).

The SARIMA model can be represented by the formula:

$$\begin{aligned} Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \\ + \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2s} + \cdots + \Phi_p Y_{t-Ps} + \Theta_1 \varepsilon_{t-s} + \Theta_2 \varepsilon_{t-2s} + \cdots + \Theta_Q \varepsilon_{t-Qs} + \varepsilon_t, \end{aligned} \quad (2)$$

where: Φ_i– coefficients of seasonal autoregression ,
Y_{t-s} is the value of the time series with observations separated by *s* (seasonal lag),
Θ_i– seasonal moving average coefficients,
ε_{t-s} is the prediction error at the moment of time *t*–*s*.

Both of these models help to analyze and forecast time series taking into account autocorrelation, seasonality and changes in the time series.

After exploring autoregressive models for forecasting peak electricity consumption, let's turn our attention to the use of more complex and powerful neural network architectures, in particular LSTM, which allow us to better avoid the limitations of traditional approaches and obtain more accurate and realistic forecasts.

LSTM model (Long Short-Term Memory) is a subtype of recurrent neural networks designed to process and model data sequences such as time series [11]. One of the key advantages of LSTM is its ability to efficiently deal with long-term dependencies in data. LSTM includes special mechanisms for storing, retrieving, and updating information from previous time steps. The basic idea is to use an internal state that can store information for a long period of time, and use gates to adjust the internal state and output the information to the outer layer.

The LSTM structure includes the following components [14]:

– Forget Gate:

The building gate, responsible for deciding which information from the previous state should be forgotten, is characterized by the formula (3) used to construct these gates.

$$f_t = \sigma\big(W_f \cdot [h_{t-1}, x_t] + b_f\big), \quad (3)$$

where:
$f_t$is the building gate vector at step *t*,
W_f– matrix of weights,
$h_{t-1}$is the vector of the hidden state in the previous step,
$x_t$is the input vector at step *t*,
b_f- displacement,
σ– activation function (sigmoid).

– Input Gate:

The update gate, which determines the incorporation of new information into the internal state, is defined by the formula (4) for the update gate.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \quad (4)$$

where:
$i_t$is the update gate vector at step *t*,

W_i– matrix of weights,
$h_{t-1}$is the vector of the hidden state in the previous step,
$x_t$is the input vector at step *t*,
b_i- displacement.

– New Cell State:

The formula for the new internal state is:

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C), \quad (5)$$

where:
$\tilde{C}_t$is the new internal state at step *t*,
W_C– matrix of weights,
$h_{t-1}$is the vector of the hidden state in the previous step,
$x_t$is the input vector at step *t*,
b_C- displacement,
$tanh$– activation function (hyperbolic tangent).

– Output Gate:

The output gate, responsible for selecting the output from the internal state, is defined by the formula (6) for the output gate.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \quad (6)$$

where:
$o_t$is the output gate vector at step *t*,
W_o– matrix of weights,
$h_{t-1}$is the vector of the hidden state in the previous step,
$x_t$is the input vector at step *t*,
b_o- displacement.

– Hidden State:

The formula for calculating the new hidden state at step t is:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t,$$
$$h_t = o_t \cdot \tanh(C_t), \quad (7)$$

where:
$C_t$is the new internal state at step *t*,
$f_t$is the building gate vector at step *t*,
$C_{t-1}$is the internal state at the previous step *t*,
$i_t$is the update gate vector at step *t*,
$\tilde{C}_t$,is the new internal state at step *t*,
$o_t$is the output gate vector at step *t*,
$h_t$is the new hidden state at step *t.*

LSTM can be applied to predict peak load by learning from historical data and using the acquired knowledge to predict future values. Its ability to model long-term dependencies and account for a variety of input parameters makes it a powerful tool for time series analysis and forecasting.

After a detailed consideration of LSTM, it is worth turning to another important type of recurrent neural networks - Gated Recurrent Unit (GRU). Following this, the Nonlinear Auto Regressive model with exogenous inputs (NARX) [13] will be discussed, offering efficient modeling and forecasting of time series while accounting for external influences.

Gated Recurrent Unit (GRU) is an improved version of LSTM that has fewer parameters and may be less prone to overtraining on small datasets [15]. The GRU also uses gates to control the flow of information. It has two gates: the update gate (update gate) and priority gate (reset gate). An update gate decides what information should be transferred to a future state, while a preference gate helps decide what information should be forgotten from a previous state.

Evaluating the performance of forecasting models is a critical step in the process of developing forecasting algorithms, which helps determine how well the model fits real data and how accurately it can predict future values.

For the evaluation of forecasting models, an initial step involves partitioning the accessible data into distinct training and test subsets. Within this framework, the training set assumes the role of facilitating model training, essentially fine-tuning its parameters in accordance with the input data. Conversely, the test set is integral in gauging the predictive precision of the model when applied to novel data instances that remain unfamiliar to the model.

Various metrics are employed to evaluate the precision and effectiveness of a model's forecasting performance, with the selection of appropriate metrics contingent upon the specific characteristics of the forecasting task at hand [16]. These evaluation metrics serve as essential tools for quantifying the level of agreement between the predicted outcomes and the actual observations, thereby shedding light on the model's capability to capture underlying patterns, trends, and fluctuations within the data [13]. For instance, popular metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE) are commonly utilized to measure the extent of deviation between predicted values and observed data.

The selection of evaluation metrics may vary depending on whether the task involves point forecasts, interval forecasts, probabilistic forecasts, or the assessment of accuracy across multiple time horizons. These metrics play a pivotal role in enabling researchers and practitioners to comprehensively gauge the quality of forecasting results, allowing for informed decisions, model comparisons, and the identification of potential areas for refinement in the predictive models under investigation. Graphs can be effectively employed to visually represent the outcomes of estimation, encompassing elements like comparative plots showcasing forecasted versus actual values, as well as graphical representations of error distributions. These visual aids play a pivotal role in conveying the extent of alignment or divergence between projected and observed data points. In the selection process of an optimal model for a particular peak load forecasting endeavor, meticulous consideration should be given to identifying the paramount metric that aligns with the primary objectives of the task.

Experiments

This section provides a comprehensive overview of the conducted experiments to assess various forecasting techniques on the dataset for peak electricity consumption. The dataset encompasses hourly records of electricity usage from a two-story building situated in Houston, Texas, USA. The temporal span of the data spans from June 01, 2016 to August 2020.

The dataset contains a variety of parameters, including electricity consumption in kWh, as well as notes indicating the type of day (working, weekend, quarantine due to COVID, holiday). In addition, the dataset contains information about weather conditions, including temperature, humidity, pressure, etc.

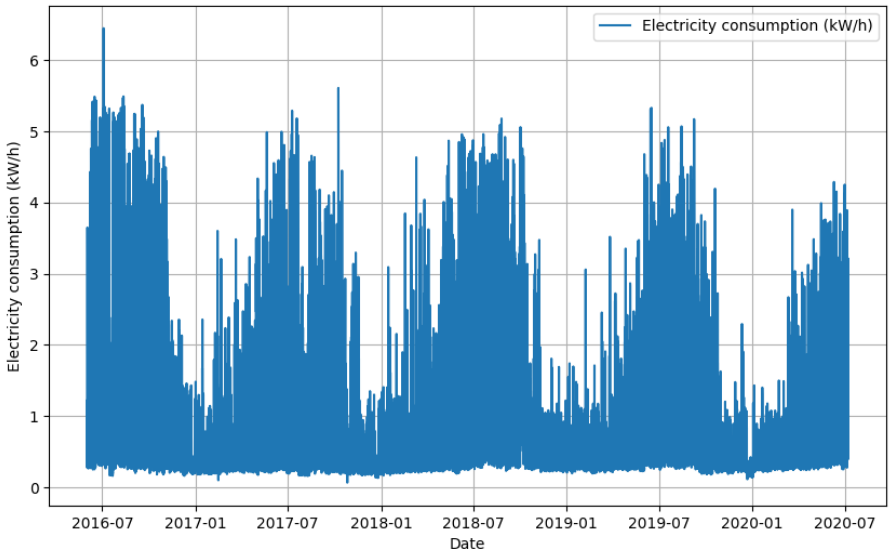

Fig. 1. Electricity consumption data set

The next stage was data processing. It included removing possible anomalies and missing values, normalizing the data and grouping it according to some parameters, such as days of the week, time intervals, etc. In addition, work was carried out to combine data on electricity consumption and weather conditions to create a connection between these factors. According to research, the peak consumption of electricity in a private house is usually observed in the evening period from 17:00 to 21:00. During this time, households actively use electrical appliances for cooking, lighting, working with electronics, as well as for the comfortable use of air conditioners and other appliances.

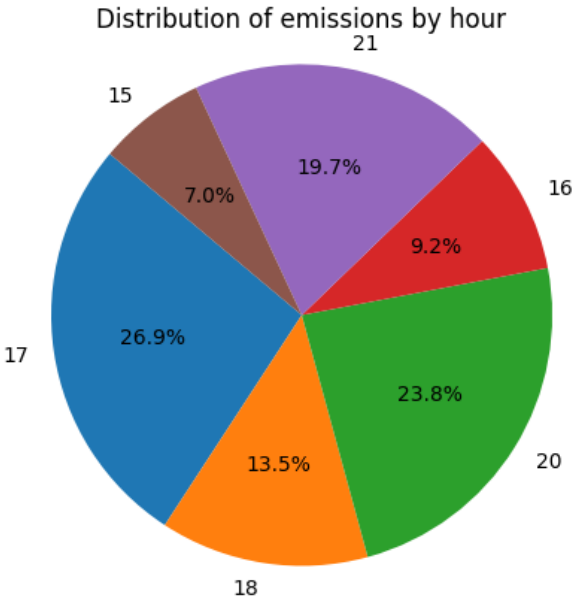

Fig. 2. The largest amount of peak value of electricity consumption

For this purpose, a range of forecasting techniques was employed, encompassing both traditional models like ARIMA and SARIMA, and more advanced approaches like LSTM and GRU.During the experiments, the dataset was split into training and test subsets. Each selected prediction model was trained on the training data, and model-specific techniques were employed to optimize and fine-tune hyperparameters. The accuracy and performance of each model were subsequently evaluated using the test data.

The initial approach involved utilizing a SARIMAX statistical analysis model. This type of model is well-suited for analyzing and predicting time series data, incorporating autoregressive, moving average, and external exogenous variables. Peak electricity consumption can be related to various factors such as weather, time of day, working hours, etc. For this code uses exogenous parameters such as 'length_of_day', ' Hour ', ' day_of_week ' which may affect power consumption. These metrics add additional context for analyzing and predicting peak values.

The model parameters (p, d, q, P, D, Q, s) are adjusted taking into account the properties of the time series and the specifics of peak consumption. Parameter selection involved trying different combinations and testing their performance on the training data using criteria such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion). Using ACF, PACF plots, and trying different parameter values helped to find the best combination for a particular time series.

SARIMAX model parameters are the following: p: Autoregressive order (AR) – 0; d: Degree of difference – 1, q: The order of the moving average (MA) - 1, P: Order of seasonal autoregression (Seasonal AR) – 1, D: The degree of seasonal difference - 0, Q: The order of the seasonal moving average - 0, s: Seasonality period - 24 (one day)

The final metrics (MAE, RMSE, MAPE) provide quantitative insight into the accuracy of model predictions. This helps to determine how effective the model is in predicting peak electricity consumption in different time frames (Table 1).

Table 1

Estimates of the accuracy of the SARIMAX model using the best parameters

Accuracy score/ prediction interval	Week	Month	Year
MAE	0.859	1.185	7.36
RMSE	1.001	1.47	8.58
MAPE	136.37	85.5	102.2

A graph is plotted comparing the actual data and the predicted values for the week using the SARIMAX model. The red points on the graph indicate the maximum predicted values that meet the given condition. A shadow range is also used for confidence intervals around predicted values (see Fig. 3).

Random Forest model was chosen next. This model is an ensemble of decisions based on decision trees, which allows to predict the peak load of electricity consumption. The data set was divided into training and test parts, where the training part contains 80% of the total amount of data. Next, important features (parameters) are selected for the model, such as 'month', 'day_of_week', 'length_of_day', 'Night ', ' Winter ' and ' Hour ', which are used to predict the peak load.

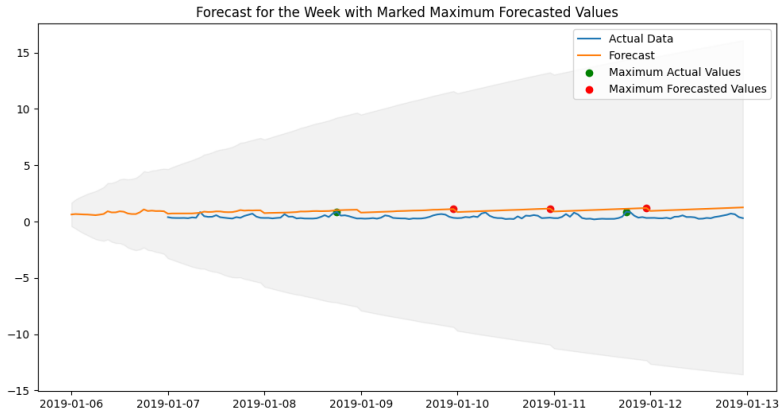

Fig. 3. Indicators of forecasting peak loads using the SARIMAX model

With the help of the trained model, peak load forecasting was carried out at different time horizons: week, month and year. The predicted values were compared with real data (Table 2).

Table 2

Accuracy estimates of the Random Forest model

Accuracy score/prediction interval	Week	Month	Year
MAE	0.802	0.835	0.48
RMSE	1.13	1.151	0.766
MAPE	93.87	58.59	50.75

A graph is plotted comparing the actual data and the predicted values for the week using the Random Forest model (see Fig. 4).

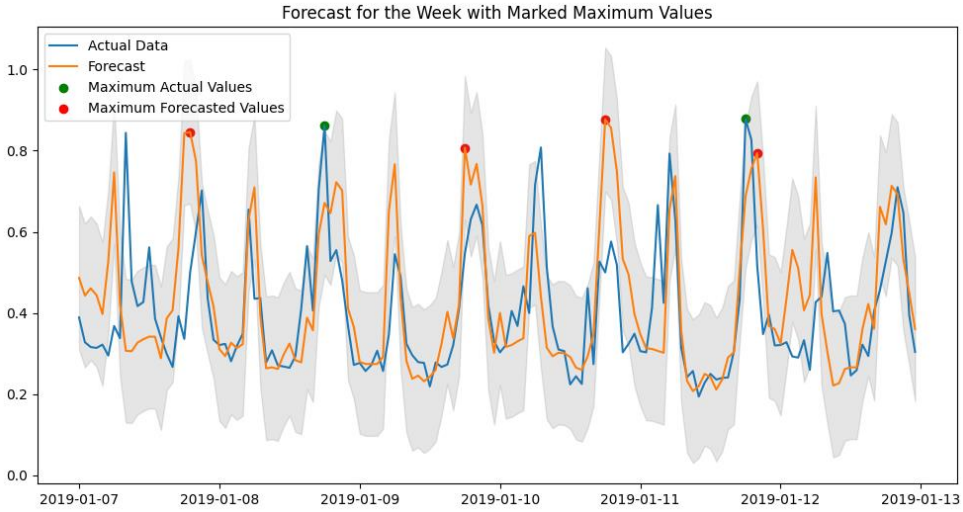

Fig. 4. Indicators of forecasting peak loads using the Random Forest model

Consider the results of the LSTM model. Input data included information on electricity consumption (' Value (kWh)') and various factors that may affect it, such as month, day of the week, length of day, time of day, etc. To improve model performance, the data were normalized to a range of 0 to 1.

After dividing the data into training and test sets, where the latter was selected for testing, an LSTM model was built. It had one LSTM layer with 50 neurons that helped detect dependencies in time series. The model training process took five epochs, and each epoch used packets of size 168.

After the training was completed, a prediction was made on the test data set. The resulting predicted values were transformed back to the original measurement scale. The predicted values were compared with the real data (Table 3).

Table 3

LSTM model accuracy estimates

Accuracy score/prediction interval	Week	Month	Year
MAE	0.10	0.13	0.29
RMSE	0.13	0.17	0.53
MAPE	25.1	30,12	31.52

When displaying the graphs, the graph displayed the observed values and the predicted data of electricity consumption per week using the LSTM model (see Fig. 5).

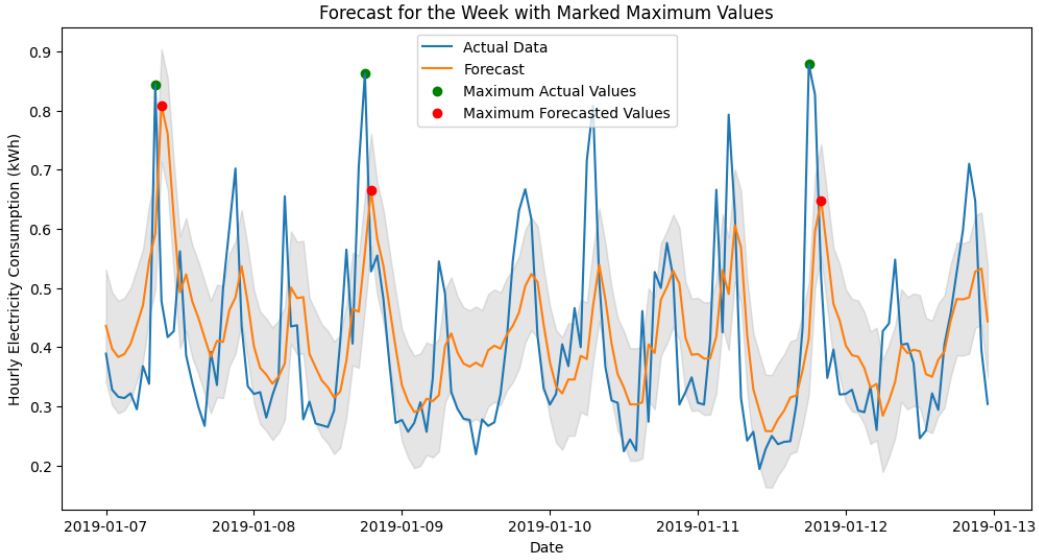

Fig. 5. Prediction of peak load indices using models LSTM

Deep recurrent neural networks (RNNs) include a variety of architectures such as LSTM and GRU (Gated Recurrent Unit), which are designed to process sequential data. These two architectures have some differences in the way they work, which makes them effective for different tasks.

Parameters of the GRU model are the following: the number of neurons in GRU layer –50; Activation function – default (of course sigmoid and tanh for GRU gates); Optimizer – Adam; Loss function – Mean Squared Error (MSE); Number of learning epochs – 20; Pack size – 168 (manually selected number); Length of incoming sequence – 24 hours.

The GRU uses two internal blocks - an update block and a transfer block. The update block specifies how much information will be updated, while the carry block specifies how much information will be passed to the next step. This allows the GRU to control the information flow in a simpler way.

Predicted values, as in other models, were compared with real data (Table 4).

Table 4

Estimates of the accuracy of the GRU model

Accuracy score/prediction interval	Week	Month	Year
MAE	0.12	0.18	0.30
RMSE	0.15	0.22	0.54
MAPE	32.94	45.50	36.07

A graph is drawn showing the observed values and predicted data of electricity consumption per week using the GRU model (see Fig. 6).

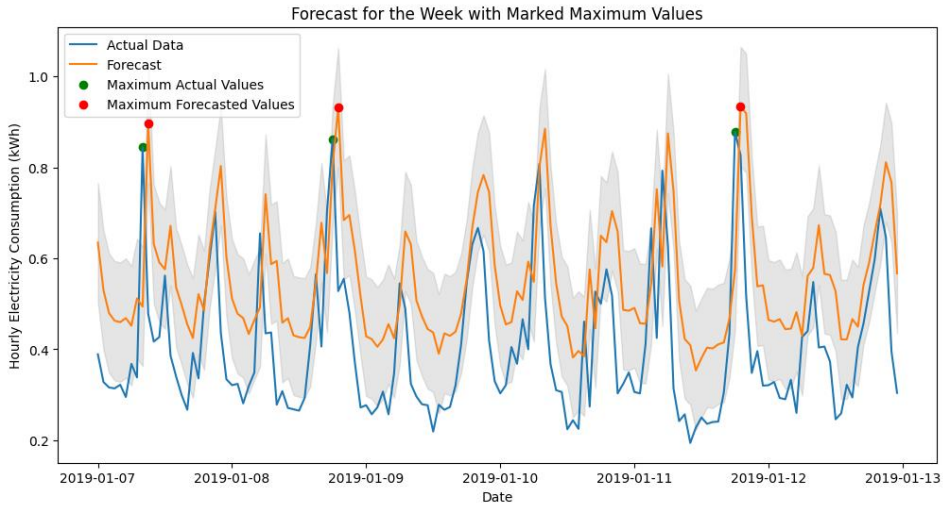

Fig. 6. Prediction of peak load indices using models GRU

As a result of the study, a consolidated (table 5) was compiled, encompassing the predictions of peak electricity consumption generated by all the employed models. Notably, among the models examined, LSTM and GRU exhibited the highest accuracy in predicting peak load, with a discrepancy margin of just one hour.

Table 5

Actual and forecasted time of peak electricity consumption indicators

Actual	LSTM	GRU	Random Forest	SARIMAX
2019-01-07 08:00:00	2019-01-07 09:00:00	2019-01-07 09:00:00	2019-01-07 19:00:00	2019-01-09 23:00:00
2019-01-08 18:00:00	2019-01-08 19:00:00	2019-01-08 19:00:00	2019-01-09 18:00:00	2019-01-10 23:00:00
2019-01-11 18:00:00	2019-01-11 20:00:00	2019-01-11 19:00:00	2019-01-11 20:00:00	2019-01-11 23:00:00

Conclusions

In this study the different forecasting methodologies, including SARIMAX, Random Forest, LSTM, and Gated Recurrent Unit, which collectively demonstrated efficacy in forecasting peak electricity consumption, were studied. Each model exhibited distinct strengths and limitations that warrant careful consideration in selecting an optimal approach.

Starting with the SARIMAX model, its suitability for long-term peak electricity consumption forecasting proved limited due to its inability to effectively capture the intricate dynamics and transformations inherent in energy systems, particularly over extended forecast horizons.

The Random Forest model showcased its versatility by efficiently accommodating the complex dynamics of electricity data. This model autonomously identified dependencies, nonlinearities, and seasonality within input data while considering external factors influencing consumption.

Deep neural models, namely LSTM and GRU, emerged as formidable tools for managing trends, seasonality, and non-linearities within electricity consumption time series. Of particular significance is the remarkable performance of LSTM and GRU in accurately forecasting long-term peak values.

The SARIMAX model serves as a viable tool for predicting general trends and standard changes in electricity consumption but lacks optimal performance for long-term peak value forecasting. Random Forest, LSTM, and GRU models demonstrated their prowess in addressing complex data variations and offering accurate peak electricity consumption forecasts.

As the result, using the LSTM model the highest forecasting accuracy across all time intervals was achieved. With M.A.E values of 0.10 kW/h for weekdays, 0.13 kW/hr for weekends, and 0.29 kW/h for holidays, the LSTM model showcased its robust performance. Additionally, both the LSTM and GRU models exhibited the capacity to identify all peak electricity consumption instances within a few hours, thus solidifying their role in the task of assessing the state of energy microgrids. The developed models will serve as integral components in the ongoing evaluation of energy microgrid conditions, contributing to the enhancement of energy distribution system assessment and management.

References

1. Load factor improvement of the electricity grid considering distributed resources operation and regulation of peak load / F. V. Cerna et al. *SSRN electronic journal*. 2022. URL: https://doi.org/10.2139/ssrn.4293004 (date of access: 23.08.2023).
2. Research on peak load shifting based on energy storage and air conditioning load in power grid / P. Xiao et al. *IOP conference series: earth and environmental science.* 2020. Vol. 546. P. 022021. URL: https://doi.org/10.1088/1755-1315/546/2/022021 (date of access: 23.08.2023).
3. Saracoglu B. O. An experimental fuzzy inference system for global grid electricity peak power load forecasting third core module of first console on G2P3S. *Journal of energy systems.* 2017. Vol. 1, no. 2. P. 75–101. URL: https://doi.org/10.30521/jes.338575 (date of access: 23.08.2023).
4. Seemingly unrelated time series model for forecasting the peak and short-term electricity demand: Evidence from the Kalman filtered Monte Carlo method / F. K. Owusu et al. Heliyon. 2023. P. e18821. URL: https://doi.org/10.1016/j.heliyon.2023.e18821 (date of access: 23.08.2023).
5. Study on dynamic adjustment method of power grid critical peak periods considering load characteristics / D. Liu et al. IOP conference series: earth and environmental science. 2021. Vol. 804, no. 3. P. 032004. URL: https://doi.org/10.1088/1755-1315/804/3/032004 (date of access: 23.08.2023).
6. Review of research on load forecasting methods for smart grid / P. Li et al. 2022 4th international conference on communications, information system and computer engineering (CISCE), Shenzhen, China, 27–29 May 2022. 2022. URL: https://doi.org/10.1109/cisce55963.2022.9851039 (date of access: 23.08.2023).
7. Short-term load forecasting: A recurrent dynamic neural network approach using NARX / S. Kumar et al. *Advances in data-driven computing and intelligent systems.* 2023. P. 509–522. URL: https://doi.org/10.1007/978-981-99-3250-4_39 (date of access: 23.08.2023).
8. Data-Driven copy-paste imputation for energy time series / M. Weber et al. *IEEE transactions on smart grid.* 2021. P. 1. URL: https://doi.org/10.1109/tsg.2021.3101831 (date of access: 23.08.2023).
9. Intelligent techniques for forecasting electricity consumption of buildings / K. P. Amber et al. Energy. 2018. Vol. 157. P. 886–893. URL: https://doi.org/10.1016/j.energy.2018.05.155 (date of access: 23.08.2023).
10. Ünal F., Almalaq A., Ekici S. A novel load forecasting approach based on smart meter data using advance preprocessing and hybrid deep learning. *Applied sciences.* 2021. Vol. 11, no. 6. P. 2742. URL: https://doi.org/10.3390/app11062742 (date of access: 23.08.2023).
11. Comparison of short-term forecasting methods of electricity consumption in microgrids / Y. V. Parfenenko et al. *Radio electronics, computer science, control*. 2023. No. 1. P. 14. URL: https://doi.org/10.15588/1607-3274-2023-1-2 (date of access: 23.08.2023).

12. Development of expert assessment methods in planning energy supply of buildings with renewable energy sources / O. Boiko et al. *Technology audit and production reserves*. 2021. Vol. 2, no. 2(58). P. 51–54. URL: https://doi.org/10.15587/2706-5448.2021.230230 (date of access: 23.08.2023).
13. Time series forecasting of domestic shipping market: comparison of SARIMAX, ANN-based models and SARIMAX-ANN hybrid model / A. G. Cerit et al. *International journal of shipping and transport logistics*. 2022. Vol. 14, no. 3. P. 193. URL: https://doi.org/10.1504/ijstl.2022.10046664 (date of access: 23.08.2023).
14. Fu R. J. CCG Realization with LSTM Hypertagging. 2018. URL: http://rave.ohiolink.edu/etdc/view?acc_num=osu1534236955413883 (date of access: 23.08.2023).
15. Salem F. M. Gated RNN: the gated recurrent unit (GRU) RNN. Recurrent neural networks. Cham, 2021. P. 85–100. URL: https://doi.org/10.1007/978-3-030-89929-5_5 (date of access: 23.08.2023).
16. Oberkampf W. L., Roy C. J. Model accuracy assessment. Verification and validation in scientific computing. Cambridge. P. 469–554. URL: https://doi.org/10.1017/cbo9780511760396.017 (date of access: 23.08.2023).

Євген Холявка **Yevhen Kholiavka**	PhD student, Information Technology Department, Sumy State University, Sumy, Ukraine, e-mail: yevhen.kholyavka@student.sumdu.edu.ua, https://orcid.org/0000-0002-9841-7909	аспірант кафедри інформаційних технологій, Сумський державний університет, Суми, Україна.
Юлія Парфененко **Yuliia Parfenenko**	PhD, Associate Professor of Information Techology Department, Sumy State University, Sumy, Ukraine, e-mail: yuliya_p@cs.sumdu.edu.ua. https://orcid.org/0000-0003-4377-5132	кандидат технічних наук, доцент кафедри інформаційних технологій, Сумський державний університет, Суми, Україна.

https://doi.org/10.31891/csit-2023-3-3
UDC 004.89: 004.3

Olga PAVLOVA, Andrii BASHTA, Andrii KUZMIN
Khmelnytskyi National University

APPROACHES OF BUILDING A REAL-WORLD OBJECT DETECTOR DATA SOURCE

In our constantly developing world virtual, augmented, and mixed reality technologies are becoming integral parts of our daily lives. In the current stage of Information Technology field development, technologies of virtual, augmented and mixed reality can be seen in almost all areas of human life. Nowadays AR is used in Marketing and Advertising, Education, Medicine, Automotive, Game Development, Navigation and other areas of our everyday life. Therefore, object detection is a crucial task in computer vision and AI applications, enabling machines to identify and locate objects within images or video frames. The accuracy and performance of an object detector heavily rely on the quality and diversity of the training data. This paper is aimed at finding the approaches of building a real-world object detector data source to be able to create a model for detecting a sport games surfaces using the Action & Vision App. During this research several structured approaches of building an object detector data source have been built, drawing inspiration from Apple's Create ML documentation on the topic. Additionally, real-world applications available on both the App Store and Google Play that leverage object detection technology were showcased and analyzed. In the course of study a dataset of objects has been collected and then utilized to build a robust detection model, tailored to function seamlessly with Vision and Core ML frameworks on iOS devices. The trained object detection model, informed by the diverse dataset and robust training process, is employed to identify and outline tables and rectangles in each frame of the video stream. The model and the proposed approaches will be further applied to develop the method of object detection in the real world and create a mobile application for sport games simulation, that would help players to practice their skills out of the training field.

Keywords: object detection, computer vision, AI applications, training data, Create ML, structured data source, real-world applications, App Store, Google Play.

Ольга ПАВЛОВА, Андрій БАШТА, Андрій КУЗЬМІН
Хмельницький національний університет

ПІДХОДИ ДО ПОБУДОВИ ДЖЕРЕЛ ДАНИХ ДЛЯ ВИЗНАЧЕННЯ ОБ'ЄКТІВ У РЕАЛЬНОМУ СВІТІ

На сучасному етапі розвитку сфери інформаційних технологій технології віртуальної, доповненої та змішаної реальності можна побачити практично у всіх сферах життя людини. Сьогодні AR використовується в маркетингу та рекламі, освіті, медицині, автомобілебудуванні, розробці ігор, навігації та інших сферах нашого повсякденного життя. Тому виявлення об'єктів є ключовим завданням у програмах комп'ютерного зору та штучного інтелекту, що дозволяє машинам ідентифікувати та знаходити об'єкти в зображеннях або відеокадрах. Точність і продуктивність детектора об'єктів значною мірою залежать від якості та різноманітності навчальних даних. Ця стаття спрямована на пошук підходів до побудови реального джерела даних детектора об'єктів, щоб мати можливість створити модель для виявлення поверхонь для спортивних ігор за допомогою програми Action & Vision. Під час цього дослідження було створено кілька структурованих підходів до створення джерела даних детектора об'єктів, черпаючи натхнення з документації Apple Create ML на цю тему. Крім того, були продемонстровані та проаналізовані реальні програми, доступні як в App Store, так і в Google Play, які використовують технологію виявлення об'єктів. У ході дослідження було зібрано набір даних об'єктів, який потім використано для побудови надійної моделі виявлення, адаптованої для бездоганної роботи з фреймворками Vision і Core ML на пристроях iOS. Навчена модель виявлення об'єктів, яка базується на різноманітних наборах даних і надійному процесі навчання, використовується для ідентифікації та окреслення таблиць і прямокутників у кожному кадрі відеопотоку. Модель і запропоновані підходи будуть надалі застосовані для розробки методу виявлення об'єктів у реальному світі та створення мобільного додатку для симуляції спортивних ігор, який допоможе гравцям відпрацьовувати свої навички поза тренувальним полем.

Ключові слова: виявлення об'єктів, комп'ютерний зір, програми ШІ, навчальні дані, Create ML, джерело структурованих даних, реальні програми, App Store, Google Play.

Introduction

Currently, virtual, augmented and mixed reality technologies can be increasingly found in various areas of human life and even in everyday use. These technologies are already actively used in such fields as education, medicine, marketing and advertising, automotive industry, and navigation. Therefore, involving these technologies in application development is a relevant task. Object detection has a crucial part in various applications, from computer vision to autonomous vehicles. One of the key factors that influence the accuracy and robustness of an object detector is the quality of the training data. In our previous works we proposed the application of augmented reality for navigation purposes [1] and for Objects 3D Models visualization using Augmented Reality [2].

In this paper, we will explore several structured ways to build an object detector data source, taking inspiration from Apple's Create ML documentation on the topic. Additionally, we will showcase real-world applications available on both the App Store and Google Play that leverage object detection technology.

Domain analysis

Object detection is a fundamental task within the domain of computer vision and artificial intelligence. It finds application in various industries and domains, driving innovations and advancements in safety, efficiency, and user experiences [3-5].

Autonomous Vehicles: In the domain of autonomous vehicles, object detection plays a pivotal role in enabling self-driving cars to perceive their surroundings and make real-time decisions. The ability to detect pedestrians, vehicles, traffic signs, and obstacles is critical for ensuring safe and reliable autonomous navigation.

Surveillance and Security: Surveillance systems heavily rely on object detection to monitor and identify suspicious activities, intruders, or potential threats in real-time. This technology has significant implications in enhancing security measures in public spaces, airports, and critical infrastructure.

Augmented Reality: Object detection is a core technology in augmented reality (AR) applications. By recognizing and tracking objects in the real world, AR systems can overlay virtual objects or information seamlessly, enriching user experiences in gaming, education, and marketing.

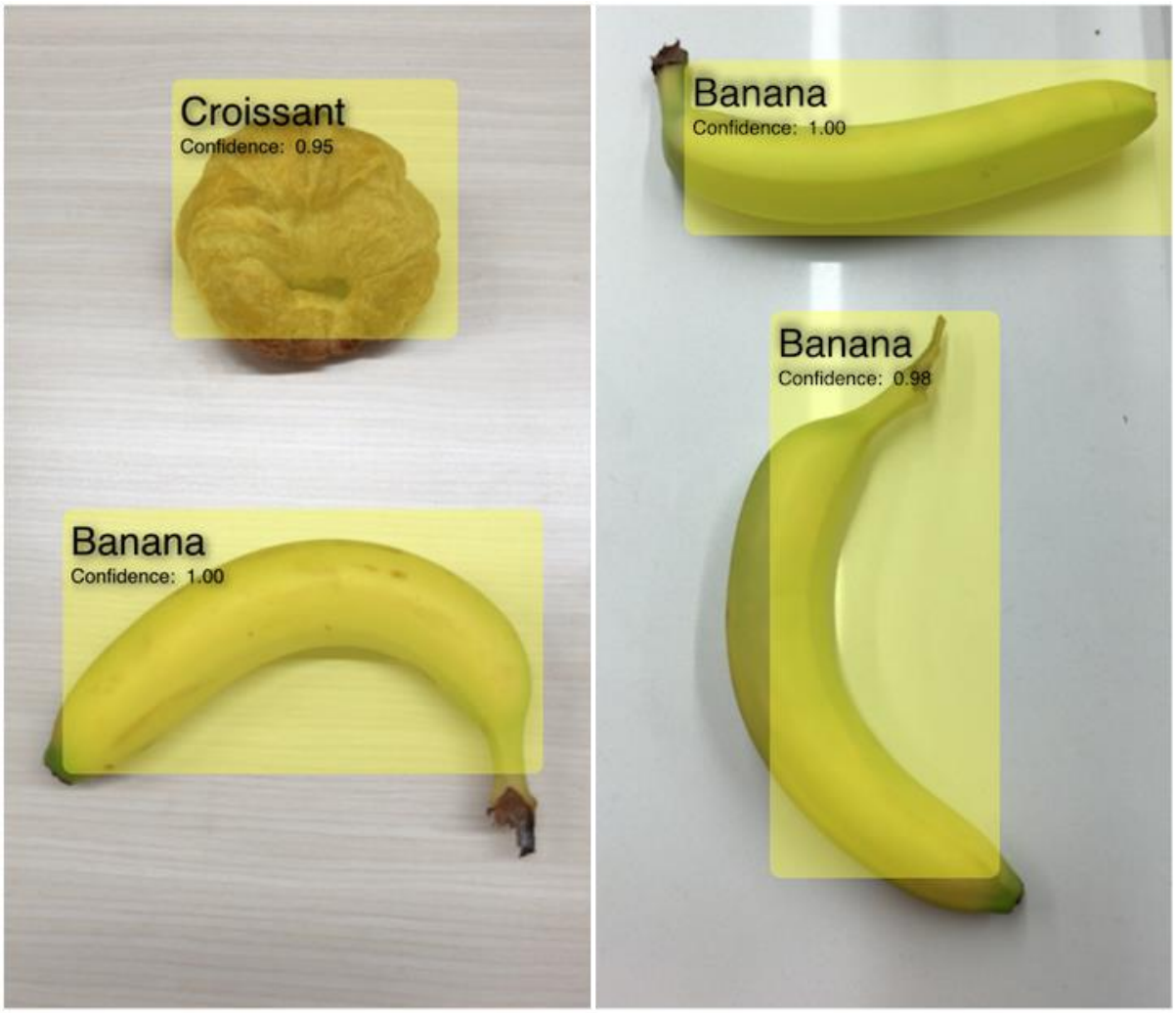

Fig. 1. Bananas detection

Medical Imaging: In medical imaging, object detection is utilized to identify and locate anomalies or specific structures within medical images, assisting in diagnosis and treatment planning. Applications range from detecting tumors in MRI scans to identifying cells in microscopy images.

E-commerce and Retail: In e-commerce and retail, object detection facilitates various applications such as visual search, product recommendation, and inventory management. By recognizing objects in images or videos, retailers can offer personalized shopping experiences and optimize supply chain operations.

Robotics: In robotics, object detection is crucial for enabling robots to interact with their environment and perform tasks autonomously. From industrial robots identifying objects on assembly lines to service robots navigating in dynamic environments, object detection enhances the capabilities of robotic systems.

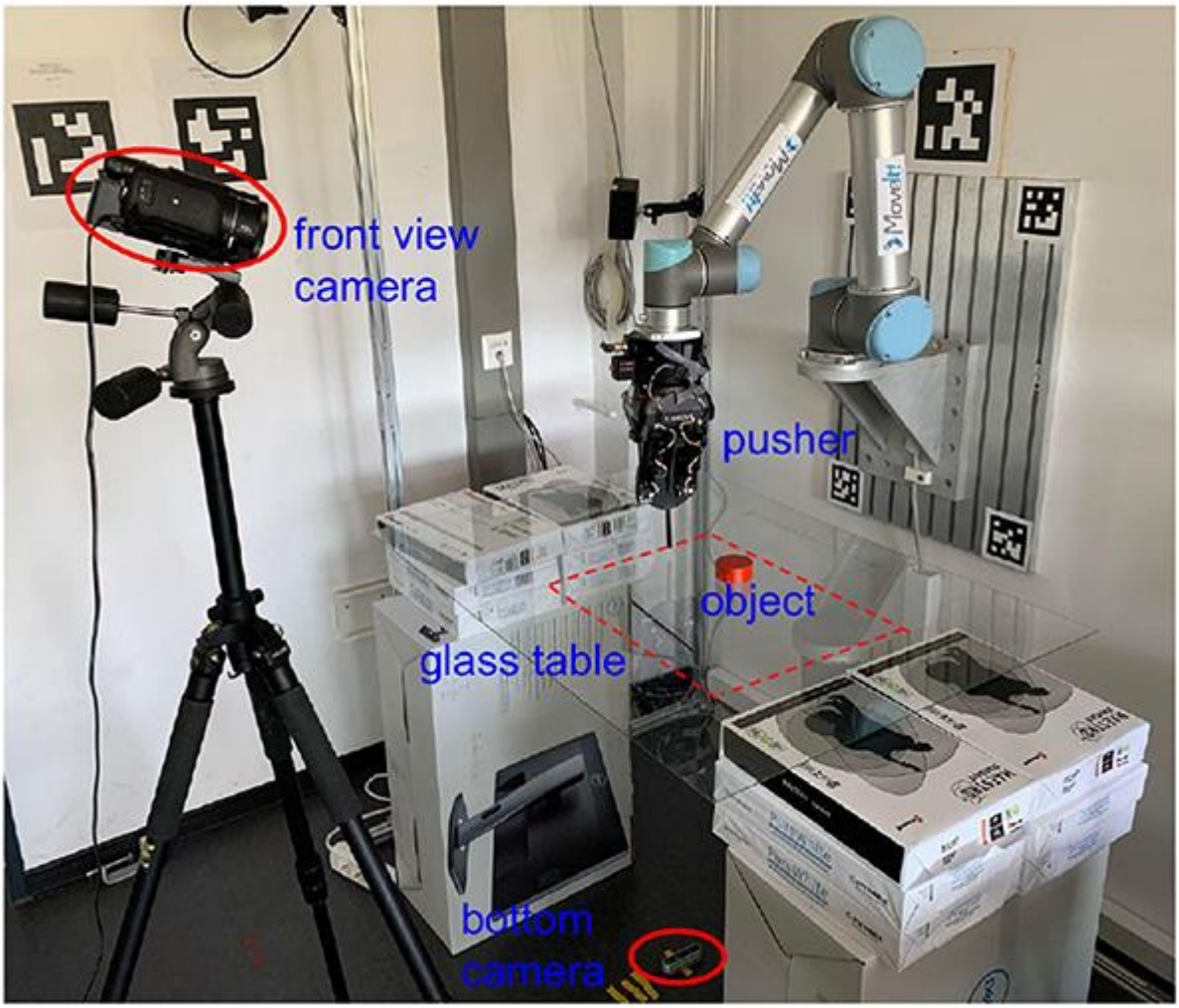

Fig. 2. Robot detects the objects and works with them

Accessibility: Object detection technologies have also been leveraged to enhance accessibility for individuals with visual impairments. By identifying and describing objects in real-time, these applications empower users to navigate their surroundings more effectively.

Understanding the specific domains and applications of object detection helps researchers, developers, and practitioners tailor their approaches and data collection methods to address the unique challenges and requirements of each use case. The quality and relevance of the training data are paramount in achieving accurate and efficient object detection systems across diverse applications.

Analysis of existing solutions and technologies

Object detection technology has found its way into numerous real-world applications, revolutionizing various industries and enhancing user experiences. Leveraging computer vision and AI algorithms, these applications have demonstrated the practicality and effectiveness of object detection in addressing real-life challenges. In this section, we will explore three prominent real-world applications that utilize object detection to provide valuable insights, improve accessibility, and transform the way we interact with our surroundings.

Application "Google Lens" (Available on Google Play). Google Lens is an innovative application that leverages object detection to provide users with instant information about the world around them. By simply pointing the smartphone camera at objects or scenes, Google Lens can identify landmarks, plants, animals, and a wide range of everyday objects. The app then delivers relevant search results, detailed information, and even language translation based on the recognized objects. Whether users are exploring a new city, trying to identify a particular species of flora or fauna, or need to understand foreign language text, Google Lens offers an intuitive and interactive experience. By incorporating object detection technology, Google Lens has revolutionized how we interact with the environment, making information readily available at our fingertips.

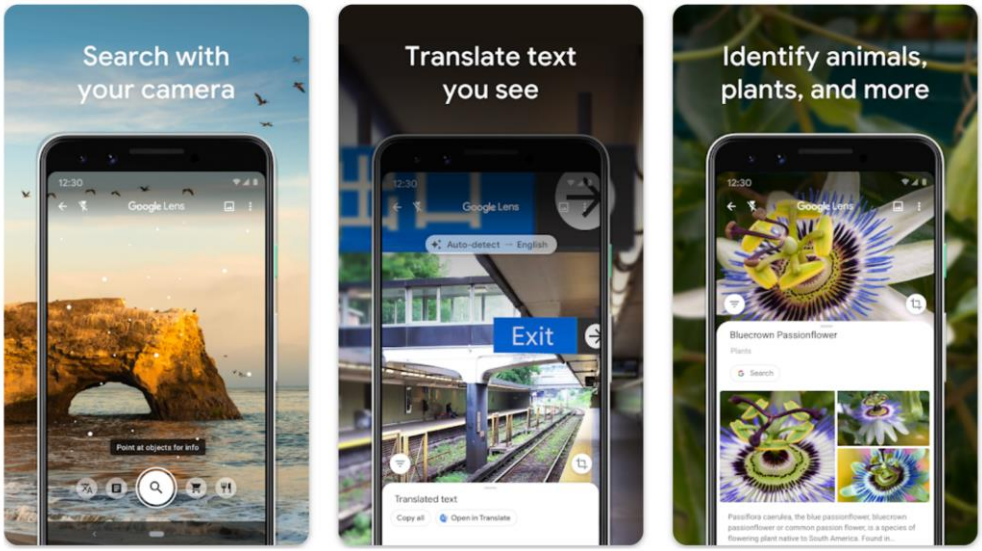

Fig. 3. Google lens in Google Play

Application "CamFind" (Available on App Store and Google Play). CamFind is a powerful image recognition app that utilizes object detection to identify objects captured in photos. Whether it's household items, fashion accessories, or artworks, CamFind can accurately recognize and classify objects, allowing users to learn more about them. The app provides detailed information, shopping links, and related content based on the identified objects. CamFind's seamless integration of object detection technology has bridged the gap between the physical and digital worlds, enabling users to explore and discover information effortlessly. From identifying unfamiliar objects to finding similar products online, CamFind exemplifies the practical applications of object detection in everyday life.

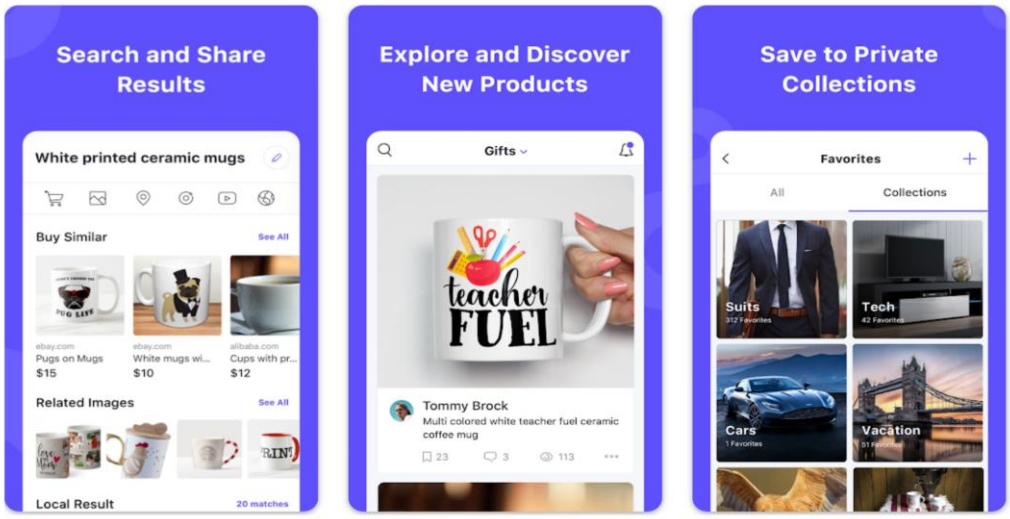

Fig. 4. CamFind in Google Play

Application "TapTapSee" (Available on App Store and Google Play). TapTapSee is an exceptional accessibility application that utilizes object detection to empower visually impaired users. By capturing images through their smartphone cameras, users can rely on TapTapSee to identify objects and receive auditory descriptions in real-time. The app is capable of recognizing a wide range of objects, including everyday items, products, and even scenes. By incorporating object detection, TapTapSee enhances the independence and confidence of visually impaired individuals, enabling them to navigate their surroundings more effectively. This application exemplifies the transformative impact of object detection in promoting inclusivity and accessibility.

Fig. 5. TapTapSee in AppStore

Now let’s compare these three apps by platform availability, functionality and benefits for users.

Application	Platform	Main Functionality	Key Benefit
Google Lens	Google Play	Instant object identification	Provides detailed information and search results based on recognized objects
CamFind	App Store,	Image recognition and object classification	Offers detailed information and shopping links for identified objects
TapTapSee	App Store,	Object recognition for visually impaired	Provides auditory descriptions of objects in real-time to enhance accessibility for users with visual impairments

Real-world applications using object detection technology have showcased its versatility and significance in various domains. From the interactive and informative experience offered by "Google Lens" to the seamless shopping and discovery facilitated by "CamFind," object detection has revolutionized how we interact with the world around us. Moreover, applications like "TapTapSee" have demonstrated the profound impact of object detection in enhancing accessibility for visually impaired individuals. As the technology continues to advance, we can expect even more innovative and transformative applications leveraging object detection to address real-life challenges and improve user experiences across diverse industries.

Structured Ways of Building an Object Detector Data Source

Object detection models heavily rely on high-quality training data for accurate and reliable performance. Building a structured object detector data source involves various approaches and techniques to ensure the dataset's diversity, precision, and scalability. One of the primary methods is hand-annotated datasets, where human experts meticulously label objects in images. Although this process can be time-consuming, it offers precise annotations, making it suitable for specialized use cases. To streamline the manual annotation process, developers can utilize tools like Labelbox, RectLabel, and VoTT.

Another effective strategy for building a robust dataset is leveraging pre-trained datasets such as COCO, ImageNet, and PASCAL VOC for transfer learning. By starting with pre-trained models and fine-tuning them on specific datasets, developers can significantly speed up the training process and improve object detection performance. Create ML, a popular tool in the Apple ecosystem, supports this technique, making it accessible to developers.

Data augmentation techniques also play a crucial role in enhancing the diversity of the dataset. By applying various transformations like rotation, flipping, scaling, and color changes, developers can create multiple variations of the original data, thereby improving the detector's generalization and reducing overfitting.

In some cases, real-world data may not be sufficient to cover all possible scenarios. Here, synthetic data generation comes into play. Developers can create synthetic data using 3D rendering engines or generative models,

allowing them to control object placements, backgrounds, and lighting conditions. This approach helps address the limitations of real-world data collection and further augments the dataset's diversity.

To optimize the training dataset and reduce annotation efforts, active learning strategies can be implemented. Active learning involves selecting the most informative samples for manual annotation, effectively focusing the annotation efforts where it matters the most. By doing so, developers can achieve higher detection accuracy while minimizing the annotation workload.

To ensure the dataset remains up-to-date and relevant, it is crucial to follow best practices for maintaining a structured object detector data source. Regularly updating the dataset based on new scenarios, edge cases, and emerging trends is vital for adapting the object detector to changing real-world conditions.

In conclusion, building a structured object detector data source involves a combination of various techniques and strategies. Hand-annotated datasets, transfer learning with pre-trained models, data augmentation, synthetic data generation, active learning, and semi-supervised learning all contribute to creating a comprehensive and diverse training dataset. Following best practices for dataset maintenance ensures the object detector remains accurate and performs effectively in real-world applications.

Creating a model for detecting a sport games surfaces using the Action & Vision App

In the realm of modern technology, the Action & Vision app stands as a remarkable testament to the synergistic capabilities of Vision and Core ML technologies. By seamlessly integrating these powerful tools, the app ventures into the intricate world of sports environments, deciphering player movements and interactions with objects with unprecedented precision.

Central to this innovative endeavor is the development of an app that not only captures sports scenes but also comprehends them. The app's underlying framework involves the meticulous curation of a diverse image dataset, meticulously annotated to facilitate the training of a robust model. This pivotal phase is elevated with the assistance of Roboflow, a specialized service that streamlines data collection and annotation, ensuring the model's accuracy and efficacy.

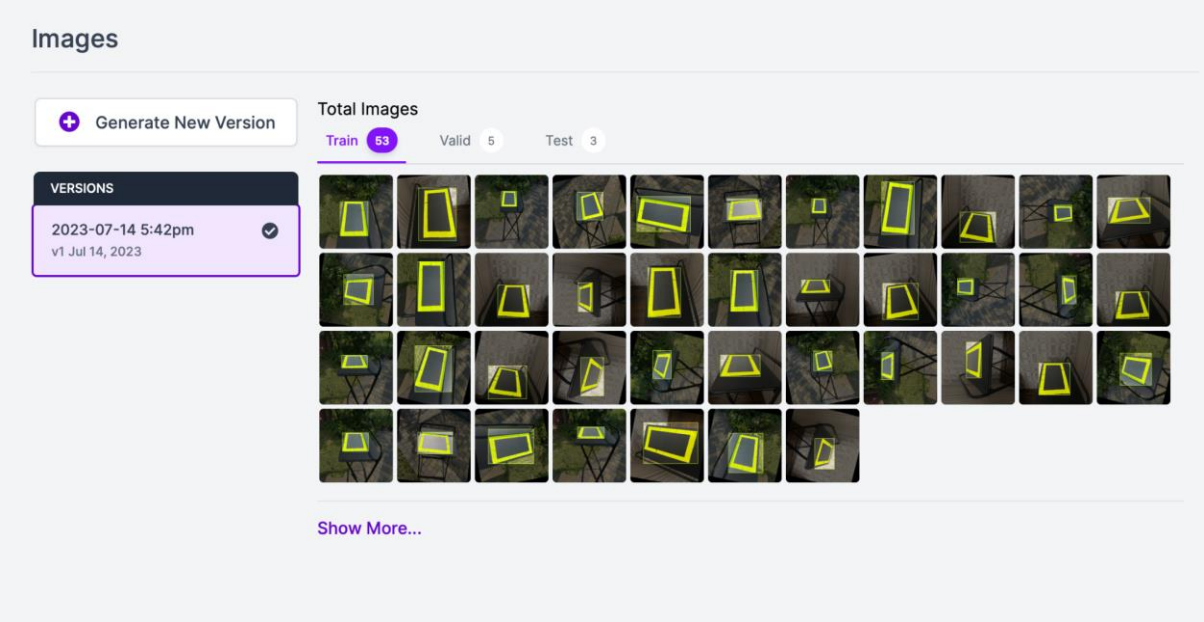

Fig 6. Annotating Images on Roboflow

Once the dataset is curated and annotated via Roboflow, it is transformed into a comprehensive training dataset. This dataset is then utilized to build a robust detection model, tailored to function seamlessly with Vision and Core ML frameworks on iOS devices.

The core functionality of the app hinges upon its ability to perform real-time scene analysis and accurately identify tables and rectangles within the video frames. This is achieved through the intelligent application of VNCoreMLRequest, a Vision framework component. The trained object detection model, informed by the diverse dataset and robust training process, is employed to identify and outline tables and rectangles in each frame of the video stream.

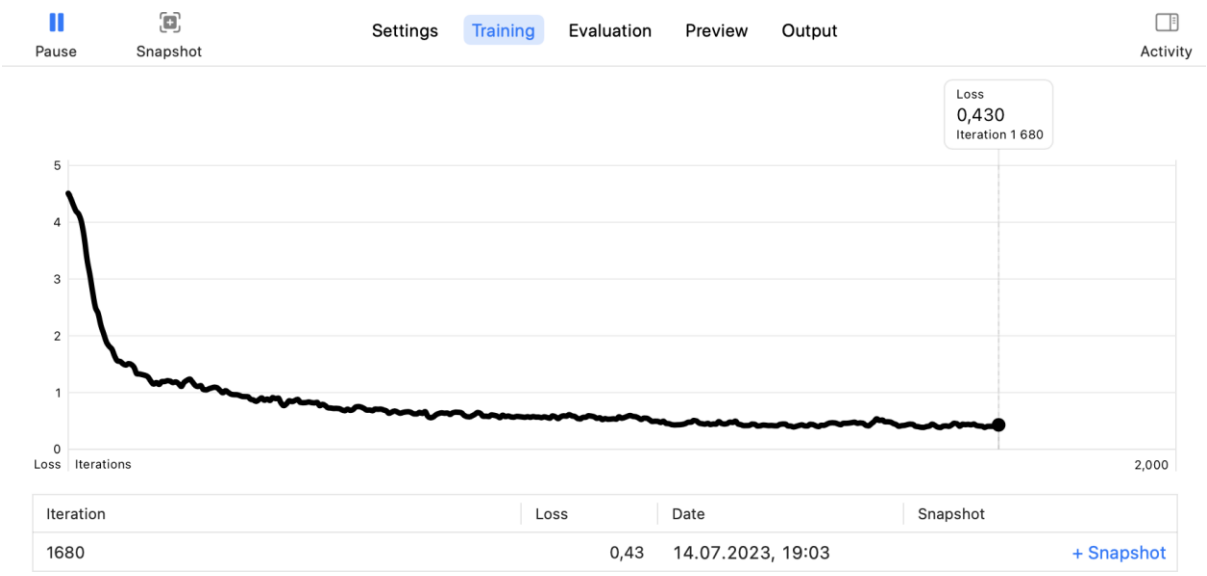

Fig 7. Generating Model with Create ML

Subsequent to the successful generation of the detection model, the app harnesses the capabilities of Core ML Preview mode. This feature enables the integration of the model into the application and allows for real-time testing and validation of the model's performance on sample video frames. This iterative testing process ensures the model's accuracy and reliability in identifying tables and rectangles across various scenarios.

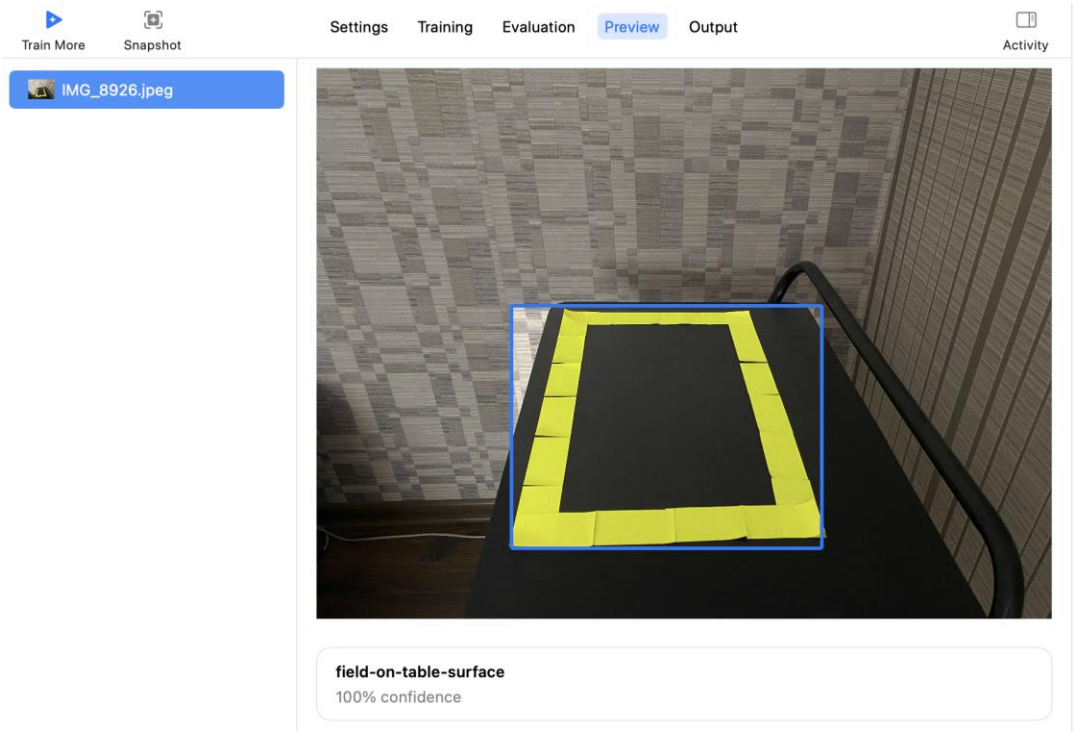

Fig 8. Using Core ML Preview Mode

Fig 9. Using the model in the iOS app

In essence, the integration of machine learning models via Core ML within the Action & Vision app showcases the profound impact of Vision and Core ML on advancing object detection capabilities. This marriage of technologies not only augments the functionality of the app but also serves as a testament to the transformative potential of machine learning and computer vision in the modern technological landscape.

Experiments and directions of further work

As object detection technology continues to advance, there are several directions for future work to enhance the capabilities and applications of object detectors. First, exploring more advanced data augmentation techniques and generative models can further improve dataset diversity and expand the training data pool.

Additionally, the integration of domain adaptation methods can help improve the generalization of object detectors to new environments and domains. This is particularly relevant in scenarios where the training data may differ significantly from the test data.

Furthermore, research on active learning strategies tailored specifically for object detection can lead to more efficient annotation processes and better utilization of limited labeled data.

Lastly, investigating methods to address potential biases in the training data and ensuring the fairness and ethics of object detection models is essential, especially in applications with high social impact.

By continuously exploring and advancing these areas, object detection models can continue to evolve and provide valuable insights and solutions in a wide range of real-world applications.

Conclusions

In our constantly developing world virtual, augmented, and mixed reality technologies are becoming integral parts of our daily lives. The study has highlighted the significance of object detection in the domains of computer vision and AI, as it empowers machines to identify and locate objects in images or video frames. The accuracy and performance of object detectors are intricately linked to the quality and diversity of the training data, making the establishment of a reliable data source a critical endeavor.

Through a systematic exploration of methodologies and insights derived from Apple's Create ML documentation, this research has successfully constructed structured approaches for building a real-world object detector data source. Additionally, the paper has shed light on various real-world applications that leverage object detection technology, available on platforms like the App Store and Google Play. The culmination of this work is the development of a robust object detection model, tailored to seamlessly integrate with Vision and Core ML frameworks on iOS devices. This model excels in identifying and outlining tables and rectangles within video streams, particularly in the context of sports game surfaces. The practical implications of this research extend beyond the academic realm and pave the way for the creation of a mobile application, the Action & Vision App, which will simulate sports games. This application holds the potential to revolutionize the way players practice and enhance their skills, extending the boundaries of traditional training fields.

References

1. Pavlova, O., Bashta, A., Kravchuk, S., Hnatchuk, Y., Bouhissi, H.E. Augmented Reality Based Technology and Scenarios For Route Planning and Visualization. CEUR Workshop Proceedings, 2022, 3156, pp. 613–623
2. Pavlova O., Bashta A., Kostiuk M. and El Bouhissi H. Technology and Scenarios for Objects 3D Models visualization using Augmented Reality CEUR Workshop Proceedings, 2023, 3373, (IntelITSIS, Khmelnytskyi, March 23-25) pp. 343-353
3. Kang, J., Tariq, S., Oh, H., & Woo, S. S. (2022). A survey of deep learning-based object detection methods and datasets for overhead imagery. IEEE Access, 10, 20118-20134.
4. Markov, T., Zhang, C., Agarwal, S., Nekoul, F. E., Lee, T., Adler, S., ... & Weng, L. (2023, June). A holistic approach to undesired content detection in the real world. In Proceedings of the AAAI Conference on Artificial Intelligence .Vol. 37, No. 12, pp. 15009-15018
5. VS, V., Oza, P., & Patel, V. M. (2023). Towards online domain adaptive object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 478-488.

Olga Pavlova **Ольга Павлова**	PhD, Associated professor of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: olya1607pavlova@gmail.com https://orcid.org/0000-0001-7019-0354	доктор філософії, доцент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.
Andriy Bashta **Андрій Башта**	PhD Student of Computer Engineering & Information Systems, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: andreybashta@gmail.com https://orcid.org/0000-0002-0775-1347	аспірант кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.
Andrii Kuzmin **Андрій Кузьмін**	Master Student of Computer Engineering & Information Systems, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: andriy1731@gmail.com	студент-магістр кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.

https://doi.org/10.31891/csit-2023-3-4
UDC 004.89: 004.3

Olga PAVLOVA, Ada BILINSKA, Andrii HOLOVATIUK,
Yaroslav BINKOVSKYI, Denys MELNYCHUK
Khmelnytskyi National University

AUTOMATED SYSTEM FOR DETERMINING SPEED OF CARS AHEAD

Road accidents and speeding violations are pervasive issues that pose substantial threats to road users on a daily basis. In an ongoing effort to improve road safety and reduce the frequency of accidents, researchers and engineers have been dedicated to the development and implementation of new technologies. One such significant innovation is the utilization of speed control systems based on traffic cameras.

This paper delves into a thorough exploration of the pivotal role and significance of speed control systems on our roadways. It investigates the operational principles, advantages, and various strategies employed to enhance the efficiency of these systems, with the ultimate goal of achieving optimal results in speed control and ensuring road safety. Speeding remains a widespread concern that significantly contributes to road accidents. Such incidents lead to injuries, fatalities, and extensive property damage, underscoring the urgent need for effective speed control measures. Among the arsenal of solutions available, speed control systems utilizing traffic cameras have emerged as a prominent and promising approach. These systems function by monitoring and recording the speed of vehicles at specific locations, which is later used to enforce speed limits and penalize offenders. The advantages of speed control systems based on traffic cameras are multifaceted. They offer an objective and reliable method for detecting and documenting speeding violations, eliminating the need for law enforcement personnel to be present at all times. This aspect not only frees up law enforcement resources but also ensures consistent and unbiased enforcement of speed limits. Additionally, the data collected by these systems can serve as a valuable resource for traffic management, accident analysis, and road safety research.

Keywords: speed control system, traffic camera, road safety, speeding, road accidents.

Ольга ПАВЛОВА, Ада БІЛІНСЬКА, Андрій ГОЛОВАТЮК,
Ярослав БІНЬКОВСЬКИЙ, Денис МЕЛЬНИЧУК
Хмельницький національний університет

АВТОМАТИЧНА СИСТЕМА ВИЗНАЧЕННЯ ШВИДКОСТІ АВТОМОБІЛІВ, ЩО РУХАЮТЬСЯ ПОПЕРЕДУ

Дорожньо-транспортні пригоди та порушення швидкісного режиму є поширеними проблемами, які щодня становлять серйозну загрозу для учасників дорожнього руху. У постійних зусиллях щодо покращення безпеки дорожнього руху та зменшення частоти аварій дослідники та інженери присвятили себе розробці та впровадженню нових технологій. Одним із таких важливих нововведень є використання систем контролю швидкості на основі камер руху.

Ця стаття детально досліджує ключову роль і значення систем контролю швидкості на наших дорогах. Він досліджує принципи роботи, переваги та різні стратегії, що використовуються для підвищення ефективності цих систем, з кінцевою метою досягнення оптимальних результатів у контролі швидкості та забезпеченні безпеки дорожнього руху. Перевищення швидкості залишається широко поширеною проблемою, яка значною мірою сприяє дорожньо-транспортним пригодам. Такі інциденти призводять до травм, смертельних випадків і значної матеріальної шкоди, що підкреслює нагальну потребу в ефективних заходах контролю швидкості. Серед арсеналу доступних рішень системи контролю швидкості, які використовують камери дорожнього руху, стали видатним і багатообіцяючим підходом. Ці системи функціонують шляхом моніторингу та реєстрації швидкості транспортних засобів у певних місцях, що згодом використовується для дотримання обмежень швидкості та покарання порушників. Переваги систем контролю швидкості на основі відеокамер багатогранні. Вони пропонують об'єктивний і надійний метод виявлення та документування порушень швидкісного режиму, усуваючи необхідність постійної присутності правоохоронних органів. Цей аспект не тільки звільняє ресурси правоохоронних органів, але й забезпечує послідовне та неупереджене дотримання обмежень швидкості. Крім того, дані, зібрані цими системами, можуть служити цінним ресурсом для управління дорожнім рухом, аналізу аварій і дослідження безпеки дорожнього руху.

Ключові слова: система контролю швидкості, камера дорожнього руху, безпека дорожнього руху, перевищення швидкості, дорожні пригоди.

Introduction

On the modern roads, where road transport has become an integral part of life, ensuring the safety of all road users is one of the key tasks for states and bodies responsible for road traffic. Given the growing number of motor vehicles, it is necessary to constantly develop and implement new technologies that contribute to reducing the risk of traffic accidents and violations of traffic rules [1].

One of the promising solutions for controlling compliance with the speed limit on roads is the use of traffic camera systems. These systems, based on advanced technologies, provide effective traffic monitoring and detect violations such as speeding, which are certainly one of the most common causes of road accidents [2].

A continued focus on road safety is an important task, and effective speed control systems play an important role in achieving this goal. The use of advanced technologies and systems that allow accurate measurement of the speed of vehicles reduces the risk of accidents and contributes to the improvement of road discipline.

Domain analysis

Road speed is an important aspect of road safety, but the lives and safety of thousands of people are put at risk every day because of the unconscious attitude of some drivers to this aspect. Speeding is one of the most common and dangerous causes of road accidents, which leads to serious and tragic consequences for road users. [3].

According to the collected statistics (2011 - 2021), at least 25 percent of the total number of deaths and 10 percent of people injured in road accidents are due to speeding incidents. These statistics are detailed in Table 1 [4-5].

The data from the table were structured and displayed in the form of two diagrams in Figures 1 and 2 [5]. According to the charts presented about the number of people killed and injured in the world from 2011 to 2021 due to speeding, excessive speed has quite serious consequences. During the last decade, the number of people killed due to speeding has increased by 23% - from 10,001 people in 2011 to 12,330 people in 2021. In general, during the studied period, the loss of life due to speeding is approximately 29% of the total number of road accident fatalities, and the number of injured persons was also 13%. These are very alarming indicators that require immediate attention and action.

Table 1

Killed and injured people due to speeding in 2011-2021

Year	Number	Percent	Total	Number	Percent	Total
	Killed			Injured		
2011	10 001	31	32 479	459 776	21	2 227 209
2012	10 329	31	33 782	502 846	21	2 369 083
2013	9 696	29	32 893	383 137	17	2 318 992
2014	9 283	28	32 744	339 189	14	2 342 621
2015	9 723	27	35 484	348 16	14	2 454 778
2016	10 291	27	37 806	376 914	12	3 061 885
2017	9 947	27	37 473	361 95	13	2 745 268
2018	9 579	26	36 835	358 924	13	2 710 059
2019	9 592	26	36 355	326 554	12	2 740 141
2020	11 258	29	38 824	308 013	13	2 282 015
2021	12 330	29	42 939	328 946	13	2 497 657

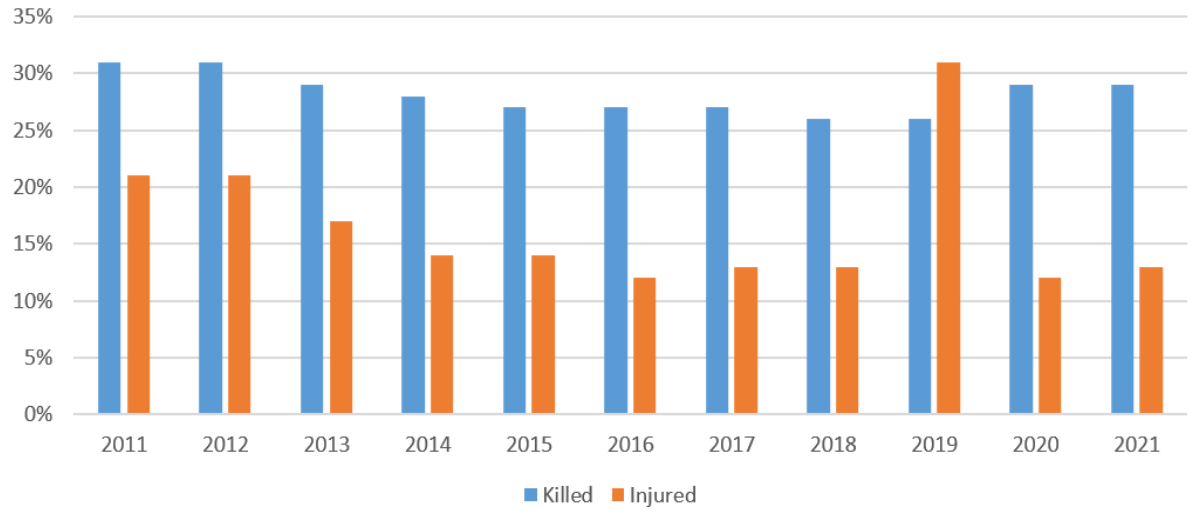

Fig. 1. Percentage representation of killed and injured people due to speeding in 2011-2021

Using the speed control system in a car is a key component of safe road operation and reducing road accidents. This system helps drivers maintain a safe speed and a safe distance from the vehicle in front by providing timely warning of dangerous distance, speeding or sudden braking.

Such speed control systems are becoming increasingly popular in the automotive industry as they demonstrate their effectiveness in preventing accidents and reducing injuries. They help drivers to become more attentive and responsible road users, and also contribute to the preservation of life and health of everyone on the roads [6]. The application of speed control systems is an important step towards achieving a safer and more stable road traffic, which is important for humanity.

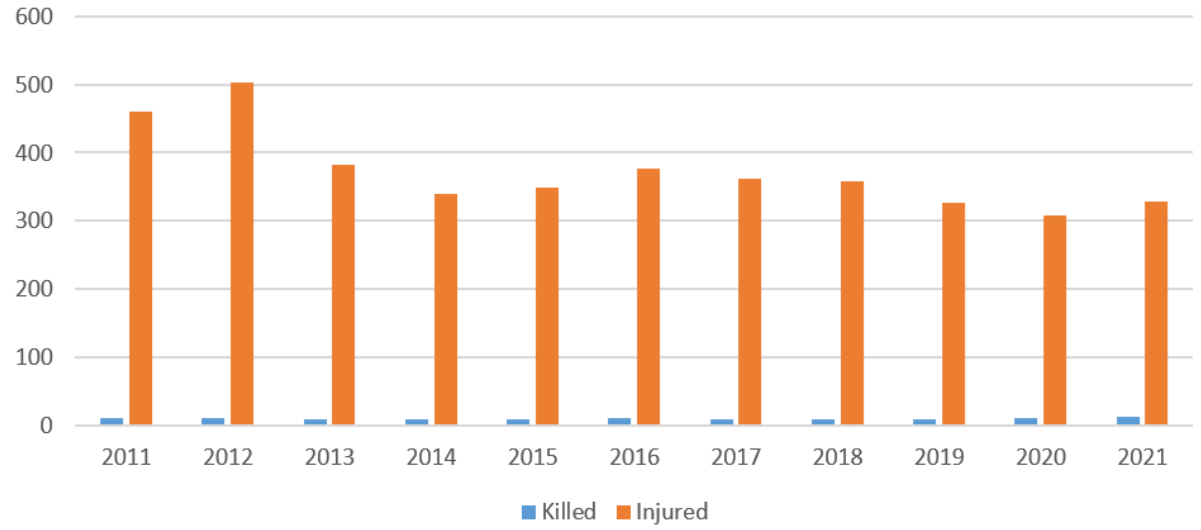

Fig.2. Quantitative representation of killed and injured people due to speeding in 2011-2021

Analysis of existing solutions and technologies

Preventing road accidents and ensuring road safety is a priority for every driver. In this regard, within the framework of the development and improvement of vehicle safety systems, great attention is paid to ready-made solutions that are available on the market. Modern technological progress offers us a variety of systems that help monitor compliance with speed limits, record traffic situations and ensure the preservation of video with the prospect of use as evidence [7].

Overview of the ready-made systems will allow to understand their advantages and make a significant contribution to improving road safety and reducing the risk of road accidents:

1. State speeding video recording system

State speeding video recording systems, located on roads and highways, work on the basis of special cameras and sensors that record the movement of vehicles. Cameras can be placed on stationary posts or mobile devices. When a vehicle exceeds the set speed, the system automatically registers its license plate and time, and then generates a special ticket, which is sent to the owner of the vehicle by mail or email. The owner of the vehicle receives a notification of the violation and instructions to pay the fine. These systems help monitor compliance with speed limits on roads and ensure the safety of road users [8].

2. Dashcam - an application for recording traffic violations

Dashcam is an application that provides video from the car camera recording and stores it in the cloud. The application allows you to view recorded videos, download them to your computer or phone, and share them with others. To use Dashcam, you must first install the app on your phone or tablet. After installation, you need to create an account and connect the car camera to the application. After that, you can start recording the video. Dashcam allows you to record videos in Full HD (1080p) or HD (720p) quality. The app also allows you to adjust recording parameters such as video duration, resolution, and frame rate [9].

3. Speedometer Pro

Speedometer Pro is a mobile application that provides car speed tracking. The app is available for iOS and Android devices and displays the current speed, maximum speed, average speed, distance traveled and driving time. It is possible to view the history of tracking your own speed over a certain period of time [10].

After researching the available driver assistance systems and conducting the analysis, we can proceed to an in-depth comparison of these solutions. This process will allow us to uncover the unique characteristics of such systems, identify key benefits, and consider potential limitations. A comparison of the characteristics of the considered systems is presented in Table 2.

Table 2

Comparison of already existing solutions for speeding control

Characteristic	State speeding video recording systems	Dashcam	Speedometer PRO
Principle of the operation	Video recording of traffic violations using cameras installed on the roads	Video recording of traffic violations using cameras installed in the car	Tracking car speed using GPS
Accessibility	Available in all regions of Ukraine	Available for both - iOS and Android-based devices	Available for both - iOS and Android-based devices
Functionality	Tracking speeding, recording traffic violations, photographing traffic violations	Tracking speeding, recording traffic violations, photographing traffic violations, video recording	Speed tracking, speeding capture
Resource consumption	High	Medium	Low
Video quality	resolution up to 720p, frame rate up to 30 fps	resolution up to 1080p, frame rate up to 30 fps	Information is unavailable
Resource management	Automatic	Manual	Automatic
Advantages	High accuracy, the possibility of bringing violators to justice	High accuracy, the possibility of bringing violators to justice, the possibility of recording violations of traffic rules on roads where there are no cameras	High accuracy, the ability to control the speed of the car
Disadvantages	The cost of cameras installing	The cost of cameras installing	Impossibility of traffic rules violations recording
General impression	A successful system that helps make Ukrainian roads safer	A successful system that helps make Ukrainian roads safer	A successful system that helps drivers to control the car speed

Based on the comparison of existing systems, the authors propose to develop their own device, which will combine various functions from these systems, to increase safety and comfort on the road. Combining the features of existing driver assistance systems, our device will determine the speed of the vehicle in which the system is located, determine the speed of the vehicle ahead, collect information about cars violating traffic rules and take specific measures to punish offenders and keep other drivers on the road safe.

The operation of automated system for determining speed of cars ahead

The system for determining the speed of the car in front can work as a separate link or as part of another system. As a separate component of the system, it increases vehicle safety by providing the driver with information about the speed of vehicles moving in front of him. This information helps the driver to respond in time to changes in the speed of other cars and ensures safer driving.

The system works according to the following algorithm: when the car engine is turned on, this system is turned on together with it. The first step is to find the cars ahead. At this stage, cars are recognized using the Gaar classifier method. The search is conducted until at least one vehicle is found. When a car is found, the system determines its coordinates and assigns an index (serial number) to it. For a given car, the distance traveled in a certain period of time is calculated and, based on this, its speed is calculated.

The next step is to track the detected car and its speed. This stage works while he is in the camera's field of view. If the car begins to brake and is at a short distance from the car in which the system is installed, the user will be notified of this by a special sound signal played from the speaker. This will focus the driver's attention on the situation on the road and give additional time to prevent a possible accident.

Also, when following a car, it may happen that it exceeds the permitted speed. In this case, the system will collect data about the violator and save it in the storage. At the same time, the user will have the opportunity to inform the law enforcement authorities about this event.

Every time the system identifies a new car, this algorithm is restarted. It is also provided for the simultaneous operation of the algorithm when recognizing more than one car.

The graphic representation of the abovementioned algorithm is presented in Figure 3.

For a detailed introduction to the operation of the car speed detection system, a parametric diagram was created (Figure 4). After recognizing the car in front, with the help of a cascade classifier, the coordinates of the car in the image are determined [11]. An integral step will be to determine the Y coordinate of the center of the car, to compare with the Y coordinates of lines A and B to determine the speed of the car. You also need to determine the time of crossing lines A and B, which will help determine the time the car travels the distance between them. The speed of the car will be calculated using the determined parameter of the distance between the lines and the passing time.

Experiments and directions of further work

During the development of the speed detection subsystem, it was decided to conduct testing on video footage from a traffic camera.

During the experiments, it was found that for a more accurate measurement of the speed, it is necessary to reduce the resolution of the video stream. This will reduce the load on the microcontroller and allow more accurate determination of the car's position.

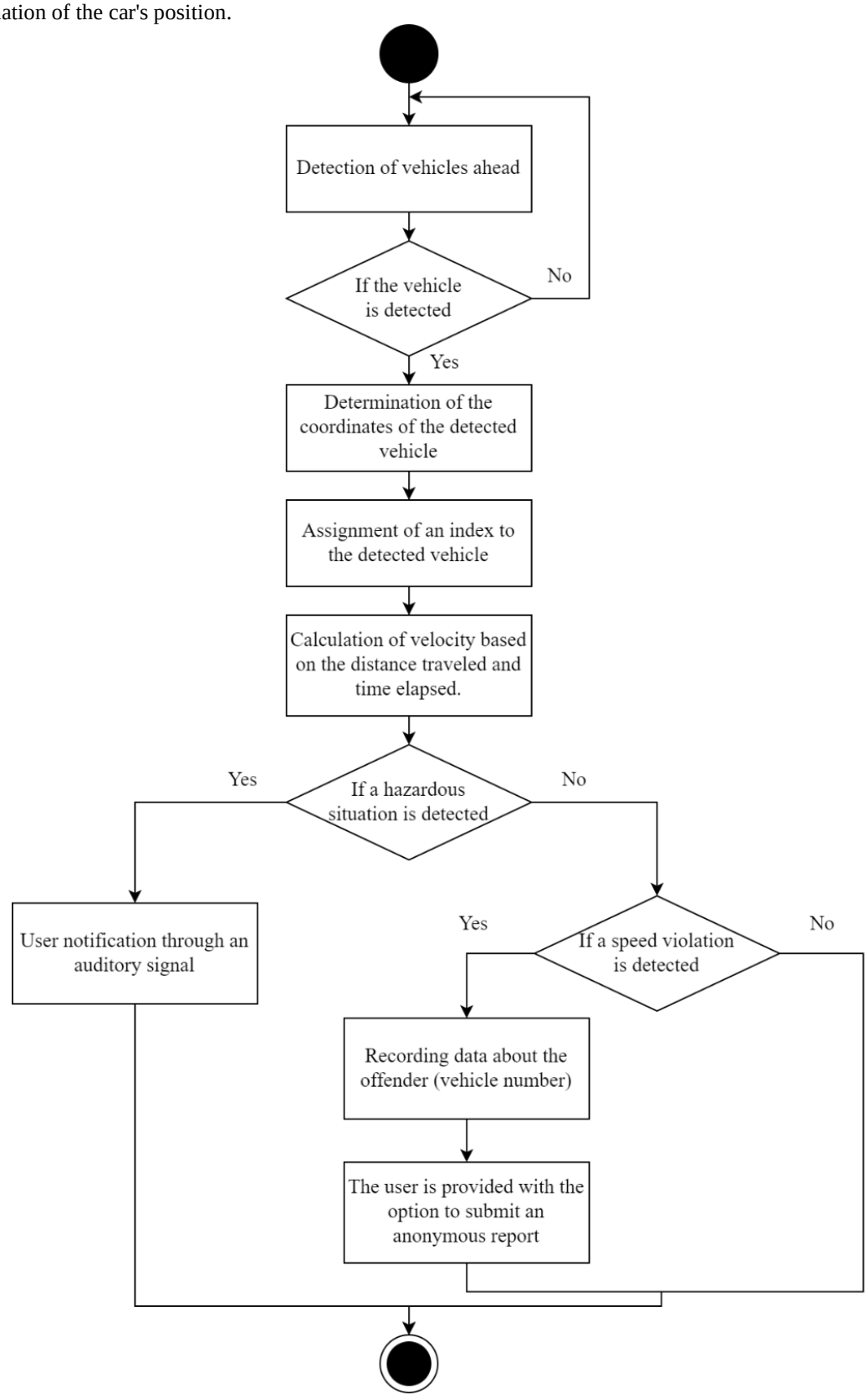

Fig.3. Graphic representation of the algorithm of the automated system for determining the speed of cars moving ahead

The test results are shown in Figures 5 and 6. They show the operation of the system, namely:

- display found cars;
- current state of the system ("Calculating");
- the position of two lines for measuring the time during which the car traveled the distance between them, on the basis of which the speed is measured;
- display of line B in green when a car crosses it;
- speed of the car when crossing line B.

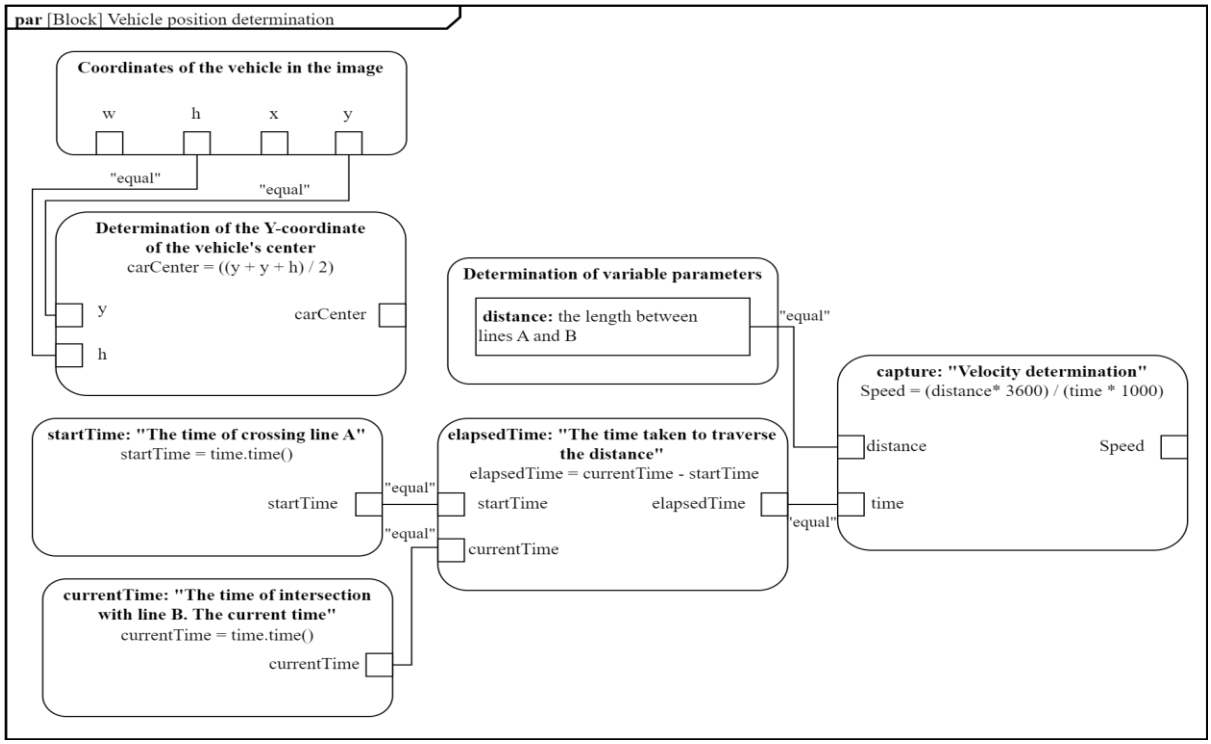

Fig.4. Parametric diagram of car speed determining

To develop a vehicle speed detection device, the first step is to research the necessary functionality and requirements that are important for such a device. One of the important aspects of device development is ensuring high accuracy of speed determination. Accuracy in such a device is a critical factor, as inaccurate speed information can lead to dangerous situations on the road or false alerts to law enforcement.

To develop a vehicle speed detection device, the first step is to research the necessary functionality and requirements that are important for such a device. One of the important aspects of device development is ensuring high accuracy of speed determination. Accuracy in such a device is a critical factor, as inaccurate speed information can lead to dangerous situations on the road or false alerts to law enforcement.

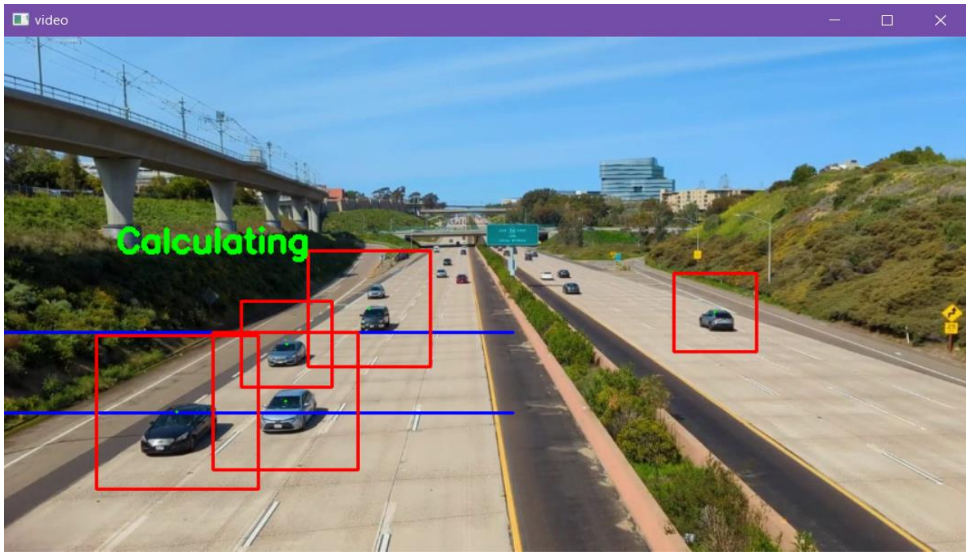

Fig.5. Cars recognition in a video frame

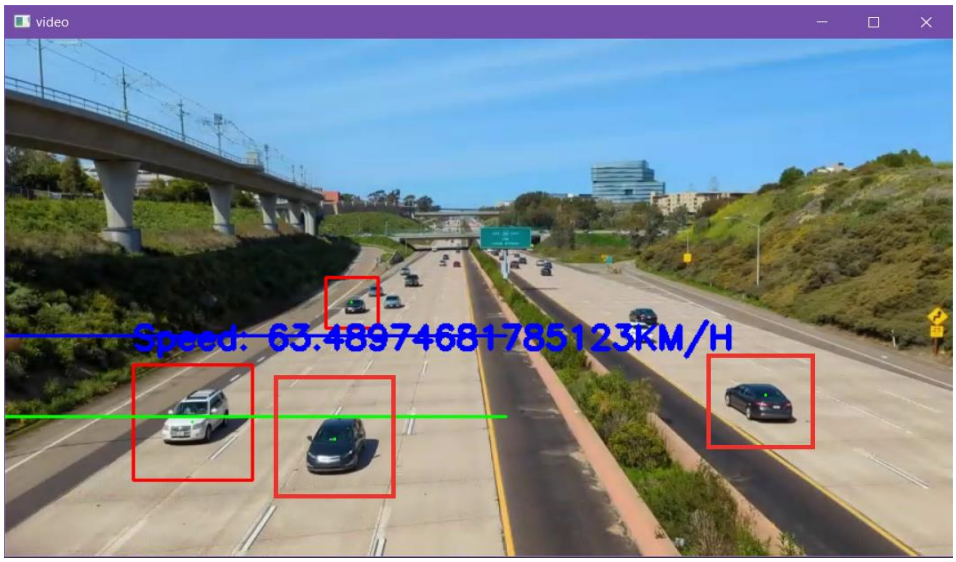

Fig.6. Crossing line B by car

As a result, certain requirements for the further development of the device were formed based on the main factors mentioned above:

1) increase in accuracy - research of new methods of speed measurement;

2) integration with other safety systems - compatibility of the device with other safety systems, such as stability control, automatic emergency braking, etc.;

3) data analysis and forecasting - using data analysis and AI to track various driver behaviors, assess risks and improve systems;

4) cooperation with law enforcement agencies - enabling the user to send information about violators of traffic rules to law enforcement agencies, which can potentially reduce the number of violators on the roads;

5) anonymity - ensuring the anonymity of the user when sending a message about an offense.

Conclusions

Considering the statistics of the number of people killed and injured on the world's roads from 2011 to 2021, related to speeding, it is clear that excessive speed is a serious problem and affects the safety of road users. Over the past ten years, due to the increase in road traffic, the number of victims and victims of speeding accidents has remained high.

After analyzing the ready-made solutions, such as the state video recording system of speeding, Dashcam, Speedometer PRO, the disadvantages and advantages of these systems were taken into account and the work of the own system for determining the speed of the car in front was developed, which will combine various functions from these systems to increase the safety and comfort on the road. Combining the features of existing driver assistance systems, our device will determine the speed of the vehicle in which the system is located, the speed of the vehicle in front, collect information about cars violating traffic rules and will allow the user to send information to law enforcement agencies for the safety of other drivers on the road.

Taking into account the negative impact of speeding on road safety, it is important to actively promote the introduction of speed control systems in cars. These systems allow the car to determine the optimal speed and safe distance to the vehicle in front, as well as provide the driver with appropriate warnings of dangerous situations, such as speeding by the vehicle in front or sudden braking.

Implementation of speed control systems is a mandatory measure to improve road safety and reduce the number of traffic accidents related to non-observance of speed limits. Such technologies will help reduce the number of victims and injured on the roads, save the lives and health of road users and make the roads safer for everyone. Therefore, the use of a speed control system in cars is an important step in improving road safety.

References

1. H. T. Mouftah, M. Erol-Kantarci, S. Sorour,Connected and Autonomous Vehicles in Smart Cities. CRC Press, 2020. pp. 174-177.
2. M. F. Kabir, Roy, S. Real-time vehicular accident prevention system using deep learning architecture. Expert Systems with Applications, 2022.
3. R. S. Job, C. Brodie, Road safety evidence review: Understanding the role of speeding and speed in serious crash trauma: A case study of New Zealand. Journal of road safety, 2022. pp. 5-25.
4. Страшна статистика аварій (accessed July 15, 2023) URL: https://ingo.kiev.ua/tpost/kxgvnta5x1-strashna-statistika-avari.

5. National Center for Statistics and Analysis. Speeding: 2021 data (Traffic Safety Facts. Report No. DOT HS 813 473). National Highway Traffic Safety Administration, 2023. pp. 2. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813473.
6. M. B. A. Rabbani, M. A. Musarat, W. S. Alaloul, S. Ayub, H. Bukhari, M. Altaf, Road accident data collection systems in developing and developed countries: a review. International Journal of Integrated Engineering, 2022. pp. 336-352.
7. D. N. Venu, G. Rakesh, K. Maneesha, K. Anusha, S. Merugu, A. Mohammad, Smart Road Safety and Vehicle Accidents Prevention System for Mountain Road. International Journal from Innovative Engineering and Management Research (IJIEMR), 2022. pp. 38-47.
8. Запуск системи автоматичної фіксації порушень Правил дорожнього руху (accessed July 19, 2023) URL: https://mvs.gov.ua/uk/ministry/projekti-mvs/avtofotovideofiksaciya-porusen-pdr/zapusk-sistemi-avtomaticnoyi-fiksaciyi-porusen-pravil-doroznyogo-ruxu-1.
9. Shift the way you drive. DashcamUA (accesses July 19,2023) URL: https://dashcam.in.ua/.
10. Speedometer PRO (accessed July 19, 2023) URL: https://play.google.com/store/apps/details?id=pl.mobiem.android.speedometer2.
11. R. Singh, A. Gehlot, V. Jain, P. K. Malik. Handbook of research on the internet of things applications in robotics and automation. IGI Global, 2019.
12. T. Hovorushchenko, O. Pavlova, and M. Kostiuk, "Method of Increasing the Security of Smart Parking System", JCSANDM, vol. 12, no. 03, pp. 297–314.

Olga Pavlova **Ольга Павлова**	PhD, Associated professor of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: olya1607pavlova@gmail.com https://orcid.org/0000-0001-7019-0354	доктор філософії, доцент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.
Ada Bilinska **Ада Білінська**	Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: bilinska.ada5@gmail.com	студентка кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.
Andrii Holovatiuk **Андрій Головатюк**	Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: golovatiukao@gmail.com	студент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.
Yaroslav Binkovskyi **Ярослав Бінькowський**	Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: binkovsky22@gmail.com	студент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.
Denys Melnychuk **Денис Мельничук**	Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University, Khmelnytskyi, Ukraine, e-mail: deniska.melnychuk@gmail.com	студент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет, Хмельницький, Україна.

https://doi.org/10.31891/csit-2023-3-5
UDC 004.94

Oleksandr RUDSKYI, Andrii KOPP
National Technical University «Kharkiv Polytechnic Institute»

ALGORITHM AND SOFTWARE TO ASSESSS THE COMPLIANCE OF BUSINESS PROCESS MODELS WITH THEIR TEXTUAL DESCRIPTIONS

This paper is devoted to solve the problem of analyzing the compliance of business process models with their textual descriptions. The problem being business process models describe re-designed or completely new organizational activities, but "wrong" models that do not reflect correctly business process requirements may mislead involved business process participants and other stakeholders, and cause workflow errors followed by extra costs. Therefore, the research goal is to ensure the correctness of business process models by analyzing their compliance with textual descriptions formulated by business process owners or business analysts. In the work, a review of existing tools for modeling and analysis of business processes is outlined, as well as the main technologies of natural language processing are considered, including tokenization, search for stop words, and stemming. These technologies are proposed to be used to analyze the compliance of business process models with their textual descriptions. An approach to solving the problem of analyzing the compliance of business process models with their textual descriptions, using the selected natural language processing tools is proposed and the respective algorithm is developed. The process of analyzing the compliance of business process models with their textual descriptions is formalized using data flow modeling. The corresponding software that implements this process is developed. Sample calculations are demonstrated that confirm the performance of the proposed approach by analyzing the model of the goods dispatch business process and the corresponding textual description of this business process. Finally, conclusions are given and the directions for further work are determined.

In the future, it is necessary to elaborate the software that will help business users analyze the compliance of BPMN models with textual descriptions of depicted workflows, as well as to elaborate the developed algorithm by using advanced artificial intelligence methods, e.g. neural networks, trained on the collection of real business process models.

Keywords: business process model analysis, business process model compliance with textual description, natural language processing, analysis of activity text labels.

Олександр РУДСЬКИЙ, Андрій КОПП
Національний технічний університет «Харківський політехнічний інститут»

АЛГОРИТМ ТА ПРОГРАМНЕ ЗАБЕЗПЕЧЕННЯ ДЛЯ ОЦІНЮВАННЯ ВІДПОВІДНОСТІ МОДЕЛЕЙ БІЗНЕС-ПРОЦЕСІВ ЇХ ТЕКСТОВИМ ОПИСАМ

Дана робота присвячена вирішенню проблеми аналізу відповідності моделей бізнес-процесів їх текстовим описам. Проблема полягає в тому, що моделі бізнес-процесів описують перепроектовані або абсолютно нові процеси діяльності організації, але "неправильні" моделі, які некоректно відображають вимоги до процесів, можуть вводити в оману учасників бізнес-процесів та інших зацікавлених сторін, а також спричиняти помилки в ході виконання бізнес-процесів, що призводять до додаткових витрат. Тому метою дослідження є забезпечення коректності моделей бізнес-процесів шляхом аналізу їх відповідності текстовим описам, сформульованим власниками бізнес-процесів або бізнес-аналітиками. У роботі здійснено огляд існуючих інструментів моделювання та аналізу бізнес-процесів, а також розглянуто основні технології обробки природної мови, серед яких токенізація, пошук стоп-слів та стеммінг. Ці технології пропонується використовувати для аналізу відповідності моделей бізнес-процесів їх текстовим описам. Запропоновано підхід до розв'язання задачі аналізу відповідності моделей бізнес-процесів їх текстовим описам з використанням обраних засобів обробки природної мови та розроблено відповідний алгоритм. Формалізовано процес аналізу відповідності моделей бізнес-процесів їх текстовим описам за допомогою моделювання потоків даних та розроблено відповідне програмне забезпечення, яке реалізує цей процес. Продемонстровано приклади розрахунків, які підтверджують працездатність запропонованого підходу на прикладі аналізу моделі бізнес-процесу доставки продукції та відповідного текстового опису даного бізнес-процесу. На завершення зроблено висновки та визначено напрямки подальших досліджень.

У майбутньому необхідно розробити програмне забезпечення, яке допоможе бізнес-користувачам аналізувати відповідність BPMN-моделей текстовим описам зображених робочих процесів, а також доопрацювати розроблений алгоритм за допомогою сучасних методів штучного інтелекту, наприклад, нейронних мереж, навчених на колекції реальних моделей бізнес-процесів.

Ключові слова: аналіз моделей бізнес-процесів, відповідність моделі бізнес-процесів текстовому опису, обробка природної мови, аналіз текстових міток дій.

Introduction

Business process models have proven themselves to be effective tools for the visualization and improvement of complex organizational operations. Business process models are used to find inefficient places in the described business processes and to eliminate the identified shortcomings by automation with the help of customizable software solutions or unified software environments for the execution of business processes [1].

BPM (Business Process Management) is the concept of managing an organization at the level of processes, which are considered as a business resource that is constantly changing and adapting to changes within and in the environment. The main principles of this concept are transparency and comprehensibility of business processes. To achieve this goal, process modeling is resorted to using notations of a certain stable standard. The most common standard of such notation is BPMN 2.0 (Business Process Model and Notation) [2]. The notation defines a constant

list of elements that are used to build diagrams – models of business processes. Such models, as a rule, include a set of actions and events.

However, the creation of business process models is a time-consuming task that requires significant human resources, so there may be situations in which the business process model does not correspond to the textual description of the business process it is supposed to represent. This can lead to errors in the execution of the business process, loss of time, and, accordingly, unforeseen monetary costs.

Thus, the task of analyzing the compliance of business process models with their textual descriptions is relevant [1] and requires the usage of computational intelligence methods and techniques to estimate the correspondence of BPMN models to workflow descriptions, and assume business process models' correctness and adequacy.

Review of existing business process modeling and analysis solutions

The first among the considered tools is Bonita BPM. Bonita BPM is a business process management software. This software that should be installed on the computer of an analyst or developer. With the help of this tool, users can edit BPMN diagrams, create data models, download user manuals, and design forms [3].

Next among the analogs is another application – Signavio. Signavio Process Manager is a web-based solution for design, analysis, and documenting (modeling) of business processes. This solution allows to create process models as the flowcharts directly in the browser, link any document with a business processes (work procedures, regulations, instructions for the provision of services, etc.), document decisions within processes in a graphical form, export processes in various formats ("png", "svg", "pdf", "xml" for BPMN 2.0) [4].

Another alternative is ProcessMaker, which is the software for business process management. It allows users to effectively model their business processes. The software is fully accessible over the Internet and accessible through any web browser, simplifying the management and coordination of business processes throughout the organization [5].

Bizagi is the software for building business process maps and models in BPMN notation. It allows users to create, interpret, and optimize workflow diagrams using BPMN notation, and publish business process documentation in Word, PDF, Excel, and Wiki formats [6].

As for the brief conclusion on the considered business process modeling and analysis tools, we can formulate the following:

1) all of the considered software tools are relatively easy to use and require only the knowledge of BPMN notation without special training in information technologies;

2) Signavio Process Manager and ProcessMaker are web-based tools, so users do not need to download and install the application;

3) Bonita BPM and Bizagi are less convenient since users have to install the software on their workstations;

4) none of these most well-known and widely-used software tools for business process modeling and analysis allow checking the compliance of BPMN models with the initial descriptions of depicted business processes.

After the analysis of the existing software tools, it can be concluded that all applications greatly simplify the process of building business process models, but these applications do not provide an opportunity to analyze the models for errors. First of all, they do not have the opportunity to compare business process models with their textual descriptions, which can lead to errors in the execution of the business process, loss of time, and significant financial costs when implementing models that are inadequate to the subject area.

Suggested natural language processing techniques

In order to analyze the compliance of business process models with their textual descriptions, it is suggested to apply the following NLP (Natural Language Processing) technologies, briefly described in Fig. 1:

1) tokenization;

2) search for stop words;

3) stemming.

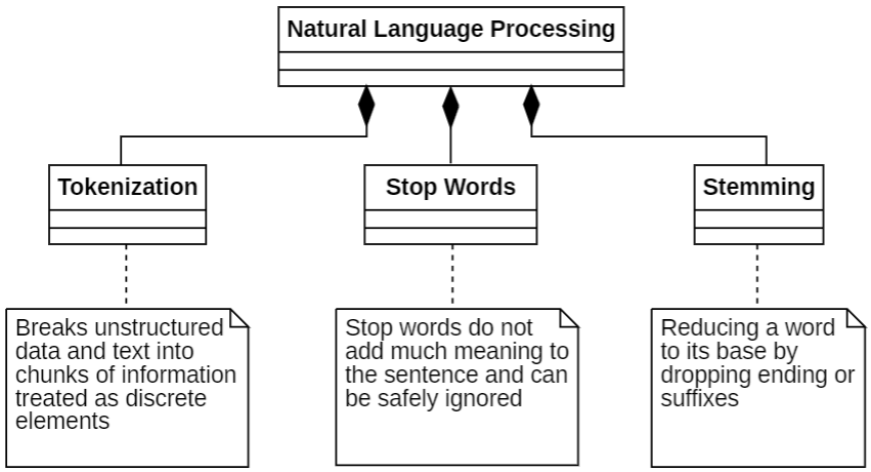

Fig. 1. Considered NLP techniques

Tokenization is the first step in any NLP process. The tokenizer breaks unstructured data and text written in natural language into blocks of information that can be treated as discrete elements. It allows to convert an unstructured string (a text document) into a numerical data structure suitable for machine learning. Tokenization can be used directly as a separate operation or in the process of machine learning as a step followed by more complex actions [7].

Tokenization can be performed for a word, a symbol, or a part of a word [8]:

1) word tokenization is the most commonly used tokenization algorithm, which breaks a piece of text into separate words, taking into account a certain separator; depending on the separators, a variety of word-level markers are formed [8];

2) character tokenization breaks a part of the text into a set of characters [8];

3) tokenization of subwords breaks a text fragment into subwords (or n-gram symbols); for example, such words as “lower” can be segmented as “low-er”, “smartest” as “smart-est” and so on [8].

Once the text is tokenized, it is often clear that not all words carry the same amount of information, if any. Common words that carry little meaningful information are called stop words. Stop words are words in any language that do not add significant meaning to a sentence. We should ignore them without sacrificing the meaning of the sentence. These are some of the most common, short function words such as “the”, “is”, “at”, “which” and “on”. In this case, stop words can cause problems when searching for phrases that include them [9].

If there is a task of text classification or tonality analysis, stop words should be removed because they do not provide any information to the model. That is, to exclude unwanted words from the corpus. But if there is language translation, then stop words should be left, since they should be translated together with other words [9].

Stemming is one of the most common data preprocessing operations performed in almost all NLP projects. Stemming is the process of reducing a word to its base by dropping auxiliary parts such as endings or suffixes. The results of stemming are sometimes very similar to determining the root of a word, but its algorithms are based on different principles. Therefore, the word after processing by the stemming algorithm may differ from the morphological root of the word [10].

There are several variants of stemming algorithms, which differ in their accuracy and performance:

1) search by a table – this algorithm uses a principle of searching by a table in which all possible variants of words and their forms after stemming are collected [10];

2) cutting off endings and suffixes – these algorithms are based on the rules according to which a word can be shortened [10];

3) lemmatization is a more complex approach based on determining the base of the word through lemmatization; the first step of this algorithm is the determination of parts of speech (POS) in a sentence, i.e. “POS tagging”, in the second step, stemming rules are applied to the word according to the part of speech [10];

4) stochastic algorithms – these algorithms are based on the probability of determining the basis of a word [10];

5) hybrid approach – when building a hybrid stemming algorithm, a combination of the above algorithms can be used; for example, the algorithm can use the method of cutting off endings and suffixes, but at the first stage perform a table search [10];

6) matching search – these algorithms use a knowledge base that contains only the bases of words, that is, this knowledge base consists of those words into which ordinary words are transformed after stemming [10].

Object, subject, and methods of research

Existing software tools provide for the construction of business process models but do not provide an opportunity to analyze these models from the point of view of their adequacy to real business processes, namely, to

compare business process models with their textual description. Therefore, the development of an approach to the analysis of the compliance of business process models with their textual descriptions is relevant.

The research object is the process of analyzing the compliance of business process models with their textual descriptions. The research subject is an algorithm for analyzing the compliance of business process models with their textual descriptions. The research purpose is to ensure the adequacy of business process models by analyzing their compliance with textual descriptions.

Thus, it is necessary to form the text T_1 from all names of elements of type "task" and related elements of actions. The following algorithm that can be used for this task is represented by the UML activity diagram in Fig. 2.

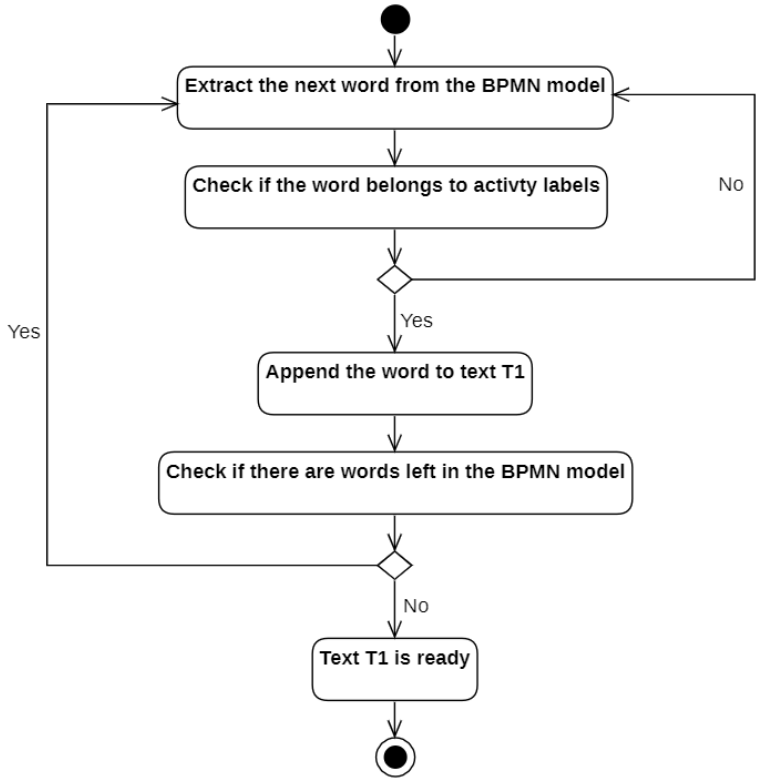

Fig. 2. Algorithm for generating text from task names of the business process model

The algorithm and respective software should extract from the BPMN file all the names of elements of type "task" and related elements of actions:

1) "Service Task" is a task that uses some service, which can be a web service or an automated application [11];

2) "Send Task" is a simple task designed to send a message to an external participant. As soon as the message is sent, the task is completed [11];

3) "Receive Task" is a simple task designed to wait for receiving a message from an external user [11];

4) "User Task" is a typical task of a business process in which a human executor performs a task with the help of a software application and is scheduled through some task list manager [11];

5) "Manual Task" is a task that is supposed to be performed without the help of any business process execution mechanism or any program [11];

6) "Business Rule Task" is a task that provides a mechanical process to provide input data for the business rule mechanism and obtain the output data of calculations that the business rule mechanism can provide [11];

7) "Script Task" is a task that is executed by the business process engine. When the task is ready to run, the engine will execute the script. After completing the script, the task will also be executed [11].

Along with the BPMN file of the business process model, the input is also a textual description T_2 of the business process, which this model should represent.

Thus, to solve the problem of analyzing the compliance of business process models with their text descriptions using the received texts T_1 and T_2, the following actions must be performed:

1) split the input texts T_1 and T_2 into separate words (tokenize) and obtain the corresponding multisets of words:

$$(W_1, m_1) = \left\{\left(t_i^1, m_1(t_i^1)\right), t_i^1 \in W_1 \wedge i = \overline{1,n}\right\},$$
$$(W_2, m_2) = \left\{\left(t_j^2, m_2(t_j^2)\right), t_j^2 \in W_2 \wedge j = \overline{1,q}\right\}, \quad (1)$$

where:

– W_1 is the set of words obtained as a result of the tokenization of text T_1;

– W_2 is a set of words obtained as a result of the tokenization of text T_2;

– $t_i^1 \in W_1$, $i = \overline{1,n}$ – word obtained as a result of tokenization of text T_1;

– $t_j^2 \in W_2$, $j = \overline{1,q}$ – word obtained as a result of tokenization of text T_2;

– $m_1(t_i^1)$ – mapping $m_1: W_1 \to \mathbb{Z}^+$, which for each word $t_i^1 \in W_1$, $i = \overline{1,n}$ sets the number of its repetitions in the text T_1;

– $m_2(t_j^2)$ – mapping $m_2: W_2 \to \mathbb{Z}^+$, which for each word $t_j^2 \in W_2$, $j = \overline{1,q}$ sets the number of its repetitions in the text T_2;

– n is the number of words obtained as a result of the tokenization of text T_1;

– q is the number of words obtained as a result of the tokenization of text T_2;

2) remove stop words from sets W_1 and W_2 to obtain sets of only meaningful terms related to the subject area of the business process:

$$stop: \{W_k, k = \overline{1,r}\} \to \{W_k', k = \overline{1,r}\}, \quad (2)$$

where:

– W_k, $k = \overline{1,r}$ is the set of words obtained as a result of tokenization of the source text, which also contains stop words;

– W_k', $k = \overline{1,r}$ – set of words obtained as a result of tokenization of the source text, from which stop words are removed;

– $stop$ is a mapping that, for each set W_k, $k = \overline{1,r}$, which contains stop words, matches the set W_k', $k = \overline{1,r}$, which does not contain stop words;

– r is the number of sets of words processed, $r = 2$;

3) perform stemming of words in sets W_1' and W_2', remaining after removing stop words:

$$stemm: \{W_k', k = \overline{1,r}\} \to \{W_k'', k = \overline{1,r}\}, \quad (3)$$

where:

– W_k'', $k = \overline{1,r}$ is the set of words obtained as a result of stemming the words remaining after the removal of stop words;

– $stemm$ is a mapping which, for each set W_k', $k = \overline{1,r}$, from which the stop words are removed, matches the set W_k'', $k = \overline{1,r}$ in which words remained after the removal of the stop words, are shortened to the base.

Thus, as a result of the previous actions (1) – (3), two sets of words W_1'' and W_2'' will be obtained:

$$W_1'', W_2'' \in \{W_k'', k = \overline{1,r}\}, \quad (4)$$

where:

– W_1'' is the set of words obtained after the processing steps (1) – (3) are completed for the text T_1 built using all names of elements of type "task" and related elements of BPMN business process model actions;

– W_2'' is a set of words obtained after the processing steps (1) – (3) are completed for the textual description T_2 of the business process, which the BPMN model should represent.

The similarity of these two sets of words W_1'' and W_2'' (4) can be calculated using one of the distance metrics [12]. The Jacquard coefficient is suggested for use in the problem-solving algorithm because it gives an accurate estimate in the range between 0 and 1, and at the same time is simple to implement.

Hence, the obtained value of the Jaccard coefficient [12] can be interpreted as the degree of compliance of the business process model with its textual description:

$$K_J = \frac{|w_1'' \cap w_2''|}{|w_1''| + |w_2''| - |w_1'' \cap w_2''|}. \quad (5)$$

Thus, the paper elaborates the following algorithm, earlier proposed by authors in [13], for solving the task of analyzing the compliance of business process models with their textual descriptions, represented by the UML activity diagram in Fig. 3.

Since the Jaccard similarity coefficient [12] produces values in the 0 – 1 range, it is proposed to measure the degree of correspondence of BPMN workflow diagrams to the initial textual descriptions of ongoing or planned business processes. For this purpose, we suggest using the psychophysical Harrington's scale of quality [14]:

1) 0.00 – 0.20 for the "very bad" compliance;

2) 0.21 – 0.37 for the "bad" compliance;

3) 0.38 – 0.63 for the "satisfactory" compliance;

4) 0.64 – 0.80 for the "good" compliance;

5) 0.81 – 1.00 for the "very good" compliance.

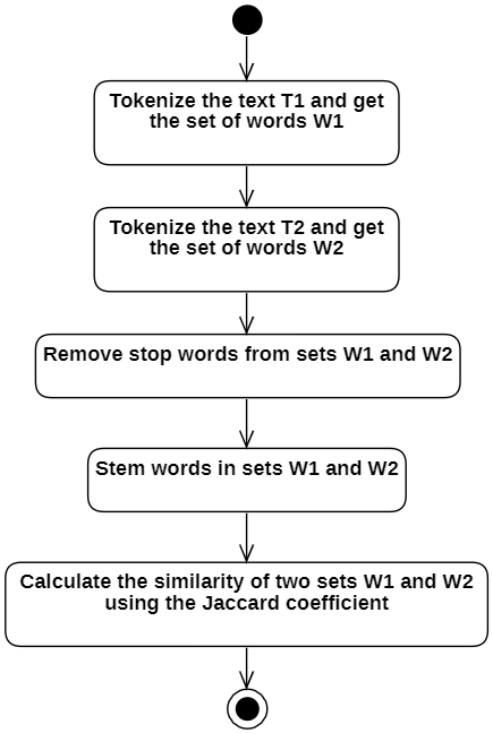

Fig. 3. Algorithm for solving the problem of analyzing the compliance of business process models with their textual descriptions

Results and discussion

In the process of analyzing the compliance of business process models with their text descriptions, the BPMN file of the model and the corresponding text description of the real business process are used. Compliance check is carried out on the basis of the developed algorithm. Future software based on the proposed algorithm should generate a report based on the verification result. It is intended that the software will be used by both registered and non-registered users. The software requirements were earlier formulated by authors in [15].

Using the DFD (Data Flow Diagram) modeling method, a context diagram of the process of business process models compliance analysis with their textual descriptions is designed, which is shown in Fig. 4.

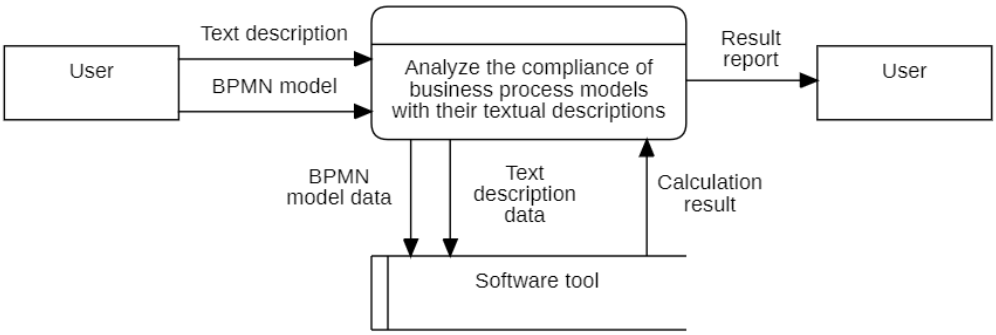

Fig. 4. Process context diagram

This diagram (Fig. 4) shows inputs and an output, the software tool implementing the algorithm given in Fig. 3, and involved participants (users) that analyze the compliance of business process models with their textual descriptions.

A decomposition diagram of the process of analyzing the compliance of business process models with their textual descriptions is designed, which is shown in Fig. 5.

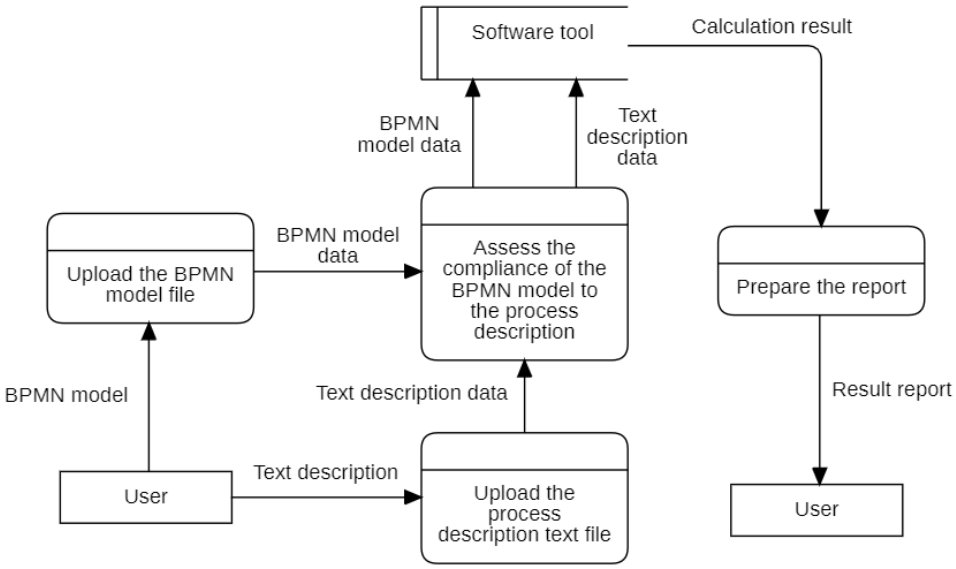

Fig. 5. Process decomposition diagram

This diagram (Fig. 5) shows the activities of BPMN models and text descriptions uploading, analysis of the compliance of business process models with their textual descriptions using the proposed algorithm (Fig. 3), and report generation with the provided evaluation and recommendations.

Let us consider the sample business process model of the goods dispatch process that happens in a small hardware store (Fig. 6) [16].

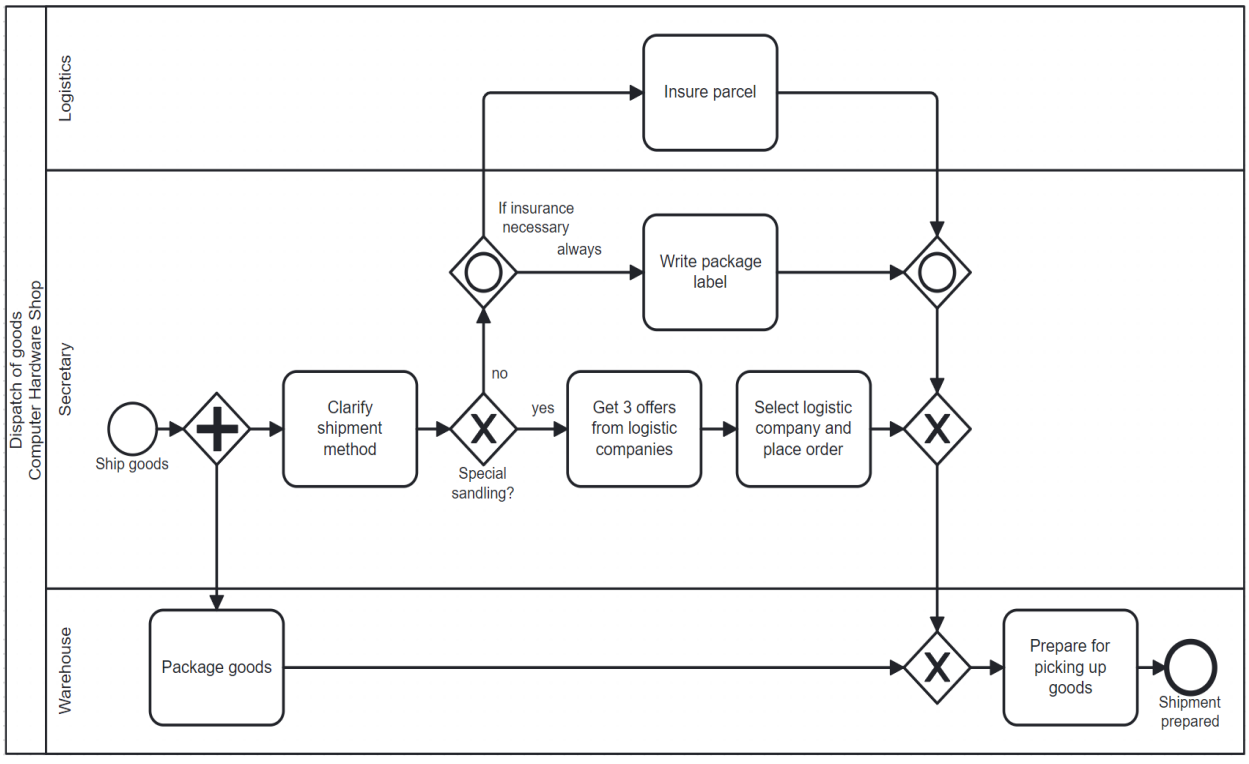

Fig. 7. Goods dispatch BPMN business process model

According to [16], this business process model represents the process described using the following text: “If goods shall be shipped, the secretary clarifies who will do the shipping. If you have large amounts, special shipping will be necessary. In these cases, the secretary invites three logistic companies to make offers and she selects one of them. In case of small amounts, normal post shipment is used. Therefore, a package label is written by the secretary and a parcel insurance taken by the logistics department head if necessary. In the meantime, the goods can be already packaged by the warehousemen. If everything is ready, the packaged goods are prepared for being picked up by the logistic company.”

Let us use Python and NLTK (Natural Language Toolkit) [17] to implement the proposed approach and verify its efficiency. Input texts T_1 and T_2 are given in Fig. 8.

```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer

T2 = 'If goods shall be shipped, the secretary clarifies who will do the shipping. If you have large amounts, special shipping wi
print('\nT2:', T2)

bpmn_activities = ['Clarify shipment method', 'Package goods', 'Get 3 offers from logistic companies', 'Insure parcel', 'Write pa
print('\nBPMN activities:', bpmn_activities)
```

```
T2: If goods shall be shipped, the secretary clarifies who will do the shipping. If you have large amounts, special shipping wi
ll be necessary. In these cases the secretary invites three logistic companies to make offers and she selects one of them. In c
ase of small amounts, normal post shipment is used. Therefore a package label is written by the secretary and a parcel insuranc
e taken by the logistics department head if necessary. In the meantime the goods can be already packaged by the warehousemen. I
f everything is ready, the packaged goods are prepared for being picked up by the logistic company.

BPMN activities: ['Clarify shipment method', 'Package goods', 'Get 3 offers from logistic companies', 'Insure parcel', 'Write p
ackage label', 'Select logistic company and place order', 'Prepare for picking up goods']
```

```
T1 = ' '.join(bpmn_activities).strip()
print('\nT2:', T1)
```

```
T2: Clarify shipment method Package goods Get 3 offers from logistic companies Insure parcel Write package label Select logisti
c company and place order Prepare for picking up goods
```

Fig. 8. Input texts declared in the Python notebook

Fig. 9 demonstrates steps of the proposed algorithm (Fig. 3) sequentially applied to input texts T_1 and T_2, and obtained sets of words W_1'' and W_2'' as the result.

```
stop_words = set(stopwords.words('english'))
porter = PorterStemmer()
```

```
W1 = word_tokenize(T1)
W1 = [word.lower() for word in W1]
W1 = [word for word in W1 if word.isalpha()]
W1 = [word for word in W1 if not word in stop_words]
W1 = [porter.stem(word) for word in W1]
W1 = list(dict.fromkeys(W1))
print('\nW1:', W1)
```

```
W1: ['clarifi', 'shipment', 'method', 'packag', 'good', 'get', 'offer', 'logist', 'compani', 'insur', 'parcel', 'write', 'labe
l', 'select', 'place', 'order', 'prepar', 'pick']
```

```
W2 = word_tokenize(T2)
W2 = [word.lower() for word in W2]
W2 = [word for word in W2 if word.isalpha()]
W2 = [word for word in W2 if not word in stop_words]
W2 = [porter.stem(word) for word in W2]
W2 = list(dict.fromkeys(W2))
print('\nW2:', W2)
```

```
W2: ['good', 'shall', 'ship', 'secretari', 'clarifi', 'larg', 'amount', 'special', 'necessari', 'case', 'invit', 'three', 'logi
st', 'compani', 'make', 'offer', 'select', 'one', 'small', 'normal', 'post', 'shipment', 'use', 'therefor', 'packag', 'label',
'written', 'parcel', 'insur', 'taken', 'depart', 'head', 'meantim', 'alreadi', 'warehousemen', 'everyth', 'readi', 'prepar', 'p
ick']
```

Fig. 9. Results obtained using the proposed algorithm

The following figure (Fig. 10) demonstrates the computed value of the Jaccard coefficient [12] together with the cardinalities of sets of words W_1'' and W_2'', and their intersection $|W_1'' \cap W_2''|$.

```
card_W1 = len(W1)
print('\ncard(W1):', card_W1)

card_W2 = len(W2)
print('\ncard(W2):', card_W2)

card_W1_intersect_W2 = len(set(W1).intersection(set(W2)))
print('\ncard(W1 ∩ W2):', card_W1_intersect_W2)

jaccard = card_W1_intersect_W2 / (card_W1 + card_W2 - card_W1_intersect_W2)
print('\njaccard:', round(jaccard, 2))
```

```
card(W1): 18

card(W2): 39

card(W1 ∩ W2): 13

jaccard: 0.3
```

Fig. 10. Computed value of the Jaccard coefficient

Results given in Fig. 10 confirm that the similarity between the BPMN model representing goods dispatch business process (Fig. 7) and its textual description [16] is 0.3, according to the Jaccard coefficient [12].

Therefore, the compliance of the business process model (Fig. 7) with its text description can be estimated as 30%, which means that some information about the business process given in its description is missing in the BPMN model (i.e. $|W_1''| = 18 < |W_2''| = 39$).

The software solution for analyzing the compliance of business process models with their textual descriptions was developed using the .NET platform and the C# programming language [18]. It is proposed to choose a three-tier architecture (Fig. 11) to build the software for analyzing the compliance of business process models with their textual descriptions. The 3-tier architecture assumes that a software application consists of three components: a client application that interacts with an application server, which is connected to a database server (we use Microsoft SQL Server) [19].

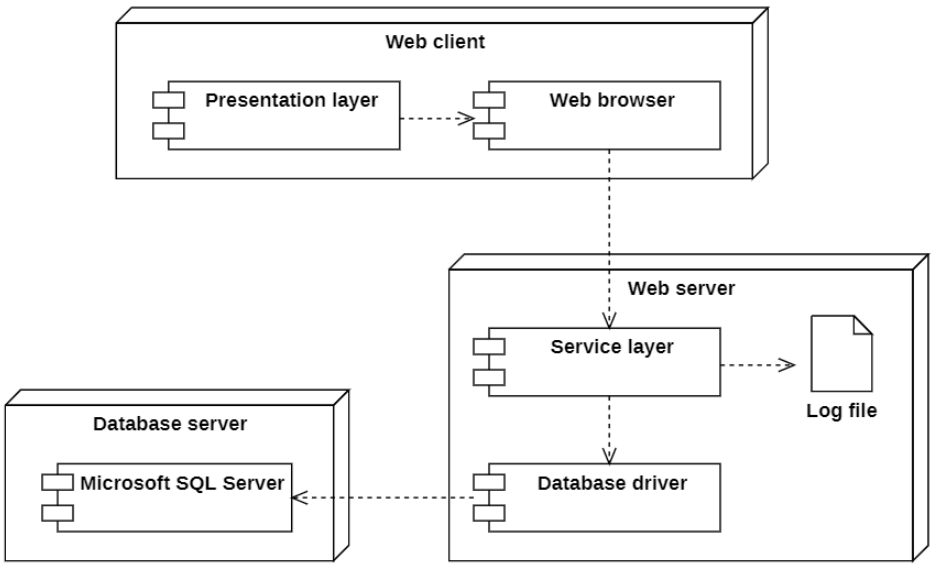

Fig. 11. Software architecture design

When the users open the application, they are taken to the main page of the website. The user should specify the name of a business process, upload a BPMN file and a text file describing the model. Then the user is redirected to the results page. It demonstrates the compliance of the business process model with its textual description, and the number of words in the result sets of words W_1'' and W_2'' are demonstrated on the results page as well (Fig. 12).

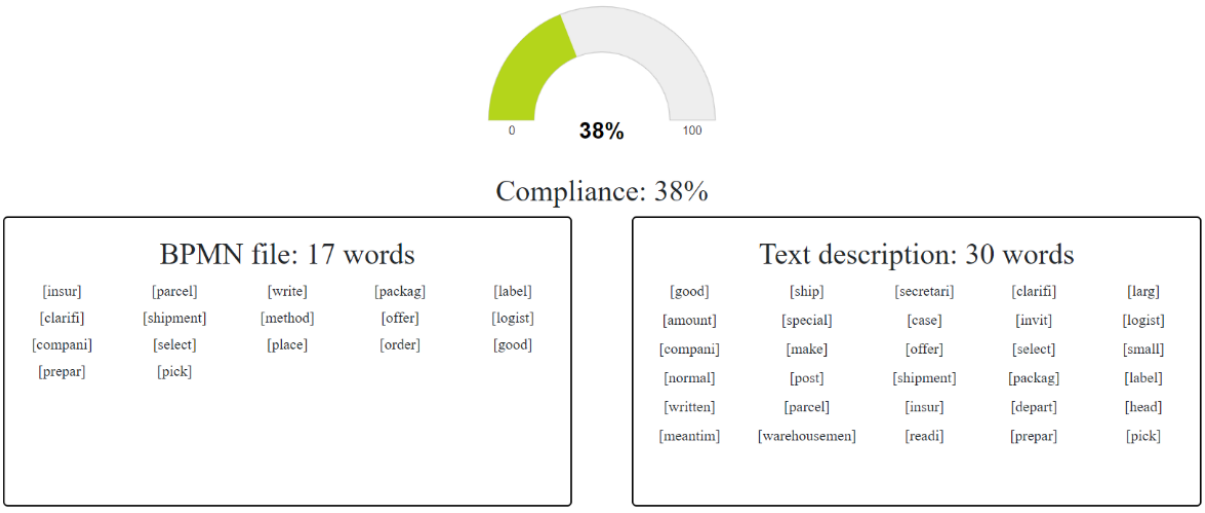

Fig. 12. Results page demonstrates obtained results

However, this approach has the limitation connected to the fact that rich business process descriptions may cause lower similarity values and, thus, signalize the false lower compliance of business process models. Hence, this limitation can be bypassed by the inclusion of other BPMN elements, such as events and gateways, into consideration when measuring the similarity index. Another limitation of the proposed approach is connected with the possible presence of synonyms and phrases that should not be divided into different words to keep their meaning.

Conclusions

In this research, the business process modeling is considered and the relevance of the problem of analysis of the compliance of business process models with their textual descriptions is defined. The following conclusions can be made:

1) the analysis and comparison of existing software tools for business process modeling and analysis have shown that widely-used and well-known tools do not allow checking the compliance of BPMN workflow diagrams with their textual descriptions;

2) an overview of NLP techniques, such as tokenization, removal of stop words, and stemming are discussed and considered as methods that can be used to solve the problem of analyzing the compliance of business process models with their textual descriptions;

3) the proposed algorithm based on NLP techniques solves the formulated problem of the analysis of business process models compliance with respective textual descriptions and produces estimations in the range between 0 and 1 for the very bad and very good compliance respectively.

In the future, it is necessary to elaborate the software that will help business users analyze the compliance of BPMN models with textual descriptions of depicted workflows, as well as to elaborate the developed algorithm by using advanced artificial intelligence methods, e.g. neural networks, trained on the collection of real business process models.

References

1. Suchenia A. Overview of verification tools for business process models / A. Suchenia, P. Wiśniewski, A. Ligęza // Annals of Computer Science and Information Systems. – 2017. – Vol. 13. – P. 295-302.
2. Allweyer T. BPMN 2.0: introduction to the standard for business process modeling / T. Allweyer // BoD – Books on Demand. – 2016. – 172 p.
3. Be Efficient: Bonitasoft Introduces New Bonita BPM 6 Platform, https://www.businesswire.com/news/home/20130605006087/en/Efficient-Bonitasoft-Introduces-Bonita-BPM-6-Platform, 30.05.2023.
4. SAP Signavio, https://www.signavio.com/, 30.05.2023.
5. ProcessMaker Review, https://comparecamp.com/processmaker-review-pricing-pros-cons-features/, 30.05.2023.
6. Process Documentation – Bizagi Models, https://help.bizagi.com/process-modeler/en/index.html?where_to_share2.htm, 30.05.2023.
7. Tokenization in NLP: Types, Challenges, Examples, Tools, https://neptune.ai/blog/tokenization-in-nlp, 30.05.2023.
8. Tokenization in NLP, https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/, 30.05.2023.
9. Stop Words in NLP, https://medium.com/@saitejaponugoti/stop-words-in-nlp-5b248dadad47, 30.05.2023.
10. Jongejan B. Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike / B. Jongejan, H. Dalianis // Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. – 2009. – P. 145-153.
11. Business Process Model and Notation (BPMN), Version 2.0, https://www.omg.org/spec/BPMN/2.0/PDF, 30.05.2023.
12. Warrens M. J. Similarity coefficients for binary data: properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients / M. J. Warrens // Leiden University. – 2008. – 254 p.

13. Rudskyi O. V. Towards the analysis of the compliance of business process models with their textual descriptions / O. V. Rudskyi, A. M. Kopp, D. L. Orlovskyi // Information technologies: science, technology, technology, education, health: abstracts of reports of XX international scientific and practical conference MicroCAD-2022. – 2022. – P. 822.
14. Tielietov O. S. Four-vector efficiency of infrastructure in the system of providing regional socially significant needs taking into account the concept of marketing of changes / O. S. Tielietov, N. Y. Letunovska, Y. M. Melnyk // Bioscience Biotechnology Research Communications. – 2019. – Vol. 12, N. 3. – P. 637-645.
15. Rudskyi O. V. Software requirements elicitation for the analysis of the compliance of business process models with their textual descriptions / O. V. Rudskyi, A. M. Kopp // Materials of the 5th All-Ukrainian scientific and practical Internet conference of students, postgraduates and young scientists on the topic "Modern computer systems and networks in management". – 2022. – P. 158-161.
16. BPMN for research, https://github.com/camunda/bpmn-for-research, 30.05.2023.
17. NLTK :: Natural Language Toolkit, https://www.nltk.org/, 30.05.2023.
18. A tour of the C# language // URL: https://learn.microsoft.com/uk-ua/dotnet/csharp/tour-of-csharp/, 30.05.2023.
19. Fowler M. Patterns of Enterprise Application Architecture / M. Fowler // CreateSpace Independent Publishing Platform. – 2017 – 560 p.

Oleksandr Rudskyi **Олександр Рудський**	Master's Student of Software Engineering and Management Intelligent Technologies Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, e-mail: oleksandr.rudskyi@cs.khpi.edu.ua. https://orcid.org/0009-0001-1130-9957	магістрант кафедри програмної інженерії та інтелектуальних технологій управління, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна
Andrii Kopp **Андрій Копп**	PhD, Associate Professor, Associate Professor of Software Engineering and Management Intelligent Technologies Department, National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine, e-mail: andrii.kopp@khpi.edu.ua https://orcid.org/0000-0002-3189-5623	доктор філософії, доцент, доцент кафедри програмної інженерії та інтелектуальних технологій управління, Національний технічний університет «Харківський політехнічний інститут», Харків, Україна

https://doi.org/10.31891/csit-2023-3-6
UDC 004.94

Olha KOVALCHUK
West Ukrainian National University

A MACHINE LEARNING CLUSTER MODEL FOR THE DECISION-MAKING SUPPORT IN CRIMINAL JUSTICE

In today's digital society, information technology plays a crucial role in supporting decision-making in the development of national security policy. The growing number of criminals and the expanding range of crimes they commit, observed around the world, pose serious risks to the personal safety of citizens, internal security of the country and international security. Advanced technologies are transforming the functioning of the criminal justice system, penitentiary system and police. The use of intelligent technologies to optimize the work of criminal justice agencies is the basis of the concept of "smart criminal justice". Computer models are used to support crime investigation, automate court proceedings, identify potentially dangerous individuals and predict crimes.

This paper is a part of the research on the development of information and analytical support for a decision support system in criminal justice. This paper presents a new analytical approach to criminal profiling. The empirical analysis was conducted on the basis of real data from criminal records of 13010 prisoners serving their sentences in penitentiary institutions of Ukraine. The k-means clustering technique was used to identify significant indicators (individual characteristics of prisoners) that determine the propensity of convicts to commit repeated criminal offenses. The built computer model can explain the connection between the propensity of convicts to recidivate and the following elements of the offender's profile: the number of previous convictions, age at the time of the first conviction, the presence of suspended sentences and early release.

The judicial system needs to understand such non-obvious relationships to effectively prevent and solve criminal offenses. The developed computer model can be applied to new datasets and ensure the elimination of subjectivity and bias.

Keywords: decision-making support, machine learning; criminal profiling, k-means clustering, recidivism, information, analytical support

Ольга КОВАЛЬЧУК
Західноукраїнський національний університет

КЛАСТЕРНА МОДЕЛЬ МАШИННОГО НАВЧАННЯ ДЛЯ ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ У КРИМІНАЛЬНОМУ ПРАВОСУДДІ

У сучасному цифровому суспільстві інформаційні технології відіграють вирішальну роль у підтримці прийняття рішень при формуванні національної політики безпеки. Зростанні кількості злочинців та розширення спектру скоєних ними злочинів, що спостерігається по всьому світу, складає серйозні ризики для особистої безпеки громадян, внутрішньої безпеки країни та міжнародної безпеки. Передові технології трансформують функціонування системи кримінального правосуддя, пенітенціарної системи та поліції. Використання інтелектуальних технологій для оптимізації роботи органів кримінальної юстиції є основою концепції створення "розумного кримінального правосуддя". Для підтримки розслідування злочинів, автоматизації судових процесів, ідентифікації потенційно небезпечних осіб та передбачення злочинів застосовують комп'ютерні моделі.

Представлена робота є частиною досліджень з питань розробки інформаційно-аналітичного забезпечення системи підтримки прийняття рішень у кримінальний юстиції. У цьому документі представлено новий аналітичний підхід до кримінального профілювання. Емпіричний аналіз проведено на основі реальних даних кримінальних записів 13010 ув'язнених, які відбувають покарання у пенітенціарних закладах України. Використано техніку кластеризації k середніх для визначення значущих індикаторів (індивідуальних характеристик ув'язнених), які визначають схильність засуджених до вчинення повторних кримінальних злочинів. Побудована комп'ютерна модель може пояснити зв'язок між схильністю засуджених до кримінального рецидиву та наступними елементами профіля злочинця: кількістю попередніх судимостей, віком на момент першого засудження, наявністю умовних судимостей та дострокових звільнень.

Судова система потребує розуміння таких неочевидних взаємозв'язків для ефективного запобігання та розкриття кримінальних злочинів. Розроблена комп'ютерна модель може бути застосована до нових наборів даних та забезпечити усунення суб'єктивності та упередженості.

Ключові слова: підтримка прийняття рішень, машинне навчання, кримінальне профілювання, кластеризація k-середніх, рецидивізм, інформаційно-аналітичне забезпечення

Introduction

The efficiency of the internal security system is an integral aspect of the country's external security, international security in general and attracts the attention of many politicians and officials. The effectiveness of the criminal justice system is increasingly the subject of discussion in academic circles. The justice of legal decisions is causing more and more disputes in society, as a part of the judicial system is under pressure from the government and politics. Thus, as the number of prisoners grows rapidly, it is logical that the police and justice authorities increasingly use decision-support systems to support decision-making processes. A decision support system (DSS) is an interactive computer system designed to support various activities when making decisions on unstructured and semi-structured decision problems. Such a system has the ability to work with interactive requests [1].

Most decision support systems, used in criminal justice (decision support systems in justice, DSSJ), are based on machine learning systems that use algorithms capable of learning from previous data and making predictions [2]. This increases the predictive probability of committing repeated criminal offenses, which is much more effective than

the analysis of significant factors affecting the propensity of criminals to recidivism in each individual case. In addition, such computer algorithms are devoid of bias and subjectivity. Information systems that use such algorithms can provide reliable support for decision-making regarding the assessment of the risks of committing repeated criminal offenses by individuals and contribute to the improvement of the effectiveness of intelligent policing.

With the increase in the range of tasks that are solved with the help of DSSJ, the requirements for their use are increasing. Information technologies are constantly being improved. However, their application to support the decision-making process concerning the safety of the individual and society, the freedom and even the life of a person who may be falsely accused, imposes high requirements on the accuracy, reliability, and transparency of algorithms. The results of using such computer models should be comparable and verified on a large number of cases. The use of decision-making systems in the criminal justice system should help reduce the level of crime, and increase the level of personal safety of citizens and the safety of society.

Related works

The effectiveness of the use of DSS, and computer technologies, in particular big data and machine learning, in policing and the criminal justice system is one of the important and debatable topics in scientific literature today. B. Simmler studied the degree of distribution, implementation possibilities, technical features, institutional implementation and psychological aspects of the use of algorithms in the criminal justice system on the example of Switzerland [3]. O. Doyle studied the issue of efficiency in applying the DSS for reducing the level of crime [4]. A. Završnik studied the influence of big data, algorithmic analytics, and machine learning tools on knowledge production in criminal justice settings [5]. B. Benbouzid conducted a detailed analysis of the content of predictive policing applications [6]. R. Berk assessed the risks of using artificial intelligence in law enforcement agencies [7]. S. Brayne studied the problems of using big data by law enforcement agencies [8]. M. Cavelty reviewed various analytical tools used in Switzerland to identify the recursive and non-linear relationship between security policy and technologies such as predictive policing, artificial intelligence and spyware [9]. S. Egbert et al investigated crime forecasting technologies based on data analysis and algorithmic detection of patterns, which are used to prevent criminal offenses as elements of preventive strategies in German-speaking countries [10]. A. Sandhu and P. Fussey analyzed the advantages of predictive policing, focused on the automation of police decisions, and the ability of predictive computer software to neutralize the subjectivity of police work [11]. B. Cheng et al. applied the FP-growth algorithm to find association rules between criminals and innocent people in order to identify persons suspected of crimes. [12]. K. Kotsoglou et al. investigated the possibilities of automated facial recognition for identifying suspects in order to facilitate the detection of crimes and eliminate false convictions [13]. A. Rummens et al. evaluated the effect of changing the spatio-temporal parameters of the predictive police model on the effectiveness of crime prediction based on data on apartment burglaries in a large city in Belgium [14]. M. Simmler et al. presented and discussed recommendations for assessing the usefulness and legitimacy of technical innovations in the criminal justice system [15]. P. Ugwudike conducted a critical analysis of the relationship between race and the use of risk prediction technologies used in justice systems in Western jurisdictions such as the UK and the USA. [16]. F. Miro-Llinares established that the complexity of an algorithmic tool can cause misunderstanding of the decision-making process by users [17]. R. Yu et al. used longitudinal convictions data from district courts and assessed recidivism rates among individuals released from Swedish prisons in three security levels [18].

The use of a DSSJ has both advantages and disadvantages, which have not yet been sufficiently explored in scientific and legal circles. In addition, each country has specific features of development, validation, and use of DSS for the implementation of smart policing and smart criminal justice. In Ukraine, the implementation of the concept of implementing complex DSS for criminal justice is only at the stage of development and testing. Therefore, complex, multifaceted research in this field at the national level is expedient.

Research Methodology

This work is a continuation of a series of studies on the problems of applying quantitative methods and tools to support decision-making in criminal justice. Section 3.1 presents the main results obtained in previous research. Section 3.2 presents the descriptions of the dataset, attributes and attribute values used in the analysis. Section 3.3 presents a new analytical approach to criminal profiling – applying a cluster model to determine significant indicators (individual characteristics of prisoners) of criminal recidivism.

The previous results

In the previous articles, we obtained the following results:

1. The machine learning models (Generalized Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees, and Support Vector Machine) were built for predicting the propensity of convicts to commit criminal recidivism. It was found that the presence of conditional convictions and the number of convictions to the actual punishment are significant factors that affect the tendency of customers of penitentiary institutions to commit

repeated criminal offenses in the future. Decision Trees models for the classification of convicts into "prone" and "non-prone" to recidivism were built [19].

2. The scoring model was created to assess the risk of repeated criminal offenses by convicts based on their individual statistical and dynamic characteristics. An optimal model based on Machine Learning systems was built to determine important factors that influence the propensity of convicted criminals to repeat criminal offenses and prisoners with a high level of recidivism [20]

3. The logistic regression model to predict the probability of convicted criminal recidivism in the future was built based on the analysis of individual characteristics of prisoners. It has been proven that age at the time of the first conviction, type of employment at the time of conviction, number of suspended convictions, number of minor crimes, number of crimes of medium gravity, and availability of early dismissals are determinants of the propensity of prisoners to criminal recidivism [21].

4. Applied the associative rule mining for the extract correlations and co-occurrences between the historical crime information of convicted. An associative rule mining model was built to search for non-obvious interesting connections between historical crime information of convicted and repeated offenses. The frequent item sets, which are combinations of individual characteristics of prisoners who commit criminal recidivism, and the strong association rules have been revealed. It was established that early dismissals and suspended convictions are the significant factors that cause the risk of recidivism [22].

5. Using the statistical survival analysis we computed the probability of the accused confessing to the commission of a criminal offense at a specific stage of the duration of the investigation. Assessment and forecasting of the risks of admitting guilt in committing criminal offenses under conditions of incomplete data was carried out. The Kaplan-Meier model was built for calculating the chances of obtaining evidence of a confession after the end of the trial in criminal proceedings. Created the Cox regression model to establish the relationship between the stages of the pre-trial investigation, at which the accused gives a confession, with the duration of the investigation and the method of prosecution (a crime committed by one person or a crime committed by a group of persons) [23].

The formation of reliable information and analytical support for complex DSSJ requires multi-faceted research and the construction of various effective models, the results of which confirm previously obtained assessments.

Data selection and description

It is a comprehensive case study of a unique real-world of 13,010 criminal convicted dataset. Our study uses data on individual statistical and dynamic characteristics of prisoners serving their sentences in penitentiary institutions of Ukraine (Table 1).

Table 1

The crime records data

Attribute	Value	Meaning
Recidivism	1	yes
	0	no
Sex	1	male
	2	female
Age	1	to 18 years
	2	from 18 to 30 years
	3	from 30 to 45 years
	4	older than 45 years
Age1 (age at the time of the first conviction (to the actual degree of punishment))	1	to 18 years
	2	from 18 to 30 years
	3	from 30 to 45 years
	4	older than 45 years
Age2 (age at the time of the first conviction (conditional or actual sentence))	1	to 18 years
	2	from 18 to 30 years
	3	from 30 to 45 years
	4	older than 45 years
Marital status	1	single
	2	married
Education	0	incomplete secondary
	1	secondary
	2	special secondary
	3	incomplete higher
	4	higher
Place of residence (place of residence to the actual degree of punishment)	0	rural area
	1	urban area
Type of employment (type of employment at the time of conviction (up to actual punishment))	0	unemployed
	1	part-time
	2	full-time
Early dismissals (availability of early dismissals)	0	no
	1	yes
Motivation for dismissal	0	no
	1	yes
Suspended convictions		number of suspended convictions

Criminal profiling strategy

One of the decision-support strategies in criminal proceedings is criminal profiling [24]. A criminal profile is a set of conclusions about the qualities of a person responsible for committing a crime or a series of crimes. To date, there is no single effective method of forming accurate, substantiated conclusions regarding criminal profiling. In this work, we used one of the methods of intelligent data analysis to create typical criminal profiles of convicts. The k-means method [25] was used for clustering convicts (distribution into relatively homogeneous groups) according to their individual characteristics. The empirical analysis was carried out in the RapidMiner predictive analytics environment [26]. To create a cluster model, a process consisting of operators is presented in Fig. 1–3 and Tables 2–4.

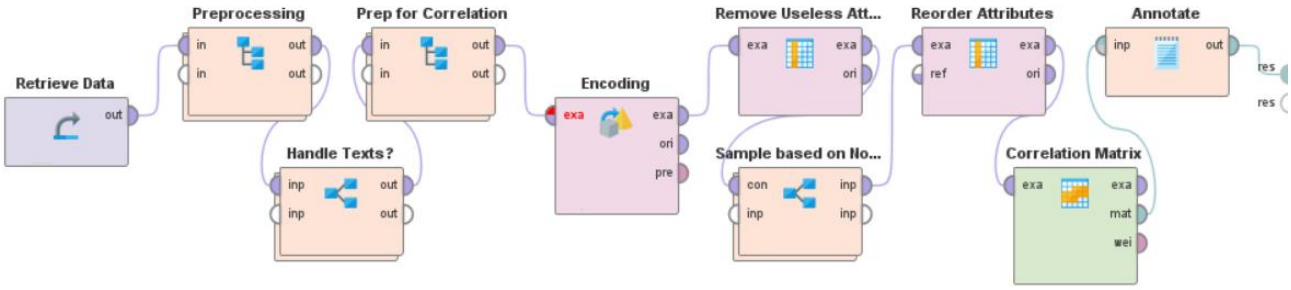

Fig. 1. Operators of the process of creating a cluster model for dividing convicts into groups based on similar individual characteristics

Table 2

***K*-means clustering process operates**

Operator	Description
Retrieve data	Loads a RapidMiner object (dataset) into the process
Preprocessing	Introduces a process subprocess (chain of operators that will be applied later) within a process
Handle text?	The root operator “Should handle text columns?” is the outermost operator of every process
Prep for correlation	Prepares dataset for correlation calculation
Encoding	Performs one-hot encoding on the data and removes columns with too many nominal values
Remove Useless Attributes	Removes useless columns like constants
Sample-based on No Attributes	Samples data down based on the number of attributes
Recorder Attributes	Orders columns alphabetically
Correlation Matrix	Creates the actual correlation matrix
Annotate	Defines a result name

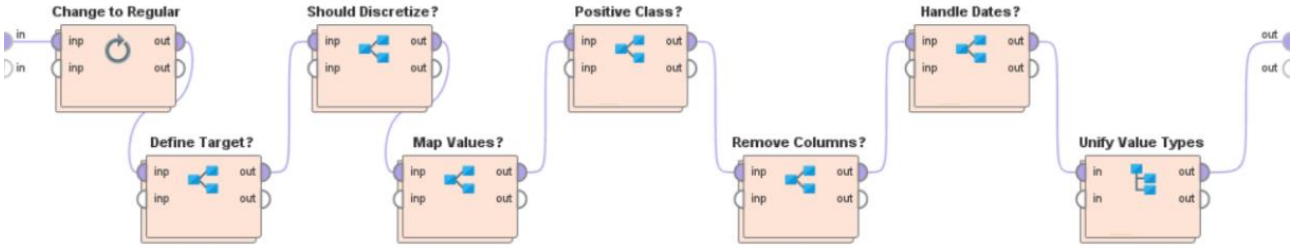

Fig. 2. Preprocessing operators

Table 3

Nested statements of the preprocessing

Operator	Description
Change to Regular	Changes the role to 'regular' for all columns.
Define Target	Selects Subprocess
Should Discretize?	Introduces a process within a process
Map Values	Switches options
Positive Class?	Consists of the nested operators: Nominal to Binominal, which changes the type of selected nominal attributes to a binominal type, and Define Positive Class, which modifies the internal value mapping of binominal attributes according to the specified negative and positive values
Remove Columns	Consists of the nested operator Remove Columns, which removes columns
Handle Dates	Consists of the nested operator Remove Dates
Unify Value Types	Consists of the nested operators: Remove Unused Values, which removes all unused values and orders the value mappings alphabetically; Nominal to Text, which transforms all nominal columns to text; Text to Nominal, which transforms all text columns into polynomial columns; Numerical to Real, which turns all numerical columns; Set Text Column, which defines the value type of all texts column

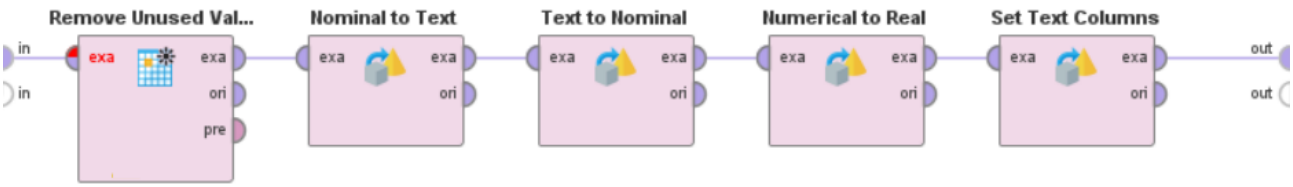

Fig. 3. Preprocessing: Unify Value types

Table 4

Nested statements of the Unify Value Types Operator

Operator	Description
Remove Unused Value	Removes each nominal value that is not assigned to an example
Nominal to Text	Changes the type of selected nominal attributes to text
Text to Nominal	Changes the type of selected text attributes to nominal
Numerical to Real	Changes the type of the selected numerical attributes to the real type
Set Text Columns	Sets text columns

Experiments

Two clusters of convicts were identified, which included 6,266 and 6,744 people, respectively (Fig. 4, 5). The Real convictions and Early dismissals attributes had the greatest influence on the formation of cluster 0, and the Recidivism attribute (the presence of criminal recidivism) on the formation of cluster 1.

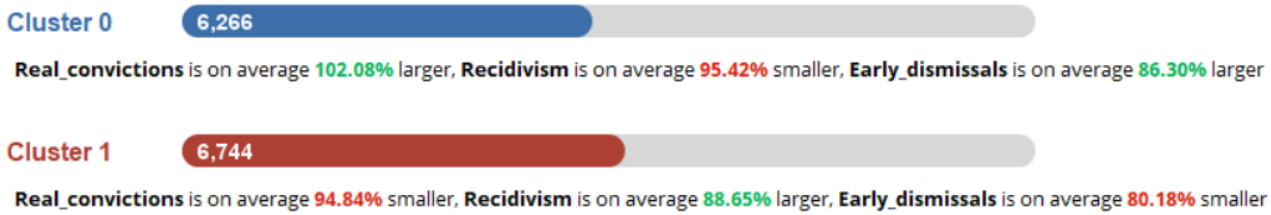

Fig. 4. K-means summary

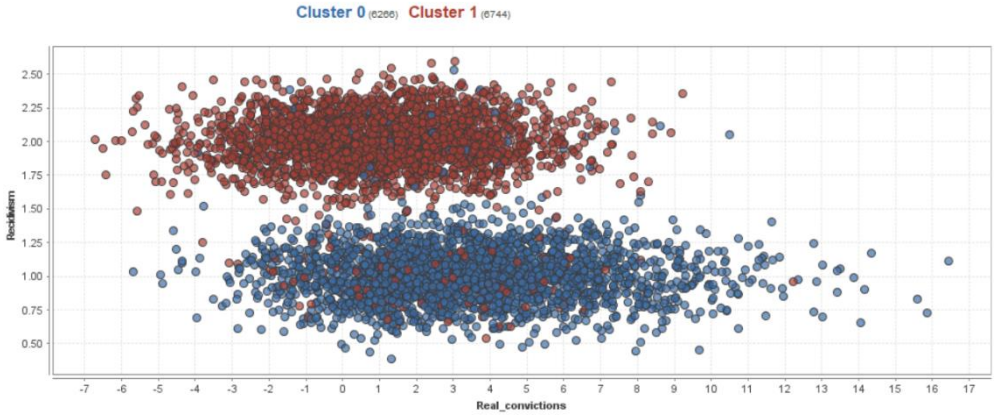

Fig. 5. Scatter Plot

The average values of the analyzed attributes (centroids) were calculated for each of the selected clusters of convicts (Table 5). Cluster 0 includes criminals who committed twice as many (on average) repeated criminal offenses as those convicted from cluster 1. Criminals from cluster 0 committed a first crime and were convicted for the first time (on average) at an earlier age than prisoners, which make up cluster 1. Convicts who formed cluster 0 served (on average) three times more actual sentences compared to individuals from cluster 1. Almost twice as many (on average) suspended sentences were given to prisoners included in cluster 0 than to convicts from cluster 1.

Table 5

Centroid table (fragment)

Cluster	Age	Age1	Age2	Early dismissals	Motivation for dismissal	Real convictions	Recidivism	Suspended convictions
Cluster 0	3.14	1.86	1.72	0.55	0.85	3.48	1.02	0.08
Cluster 1	3.92	2.52	2.42	0.06	0.90	1.06	1.06	0.50

The graph of cluster averages (Fig. 6) gives reason to conclude that the biggest differences between the selected groups of convicts are observed among the average values of the following attributes: Real convictions (number of convictions before the actual punishment), Recidivism (presence of relapses), Age1, Age2 (age at the time of the first convictions to the actual and conditional sentence), Suspended convictions (the presence of conditional convictions) and Early dismissals (the presence of early dismissals). The obtained results give an idea of the regularities revealed between the analyzed individual characteristics of criminals.

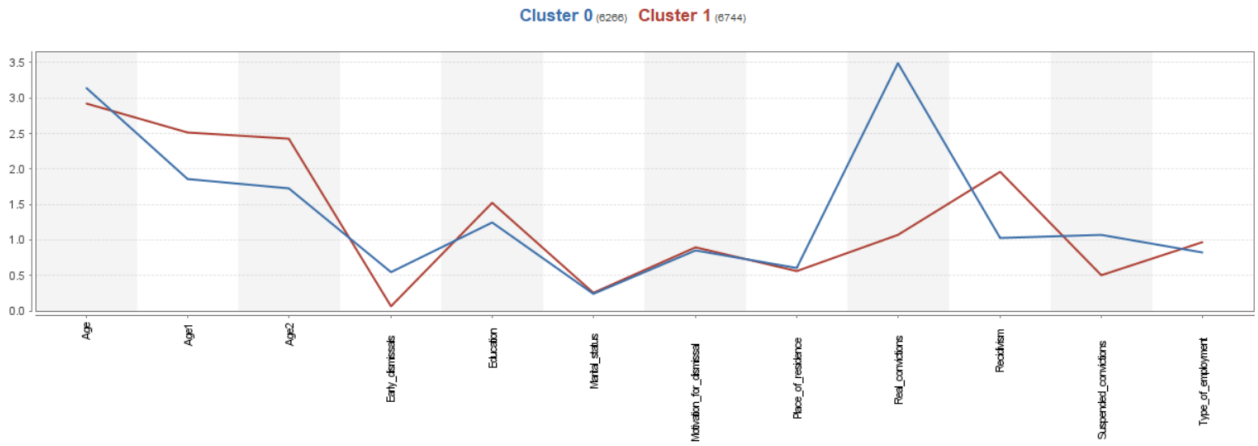

Fig. 6. K-means centroid chart

A correlation matrix was constructed to assess the relationships between the investigated features (Table 6). The Recidivism attribute has the strongest correlations with Early dismissals (0.41), Age2 (0.39) and Age1 (0.38). Therefore, it can be argued that early release and age at the time of the first conviction (both before actual and conditional punishment) are risk factors for convicts committing new criminal offenses.

Table 6

Correlation table (fragment)

Attributes	Age	Age1	Age2	Early dismissals	Motivation for dismissal	Real convictions	Recidivism	Suspended convictions
Age	1	0.37	0.35	0.14	-0.01	3.48	0.25	0.08
Age1	0.37	1	0.88	-0.22	0.06	1.06	1.06	0.50
Age2	0.35	0.88	1	-0.23	0.05	-0.35	0.40	-0.22
Early dismissals	0.14	-0.22	-0,23	1	0.02	0.41	-0.50	0.19
Motivation for dismissal	-0.00	0.06	0.01	0.02	1	-0.08	0.08	0.00
Real convictions	0.25	-0.36	0.41	0.41	-0.08	1	-0.68	0.12
Recidivism	-0.2	0.38	-0.40	-0.50	0.08	-0.68	1	-0.19
Suspended convictions	-0.00	-0.13	-0.22	0.19	0.00	0.12	-0.19	1

To simplify the understanding of the algorithm for the distribution of convicts into selected clusters based on the analyzed attributes, a decision tree for the cluster model was built (Fig. 7).

The results of the built machine learning and artificial intelligence cluster model confirmed the estimates obtained in previous works of this series. The penitentiary system does not perform a correctional function yet. Rather, the opposite is true: the earlier a person enters correctional institutions for the first time, the more likely he is to reoffend. However, the chance for correction provided by the judicial system in the form of conditional convictions and early releases also provokes relapses in most cases.

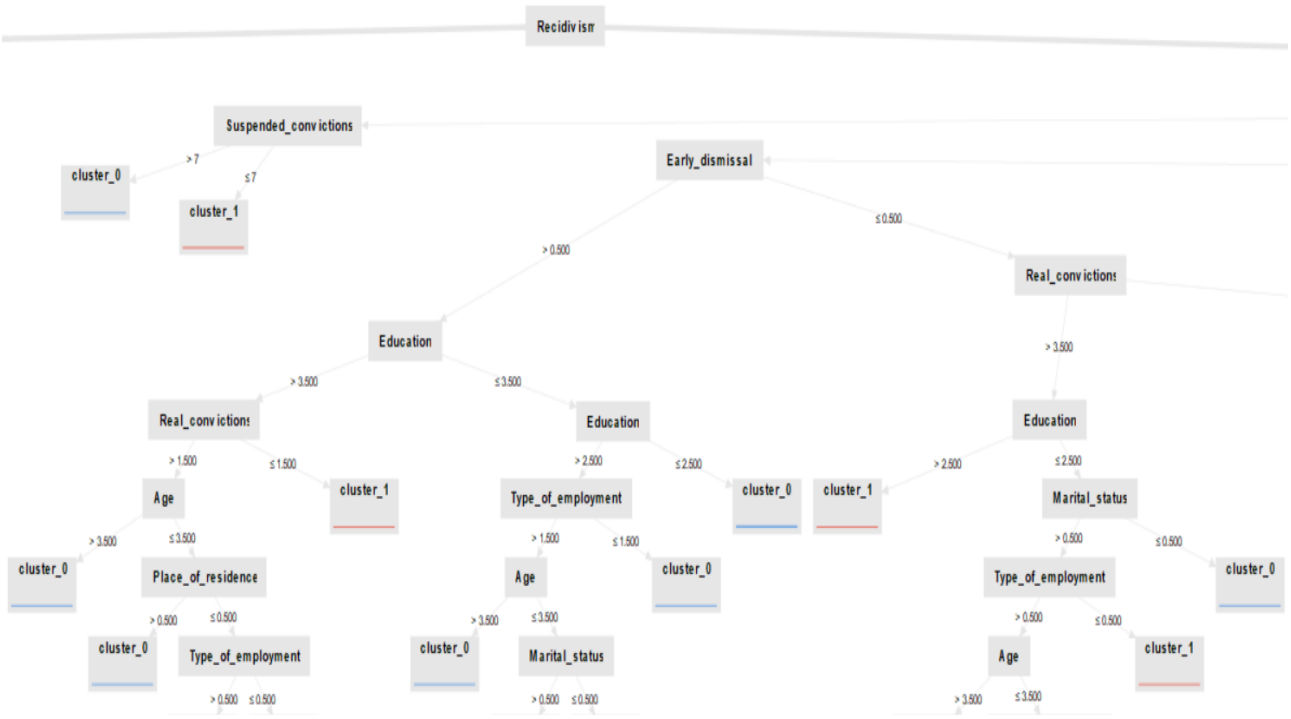

Fig. 7. Cluster Tree (fragment)

Conclusions

Criminal offenses pose a significant threat to Ukraine's internal security. The study of individual characteristics of criminals, which are risk factors for criminal recidivism, requires special attention. The public danger is not the fact of committing a repeated offense, but the personal qualities of the criminal. Identifying links between the individual characteristics of prisoners and their criminal recidivism can help to solve serial crimes, develop new crime prevention strategies, and provide reliable support for public safety decisions. In the context of russia's war against Ukraine, the problem of crime poses a serious challenge to Ukraine's external security, since a significant proportion of those mobilized into the russian army are prisoners. They commit repeated crimes already on the territory of Ukraine. Criminal profiling does not provide unequivocal indisputable evidence to solve a criminal case, but it is an effective tool in the investigation of serial crimes, hostage taking, rape, and sexual murders and in establishing the authorship of texts, such as threatening letters.

It is a case study of a unique real-world dataset of 13,010 criminal convicts. We applied the Rapid Miner tool to the machine learning k-clustering algorithm and built a cluster model. The relationship between the number of previous convictions of prisoners, the age at the time of the first conviction, the presence of conditional convictions and early releases with the risk of criminal recidivism in the future has been proven. The built computer model makes obvious the relationship between the fact of criminal recidivism and the elements of the profile of the criminal (individual characteristics of the person), providing reliable support when making decisions in criminal proceedings. The obtained results confirmed the estimates obtained at the previous stages of a series of studies on the application of quantitative tools and the construction of computer models for the development of informational and analytical support of the decision-making in criminal justice, to simplify the understanding of criminal behavior and to provide effective support for judicial decision-making. The models of computation developed in this series can be applied to new criminal convicted datasets and become the basis for the formation of information and analytical support for complex DSS and provide reliable information support when making effective decisions in criminal proceedings and developing effective strategies for crime prevention and ensuring internal security.

The next stage of our research will be the corresponding analysis to study the relationship between individual characteristics of convicts and the fact of criminal recidivism.

References

1. Blokdyk G. Decision Support System. A Complete Guide. Publisher: 5STARCooks. 2020. URL: https://www.scribd.com/book/426601046/Decision-Support-System-A-Complete-Guide-2020-Edition (Last accessed: 04.02.2023).

2. Reforming Benefits Decision-Making. A Report by Justice and the Administrative Justice Council. Carter Lane. London. 2021. URL: https://justice.org.uk/wp-content/uploads/flipbook/50/book.html (Last accessed: 22.01.2023).

3. Simmler M., Brunner S., Canova G., et al. Smart Criminal Justice: exploring the use of Algorithms in the Swiss Criminal Justice System. *Artif Intell Law*. 2022. Vol. 31. Pp. 213–237. https://doi:10.1007/s10506-022-09310-1.
4. Doyle O. The Role of Decision Support Systems in the Criminal Justice System. Department of Political Science. 2020.
5. Zavrsnik, Algorithmic justice: algorithms and big data in criminal justice settings. *Eur J Criminol*. 2019. Vol. 18(35). https://doi:10.1177/1477370819876762.
6. Benbouzid, To Predict and to Manage: Predictive Policing in the United States. Big Data Soc. 2019. Vol. 6(1). Pp. 1–13. https://doi:10.1177/2053951719861703.
7. Berk R. A. Artificial Intelligence. Predictive Policing, and Risk Assessment for Law Enforcement. *Ann Rev Criminol*. 2021. Vol. 4(1). Pp. 209–237. https://doi:10.1146/annurev-criminol-051520-012342.
8. Brayne S. Predict and Surveil: Data, Discretion, and the Future of Policing. Oxford University Press. 2021. https://doi:10.1093/oso/9780190684099.001.0001.
9. Cavelty M. D., and Hagmann J. The Politics of Security and Technology in Switzerland. *Swiss Political Sci Rev*. 2021. Vol. 27(1). Pp. 128–138. https://doi:10.1111/spsr.12430.
10. Egbert S., and Krasmann S. 2020. Predictive Policing: Not Yet, but Soon Preemptive? *Policing and Society*. 2020. Vol. 30(8), Pp. 905–919. https://doi:10.1080/10439463.2019.1611821.
11. Sandhu A. The "Uberization of Policing"? How Police Negotiate and Operationalize Predictive Policing Technology. *Polic Soc*. 2021. Vol. 31(1). Pp. 66–81. https://doi:10.1080/10439463.2020.1803315.
12. Cheng W. Li., Tong H. Prediction of Criminal Suspects Based on Association Rules and Tag Clustering. *Journal of Software Engineering and Applications*. 2019. Vol. 12(3). Pp. 35–50. https://doi:10.4236/jsea.2019.123003.
13. Kotsoglou K. N., Oswald M. The Long Arm of the Algorithm? Automated Facial Recognition as Evidence and Trigger for Police Intervention. *For Sci Int Synergy*. 2020. Vol. 2. Pp. 86–89. https://doi:10.1016/J.FSISYN.2020.01.002.
14. Rummens A, Hardyns W. The Effect of Spatiotemporal Resolution on Predictive Policing Model Performance. *Int J Forecast*. 2021. Vol. 37(1). Pp. 125–133. doi: 10.1016/j.ijforecast.2020.03.006.
15. M. Simmler, G. Canova, and K. Schedler, Smart Criminal Justice: Phenomena and Normative Requirements. Int Rev Admin Sci, 2021. https://doi:10.1177/00208523211039740.
16. Ugwudike P., Digital Prediction Technologies in the Justice System: the Implications of a 'Race-Neutral' Agenda. *Theor Criminol*. 2020. Vol. 24(3), Pp. 482–501.
17. F. Miro-Llinares F., Predictive Policing: Utopia or Dystopia? On Attitudes Towards the Use of Big Data Algorithms for Law Enforcement. *Idp-Internet Law Polit*. 2020. Vol. 30. Pp. 1–18. https://doi:10.7238/idp.v0i30.3223.
18. Yu R., Långstro N., Forsman M., Sjo A., Fazel S., Molero Y. Associations between prisons and recidivism: A nationwide longitudinal study. *PLoS ONE*. 2022. Vol. 17(5):e0267941. https://doi:10.1371/journal.pone.0267941.
19. Kovalchuk O., Karpinski M., Banakh S., Kasianchuk M., Shevchuk R., and Zagorodna N., Prediction Machine Learning Models on Propensity Convicts to Criminal Recidivism. *Information*. 2023. Vol. 14(3). P. 161. https://doi:10.3390/info14030161.
20. Kovalchuk O., Banakh S., Masonkova M., Burdin V., Zaverukha O., and Ivanytskyy R. A Scoring Model for Support Decision Making in Criminal Justice. *In Proceedings of the 12th International Conference on Advanced Computer Information Technologies (ACIT-2022)*. IEEE**,** Ruzomberok, Slovakia. 2022. Pp. 116–120. https://doi:10.1109/ACIT54803.2022.9913182.
21. Berezka K. M., Kovalchuk O. Ya., Banakh S. V., Zlyvko S. V., and Hrechaniuk R. A Binary Logistic Regression Model for Support Decision Making in Criminal Justice. *Folia Oeconomica Stetinensia*, 2022. Vol. 22 (1). Pp. 1–17. https://doi:10.2478/foli-2022-0001.
22. Kovalchuk O., Banakh S., Kasianchuk M., Moskaliuk *N., and Kaniuka V. Associative Rule Mining for the Assessment of the Risk of Recidivism. In Proceedings of the 4th International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS 2023).* CEUR Workshop Proceedings, Khmelnytsky, Ukraine. 2023. Vol. 3373. Pp. 376–387. URL: https://ceur-ws.org/Vol-3373/.
23. Kovalchuk O. Modeling the risks of the confession process of the accused of criminal offenses based on survival concept. *Scientific Journal of TNTU*. 2022. Vol. 108(4). Pp. 27–37.
24. Petherick W., and Brooks N. Reframing Criminal Profiling: a Guide for Integrated Practice. *Psychiatr Psychol Law*. 2020. Vol. 28(5). Pp. 694–710. https://doi:10.1080/13218719.2020.1837030.
25. Cluster Analysis. Science Direct. URL: https://www.sciencedirect.com/topics/medicine-and-dentistry/cluster-analysis (Last accessed: 13.06.2023).
26. RapidMiner Documentation. URL: https://docs.rapidminer.com (Last accessed: 03.06.2023).

Olha Kovalchuk **Ольга Ковальчук**	PhD, Associate Professor of Applied Mathematics Department, West Ukrainian National University, Ternopil, Ukraine, e-mail: olhakov@gmail.com https://orcid.org/0000-0001-6490-9633	кандидат фізико-математичних наук, доцент кафедри прикладної математики, Західноукраїнський національний університет, Тернопіль, Україна

https://doi.org/10.31891/csit-2023-3-7
UDC 004.9

Tetiana HOVORUSHCHENKO, Iryna ZABELINA, Kostyantyn REI
Khmelnytskyi National University
Olha HOVORUSHCHENKO
National Pirogov Memorial Medical University

METHOD OF CREATING AN INFORMATION SYSTEM FOR MONITORING INFECTIOUS PATIENTS

In the context of the COVID-19 pandemic, infectious disease information systems are widely used and promoted to prevent the spread of the pandemic (mainly in the form of mobile applications). Many countries have offered their apps to improve contact tracing and thus reduce the number of infections. However, the level of adoption of such applications has been and remains relatively low, which, obviously, given their massive use and effectiveness, has been largely influenced by issues related to privacy and anonymity, as well as the perception of potential users of the price-benefit ratio. Thus, the task of creating information systems for monitoring infectious patients is still relevant today. Therefore, our study is devoted to the development of a method and an information system for monitoring infectious patients.

The article develops a method for creating an information system for monitoring infectious patients, which, unlike the known ones, is based on intelligent analysis of data on the geolocation of patients and contact persons, and provides the ability to design an information system for controlling infectious patients. The purpose of the information system for monitoring infectious patients is to prevent the spread of epidemics and pandemics by tracking patient contacts and reducing the number of infections. The tasks of the information system for monitoring infectious patients are to track the self-isolation of infectious patients and their contacts, identify the most "infected" buildings, districts, cities, etc. based on intelligent analysis of data on infectious patients and their contacts.

The designed mobile-oriented information system for monitoring infectious patients can be used to prevent the spread of the pandemic by tracking contacts and reducing the number of infections. The design of screen forms, reports, implementation, testing and commissioning of the information system for monitoring infectious patients will be carried out by the authors in the course of their further research.

Keywords: information system, control of infectious patients, method of creating an information system, data mining.

Тетяна ГОВОРУЩЕНКО, Ірина ЗАБЄЛІНА, Костянтин РЕЙ
Хмельницький національний університет
Ольга ГОВОРУЩЕНКО
Вінницький національний медичний університет ім. М. І. Пирогова

МЕТОД СТВОРЕННЯ ІНФОРМАЦІЙНОЇ СИСТЕМИ КОНТРОЛЮ ЗА ІНФЕКЦІЙНИМИ ХВОРИМИ

В контексті пандемії COVID-19 інформаційні системи контролю за інфекційними хворими широко використовуються і просуваються для запобігання поширенню пандемії (в основному, у вигляді мобільних додатків). Багато країн запропонували свої додатки з метою покращити відстеження контактів і таким чином зменшити кількість випадків інфікування. Однак рівень впровадження таких додатків був і залишається відносно низьким, на що, очевидно, враховуючи їх масове використання та ефективність, значною мірою вплинули питання, пов'язані з конфіденційністю та анонімністю, а також сприйняття потенційними користувачами співвідношення ціни та вигоди. Отже, задача створення інформаційних систем контролю за інфекційними хворими є актуальною і в наш час. Відтак наше дослідження присвячене розробленню методу та інформаційної системи контролю за інфекційними хворими.

У статті розроблено метод створення інформаційної системи контролю за інфекційними хворими, який, на відміну від відомих базується на інтелектуальному аналізі даних про геолокації хворих та контактних осіб, та забезпечує можливість проєктування інформаційної системи контролю за інфекційними хворими. Метою інформаційної системи контролю за інфекційними хворими є запобігання поширенню епідемії та пандемії шляхом відстеження контактів хворих та зменшення кількості випадків інфікування. Задачі інформаційної системи контролю за інфекційними хворими: відстеження перебування на самоізоляції інфекційних хворих та їхніх контактних осіб, визначення найбільш «заражених» будинків, районів, міст, тощо на основі інтелектуального аналізу даних про інфекційних хворих та їхніх контактних осіб.

Спроєктована мобільно-орієнтована інформаційна система контролю за інфекційними хворими може використовуватись для запобігання поширенню пандемії шляхом відстеження контактів і зменшення кількості випадків інфікування. Проектування екранних форм, звітів, реалізація, тестування та введення в дію інформаційної системи контролю за інфекційними хворими будуть виконуватись авторами під час їх подальших досліджень.

Ключові слова: інформаційна система, контроль за інфекційними хворими, метод створення інформаційної системи, інтелектуальний аналіз даних.

Introduction

An epidemic is a massive spread of an infectious disease among the population of a given area over a short period of time. Despite all scientific advances, the spread of infectious diseases continues to pose a significant threat to the health and prosperity of our society.

Humanity has long suffered from epidemics (smallpox, plague, cholera, malaria, typhoid, diphtheria, etc.) and for a long time could not withstand the high mortality rate, which sometimes reached 90% of the population of

the epidemic-affected region. Since ancient times, in addition to quarantine, the fight against epidemics has included the creation of contagious barracks where patients were isolated and treated, the involvement of additional medical personnel, disinfection of territories and housing, the deployment of sanitary posts, and the provision of disinfectants to the population. However, even then, the problem of effective organization of management and control over the actions of quarantine and medical services on the scale of both individual territories and the entire state was acute [1].

Currently, protecting the population from infectious diseases is one of the priority areas of activity of executive authorities and local governments. In the area of protection of the population from infectious diseases, the Cabinet of Ministers of Ukraine develops and implements relevant state targeted programs, provides funding and logistical support to healthcare institutions, institutions and facilities of the State Sanitary and Epidemiological Service, enterprises, institutions and organizations involved in activities and works related to the elimination of epidemics, coordinates these activities and works, and resolves other issues within the powers defined by law. At the same time, in this area, local governments ensure the implementation of preventive and anti-epidemic measures in the territories of settlements, in places of mass recreation and recreational areas, as well as work to eliminate epidemics and outbreaks of infectious diseases, and resolve issues of financial and logistical support for these measures and work. The specially authorized central executive body on health care in the field of protection of the population from infectious diseases develops, approves and enacts sanitary and anti-epidemic rules and regulations, methods of examination and treatment of patients, diagnosis and prevention of infectious diseases, and other regulatory acts [2].

The fundamentals of Ukrainian healthcare legislation regulate the prevention of infectious diseases dangerous to the public. Thus, persons who carry infectious diseases dangerous to the public are excluded from work and other activities that may contribute to the spread of infectious diseases and are subject to medical supervision. In respect of certain particularly dangerous infectious diseases, mandatory medical examinations, preventive vaccinations, treatment and quarantine measures may be carried out in accordance with the procedure established by the laws of Ukraine. At the same time, persons suffering from infectious diseases or being bacterial carriers are obliged to take measures recommended by medical professionals to prevent the spread of infectious diseases, comply with the requirements and recommendations of medical professionals regarding the procedure and conditions of treatment, comply with the operating hours of healthcare facilities and scientific institutions where they are treated, and undergo the necessary medical examinations and examinations within the established time limits [2].

Immunization of the population plays a significant role in reducing the level of infectious diseases, but it is possible and effective only for the prevention of known diseases. Thus, in 2019-2021, humanity faced the new COVID-19 virus, which it was completely defenseless against due to its novelty and the lack of a vaccine against this virus. Traditional epidemiological surveillance of infectious diseases failed to warn health authorities in time to intervene and mitigate and control COVID-19 before the epidemic turned into a pandemic. To stop the spread of COVID-19, humanity was forced to resort to strict quarantine measures (self-isolation, strict control of patients and their contacts, etc.)

Information systems and technologies help to monitor compliance with quarantine measures. For example, in Ukraine, compliance with quarantine measures was monitored using the Vdoma electronic service or the installed and activated mobile application of the Vdoma electronic service of the Unified State Web Portal of Electronic Services [3]. This service/application was used to counteract the spread of COVID-19 and monitor self-isolation and was mandatory for Ukrainians, foreigners and stateless persons crossing the state border to enter Ukraine. Upon arrival at the place of self-isolation, the user of the system/application was obliged to confirm his/her arrival and send a photo, and the user's geolocation was recorded when sending the photo. During self-isolation, the user was allocated 2 hours per day for personal needs. During this time, it was possible to leave the place of self-isolation to visit places of trade in food, hygiene products, medicines, and medical devices located at a distance of no more than 2 kilometers from the place of self-isolation [4].

In the paper [5], the heterogeneous network-based epidemic model with positive and negative information is proposed. This model considered three time-varying control schemes for containing the contact infection rate and enhancing the implementation rate of positive information for reducing the number of infected and the cost of control.

The surveillance and early warning system is the basis of public health emergency prevention and control. Authors of [6] attempted to develop a new pattern of integrated surveillance and early warning system for the emerging infectious disease.

The paper [7] designed the “detection-service-mobile” three-terminal geographic information system for realizing the control of diagnostic instruments and the comprehensive management of data. Machine learning is used to marked the detection results on the mobile terminal map to realize the visual display of the positive results of nucleic acid amplification detection and the early warning of infectious diseases.

The paper [8] proposed the intelligent COVID-19 early warning system using Twitter data with novel machine learning methods – the natural language processing (NLP) pre-training technique, fine-tuning BERT as a Twitter classification method. In addition, authors of [8] implemented a COVID-19 forecasting model through a Twitter-based linear regression model to detect early signs of the COVID-19 outbreak. Also, in [8] the expert system, an early warning web application based on the proposed methods were developed.

Authors of [9] proposed the principles for the development of an intelligent information system for decision-making support for epidemiological diagnostics, which is based on the mathematical tools for analyzing morbidity

data, as well as modeling of epidemic processes.

The paper [10] presented the inferring disease data management system with blockchain and machine learning. This is a solution for organizing, sharing and analyzing the disease data with trusted, privacy-preserving and interoperable methodologies to improve the outreach, time and cost-effectiveness for disease-control and treatment interventions.

So, in the context of the COVID-19 pandemic, infectious disease information systems are widely used and promoted to prevent the spread of the pandemic (mainly in the form of mobile applications). Many countries have offered their apps to improve contact tracing and thus reduce the number of infections. However, the level of adoption of such applications has been and remains relatively low, which, obviously, given their massive use and effectiveness, has been largely influenced by issues related to privacy and anonymity, as well as the perception of potential users of the price-benefit ratio. Thus, the task of creating information systems for monitoring infectious patients is still relevant today. Therefore, our study is devoted to the development of a method and an information system for controlling infectious patients. [11].

Thus, the task of creating information systems for monitoring infectious patients is still *relevant* today. Therefore, our *study is devoted* to the development of a method and an information system for monitoring infectious patients.

Information System for Monitoring Infectious Patients

The use of effective methods and tools for creating an information system and the correct construction of its creation technology can significantly reduce costs and shorten development time, ensuring the quality of the data processing system that meets the requirements of users. When creating an information system, a whole range of methods and tools are used to develop it. An information system development method is a way of creating an information system supported by appropriate design tools. The means of creating an information system are standard design solutions, application packages, standard projects or tools for designing an information system [12, 13].

Thus, the *method of creating an information system for monitoring infectious patients* consists of the following steps:

1) defining the purpose and identifying the tasks of the information system;
2) formulation of requirements for the information system:
   - identification of the required functionality of the system and the level of its adaptability to constantly changing operating conditions;
   - determination of the required system capacity and system response time to a request;
   - determination of the required level of security;
   - identification of proposals for ease of operation and maintenance of the system;
3) design of the information system:
   - designing the architecture of the information system;
   - design of modules (blocks) of the information system;
   - determination of the network topology, hardware configuration, architecture used (file-server or client-server), parallel and distributed data processing, etc. (if necessary);
   - design of data objects to be implemented in the database;
   - designing screen forms and reports that will ensure the execution of data queries;
4) implementation of the information system;
5) testing of the information system:
   - standalone testing of modules (blocks);
   - integration and system testing of the information system as a whole;
   - acceptance testing;
6) commissioning of the information system;
7) operation and maintenance of the information system.

Let's design an information system for monitoring infectious patients using the proposed method. Let's start creating an information system by defining the goal, which can be generally defined as solving a number of interrelated tasks, including ensuring the launch of the system and its operation for a certain period of time. Thus, the *purpose* of the information system for monitoring infectious patients is to prevent the spread of epidemics and pandemics by tracking patient contacts and reducing the number of infections. The *tasks* of the information system for monitoring infectious patients are to track the self-isolation of infectious patients and their contacts, identify the most "infected" buildings, districts, cities, etc. based on intelligent analysis of data on infectious patients and their contacts.

Let's formulate generalized high-level requirements for an information system for monitoring infectious patients. Taking into account the identified tasks of the information system, we identify the necessary functionality of the system: geolocation of infectious patients, geolocation of contact persons of infectious patients, tracking of self-isolation of infectious patients, tracking of self-isolation of contact persons of infectious patients, issuance of urgent notifications about violations of self-isolation by infectious patients or their contact persons, intelligent analysis of data on infectious patients and their contacts, identification of the most "infected" buildings, districts, cities, etc. The

required system bandwidth is at least 100 Mbps; the system response time to a request is no more than 3 seconds. The required security level is high; as personal data will be processed. The system should have a simple, intuitive, user-friendly interface and be accessible to people of all ages. The system should be mobile-oriented (in the form of a mobile application).

Let's design an information system for monitoring infectious patients. The architecture of the information system for monitoring infectious patients is shown in Fig. 1.

The infectious patient is registered in the system/application (Module for registering an infectious patient in the system/application) and provides information about his/her contact persons (Module for collecting information about contact persons of an infectious patient). The provided data is recorded in the system's Database. During registration, the patient must provide access to the geolocation of his/her mobile device.

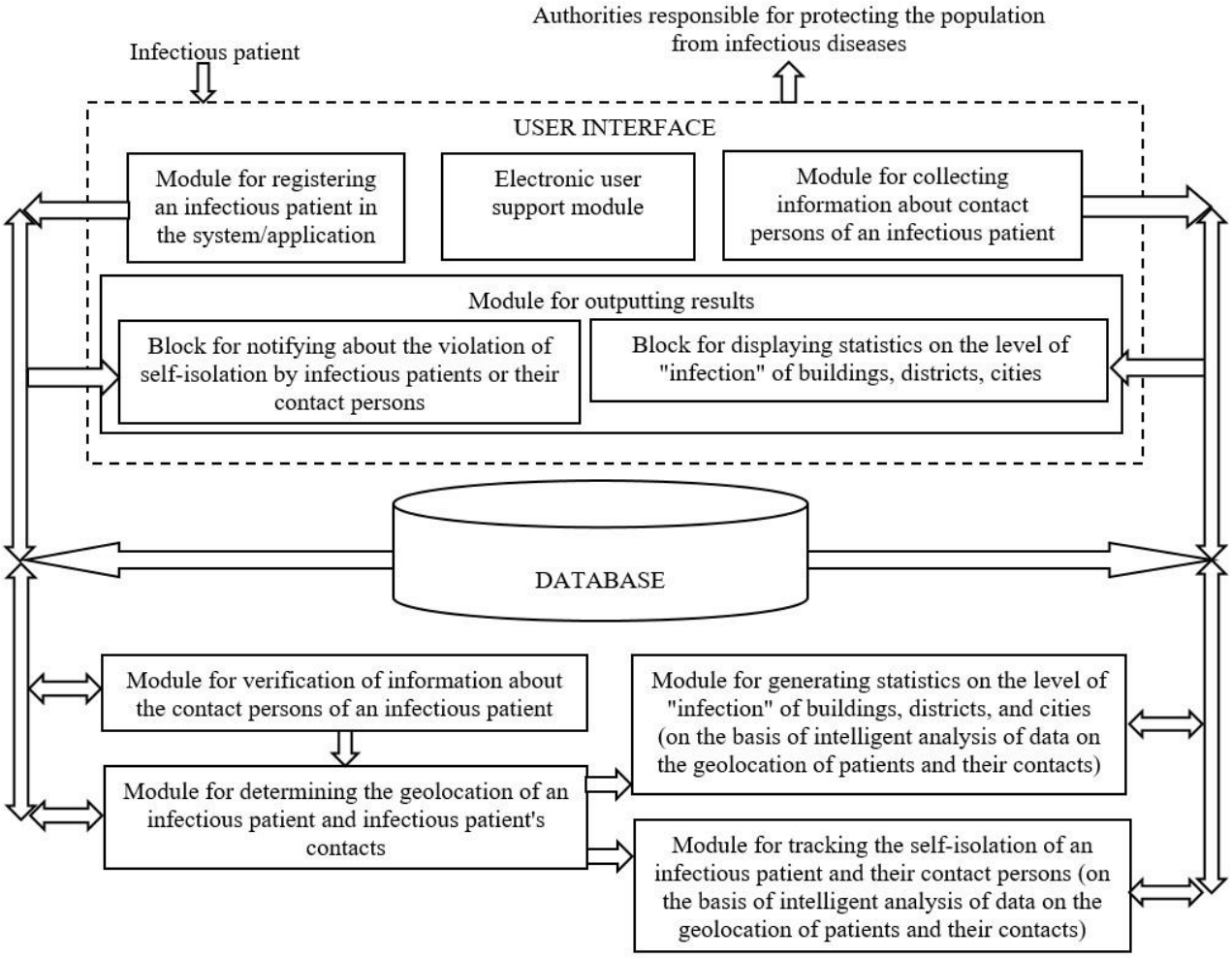

Fig. 1. Architecture of information system for monitoring infectious patients

The Module for verification of information about the contact persons of an infectious patient analyzes photos and videos from the last 3-5 days on a mobile device and in cloud environments to search for and recognize persons with whom the patient has had contact in the last 3-5 days, and also checks whether these persons are listed by the patient among the contact persons (this action is possible only if the patient has given permission to access his/her phone data - photos, videos, etc.)

The authorities responsible for protecting the public from infectious diseases contact the contact persons of the infectious patient and request them to register in the system/application. The system/application also tries to identify contact persons by contacts in the patient's mobile device (if the patient has provided such access) and send them a message that they are contact persons of the infectious patient and must register in the system/application.

The Module for determining the geolocation of an infectious patient and infectious patient's contacts records the geolocation of an infectious patient and his/her contacts and transfers them to the database, as well as to the Module for tracking the self-isolation of an infectious patient and his/her contacts and the Module for generating statistics on the level of "infection" of buildings, districts, and cities. Both of these modules work on the basis of intelligent analysis of data on the geolocation of patients and their contacts.

The Database of the information system stores registration data of infectious patients and information about their contact persons, geolocation of each infectious patient and their contact persons; in addition, the Database stores

statistics on the level of "infection" of buildings, districts, and cities, which is generated by the Data mining modules.

If the Module for tracking the self-isolation of an infectious patient and their contact persons has established that self-isolation has been violated, an appropriate urgent notification is displayed by the Block for notifying the authorities responsible for protecting the population from infectious diseases of violation of self-isolation by infectious patients or their contact persons. The Block for displaying statistics on the level of "infection" of buildings, districts, cities provides the authorities responsible for protecting the population from infectious diseases with relevant statistics.

The designed mobile-oriented information system for monitoring infectious patients can be used to prevent the spread of the pandemic by tracking contacts and reducing the number of infections.

The design of screen forms, reports, implementation, testing and commissioning of the information system for monitoring infectious patients will be carried out by the authors in the course of their further research.

Conclusions

In the context of the COVID-19 pandemic, infectious disease information systems are widely used and promoted to prevent the spread of the pandemic (mainly in the form of mobile applications). Many countries have offered their apps to improve contact tracing and thus reduce the number of infections. However, the level of adoption of such applications has been and remains relatively low, which, obviously, given their massive use and effectiveness, has been largely influenced by issues related to privacy and anonymity, as well as the perception of potential users of the price-benefit ratio.

Thus, the task of creating information systems for monitoring infectious patients is still relevant today. Therefore, our study is devoted to the development of a method and an information system for monitoring infectious patients.

The article develops a method for creating an information system for monitoring infectious patients, which, unlike the known ones, is based on intelligent analysis of data on the geolocation of patients and contact persons, and provides the ability to design an information system for monitoring infectious patients.

The purpose of the infectious disease monitoring information system is to prevent the spread of epidemics and pandemics by tracking patient contacts and reducing the number of infections. The tasks of the information system for monitoring infectious patients are to track the self-isolation of infectious patients and their contacts, identify the most "infected" buildings, districts, cities, etc. based on intelligent analysis of data on infectious patients and their contacts.

The designed mobile-oriented information system for monitoring infectious patients can be used to prevent the spread of the pandemic by tracking contacts and reducing the number of infections.

The design of screen forms, reports, implementation, testing and commissioning of the information system for monitoring infectious patients will be carried out by the authors in the course of their further research.

References

1. V. Shandra, O. Karlina. Epidemics and the Fight Against Them in Ukraine (End of the Eighteenth – Early Twentieth Century). Ukrainian Historical Journal. 2020. Vol. 5. Pp. 37-54.
2. Legal Regulation of Relations in the Field of Protection of the Population from Infectious Diseases. URL: https://minjust.gov.ua/m/str_24798.
3. Resolution of the Cabinet of Ministers of Ukraine No. 1236 of December 9, 2020 «On the establishment of quarantine and the introduction of restrictive anti-epidemic measures to prevent the spread of acute respiratory disease COVID-19 caused by the SARS-CoV-2 coronavirus in Ukraine». URL: https://ips.ligazakon.net/document/KP201236?an=1.
4. How to install the Vdoma app and stop self-isolation? Instructions from Misto.travel. URL: https://misto.travel/page/app-covid.
5. Stability analysis and optimal control of an epidemic model with multidimensional information of media coverage on networks / X. Chang et al. Mathematical Methods in the Applied Sciences. 2022.
6. Integrated surveillance and early warning system for emerging infectious diseases based on One Health concept: Structures and innovations / H.-M. Li et al. Chinese Journal of Parasitology and Parasitic Diseases. 2 Vol. 40, issue 5. Pp. 572-578.
7. A Detection-Service-Mobile Three-Terminal Software Platform for Point-of-Care Infectious Disease Detection System / X. Su et al. Biosensors. 2022. Vol. 12, no. 9. Pp. 684.
8. An intelligent early warning system of analysing Twitter data using machine learning on COVID-19 surveillance in the US / Y. Zhang et al. Expert Systems with Applications. 2022. Vol. 198. Pp. 116882.
9. Intelligent Decision-Support System for Epidemiological Diagnostics. I. A Concept of Architecture Design / K. O. Bazilevych et al. Cybernetics and Systems Analysis. 2022.
10. M. Hiwale, S. Phanasalkar, K. Kotecha. Using Blockchain and Distributed Machine Learning to Manage Decentralized but Trustworthy Disease Data. Science & Technology Libraries. 2021. P. 1–24.
11. Mobile Applications for Pandemic Monitoring : Approaches, Challenges and Opportunities / B. Alves et al. 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), Chaves, Portugal, 23–26 June 2021. 2021.
12. I. Zavushchak. Methods and tools for creating an information system. URL: http://conferences.neasmo.org.ua/uploads/article/file/563/iryna_suchasni_30-31.5.2014.zip
13. Design of information systems: General issues of theory IS design. URL: https://ela.kpi.ua/bitstream/123456789/33651/1/PIS_KL.pdf.

Tetiana Hovorushchenko **Тетяна Говорущенко**	DrSc (Engineering), Professor, Head of Computer Engineering & Information Systems Department, Khmelnytskyi National University https://orcid.org/0000-0002-7942-1857 e-mail: govorushchenko@gmail.com	Доктор технічних наук, професор, завідувач кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Iryna Zabelina **Ірина Забєліна**	Master Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University e-mail: zabelina@i.ua	Магістрантка кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Kostyantyn Rei **Костянтин Рей**	PhD Student of Computer Engineering & Information Systems Department, Khmelnytskyi National University https://orcid.org/0000-0001-9658-3148 e-mail: kostia.rei@gmail.com	Аспірант кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Olha Hovorushchenko **Ольга Говорущенко**	Student of National Pirogov Memorial Medical University https://orcid.org/0000-0001-6583-5699 e-mail: govorusenkoo@gmail.com	Студентка Вінницького національного медичного університету ім. М. І. Пирогова

https://doi.org/10.31891/csit-2023-3-8
UDC 621.317

Yuriy SKORIN, Iryna ZOLOTARYOVA
Simon Kuznets Kharkiv National University of Economics

ENHANCING THE EFFECTIVENESS OF USABILITY TESTING FOR USER INTERFACES

The paper analyses the problems of maintaining software quality, namely, usability testing as a direction of ensuring this quality, analyses publications that consider the methods of electro-oculography, electro-retinography, mouse-tracking, eye-tracking, etc. and proves that the issue of improving the quality of software products largely depends on the effectiveness of usability testing. The study is based on the analysis of traditional software testing methods and tools and the proposal of mouse-tracking and eye-tracking technologies as an alternative solution to the problem. Criteria and metrics for assessing the usability of web applications were identified, an analytical approach was used as a methodological basis, which involves a comprehensive consideration of the research object, methods of comparative analysis and classification were used to obtain data, the results were processed using standard statistical methods, and the necessity and importance of usability testing of websites was assessed. The article considers generally accepted rules and recommendations in the field of usability testing, analyses quantitative and qualitative methods of usability testing evaluation, for a more detailed study of the subject of research the ergonomic interaction of a user with an information system, namely with a web resource, studies the main categories of users depending on a number of indicators, and pays special attention to establishing quality assessment criteria based on existing standards of recommendations. the article reviews the process of usability testing of information system interfaces, its analysis and evaluation, analyses the methods of electro-oculography and electro-retinography, mouse-tracking and eye-tracking technologies, and concludes that the use of eye-tracking technology will allow collecting and systematising quantitative and qualitative data on user interaction with the system and will make it possible to optimise the process of usability testing by reducing the time of its implementation. The prerequisites for conducting an experimental study of theoretical results using eye-tracking technologies have been created.

Keywords: testing, usability, quality, website, method, electro-oculography, electroretinography, mouse-tracking, eye-tracking, system.

Юрій СКОРІН, Ірина ЗОЛОТАРЬОВА
Харківський національний економічний університет імені Семена Кузнеця

ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ USABILITY-ТЕСТУВАННЯ КОРИСТУВАЛЬНИЦЬКИХ ІНТЕРФЕЙСІВ

У статті проаналізовано проблеми підтримки якості програмного забезпечення, а саме, usability-тестування, як напрямку забезпечення цієї якості, був проведений аналіз публікацій, в якому були розглянуті методи електроокулографії, електроретинографії, mouse-tracking, eye-tracking тощо і який довів, що питання підвищення якості програмних продуктів, в значній мірі залежить від ефективності usability-тестування, проведене дослідження базується на проведеному аналізі традиційних, методів і засобів тестування програмного забезпечення і пропозиції в якості альтернативного вирішення проблеми, саме технологій mouse-tracking і eye-tracking, було визначено критерії і метрики для оцінки usability веб-додатків, в якості методологічної бази застосовувався аналітичний підхід, що передбачає комплексний розгляд об'єкта дослідження, для отримання даних використовувалися методи порівняльного аналізу, класифікації, обробка результатів проводилася за допомогою стандартних методів статистичної обробки, було проведено оцінювання необхідності і значення usability-тестування веб-сайтів, розглянуто загальноприйняті правила та рекомендації в області usability-тестування, проаналізовано кількісні та якісні методи оцінки usability-тестування, для більш детального вивчення предмета дослідження були розглянуті ергономічні взаємодії користувача з інформаційною системою, а саме з веб-ресурсом, вивчено основні категорії користувачів в залежності від цілої низки показників, особлива увага приділялася встановленню критеріїв оцінки якості виходячи з існуючих стандартів рекомендацій, було проведено огляд процесу usability-тестування інтерфейсів інформаційних систем, його аналіз та оцінку, проведено аналіз методів електроокулографії та електроретинографії, технології mouse-tracking та eye-tracking, зроблено висновок, що використання саме технології eye-tracking дозволить збирати і систематизує кількісні та якісні дані про взаємодії користувача з системою і дасть можливість оптимізувати процес usability-тестування скоротивши час його проведення, створені передумови щодо проведення експериментального дослідження теоретичних результатів за допомогою технологій eye-tracking.

Ключові слова: тестування, usability, якість, веб-сайт, метод, електроокулографія, електроретинографія, mouse-tracking, eye-tracking, система.

Introduction

Contemporary times are characterized by a notable shift within scientific and applied domains, wherein increasing emphasis is placed on the pivotal role played by the ergonomics of software products. This emphasis delineates the software's commercial allure, the extent of its market penetration, and the subjective gratification experienced by users in their interactions.

The practical import of this study is primarily intertwined with the dynamic evolution of the technological landscape, the burgeoning expanse of the Internet, and the imperative to fashion computer interfaces for a diverse user base. The burgeoning proliferation of tasks catering to a broad spectrum of users is propelled by the overarching trend of computerization and the progressive substitution of face-to-face interactions with information systems within select professional spheres.

The incorporation of standards governing software product ergonomics stands as a prerequisite to achieving a satisfactory level of accessibility within an information system. The orchestration of accessibility is seamlessly integrated into the broader spectrum of design and developmental processes, encompassing the methodical delineation of accessibility requisites, the quantitative assessment of accessibility metrics, and the formulation of a discernible criterion for verification within the purview of user engagement [5].

The endeavor to define accessibility assumes paramount significance due to the intricate interplay among users, tasks, and various other components constituting the user experience. Notably, a service, system, product, or apparatus can manifest varying degrees of accessibility across diverse use cases, an aspect that assumes heightened prominence particularly in scenarios involving distinct user segments characterized by specific disabilities [6].

To enhance the efficacy of usability testing, it is essential to perform a comparative analysis of usability testing techniques, particularly eye-tracking and mouse-tracking, and establish the necessary conditions to execute an experimental examination of theoretical outcomes utilizing these technologies.

Related works

An analysis of publications has shown that the issue of improving the quality of software products depends largely on the effectiveness of usability testing, which has been the subject of a number of studies [7-14].

The analysis and application of the electro-oculography (EOG) method is discussed in publications [15-16]. EOG records the electrical activity of the eyes at high speed using electrodes placed around the human eye. Vision control systems use this method to obtain EOG signals which, after processing, determine the direction of the user's gaze. This is particularly important for people with disabilities.

In this study [17], the measurement methods used in oculographic studies are analysed. The metrics used to quantify various psychophysiological states of the subjects are systematized by the authors. Through these metrics, it is possible to identify the elements of visual interest based on their size, brightness, colour, and location. The metrics can be divided into three groups: (1) those related to oculomotor activity, (2) technical indicators of eye-tracking performance, and (3) data tracking the actions of subjects as a result of keystrokes and mouse clicks. From the standpoint of practical tools for managing consumer attention in an information-saturated competitive environment, the paper discusses the capabilities, mechanisms of application, and possibilities for analysis of each group of metrics. The paper places particular attention on the overview of specialised paid and free software capabilities for oculographic research.

Publications [18-19] have described the use of the electroretinography (ERG) method. ERG is a method for objectively studying the functional state of the retina by recording biopotentials during light irritation. ERG enables an objective assessment of the functional state of different retinal layers and neurons. Electroretinography plays a crucial role in diagnosing retinal diseases and engages the attention of morphologists, ophthalmologists, physiologists, and electrophysiologists for scientific research beyond being a diagnostic method. Contemporary technological capabilities enable the identification of subtle disturbances in retinal bioelectrical activity, which forms the basis of initial and differential diagnosis. Electroretinography presents a characteristic diagnostic effect and opens up new aspects of the mechanisms of visual dysfunction by relying on our understanding of the neurophysiology of the visual system and molecular biology, which enables not only gene mapping and cloning but also comprehending the subtle changes in photoreception's structure, commonly causing hereditary retinal diseases. Moreover, it aids in the search for innovative initial diagnostic methods and pathogenetically-grounded treatments.

Several studies have addressed the modelling and geometry of vision [20-24]. The study has unveiled the unique features of visual perception, the understanding of which is essential in explaining many of the observed phenomena, given that 90% of information is processed by the brain through the eyes. The knowledge of the particularities of vision can aid in the analysis of the resulting image.

Publication [25] contains relevant information about the free software Mill Mouse, which uses oculography and Eye Tracker 4C to control the mouse and allow users to move the pointer to their line of sight. It can click when the user is looking at a fixed position or blinks. Additionally, Mill Mouse supports scrolling and compensates for involuntary eye movements.

The analysis and uses of the mouse-tracking method are described in reference [26]. Mouse-tracking is a method that employs software to collect data on the position of the user's mouse cursor on a computer. The objective is to automatically collect more detailed information about what users are doing, typically to enhance interface design. Reference [27] highlights that this technology enables the capture of all mouse movements of a website visitor. It is widely accepted that the mouse cursor always follows the visitor's gaze. This capability facilitates the identification of the content elements to which the user pays special attention, as well as the order in which the user studies them. The obtained data is represented on a map, highlighting the most visited areas with a bright colour. This mapping image is superimposed on a screenshot of the experimental page. This approach provides an accurate location for the most visited areas. If there is not any significant button, form or link among the hotspot areas, their design or location must be modified. According to [28], Mouse Tracking can be utilised to record every movement of the user's cursor on the site. This enables tracking of user's attention-focusing areas and their order of studying the whole page.

As per publication [29], eye-tracking is a type of technology that tracks gaze lines or points, also known as

eye-tracking technology. According to paper [30], eye-tracking has a wide range of applications such as marketing research, development of high-quality interfaces, improving usability, increasing conversion rate, determining advertising strategies, and studying user experience. Eye-tracking is considered impartial as it captures natural human reactions by studying the movements and reactions of the pupils, which cannot be imitated. This is another significant advantage of the method.

In article [31], ongoing research is described that aims to optimize human-computer interaction for emergency planning and decision support scenarios. The objective of this research is to explore the possibility of using low-cost eye tracking system with open source software in interpreting geospatial images. The project referred to in this article also includes an educational aspect.

Purpose and Task Statement

The aim of this article is to enhance software by improving the efficiency of usability testing. This study analyses conventional techniques and tools for assessing the usability of software such as 'testing in the corridor', remote usability testing, expert review, 'paper prototype testing' and automated usability testing [32]. Additionally, we propose alternative approaches to the matter. This article suggests alternative approaches such as the electro-oculography method, the electro-retinography method and technologies such as mouse-tracking and eye-tracking for further analysis. If we conclude that capturing quantitative and qualitative data on user interaction with the system via eye-tracking and mouse-tracking technologies helps systematise the data and streamline the usability testing process by reducing time, we can explore the theoretical outcomes by using these technologies. Subsequently, we can conduct an experimental study and compare the results against the other methods mentioned above.

The scope of this research aims to define a set of comprehensive research objectives that collectively lead to a deep understanding of the intricate interplay between software product ergonomics and user interfaces.

Careful exploration of ergonomic indicators in user interfaces will reveal the essential aspects of user experience in the digital realm.A comprehensive understanding of the intrinsic attributes that synergistically shape user interactions can be achieved by systematically dissecting and analyzing these indicators.Clarification of both qualitative and quantitative criteria for assessing usability. This research aims to establish a comprehensive framework that encompasses both qualitative and quantitative criteria for assessing usability testing.The dual-pronged approach aims to establish a nuanced methodology that captures both tangible metrics and intangible facets contributing to the holistic user experience. This study aims to meticulously identify and explain the key features inherent to the employed usability testing methods. The research seeks to explore the intricacies of these methodologies to understand the multifaceted layers underlying the efficacy and comprehensiveness of usability evaluations. As part of this research, software product usability will be analyzed using cutting-edge eye-tracking and mouse-tracking technologies, representing an innovative approach.

Usability testing will be executed through eye-tracking and mouse-tracking methods. In this research, a critical phase involves executing hands-on usability testing using the previously mentioned eye-tracking and mouse-tracking methods. The aim of this empirical endeavour is to bridge the gap between theoretical foundations and practical implementation. This will forge a visible link between conceptual frameworks and real-world applicability.

Comparison and analysis of usability testing results. The research will culminate in conducting a rigorous comparative analysis of usability testing results. The final phase aims to distill meaningful insights by meticulously scrutinizing and examining the data, thereby providing a conclusive panorama that highlights the efficacy and implications of the research.

The effectiveness of usability testing for user interfaces

Usability testing is a type of software testing that focuses on non-functional aspects. It is commonly classified into several categories including understandability, learnability, operability, attractiveness, and relevance. Usability testing is generally classified into three main categories: exploratory, evaluative, and comparative [32].

Modern models of software quality assessment frequently use the term "usability" to describe the ergonomics or usability of a software product.

The analysis indicated that including functions and properties that offer an additional ergonomic effect in certain usage conditions of the software product can enhance the usability indicator. In order to determine the level of ergonomics, assessing both user satisfaction and labour productivity when using the software is necessary.

The primary usability criteria are [1-4]: product effectiveness, which refers to the ratio of achieved results to resources used; productivity, which refers to the extent of resources that a user must expend to attain goals with accuracy and completeness; and satisfaction, which refers to the degree of comfort and ease of use.

To assess the quality of usability testing, it is necessary to establish testing objectives linked to the user's effectiveness, efficiency, and satisfaction, as well as the characteristics of the component's usage conditions. Figure 1 illustrates the components and dependencies between them in a schematic manner.

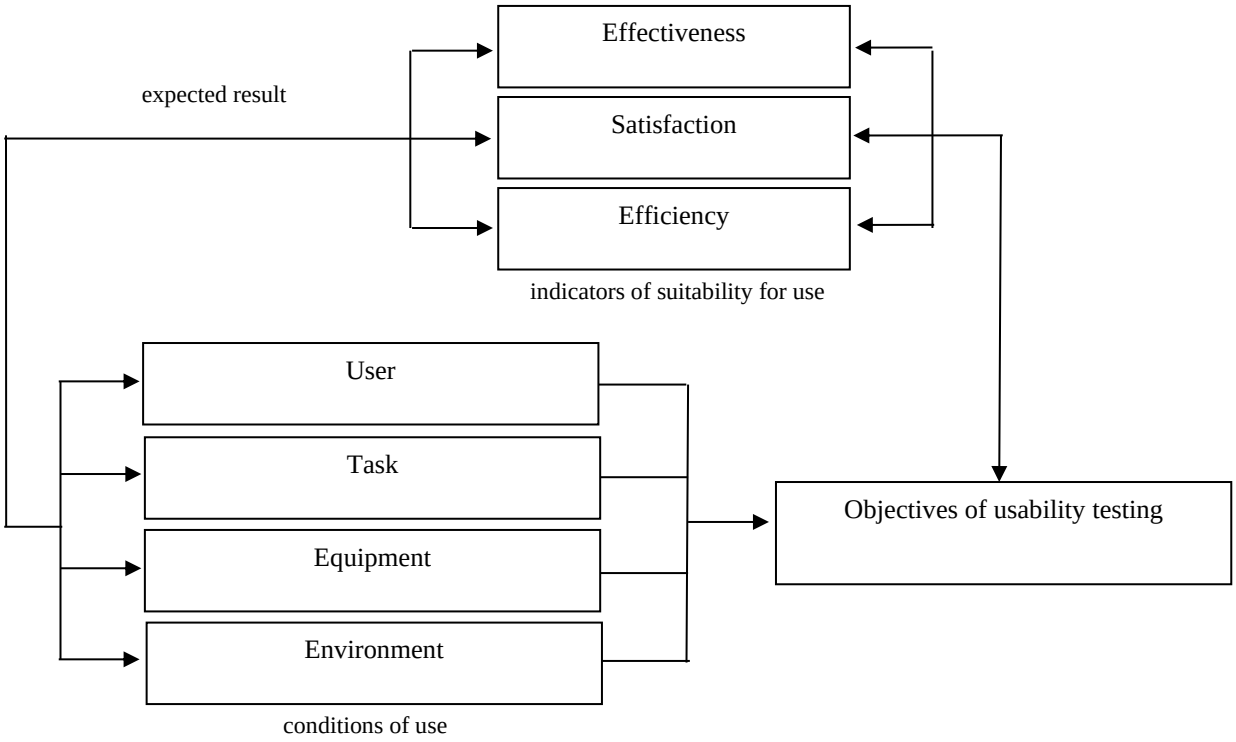

Fig. 1. Structure of suitability for usage

The design of software, services, equipment, and systems takes usability into consideration. Usability is the property that enables users to apply software products with the necessary efficiency, effectiveness, and satisfaction.

While developing a software product, it is necessary to identify usability requirements systematically. This includes usability indicators and software usage conditions. These conditions help to create project goals that are appropriate for verifying the finished software product.

When designing software products, systems, services, and equipment, it is important to consider the accessibility criterion. This criterion has an impact on the range of users who can easily utilise them. Accessibility is a characteristic of software that enables people with a wider range of abilities to achieve the set goals in specific conditions of use. Increasing accessibility can enhance the quality of usability and widen the range of users who can use a system, product, equipment, or service.

Compliance with software product ergonomics standards is necessary for achieving an acceptable level of accessibility in information systems. Planning for accessibility is an essential aspect of the design and development process, which involves systematically defining accessibility requirements, measuring accessibility, and specifying verification criteria for defined usage conditions. Defining accessibility is valuable due to the intricate interactions between users, tasks, and other elements of the use case. Different groups of users with disabilities may experience varying levels of accessibility when using a service, system, product or equipment in different use cases.

Standards in the field of usability testing for user interfaces contain guidelines for application and a description of the requirements for assessing the user interface of an information system. Such standards govern the quality of development and testing of user interfaces for software products.

The ergonomic principles for establishing a dialogue between users and an information system encompass suitability to the task, informativeness, fulfillment of user expectations, learning capacity, controllability, error tolerance, and individualization.

These principles facilitate the comprehension of software usability needs.

Human-centered design of interactive systems makes the existing research findings in ergonomics and usability accessible, enabling their application in testing the user interface. It also complements existing design methodologies and introduces the principle of a human-centered approach to designing information systems. This can be implemented in various software development and design processes.

Analysis and correct interpretation of the obtained data are important parts of carrying out usability testing. Quantitative data is suitable for identifying the most serious defects in an information system and for comparing different interfaces. Measurement results can be influenced by human factors, which leads to the need for a more detailed examination of the obtained information.

In the past, eye movements were studied using simple observations that required significant time to collect and process the information obtained. Initially, the primary field of application for eye reaction information processing

was the investigation of the psychological principles of human visual perception. Currently, the application of eye reaction information processing has been rapidly expanded, following the development of powerful software tools for processing various types of influences.

With regard to eye movement registration methods, two main groups can be distinguished: contact methods that register the impact directly around or on the cornea with the aid of sensors, such as electro-oculography, photo-optical, and electromagnetic methods; and non-contact methods, such as photoelectric, film, and video registration.

The electrooculography (EOG) method utilizes the inherent electrical properties of the eyeball. Due to the cornea's electrical positivity in relation to the retina, the eyeball is a dipole by its physical nature. As the electrical axis of the eyeball aligns with the optical axis, it provides a method of tracking gaze direction. Alterations in the potential difference between the cornea and the retina are identified through changes in the potential in the tissues bordering the eyeball. Electrodes placed horizontally around the ocular cavity detect movements of the eye. The method has the disadvantage of low resolution, but the advantage of low equipment cost. Eye movement recording does not disrupt the natural conditions of the subject's visual activity, and it can be performed in light or dark conditions, with eyes open or closed.

The photo-optical method involves directing a narrow beam of light at the eyeball, which is then reflected from a miniature mirror mounted on it. The reflected beam enters the input of the photo-registering device. This method's advantage is its high resolution. However, one disadvantage is the need to strictly fix the subject's head. Registration must be carried out exclusively in a darkened room when using the contact method.

The contact electromagnetic method has the same high resolution but is more convenient when registering the user's gaze. This method is based on the principle of changing the intensity of the electromagnetic field when the distance between the emitter and the receiver changes. The emitter is fixed to the eyeball using a contact lens, central suction cup or ring. The receiving coils are placed motionless around the subject's head. This method's advantage is its high resolution, but its contact nature is a significant disadvantage.

The photoelectric method converts a beam of infrared light reflected from the cornea into an electrical signal. This is a complicated and expensive method. It is worth noting that the last two methods are no longer in use.

The mouse-tracking method allows for the monitoring and recording of all user cursor movements on the website. This enables tracking of user attention on the site, including their preferences, and the sequence in which they engage with the page. Professionals can achieve impressive outcomes with this information [28]. The data is then compiled into a special map highlighting the most popular areas. A screenshot is captured of the page under study and superimposed with a map of mouse movements.

The mouse-tracking method offers several advantages, including a visual representation of user activity on the site, high-quality usability testing, identification of weaknesses in the areas of the site where users spend less time, low software cost, and increased conversion rates through targeted improvements based on reports. There are, however, some disadvantages to this method. These include a large margin of error due to a possible mismatch between the position of the mouse and the user's focus point. This can result from a delay between the user moving the mouse and them actually looking at the corresponding point on the screen. Furthermore, this method is not able to determine the emotions experienced by the user, as a visitor's attention may be attracted to something negative, even if the emotion is not explicitly negative. In addition, this method can only be used for pre-existing websites. Considering these limitations, the mouse-tracking method is recommended as one of several options for assessing website usability, but not as the main method. Furthermore, it should be noted that this method is not the primary means of evaluation. Rather, it is one of several evaluation methods that is most effective when used in conjunction with eye-tracking [27].

Eye recording is currently in active use, but due to its high labour intensity, it has not been widely promoted. Today, video motion recording has become popular as personal computers and digital video cameras are widely available. A point source of infrared radiation illuminates the subject's eye when high-speed pictures are taken by an infrared video camera. The position and size of the pupil, which appears as a dark oval in infrared rays, and the position of the corneal glare, a reflection of the infrared light source on the cornea, are determined by image software. The direction of the gaze is calculated by the system based on the vector connecting the corneal glare and the centre of the pupil. This technique's non-contact nature and ability to record the amount of opening are its advantages.

Currently, the most commonly used method is eye-tracking based on video recording of the direction of gaze. The primary components of such systems are one or more video cameras, appropriate software, and an infrared light source [27, 28].

The eye-tracker method captures a person during the usability testing process using cameras. The captured footage highlights the eyes and uses triangulation to determine each eye's position in space relative to the eye-tracker. Using high-resolution cameras ensures accurate gaze direction determination. The frame rate and delay of camera movements reveal how many images per second the camera captures and how long it takes for the images to be available for further processing [29]. Calibration is carried out before testing because each person has unique physiological eye characteristics. In order to obtain accurate and reproducible experimental data, it is necessary to use correct and reliable calibration. During calibration, the participant is instructed to consistently fixate their gaze on a series of calibration markers. Simultaneously, the eye-tracker records the pupil coordinates that correspond to the positions of each calibration marker [1].

Infrared light sources are directed towards the subject's eyes using projectors, enabling the determination of gaze direction across various conditions. Since the human eye is not sensitive to infrared light, the subject does not feel any discomfort. The cornea experiences glare or reflection due to the projector.

Two methods of infrared pupil illumination exist: the light pupil method and the dark pupil method. The distinction between these methods is their relative illumination source location in relation to the camera. When the infrared light is aligned parallel to the camera's optical axis, the eyes function as a secondary reflector of the light that is projected and reflected off the retina, resulting in a light pupil effect analogous to the red-eye effect encountered in photography. When the lighting source is not aligned with the camera's optical axis, the secondary reflection from the retina does not reach the camera, causing the pupil to appear black.

The gaze trajectory is determined by computing the vector between the pupil's center and the infrared light source's reflection from the corneal surface. By knowing the eye's position in relation to the screen and the gaze direction, the gaze point of the participant on the monitor screen is computed. The eye-tracker reads the gaze trajectory's coordinates several dozen times per second, and if unchanged, the time is accumulated. Once the threshold value, approximately 100 ms, is exceeded, the device registers a fixation. Because healthy people's eyes are consistently moving, even when looking at a specific point, the gaze coordinates include a threshold radius (30-50 pixels). If the coordinate values remain within the specified circle, fixation continues. However, if they go beyond it, a new fixation begins.

The heat map is the most commonly used type of eye-tracker data visualisation. Eye-tracker systems provide heat maps, which report the frequency of gaze movements by subjects in different parts of the software, presented as a page divided into colour-coded areas based on cursor movement statistics.

The above method enables calculating the gaze position with high accuracy, analyzing eye movement trajectory, and determining the gaze direction. Eye-tracking technology is vital for scientific research on visual perception processes or those employing visual stimulation. It is also useful for evaluating ergonomics and improving software product interfaces.

The advantages of the eye-tracking method are: the ability to identify areas of increased user attention, where the user's gaze was involuntarily delayed or returned several times, to reproduce the trajectory of gaze movement, the sequence of gaze fixations and the delay time, to select the most successful location of interface elements (buttons, images, etc.) in terms of user comfort, taking into account the user's needs, to set an algorithm that helps users navigate the site, which allows you to create a more effective site design, to process the data received and display several results. The disadvantages of eye tracking are Zones of increased attention can be formed for various reasons, eye-tracking forms zones without determining the principle on which they are formed (necessary information, design or location), misinterpretation of the received information can change the course of further development in the wrong direction, if the user's gaze repeatedly returns to a certain part of the screen or there is a long fixation, Taking into account the above-mentioned disadvantages of this method, it is proposed to combine usability testing methods with other methods in order to increase the efficiency of software quality evaluation.

The analysis showed that both mouse tracking and eye tracking are very convenient and useful tools, and it is rather pointless to oppose them. You just need to understand what you want to achieve and what parameters you want to measure [23].

The paper [23] states that if there was a cursor in a certain place on the page, the visitor looked there with a probability of 84%, and if he did not look at that place, there was no cursor there with a probability of 88%. However, this does not mean that the cursor was everywhere the visitor looked. In other words, the heatmap of mouse movements shows a part of what the user probably saw. Therefore, mouse movements cannot be used to estimate the temporal parameters of information perception, including the sequence of this perception. In other words, there is a certain correlation between cursor movement and gaze, but it is not entirely clear how significant it is and in which cases it is applicable.

Eye-tracking's primary drawback is its high cost. This includes expensive equipment, hiring highly qualified staff, and paying for respondents' recruitment and compensation. Yet, if we consider the type of data gathered and its potential cost-saving outcomes, the research price is not excessively high. When compared to other methods, eye-tracking offers a wider range of research possibilities. Mouse-tracking is limited to exploring websites and only allows restricted analysis of software interfaces. However, eye-tracking enables the study of additional materials such as design layouts and prototypes, printing, and commercials that don't require user interaction.

Testing layouts and prototypes is one of the most effective ways to employ eye-tracking. This method enables error detection at the early stages of development, thereby reducing the cost of fixing them. While mouse-tracking is limited to examining websites and application interfaces, eye-tracking can be utilised for examining design layouts, prototypes, printing, commercials, and other materials that do not require any user interaction.

Figure 2 illustrates the findings from a comparative analysis of traditional testing methods, mouse-tracking and eye-tracking.

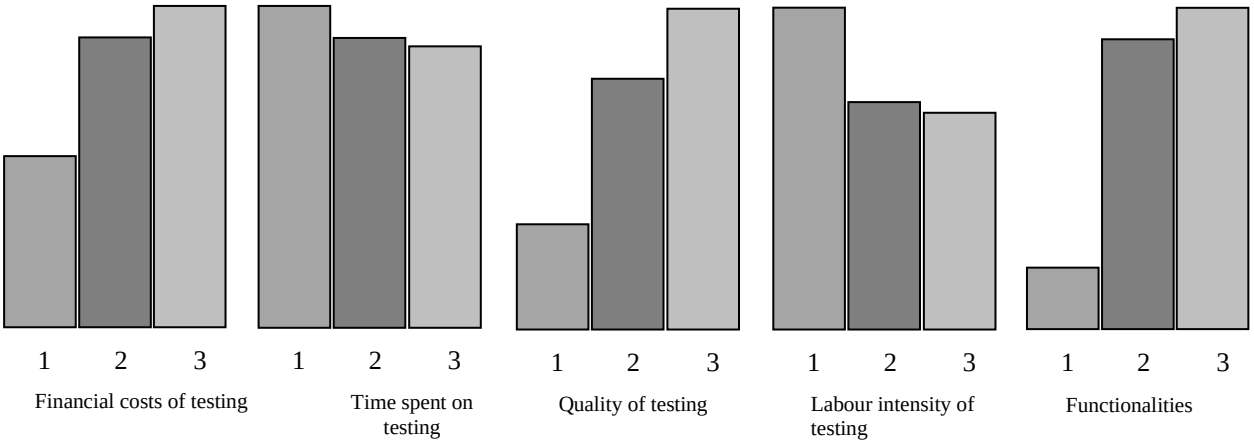

Fig. 2. Comparative analysis of software usability testing methods, where 1 - traditional testing methods; 2 - mouse-tracking technology; 3 - eye-tracking technology.

We formulated a sample usability testing task for the tax service website, based on the previously mentioned tasks:

1. Task. The task is to obtain personal information on the website. To access personal information, users must log in to the individual's account.

1.1. Find your individual tax number.

1.2. Find information about taxable objects and view information about accrued taxes.

1.3. Determine the current status of tax accounts.

1.4. Find the series and number of an identity document.

1.5. Find information about certificates.

1.6. Find information about income from the income tax return.

2. Task. The objective is to retrieve reference information from the website. To obtain more information, the user must download either a strengthened qualified electronic signature key or one of the two versions of a strengthened unqualified electronic signature to their personal account on their computer.

2.1. Send a request for information about the status of payments.

2.2. File a complaint against an act of a tax authority of an abnormal nature.

2.3. Send an application for clarification of information about the objects.

2.4. Submit an application for property tax relief.

2.5. Send a request for information about the status of tax accounts.

3. Task. Task description. Notification of changes in personal information should be provided for individuals. To report the presence of real estate or a vehicle, an unqualified electronic signature key must be created and a verification key certificate obtained.

3.1. Notification of the tax service of the change of place of registration.

3.2. Notify the tax authorities of the ownership of real estate or a vehicle.

3.3 Notify of opening an account in a bank located outside the country.

3.4. To notify the tax authorities of failure to file personal income tax returns.

4. Task. Preparation of standard documents of an individual. To fill out and send receipts/applications/declarations, the user needs to download a software program to fill them out on his/her computer. Also, the generated document must be signed with an enhanced qualified or unqualified electronic signature from the personal account before being sent to the tax authority.

4.1 Send a statement of receipt (non-receipt) of the social contribution.

4.2. Send/fill in the declaration online.

4.3. Send an application for confirmation of the right to receive property deductions.

4.4. Send an appeal to the tax authority.

5. Task. Payment of mandatory taxes by an individual. To pay bills, you are offered to download completed receipts or make electronic payment through the services of partner credit institutions.

5.1. Pay off tax arrears.

5.2. Pay real estate tax.

5.3. Pay income tax.

5.4. Pay the state fee for registration of an individual entrepreneur.

6. Task. Addressing a personal question to the employees of the tax service of an individual.

6.1. Submit a free-form request.

The developed test tasks cover some of the variations of the tasks set and show the main purposes of using the tax service website.

The following were chosen as fixed usability metrics:

the time taken to complete a test task is used to determine the system's learning curve;

the presence of user errors, such as selecting menu items, changing environment settings that are not required to solve the task, is used to determine the success of the task;

subjective user satisfaction will be assessed using a questionnaire, etc.

Additional questions asked to the test subject after completion of the test tasks form a subjective assessment and reflect the test subject's general attitude to the website under test. This score depends on the user's subsequent visits to the site. The formed test task allows to qualitatively and quantitatively evaluate the user interface of the tax service website and determine the degree of usability of this resource [33].

Conclusions

This study explores the concept and definition of user interface usability in line with the standards of ergonomics and human-system interaction. The study identifies three criteria, namely efficiency, productivity, and satisfaction, for assessing usability. Obtaining a sufficient level of usability requires the application of standards in this area.

Furthermore, the article analyses the principal standards of user interface usability, which describe the requirements for developing and testing an information system's user interface.

The study examines the characteristics of software quality assessment, namely functionality, reliability, practicality, efficiency, support, and mobility.

The standards and criteria that are considered are used to develop the user interface. The design process's primary stages and the significance of usability testing at each stage are analysed. The process comprises requirement identification for the future system, system specification, implementation, verification, and system implementation and support. Usability testing is a crucial feature of the technology being considered as it enables the direct identification of ergonomic issues during the development stage of the user interface. This, in turn, will help reduce resource costs during system design.

The article explores the criteria used to assess the usability of both qualitative and quantitative methods. In this study, we examine the quantitative method of tracking user actions, known as mouse-tracking. We identify the advantages and disadvantages of this method.

The article also examines the technology behind the pupil capture method in eye-tracking. Nowadays, the most commonly used technique relies on infrared light illumination to record a video of the gaze trajectory on the face. This technique allows the subject to remain as natural as possible when working with the system under assessment. The document covers the features of pupil capture, its quantitative indicators, the operating principle, and the technology's benefits. This study also considers different types of visualisations for eye-tracking results, including gaze graphs, heat maps, bee swarms, cluster analysis, and zones of interest. Although the method of eye-tracking has its limitations in usability testing, it is currently one of the most effective among existing methods.

This study analyses the technological aspects of the eye-tracking method. This study identifies the features and positive attributes of the eye-tracking method. The technology described enables the identification of weaknesses in the development of a user interface. Comparing the predetermined usability criteria with the test results exposes issues of usability in the design of the user interface. It has been established that it is highly important to conduct usability testing during the development process to achieve high-quality software products. Characteristics identified through quality assessment determine the final appearance of the product at the end of its development. Previously set tasks facilitated a fresh perspective on user interface usability testing process.

We employed an analytical approach as a methodological basis that involves thoroughly considering the research object. To obtain the data, we used comparative analysis and classification methods. Subsequently, we processed the results using standard statistical methods.

We examined the primary challenges that arise when evaluating usability qualities during various testing methods, including the survey, mouse-tracker, and eye-tracker methods. It was found that mouse-tracking technology enables the examination of websites and, to a certain extent, application interfaces. Eye-tracking technology, on the other hand, is useful in the study of design layouts, prototypes, printing, commercials, and other user-inactive materials.

To provide a detailed analysis of the research topic, we examined the user's ergonomic interactions with the information system, specifically with the web resource. The main user categories were analyzed according to gender, age, physical and psychological condition, and citizenship. Special attention was given to establishing quality assessment criteria based on existing standards.

According to domestic and foreign studies, the modern gaze-tracking based method for usability testing is more effective as it provides both significant quantitative and qualitative results in a single testing session.

The results of the usability testing demonstrate the qualitative differences between mouse and eye-tracking. In the usability testing, it was discovered that the use of eye-tracking method facilitates the acquisition of more qualitative and quantitative data about the software product than mouse tracking, revealing more detailed information about the test subject's attitude towards the website interface. Nonetheless, in situations where it is necessary to comprehend the intentions rather than the subconscious behaviour of customers, which can be subjective and deceitful, mouse-tracking can provide more valuable

information. The integrated eye-tracking testing method is highly efficient for software interface testing as it includes data obtained from other usability testing methods.

Empirically obtained data accurately and extensively described the eye-tracker assisted testing features, revealing the problem in the best possible way. Considering these characteristics in detail indicates that they are optimal and can be supplemented by other methods, depending on the usability testing goal. Eye-tracking technology has a high volume capacity for system usability analysis, making it relevant for designing and implementing software products.

To optimize and rationalize website development processes, it is crucial to employ modern usability technologies and tools. This research provides a comprehensive study on theoretical and practical aspects of improving software quality through the direction of usability testing. It presents a comparative analysis of the effectiveness of traditional manual approaches versus the latest advances in mouse-tracking and eye-tracking technologies in ensuring software quality.

References

1. Dibrova T. G. Evaluation of the effectiveness of advertising by methods of physiological control / T. G. Dibrova, I. I. Garanina // Ekon. vistnyk of the National Technical University of Ukraine "KPI": collection of scientific papers - 2013. - Issue 10. - P. 316-320.

2. Kirilenko O. Methods for evaluating the usability of the user interface / O. Kirilenko, Y. Kuznetsova, E. Sokolova, G. Frolova // Bulletin of the National University "Lviv Polytechnic". - 2013. - № 751. - C. 244-256..

3. Kuznetsova Y. A. Visualisation of control algorithms in an automated testing system of a complex technical complex: PhD thesis: 05.13.06 / Y. A. Kuznetsova; National Aerospace University named after M. E. Zhukovsky - Kharkiv, 2015. 20 p.

4. Datsun N. M. Increasing the efficiency of the user interface of virtual laboratory work [Electronic resource] / N. M. Datsun, O. A. Goretsky // Scientific works of Donetsk National Technical University. Series: Informatics, Cybernetics and Computer Science. - 2012. - Vol. 15. - P. 239-244. - Access mode: http://nbuv.gov.ua/UJRN/Npdntu_inf_2012_15_36.

5. New concepts of modern ergonomics [Electronic resource] / G.V. Migal. Myhal, O.F. Protasenko // Open information and computer integrated technologies. - 2018. - Issue 79, pp. 162-170. - Access mode: http://nbuv.gov.ua/UJRN/vikt_2018_79_20.

6. About the place of usability testing [Electronic resource]. - Access mode: ttp://www.gui.ru/platon/-?p=27.

7. Usability tests: how to choose the right method [Electronic resource]. - Access mode: https://blog.uxfeedback/yuzabiliti-testy-kak-vybrat-pravilnyj-metod/.

8. Methods of usability testing of sites. Making the site convenient [Electronic resource]. - Access mode: https://adverbs/blog/metody-yuzabiliti-testirovaniya-saytov-delaem-sayt-udobnym/.

9. Usability Testing від А до Я: докладний гід [Електронний ресурс]. − Режим доступу: https://dou.ua/lenta/articles/usability-testing-guide/.

10. Overview of usability testing methods [Electronic resource]: https://artw/blog/archives/3537/.

11. Types of interface usability testing [Electronic resource]. - Access mode: https://askusers/blog/obuchenie/testirovanie-usability-interfejsa/.

12. Usability audit of the site [Electronic resource]. - Access mode: https://fox-it.com.ua/usability-audit/?gclid=CjwKCAiAqIKNBhAIEiwAu_ZLDnfEALYbAfyDmUmoKnswi4Ue9nkD_foEuuX8s6tLdrzj7mGjpzwg2hoCRM0QAvD_BwE.

13. Usability testing of the site [Electronic resource]. - Access mode: https://www.fabuza/blog-post/yuzabiliti-testirovanie-sayta.

14. Quantitative and qualitative research methods [Electronic resource]. - Access mode: https://sense23.com/post/diy-gajd-po-metodam-produktovyh-issledovanij-kachestvennye-kolichestvennye-ux-issledovaniya.

15. Electrooculography [Electronic resource]. - Access mode: https://wikipedia.org/wiki/%D0%A-D%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D0%BE%D0%BA%D1%83%D0%BB%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F.

16. Gaze control system based on the method of electro-oculography - ELAKRI [Electronic resource]: https://ela.kpi.ua/handle/123456789/56788?mode=full.

17. What is neuromarketing and how it helps brands [Electronic resource]: https://blog.admixer.academy/ua/shho-take-nejromarketing-i-yak-vin-dopomagaye-brendam/.

18. Electroretinography in ophthalmology [Electronic resource]: https://ophthalmobook.com.ua/knigi/oftalmologiya/diagnostika/elektroretinografiya-v-oftalmologii.html.

19. Electroretinography [Electronic resource]. - Access mode: https://ru.wikipedia.org/wiki/%-D0%AD%D0%BB%D0%B5%D0%BA%D1%82%D1%80%D0%BE%D1%80%D0%B5%D1%82%D0%B8%D0%BD%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F.

20. The Power of Eye Tracking Research: What Types of Research Can Be Conducted in Different Industries [Electronic resource]: https://eyeware.tech/uk/blog/%D0%B4%D0%BE%D1%81%D0%BB-%D1%96%D0%B4%D0%B6%D0%B5%D0%BD%D0%BD%D1%8F%D1%81%D0%BF%D0%BE%D1%81%D1%82%D0%B5%D1%80%D0%B5%D0%B6%D0%B5%D0%BD%D0%BD%D1%8F%D0%B7%D0%B0%D0%BE%D1%87%D0%B8%D0%BC%D0%B0/.

21. Geometry of the Dawn [Electronic resource]. - Available from: https://naurok.com.ua/prezentaciya-geometriya-zoru-239725.html.22. Data analysis models for computer vision [Electronic resource]: https://mmda.ipt.kpi.ua/cp_bachelors/.

23. Eye Tracking vs. Mouse Tracking Analytics [Electronic resource]: https://contentsquare.com/blog/eye-tracking-vs-mouse-tracking-analytics/.

24. Usability results - testing of Mouse-tracking and Eye-tracking [Electronic resource]: https://thelib.info/tehnologii/1934831-rezultaty-juzabiliti-testirovaniya-mouse-tracking-i-eye-tracking/.

25. Mill Mouse - Mouse Controller with Eye Tracking by Tobii Eye Tracker 4C [Electronic resource]: https://millmouse.wordpress.com/about/%D0%BE%D0%BF%D1%80%D0%BE%D0%B3%D1%-80%D0%B0%D0%BC%D0%BC%D0%B5-mill-mouse/.

26. Mouse-tracking [Electronic resource]. - Access mode: https://en.wikipedia.org/wiki/Mouse_tracking.

27. Mouse Tracking [Electronic resource]. - Access mode: https://en.ryte.com/wiki/Mouse_Tracking.

28. Improving your website with Mouse Tracking [Electronic resource]: https://zwebra.com.ua/uluchshenie-sajta-pri-pomoshhi-mouse-tracking.html.

29. Eye-tracking [Electronic resource]. - Access mode: https://www.eyetracking.care/informatsiya-/stati/chto-takoe-aytreking-eye-tracking-i-zachem-on-nuzhen/.

30. Method of using eye-tracking technology in the design of user interfaces [Electronic resource]: file:///C:/Users/%D0%9D%D0%BE%D1%83%D1%82/Downloads/12.pdf.

31. Systems and methods of decision support [Electronic resource]. - Access mode: https://ela.kpi.ua/bitstream/123456789/48418/1/Systemy_i_metody_pidtrymky_pryiniattia_rishen.pdf.
32. Textbook of usability testing [Electronic resource]. - Access mode: https://uk.myservername.com/usability-testing-tutorial.
33. Eye Tracking Technology, or What Web Analytics Services Will Not Give You [Electronic resource]: https:// http://slaidik.com.ua/tehnologiya-eye-tracking-abo-chogo-ne-dadut-vam-servisi-veb-analitiki/.
34. Bermus A. G. [Electronic resource]. - Access mode: https://www.researchgate.net/publication-/371829475_Stata_Bermusa_AG_Hronotopolingvisticeskij_HTL_podhod_issledovaniam_transformacii_obrazovania.

Yuriy Skorin **Юрій Скорін**	PhD, Associate Professor of Information Systems Department, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine. e-mail: skorin.yuriy@gmail.com https://orcid.org/0000-0002-4613-3154	к.т.н., доц. кафедри інформаційних систем, Харківський національний економічний університет ім. С. Кузнеця, Харків, Україна.
Iryna Zolotaryova **Ірина Золотарьова**	PhD, Professor of Information Systems Department, Simon Kuznets Kharkiv National University of Economics, Kharkiv, Ukraine e-mail: iryna.zolotaryova@hneu.net https://orcid.org/0000-0002-1553-2849	к.е.н., проф. кафедри інформаційних систем, Харківський національний економічний університет ім. С. Кузнеця, Харків, Україна.

https://doi.org/10.31891/csit-2023-3-9
UDC 004.65

Valerii NIKITIN, Evgen KRYLOV
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

PRIMARY-BASED SPECTRAL BLOOM FILTER FOR THE ENSURING CONSISTENCY IN DISTRIBUTED DOCUMENT-BASED NoSQL DATABASES USING ACTIVE ANTI-ENTROPY MECHANISM

The purpose of this work is to compare the existing methods of forming the Spectral Bloom filter using hash functions and the proposed method using prime numbers. The proposed method allows obtaining snapshots from documents that can be used to maintain data consistency in distributed document-oriented NoSQL databases as part of the Active Anti-Entropy mechanism. Data consistency is an important and challenging task due to the need for horizontal scaling of information systems. Neglecting this can lead to material or even human losses, since digitalization covers absolutely all spheres of human activity and there is a need for distributed processing and storage of information.

Consistency can be ensured in various ways, including an architectural approach and Active Anti-Entropy mechanisms. The architectural approach refers to centralized write operations that are distributed to secondary nodes. Accordingly, read operations take place from secondary nodes. This approach is not flexible, as it requires stable and fast communication with the central node, which is not always possible.

The Active Anti-Entropy mechanism is a background process that checks the consistency of data between nodes using special snapshots that can be obtained using hash functions or such a data structure as a Merkle Tree. Using the latter is ideal for checking the consistency of entire data sets, but for mission-critical data, this solution is not suitable. The probability of collisions or the computational cost can lead to inconsistency of the entire data set and this requires a special solution for critical data.

The proposed method makes it possible to obtain the Spectral Bloom filter from the original data set faster. In addition, it has higher collision resistance compared to the use of hash functions, which allows faster identification of inconsistencies in documents stored on different nodes.

Keywords: NoSQL, document-oriented databases, distributed databases, Spectral Bloom filter, consistency, Active Anti-Entropy

Валерій НІКІТІН, Євген КРИЛОВ
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

СПЕКТРАЛЬНИЙ ФІЛЬТР БЛУМА НА ОСНОВІ ПРОСТИХ ЧИСЕЛ ДЛЯ ВИКОРИСТАННЯ В ACTIVE-ANTI ENTROPY МЕХАНІЗМІ УЗГОДЖЕННЯ ДАНИХ У РОЗПОДІЛЕНІЙ ДОКУМЕНТООРІЄНТОВАНІЙ НЕРЕЛЯЦІЙНІЙ БАЗІ ДАНИХ

Метою даної роботи є порівняння існуючих методів формування спектрального фільтра Блума з використанням хеш-функцій та запропонованого методу з використанням простих чисел. Запропонований метод дозволяє отримати знімки з документів, які можна використовувати для підтримки узгодженості даних в розподілених документоорієнтованих NoSQL базах даних як частину механізму Active Anti-Entropy. Узгодженість даних є важливою та складною задачею через необхідність горизонтального масштабування інформаційних систем. Нехтування цим може призводити до матеріальних або навіть людських втрат, оскільки цифровізація охоплює абсолютно всі сфери діяльності людини і є необхідність у розподіленій обробці та зберіганні інформації.

Консистентність може забезпечуватись різними шляхами, серед яких архітектурний підхід та Active Anti-Entropy механізми. Під архітектурним підходом мається на увазі централізовані операції запису, які розповсюджуються до другорядних вузлів. Відповідно операції читання відбуваються з другорядних вузлів. Даний підхід не є гнучким, оскільки вимагає стабільного та швидкого зв'язку з центральним вузлом, що не завжди можливо.

Active Anti-Entropy механізм представляє собою фоновий процес, який перевіряє узгодженість даних між вузлами використовуючи спеціальні знімки, які можуть бути отримані з використанням хеш-функцій або такої структури даних як Merkle Tree. Використання останнього ідеально підходить для перевірки узгодженості цілих наборів даних, але для критично важливих даних це рішення не підходить. Ймовірність колізій або обчислювальні витрати можуть призводити до неузгодженості цілого набору даних і це вимагає спеціального рішення для критично важливих даних.

Запропонований метод дозволяє швидше отримувати спектральний фільтр Блума з вихідного набору даних. Окрім цього, він має вищу колізійну стійкість в порівнянні з використанням хеш-функцій, що дозволяє швидше ідентифікувати неузгодженість документів, які зберігаються на різних вузлах.

Ключові слова: нереляційні бази даних, документоорієнтовані бази даних, розподілені бази даних, спектральний фільтр Блума, узгодженість даних, Active Anti-Entropy технологія

Introduction

Distributed databases need special mechanisms that could ensure the necessary level of data consistency between replicas. Some modern databases solve this through architectural solutions. For example, a database like MongoDB uses a centralized approach for all write operations, while read operations are performed through replicas.

There is also a decentralized approach, in which there is no master node, and its functions are distributed among all replicas of the system. Cassandra and Riak work according to this principle. Riak, in turn, has several

additional mechanisms: Active Anti Entropy and Read Repair. Riak's Active-Anti Entropy technology is a background process that uses Merkle Tree to identify data inconsistencies. The identifiers received on different nodes in the form of hash values are compared with each other, which allows you to start the reconciliation process if necessary.

There are also some other methods of maintaining consistency that are a kind of centralized approach. For example, a transaction clock receives all transactions and merges them into a resulting transaction that is sent to the replicas. The advantage is that the resulting transaction will have the latest update and the number of writes to the database is optimized [1].

Related Works

Bloom Filter is a probabilistic data structure that is designed to check the presence of an element in a set. A filter consists of an array of values that take the value 1 or 0. To form the filter, hash functions are used that determine the elements of the array that will have the value 1. The size of the filter and the hash functions can be configured depending on the context of the task.

There are different variations of this data structure: Distributed Bloom Filter, Spectral Bloom Filter, Space-Code Bloom Filter. Each of these variants has its own differences and specific applications.

The Distributed Bloom Filter is a probabilistic data structure designed for distributed systems that require fast synchronization both spatially and temporally. This is achieved by sending a filter by each node with its ID to another node in the system, which results in a change in the probability of data replication. It should be noted that the probability of a false positive result decreases with an increase in the total number of nodes [2].

SCBF achieves a good compromise between counting accuracy and the number of bits used for counting. It represents a multiset, extending the capabilities of the traditional Bloom filter to represent a set. Given an element x, it not only allows you to check whether x is in the multiset, but also counts the number of occurrences of x. SCBF has several important usages in network measurement. For example, traffic measurement by flow for traffic design and anomaly detection [3].

Spectral Bloom Filter replaces a vector of bits with a vector of counters. Counters are positive natural numbers that have leading zeros. When calculating the position in the data vector, the corresponding counter is increased by one. Deleting an element proceeds in a similar way, but by subtracting 1 from the related counters [4].

Any hash function can be used for hashing, but since they can be combined and there is a limit of the size of the Bloom filter, it is difficult to calculate the probability of collisions. Even using an algorithm like PH-2, which is not a cryptographic algorithm and clearly defines collision situations, does not guarantee an optimal allocation [5].

The method of forming the Spectral Bloom filter using simple numbers

The Active-Anti Entropy mechanism is a background process of data reconciliation in distributed databases. The use of the Bloom filter to solve this problem is due to a convenient algorithm for forming a snapshot, which can be used to compare critically important documents that are located on different nodes of a distributed database.

To be used in the Active-Anti-Entropy mechanism, the Bloom filter must meet certain requirements:

1) low probability of collisions;
2) high speed of calculation;
3) small size.

Each of the above-mentioned components has an impact on the others, which requires the search for an optimal solution.

In the context of the task to support consistency between nodes of a distributed document-oriented database, the spectral Bloom filter is a set of eight non-negative integers. Each number is a counter that increases by one each time it is accessed. At the very beginning, the values of all counters are 0.

Access to a certain element is as follows:

1) represent the document as bytes;
2) each byte is converted to an unsigned integer;
3) division by modulo of each number by a set of prime numbers from 2 to 17 inclusive. If the remainder during division is zero, then the corresponding element in the Bloom filter is increased by one. If the number was not divided by any prime number from the set, then the last element of the filter is incremented by one;
4) if the number is not divisible by any prime number, then the last counter is increased;
5) if an integer overflow of the counter occurs, it is reset to zero.

Figure 1 shows the filter formation algorithm for a document consisting of 3x bytes. This document size was chosen for ease of visualization.

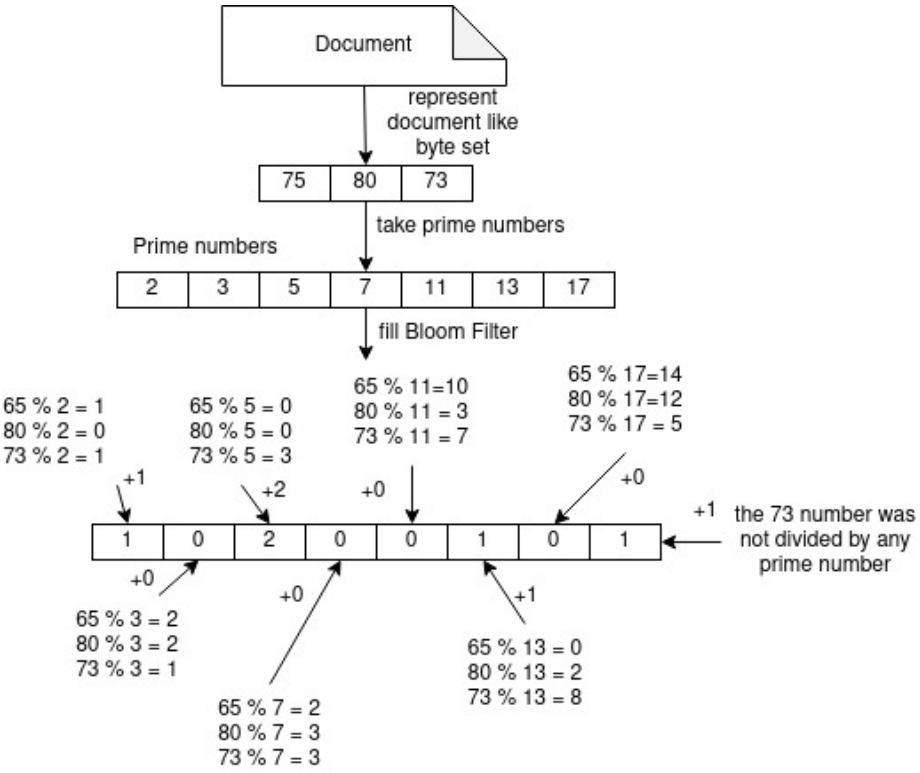

Fig. 1. Prime-based algorithm to produce Spectral Bloom filter

Experiments

The experimental procedure for comparing collision resistance consists of randomly generating input data of a fixed length and a certain step. The amount of data is 100,000. Test data are ASCII characters used to represent an information. The dependent variable in this case is the number of unique Spectral Bloom filters, and the independent variable is the amount of data. MD5, SHA1 and SHA256 were selected as hash functions, which are cryptographic functions and should ensure an even distribution of hash values.

Algorithm of forming a filter using a hash function:

1) present the document in the form of bytes;
2) hash each byte;
3) represent each hash as an unsigned integer;
4) fill the filter using the modulo division operation.

Figure 2 shows a visualization of the results obtained when generating 100,000 random data with a size of 100 to 450 bytes and a step of 50 bytes.

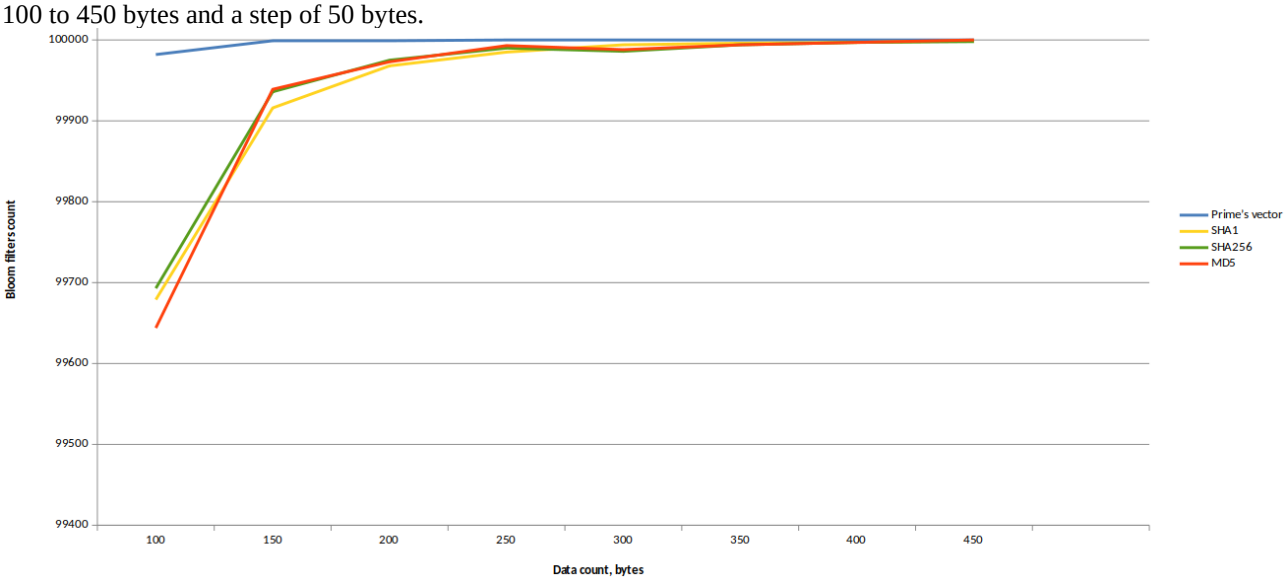

Fig. 2. Unique Spectral Bloom filters from 100,000 random data from 100 to 450 bytes and step 50

Figure 3 shows a visualization of the results obtained when generating 100,000 random data with a size from 5 to 174 bytes and a step of 1 byte.

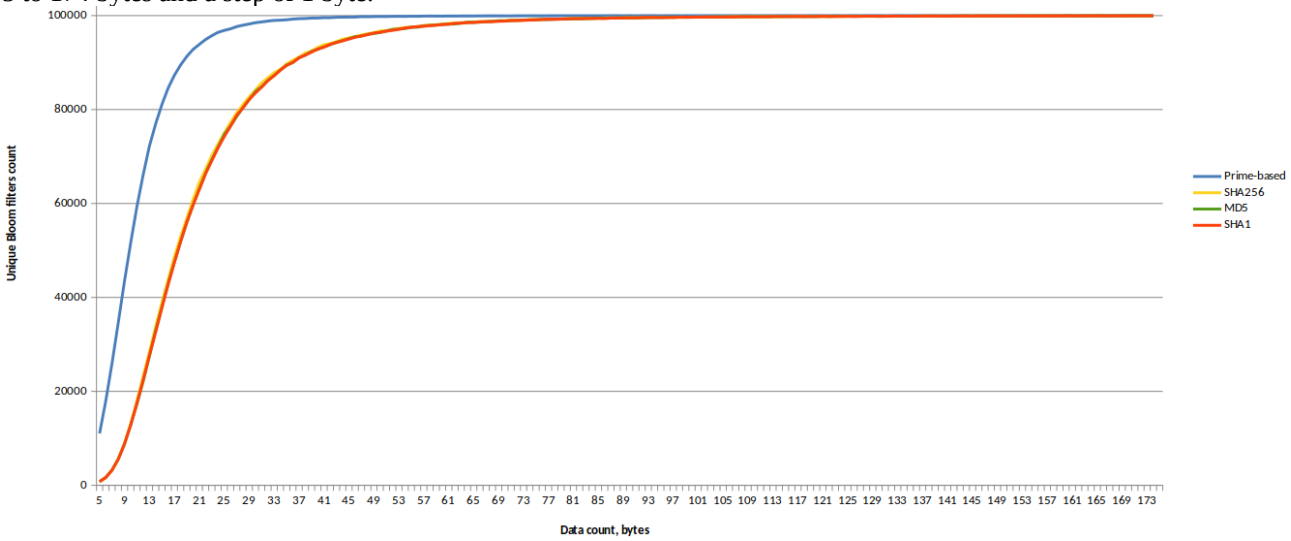

Fig. 3 - Unique Spectral Bloom filters from 100,000 random data from 5 to 174 bytes and step 1

Figures 2 and 3 show that the Prime-based Spectral Bloom filter has higher collision resistance and more quickly approaches the expected value under the given experimental conditions.

To compare the filter calculation speed, the test procedure will be the same as the previous one, but the number of filter calculation iterations for the generated test set is 100. Statistical tools such as arithmetic mean, median, minimum and maximum value will be used to evaluate the methods. The dependent variable in this case is the time required to form the filters.

Figure 4 shows a chart of the arithmetic mean value for data of different lengths.

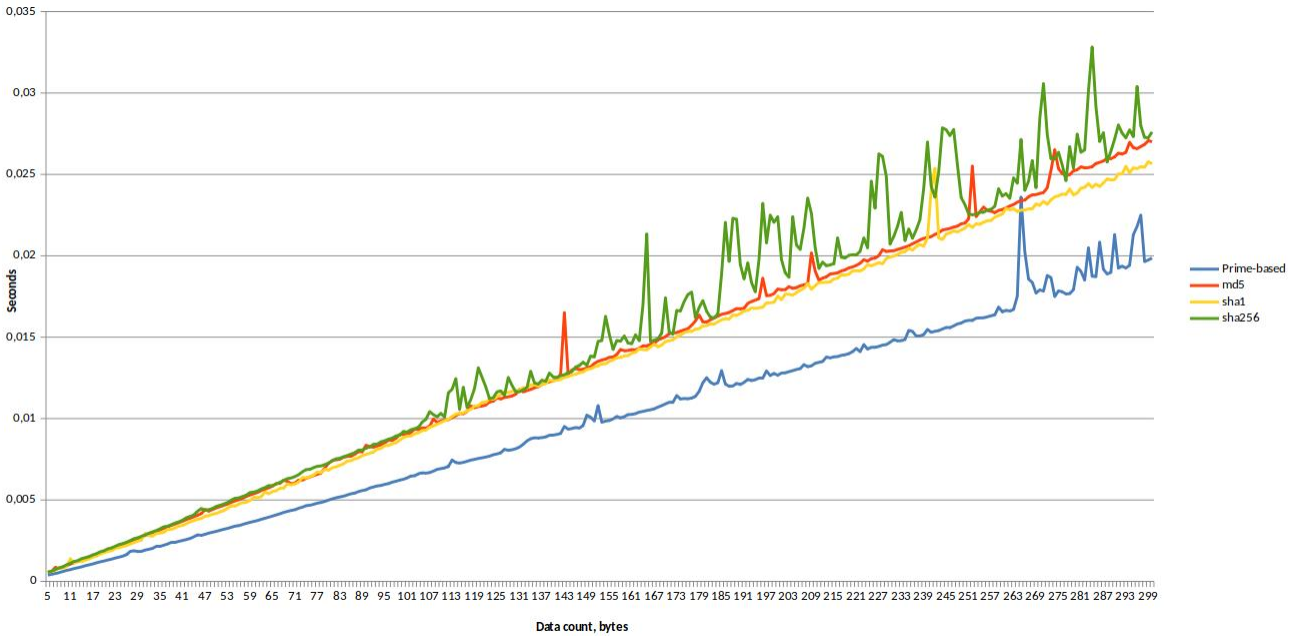

Fig. 4. Mean of elapsed time to get 100 prime-based spectral Bloom filters

Figure 5 shows a chart of the median for data of different lengths.

Fig. 5. Median of elapsed time to get 100 prime-based spectral Bloom filters

Figure 6 shows a chart of the minimum value for data of different lengths.

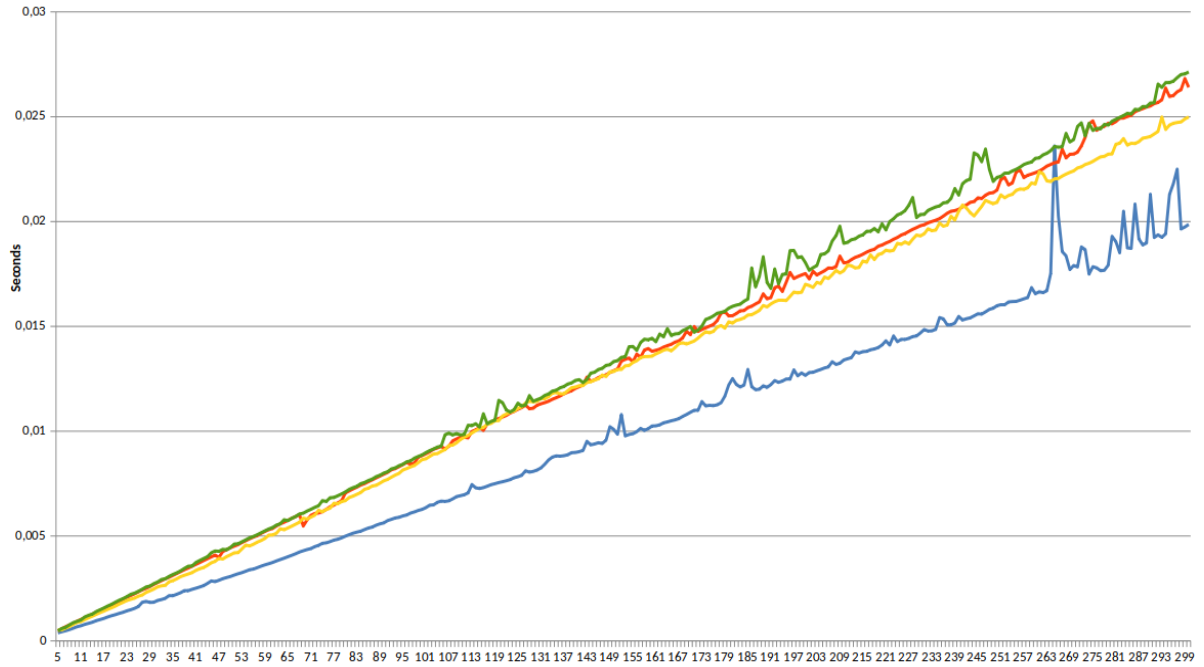

Fig. 6. Min of elapsed time to get 100 prime-based spectral Bloom filters

Figure 7 shows a chart of the maximum value for data of different lengths.

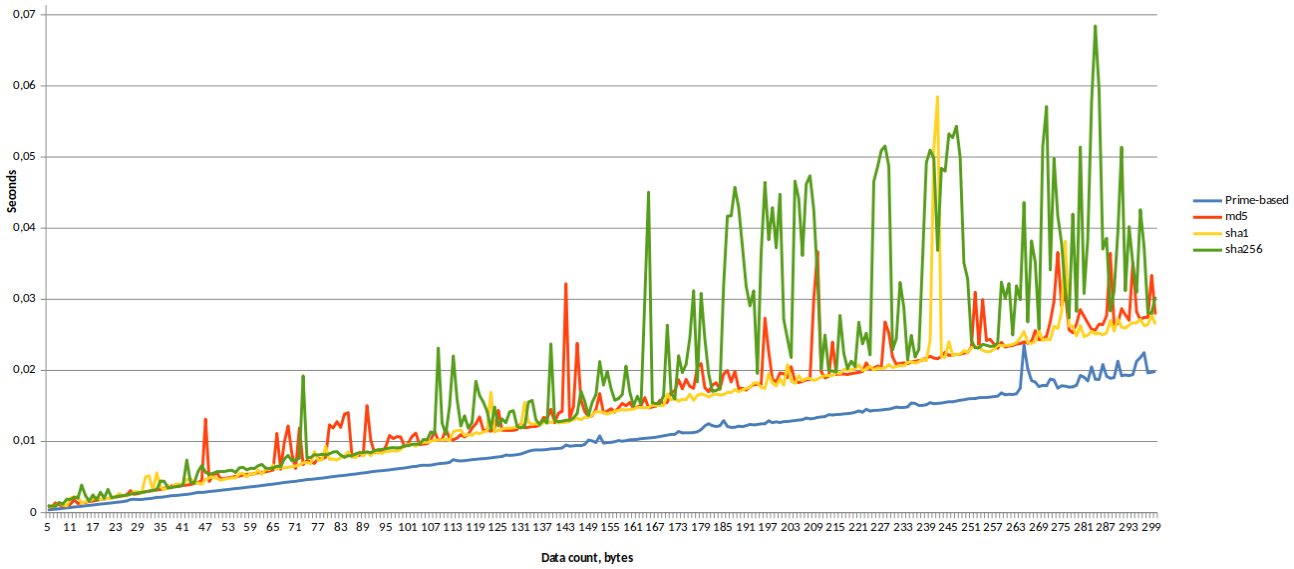

Fig. 7. Max of elapsed time to get 100 prime-based spectral Bloom filters

Figures 4, 5, 6 and 7 show that the time required for the formation of filters is shorter and varies in proportion to the values of the independent variable.

Conclusions

Thus, the proposed method of forming the Spectral Bloom filter has a number of advantages for its use in the data reconciliation process. The small size of the filter allows it to be easily transported between nodes of a distributed database to check the consistency of critical documents, and the time to create it allows for quick creation of the filter from source documents. Also, it is more resistant to collisions compared to forming a filter using hash functions.

The further direction of the work is related to the design of the Active Anti-Entropy mechanism using the proposed method for maintaining the consistency of critical data in distributed document-oriented NoSQL databases and its effectiveness evaluation.

References

1. Nikitin, V. A., & Krylov, E. V. (2023). Using a transactional clock to speed up the process of data reconciliation in distributed systems. Scientific Bulletin of Uzhhorod University. Series "Mathematics and Informatics", 42(1), 188-192. URL: https://doi.org/10.24144/2616-7700.2023.42(1).188-192

2. Ramabaja L. The Distributed Bloom Filter // arXiv. 2019. С.1-17 URL: https://doi.org/10.48550/arXiv.1910.07782

3. Kumar A., Xu J., Wang J. Space-Code Bloom Filter for Efficient Traffic Flow Measurement // IEEE Journal on Selected Areas in Communications. 2006. № 24 (12). С.1-12 URL: https://doi.org/10.1109/JSAC.2006.884032

4. Cohen S., Matias Y. Spectral Bloom Filters // Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data. 2003. С.1-12 URL: http://dx.doi.org/10.1145/872757.872787

5. Nikitin, V., & Krylov, E. (2022). A collision-resistant hashing algorithm for maintaining consistency in distributed NoSQL databases. Adaptive Systems of Automatic Control Interdepartamental scientific and technical collection, 2(41), 45–57. URL: https://doi.org/10.20535/1560-8956.41.2022.271338

Valerii Nikitin **Валерій Нікітін**	PhD student, Department of Information Systems and Technologies, Igor Sikorsky Kyiv Polytechnic Institute. e-mail: 19valeranikitin96@gmail.com https://orcid.org/0000-0002-4509-1204	аспірант кафедри інформаційних систем та технологій КПІ ім. Ігоря Сікорського.
Evgen Krylov **Євген Крилов**	PhD in Technical Sciences, Associate Professor, Department of Information Systems and Technologies, Igor Sikorsky Kyiv Polytechnic Institute. e-mail: ekrylov1964@gmail.com https://orcid.org/0000-0003-4313-938X	кандидат технічних наук, доцент кафедри інформаційних систем та технологій КПІ ім. Ігоря Сікорського.

https://doi.org/10.31891/csit-2023-3-10
UDC 004.9

Viktor SHYNKARENKO, Inna DEMIDOVICH
Ukrainian State University of Science and Technologies

CONSTRUCTIVE-SYNTHESIZING MODELING OF NATURAL LANGUAGE TEXTS

Means for solving the problem of establishing the natural language texts authorship were developed. Theoretical tools consist of a constructors set was developed on the basis of structural and production modeling. These constructors are presented in this work. Some results of experimental studies based on this approach have been published in previous works by the author, the main results should be published in the next ones.

Constructors developed: converter of natural language text into tagged, tagged text into a formal stochastic grammar and the authors style similarity degree establishment of two natural language works based on the coincidence of the corresponding stochastic grammars (their substitution rules).

In this paper, constructors are developed and presented that model a natural language text in the form of a stochastic grammar that displays the structures of sentences in it. This approach allows you to highlight the syntactic features of the construction of phrases by the author, which is a characteristic of his speech. Working with a sentence as a unit of text for analyzing its construction will allow you to more accurately capture the author's style in terms of the words use, their sequences and speech style characteristic. It allows you not to be tied to specific parts of speech, but reveals the general logic of constructing phrases, which can be more informative in terms of the author's style characteristics for any text.

The presented work is a theoretical basis for solving the problems of the text authorship establishing and identifying borrowings. Experimental studies have also been carried out. The statistical similarity of solutions to the problems of establishing authorship and identifying borrowings was experimentally revealed, which will be presented in the next article of the authors.

The proposed approach makes it possible to highlight the semantic features of the author's phrases construction, which is a characteristic of his speech. Working with a sentence as a unit of text to analyze its construction will allow you to more accurately determine the author's style in terms of the use of words, their sequences and characteristic language constructions. Allows not to be attached to specific parts of speech, but reveals the general logic of building phrases.

It is planned to use the created model in the future to determine the authorship of natural language texts of various directions: fiction and technical literature.

Keywords: natural language texts, constructive-synthesizing modeling, establishing authorship, formal grammars, stochastic grammars, text models

Віктор ШИНКАРЕНКО, Інна ДЕМИДОВИЧ
Український державний університет науки та технологій

КОНСТРУКТИВНО-ПРОДУКЦІЙНЕ МОДЕЛЮВАННЯ ПРИРОДНЬОМОВНИХ ТЕКСТІВ

Розроблені засоби для вирішення задачі встановлення авторства природньомовних текстів. Теоретичні засоби складаються з комплексу конструкторів розроблених на основі конструктивно-продукційного моделювання. Саме ці конструктори представлені в данній роботі. Деякі результати експериментальних досліджень основаних на цьому підході опубліковані в попередніх роботах авторів, основні результати мають бути опубліковані в наступних.

Розроблені конструктори: перетворювач природньомовного тексту на тегований, тегованого тексту у формальну стохастичну граматику та встановлення ступеню схожості стилю авторів двох природньомовних творів за збігом відповідних стохастичних граматик (їх правил підстановки).

У статті розроблено та представлено конструктори, які моделюють текст природною мовою у вигляді стохастичної граматики, що відображає структури речень у ньому. Такий підхід дозволяє виділити синтаксичні особливості побудови фраз автором, що є характеристикою його мовлення. Робота з реченням як одиницею тексту для аналізу його побудови дозволить точніше вловити стиль автора з точки зору вживання слів, їх послідовності та характеристики стилю мовлення. Він дозволяє не прив'язуватися до конкретних частин мови, а розкриває загальну логіку побудови фраз, що може бути більш інформативним з точки зору характеристики стилю автора для будь-якого тексту.

Представлена робота є теоретичним підґрунтям для вирішення проблем встановлення авторства тексту та ідентифікації запозичень. Також були проведені експериментальні дослідження. Експериментально виявлено статистичну схожість розв'язків задач встановлення авторства та ідентифікації запозичень, що буде представлено в наступній статті авторів.

Запропонований підхід дозволяє виділити семантичні особливості побудови фраз автором, що є характеристикою його мовлення. Робота з реченням, як із одиницею тексту для аналізу його побудови, дозволить більш точно визначити авторський стиль у частині використання слів, їх послідовностей і характерних мовних конструкцій. Дозволяє не прив'язуватись до конкретних частин мови, а виявляє загальну логіку побудови фраз.

Створену модель планується використовувати в подальшому для визначення авторства природномовних текстів різного спрямування: художньої та технічної літератури.

Ключові слова: природньомовні тексти, конструктивно-продукційне моделювання, встановлення авторства, формальні граматики, стохастические грамматики, моделі текстів

Introduction

The work develops an approach to the construction formalization proposed by the author by the constructive modeling means. This approach allows you to highlight the semantic features of the author's phrases construction,

which is a characteristic of his speech. The use of the developed model is assumed in the field of determining the texts authorship and identifying borrowed ones.

This approach to text analysis is promising, due to the presence of each person's own style and approach to constructing phrases. People communication by means of natural language texts is carried out with the use of not individual words, but expressed, sentences. A sentence as a sequence of individual words meanings is a new unit with a set of semantic values inherent only to it, among which there are also those that are not considered direct derivatives of the existing sentence composition, which is due to the peculiarity of its construction.

Modeling speech communication, in its entirety transmitted with the help of information language, is impossible without studying the features of the sentence structure. This aspect study of syntactic structures is important, in addition to purely linguistic tasks, for understanding the features and regularities of a person's mental activity.

Related works

The main text studies and methods of its formalization mostly work with words or their short sequences, and only a small part of the studies are based on the sentence as a unit of the text.

Widespread methods that investigate the test based on symbols or their sequence of some length [7] or separately words [8] and their sequences [6], work with lemmas [9] or parts of speech [10]. However, such an approach does not always adequately reflect the peculiarities of the text under study and the author's style.

The approach to the sentence as a text unit opens up new opportunities for work, since the peculiarities of its construction and the analysis of the words used within one sentence can serve as an additional source of information about the author [13]. Similar approaches were used for context research [11], working with the author's mood research [12], in terms of a deeper understanding of the text [14]. Many studies also testify to the importance of considering the sentence as a single unit, rather than a collection of individual words [15], the importance of their order [16] and the general context [17].

Currently, there are many approaches to building models proposed [18, 19]. The most popular methods are the use of trees [20] and the construction of various neural networks [23, 24]. However, the most universal tool has not yet been found [22].

Such approaches are most widespread in the field of work with artificial intelligence in terms of text recognition and understanding, which largely confirms the necessity and relevance of using a separate sentence as a unit of text structure.

The processes of texts authorship identifying using the constructive-production modeling

Generalized designer

Development of a constructive-synthesizing approach to solving the problem of technical text authorship establishing. The generalized constructor is a triple called C_G [2]

$$C_G = \langle M, \Sigma, \wedge \rangle,$$

where M – the non-homogeneous carrier of the structure, which expands during the construction process;Σ – is the operations and relations signature, that consisting of binding, substitution and derivation operations, operations on attributes, and a substitution relation; $\wedge$ – construction information support (CIS).

According to $\wedge$ [2] the $_{w}$l form with attribute w is called a set of terminals and non-terminals that are united by binding operations. The developed constructors use a single binding operation (and relationship) - concatenation. A form that contains only terminals is called a construction. Constructions are formed by derivation from the initial non-terminal, substitution operations and operations on attributes, and generalized partial and full derivation operations.

The operation of partial derivation ($| \Rightarrow \in \Sigma_P$) consists of choosing a suitable substitution rule from their set, performing this substitution and performing operations on the attributes corresponding to the selected rule in a certain sequence.

The operation of complete derivation (or simply derivation, $|| \Rightarrow \in \Sigma_P$) consists of the sequential execution of the operation of partial derivation, starting from the initial nonterminal and ending with the construction.

To form structures, it is necessary to perform several clarifying transformations of constructors:

– specialization – defines the subject area: the semantic nature of the medium, the finite set of operations and their semantics, the operations attributes, the order of their execution and restrictions on substitution rules;

– interpretation – consists in connecting the operations of the signature with the execution algorithms of some algorithmic designer;

– concretization – expansion of axiomatics by production rules set, specific sets assignment of non-terminal and terminal symbols with their attributes and, if necessary, attribute values;

– implementation – formation of a structure from the constructor carrier elements by performing algorithms related to signature operations.

Constructor-converter of natural language text into a tagged text

The purpose of construction is to convert technical text into tagged text. For each word in a sentence, its attribute is determined in the composition: part of speech (pos), number (num) and gender (gen). Consider the specialization of the designer:

$$C = \langle M, \Sigma, \Lambda \rangle \ _S \mapsto C_P = \langle M_P, \Sigma_P, \Lambda_P \rangle,$$

where M_P – a carrier that includes terminal and non-terminal alphabets, initial and tagged texts, as well as a set of production rules Ψ, separate rules ψ_i:<s_i, g_i>, where i – rule number, s_i – is a sequence of substitution relations, g_i – is a sequence of operations on attributes, Σ_P – operations and connections for elements M_P; CIS $\Lambda_P \supset \Lambda$.

The following provisions are included in CIS Λ_P.

Signature Σ_P contains the signature of specific binding operations and attribute operations.

The terminals T include symbols and words of the Ukrainian language, denoted as * - letters that can be used to form words, $\underline{*}$ – space, $\bar{*}$ – end-of-sentence symbols, ⊥ – end-of-text symbol, $W_{i,j}$, - the j-th word in the i-th sentence. The first word of each sentence will additionally store information about its length l, and the first word of the text will store the number of words in the longest sentence, max, and the total number of sentences in the text, S. Non-terminals $N = \{\sigma, \eta, \varepsilon\}$ – auxiliary elements, where ε is the symbol 'empty'.

The following attribute operations are presented.

The operation $\odot\ (word,\ ends, pos \downarrow word) \in \Sigma_P$ – defines the part of speech pos for the word 'word', which can take one of the following values: verb (v), noun (n), numeral (nume), pronoun (pron), adjective (adj), conjunction (conj), adverb (adv), preposition (prep), verb adverb (v_adv), interjection and particle (frac), verb adjective (v_adj).

The operation $\circledast$ ($word, ends, num \downarrow word$) $\in \Sigma_P$ – determines the number num for the word 'word', which can be singular (sing) or plural (plur).

The operation $\circledcirc(word, ends, gen \downarrow word) \in \Sigma_P$ – determines the gender (gen) for the word 'word', which can take one of the following values: feminine (f), masculine (m), and neuter (n).

Each of these operations compares the word with all elements of 'ends' - corresponding lists of endings [5]. If there is a match with a specific ending, a result is formed, and the parameters pos, num, and gen are assigned corresponding values.

The operation = $(a, b) \in \Sigma_P$ – assigns the value b to the variable a.

The operation $+(c, a, b) \in \Sigma_P$ – adds $c = a + b$.

The Comparison operation <>*(a,b,c,d)* $\in \Sigma_P$ – compares a with b. If a is greater, then the nested operation c is executed; if it is smaller, then d is executed.

Interpretation of the designer

Let's form a constructive system from the SR designer, as an elemental design base, and the SA algorithmic designer, as a model-executor of the design.

$$\langle \mathrm{C_P} = \langle M_P, \Sigma_P, \Lambda_P \rangle, \mathrm{C_A} = \langle M_A, V_A, \Sigma_A, \Lambda_A \rangle \rangle \ _\mathrm{I}\mapsto \langle \mathrm{C_{PAI}} = \mathrm{M_I}, \Sigma_\mathrm{I}, \Lambda_\mathrm{I} \rangle,$$

where $V_A = \left\{ A_i \ |_{X_i}^{Y_i} \right\}$ – set of forming algorithms in the basic algorithmic structure, $\mathrm{X_i}$ and $\mathrm{Y_i}$ –possible input and output data of the algorithm $A_i \ |_{X_i}^{Y_i}$, $M_A \supset \bigcup_{A_i \in V_A} (X(A_i) \cup Y(A_i))$ – carrier of the algorithmic structure, Σ_A – a set of algorithm binding operations, Λ_A – CIS, Ω($\mathrm{C_A}$) – set of algorithms constructed in $\mathrm{C_A}$[2], $\mathrm{M_I} = M_P \cup M_A, \Sigma_\mathrm{I} = \Sigma_P \cup \Sigma_A,\quad \Lambda_\mathrm{I} = \Lambda_P \cup \Lambda_A \cup \left\{ \left(A_0|_{A_i,A_j}^{A_i \cdot A_j} \downarrow " \cdot "\right); \left(A_1|_{l,s_i}^{l} \downarrow " \Rightarrow "\right); \left(A_2|_{l,\Psi}^{l} \downarrow "| \Rightarrow "\right); \left(A_3|_{\sigma}^{\Omega} \downarrow "|| \Rightarrow "\right) \left(A_4|_{word,ends}^{pos \downarrow word} \downarrow " \odot "\right); \left(A_5|_{word,ends}^{num \downarrow word} \downarrow " \circledast "\right); \left(A_6|_{word,ends}^{gen \downarrow word} \downarrow " \circledcirc "\right); \left(A_7|_{b}^{a} \downarrow " = "\right); \left(A_8|_{a,b}^{c} \downarrow " + "\right); \left(A_9|_{a,b}^{c/d} \downarrow " <> "\right) \right\}$

Structure $\mathrm{C_{PAI}}$ contains algorithms for performing operations:

- $A_0|_{A_i,A_j}^{A_i \cdot A_j}$ – is an algorithms composition, $A_i \cdot A_j$ – sequential execution of algorithm A_j after A_i;
- $A_1|_{l,s_i}^{l}$ – is a substitution, where l – current form, s_i – s the rule to be executed;
- $A_2|_{l,\Psi}^{l}$ – is a partial output, where Ψ – is the set of production rules to be executed;
- $A_3|_{\sigma,\Psi}^{\Omega}$ – is a complete derivation, where σ – is an axiom, Ψ – is a set of production rules, Ω – is a set of formed constructions;
- $A_4|_{word,ends}^{pos \downarrow word}$ – is a definition for the *word* of its language part *pos;*
- $A_5|_{word,ends}^{num \downarrow word}$ – is a definition for the *word* of its number *num;*
- $A_6|_{word,ends}^{gen \downarrow word}$ – is a definition for the *word* of its gender *gen;*
- $A_7|_{b}^{a}$ – is an assignment of the value *b* to variable *a;*
- $A_8|_{a,b}^{c}$ – is adding $c = a + b$;
- $A_9|_{a,b}^{c/d}$ – execution of action *c* or *d* based on the result of comparing *a* and *b.*

When **specifying** C_{PAI} the following is parameterized:

$$\mathrm{C}_{PAI} \ \mapsto_\mathrm{K} \mathrm{C}_{PAIK}(\mathrm{TT}) = \langle \mathrm{M_K}, \Sigma_\mathrm{K}, \Lambda_\mathrm{K} \rangle,$$

where $\Lambda_K \supset \Lambda_I \cup \{M_K = T_T \cup N\} \cup \Lambda_1$. TT is a technical text submitted for analysis.

In the substitution rules ψ_i:<s_i, g_i> the sequence of substitution relations s_i consists of the relation $s_{i,1}$ – analysis of the TT, $s_{i,2}$ – formation of a words' set $W_{i,j}$ with their attributes. Operations $g_{i,j}$ are performed after execution of $s_{i,1}$ and before $s_{i,2}$.

Initial construction conditions: σ – a non-terminal from which the derivation begins and the initial values max = 1, i = 1 and j = 1.

Construction completion condition: all incoming text is tagged.

In the first rule, the parsing of the text and the formation of the first element in the tagged text $W_{i,j}$ begins

$$s_1 = \langle \sigma \to \eta, \quad W_{i,j} \to \varepsilon \rangle.$$

Parsing occurs from one character to the next with its rewriting in $W_{i,j}$ for further tagging

$$s_2 = \langle \eta \to * \eta, \quad W_{i,j} \to * W_{i,j} \rangle.$$

When a space or end-of-sentence mark is reached, a tagging determination is made for the word and the next word is passed

$$s_3 = \langle \eta \to \underline{*} \eta, \ \varepsilon \rangle.$$

The operations $\odot$, $\circledast$, and $\circledcirc$ in attribute operations determine the part of speech, number, and gender of a word, respectively. The transition to the next word in the sentence occurs. The flag "done" for each word is set to position 0, and it will be used later for rule formation

$$g_3 = \langle \odot (W_{i,j}, pos \downarrow W_{i,j}), \ \circledast (W_{i,j}, num \downarrow W_{i,j}), \ \circledcirc (W_{i,j}, gen \downarrow W_{i,j}), = (done \downarrow W_{i,j}, 0), +(j, 1, j) \rangle.$$

The rule s_4 is applied when the end of the sentence is reached, and like the previous rule, the tagging for the word is determined and the next sentence is passed. A transition to the next word is performed. The length of each sentence is calculated and set and stored as an attribute of its first word. The maximum sentence length is determined as an attribute of the very first word in the text. Along with this, to mark the end of the sentence, $\perp(W_{i,j} = \perp)$ will be written to its final position. This is necessary for the correct operation of the following constructor

$$s_4 = \langle \eta \to \bar{*} \sigma, \ \rangle,$$

$$g_4 = \langle \odot (W_{i,j}, pos \downarrow W_{i,j}), \ \circledast (W_{i,j}, num \downarrow W_{i,j}), \ \circledcirc (W_{i,j}, gen \downarrow W_{i,j}), \ = (l \downarrow W_{i,1}, j), <>$$
$$(j, max \downarrow W_{1,1}, = (max \downarrow W_{1,1}, j), \varepsilon), = (done \downarrow W_{i,j}, 0), \ +(j, 1, j), = (W_{i,j}, \perp), = (j, 1), \ +(i, 1, i) \ \rangle.$$

The last rule is used when the end of the text is reached and is final. The *am* attribute of the first word stores the total number of sentences in the text

$$s_5 = \langle \eta \to \perp, \ \varepsilon \rangle,$$
$$g_5 = \langle = (am \downarrow W_{1,1}, i) \rangle.$$

Realization

The constructor implementation is the language constructions formation from its carrier elements through the algorithm's execution related to signature operations according to the rules of substitution:

$$\mathrm{C_{PAIK}} \ _R\mapsto \bar{\Omega}(\mathrm{C_{PAIK}}(\mathrm{TT})),$$

where $\bar{\Omega}(\mathrm{C_{PAIK}}(\mathrm{TT})) = \Omega(\mathrm{C_{PAIK}}(\mathrm{TT}))$. $\bar{\Omega}$ – all possible outcomes of the constructor, however, since the generated constructor is based on a specific text, the resulting processed text Ω will be the only possible outcome. As a result of the constructor implementation, the processed text with tagged words as $\Omega(\mathrm{C_{PAIK}}(TT))$ was received.

For example, let's take the sentence «Чорні ґрати розпанахали небо. Червоно-рожеве воно тягнуло, манило». The result of the designer's work will look like this:

$$W_{1,1} = {}_{adj,plur,-}\text{Чорні}; \ W_{1,2} =_{n,plur,-} \text{ґрати}; \ W_{1,3} =_{v,plur,-} \text{розпанахал}; \ W_{1,4} =_{n,sing,n} \text{небо};$$
$$W_{2,1} =$$
$${}_{adj,sing,n}\text{Червоно} - \text{рожеве}; \ W_{2,2} =_{pr,sing,n} \text{воно}; \ W_{2,3} =_{v,sing,n} \text{тягнуло}; \ W_{2,4} =_{v,sing,n} \text{манило}$$

Tagged text constructor-converter into formal substitution rules set with a probability measure

The purpose of construction is to build a stochastic constructor rule that formalizes the syntactic component of the technical text.

The initial construction condition is the implementation of the $\mathrm{C_P}$ constructor – the tagged text Tg obtained as a result of the constructor implementation $\mathrm{C_{PAIK}} - \Omega(\mathrm{C_P}(\mathrm{TT}))$.

Construction completion condition: each sentence of the tagged text is converted into a corresponding set of rules $\Omega(\mathrm{C_T}(R))$, which happens under the condition $\tau 5 = true$, which is set when the last word of the longest sentence in the text is reached. This will serve as an indication that all other words in the text have already been processed and the rule building process is complete.

The designer has the following specialization:

$$C = \langle M, \Sigma, \Lambda \rangle \ _S \mapsto C_T(Tg) = \langle M_T, \Sigma_T, \Lambda_T \rangle,$$

where M_T – is a carrier that includes tagged text Tg, Σ_T – operations and relationships on elements M_T and axiomatics Λ_T.

The operation $*(r, a, b)$ – checking that the attributes pos↲a, num↲a, gen↲a of element match the attributes pos↲b, num↲b, gen↲b of element b, where a and b are tagged words. If there is a complete match, the result is 1, otherwise - 0.

The operation $\&(y, x1, x2)$ – is a logical and with an unlimited number of operands y = x1 and x2 and …;

A loop operation $\circ(a, c)$ – where a – is a condition, c – is an operation performed while the condition is valid;

The operation $-(c, a, b)$ – is equal to c = a - b in infixed form;

The operation : (c, a, b) – is equal to c = a : b in infixed form, division of real numbers;

The operation ≤(r,a,b) – is a comparing a ≤ b with saving the result in r.

Interpreting the C_T constructor using the same algorithmic constructor C_A:

$$\langle C_T, C_A \rangle \ {}_{\mathrm{I}}\mapsto \langle \mathrm{C}_T = \mathrm{M}_{\mathrm{TI}}, \Sigma_{\mathrm{TI}}, \Lambda_{\mathrm{TI}} \rangle,$$

M_{TI} – algorithmic structure for the formation of a stochastic constructor from tagged text, Σ_{TI} – algorithm linking operations, $\Lambda_{\mathrm{TI}} \supset \Lambda_{\mathrm{I}} \cup \Lambda_1 \cup \Lambda_2$.

$\Lambda_2 = \{ (A_{10}|_{a,b}^{r} \text{↲ " * "}); (A_{11}|_{a,b}^{r} \text{↲ "\&"}); (A_{12}|_{a,b}^{c} \text{↲ " − "}); (A_{13}|_{a,b}^{c} \text{↲ ":"}); (A_{14}|_{a,b}^{r} \text{↲ " ≤ "}). \}$.

The $\mathrm{C}_{TA\mathrm{I}}$ structure includes the following algorithms:

- $A_0, A_1, A_2, A_3, A_7, A_8, A_9$ – are similar algorithms of the $\mathrm{C}_{\mathrm{PAI}}$ constructor;
- $A_{10}|_{a,b}^{r}$ – is a comparison of a and b for their identity;
- $A_{11}|_{a,b}^{r}$ – is logical "and";
- $A_{12}|_{a,b}^{c}$ – is a number subtraction;
- $A_{13}|_{a,b}^{c}$ – is a real numbers division;
- $A_{14}|_{a,b}^{r}$ – is a comparison of a and b.

Specification of C_T:

$$C_T \mapsto_{\mathrm{K}} C_T(\mathrm{C}_P(Tg)) = \langle \mathrm{M}_{\mathrm{K}}, \Sigma_{\mathrm{K}}, \Lambda_{\mathrm{K}} \rangle,$$

where $\Lambda_{\mathrm{K}} \supset \Lambda_{\mathrm{I}}$, $\Lambda_K \supset \{M_K = T \cup N\}$, the terminals T include all words $W_{m,j}$ with the designation of their place j in the sentence m, $\alpha_{k,j}$, which is a non-terminal of the rule being constructed, σ –is the initial non-terminal and the constructed rule ω_k, which in its attributes will have the left part of the rule L, the right part R and the probability of its operation for the given text prob. The non-terminals N: τ_{i} – is the rule availability attribute.

For each part of speech, its appearance probability (prob) in a certain place of a certain sentence in this text is calculated. The appearance probability of a certain language part in the investigated sequence will allow for a more accurately capturing of the individual author's writing style characteristic.

The probability of obtaining the entire sentence is defined as its speech parts sequences probabilities product. The resulting constructor will generate a language characteristic of the processed text and structurally similar texts of a certain author.

Thus, each sentence of the presented text will be presented in the form of a chain of rules that will reflect the sequence of used parts of speech and the probability of their appearance in the presented sequence.

Initial conditions: the initial form $W_{1,1}$ – is the first word in the text, where i = 1, n = 2 sentence numbers, j = 1, m = 2 word numbers in them. t = 0 – the number of matches with the selected pair of parameters in a sequence of two words, k = 1 – the number of the rule being built. $\tau 1 = true, \tau 2 = false \ldots \tau 5 = false$ – are conditions for the execution of the rules: if true, it is available for use, if false - not. idone = 1, jdone = 1 are variables equal to the number of the unique element in the layer and the previous layer, respectively, and u = false – a flag for marking already built rules.

Parsing begins with processing the first layer (the first words in the sentences of the text) and searching for a match by attributes among them

$s_1 = \langle W_{i,j} \ {}_{\tau 1}\!\rightarrow W_{n,j}, \varepsilon \rangle,$

$$g_{1,1} = \langle == (0, \text{done ↲ } \mathrm{W_{i,j}}, \mathrm{x1}), * (\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}) \rangle.$$

Searches for matching words with the same attributes in the current layer. If a match with the current word of this current layer is found, and no match was found for this word before, we increase the total number of similar to the searched word (t) and move to the next sentence by increasing n

$$g_{1,2} = \langle \&(\mathrm{y}, \mathrm{x1}, \mathrm{x2}), <> (y, 0, +(t, 1, t), \varepsilon), +(n, 1, n) \rangle.$$

The second rule is used when the end of the sentence is reached, in this case no calculations take place, only the sentence number n is increased to move to the next word in the layer

$$s_2 = \langle W_{i,j} \ {}_{\tau 1}\!\rightarrow W_{n,j} \perp, \varepsilon \rangle,$$
$$g_2 = \langle +(n, 1, n) \rangle.$$

The third rule is applied when it is impossible to reach the next word in the layer due to reaching its end

$$s_3 = \langle W_{i,j} \ {}_{\tau 1}\!\rightarrow \perp, \varepsilon \rangle.$$

Under this condition, the calculation of the probability of the selected sequence appearing in the text is used, the rules $s_1 - s_3$ become unreachable, and the rules $s_4 - s_7$ become available for processing

$$g_3 = \langle -(n,n,1), : (prob,t,n), = (\tau 1, false), = (\tau 2, true), = (n,1)\rangle.$$

The following rules are responsible for forming rules for the first words of each of the sentences when all corresponding words are repeated

$$s_4 = \langle W_{i,j\ \tau 2}\rightarrow W_{n,j},\ \omega_k\rangle,$$

$$g_{4,1} = \langle == (0, \text{done} \downarrow \text{W}_{\text{i,j}}, \text{x1}), * (\text{W}_{\text{i,j}}, \text{W}_{\text{n,j}}, \text{x2})\rangle.$$

If the attributes of the words match, a new rule ω_k is built. σ and $W_{i,j}\alpha_{i,j}$ are written in its left and right parts, respectively, where $\alpha_{i,j}$ – is the non-terminal of the newly formed rules, and the probability of its activation for the text $prob$ is written

$$g_{4,2} = \langle \&(\text{y}, \text{x1}, \text{x2},), \quad <> (y, 0, \cdot \left(= \left(L \downarrow \omega_{i,j}, \sigma\right), = \left(R \downarrow \omega_{i,j}, W_{i,j}\alpha_{i,j}\right), = \left(prob \downarrow \omega_{i,j}, prob\right)\right), \varepsilon)\rangle.$$

After the formation of the rule, the flag of the presence of at least one rule on this layer is set $u = true$, the rule receives the uniqueness index in the *idone* layer, and the constructor moves to the next sentence

$$g_{4,4} = \langle +(n,1,n), = \left(done \downarrow \text{W}_{\text{i,j}}, idone\right), = (u, true)\rangle.$$

The next rule is similar to rule s_2, does not make calculations and is responsible for increasing the sentence number n to move further through the layer

$$s_5 = \langle W_{i,j\ \tau 2}\rightarrow W_{n,j}\ \perp, \varepsilon\rangle,$$

$$g_5 = \langle +(n,1,n)\rangle.$$

If there is only one word in the sentence, rule s_6 is used

$$s_6 = \langle W_{i,j}\ \perp_{\tau 2}\rightarrow W_{n,j},\ \omega_k\ \rangle.$$

To form a rule in this case, the following checks will be carried out: whether the word is included in another done rule, whether the corresponding attributes of the words match

$$g_{6,1} = \langle == (0, \text{done} \downarrow \text{W}_{\text{i,j}}, \text{x1}), * (\text{W}_{\text{i,j}}, \text{W}_{\text{n,j}}, \text{x2})\rangle.$$

If the sentence is not the first, we form the corresponding rule ω_k. In its left part, write σ, respectively, and only $W_{i,m}$ in the right part

$$g_{6,3} = \langle \&(\text{z}, \text{x1}, \text{x2}, \text{x3}),\ <> \left(z, 0, \cdot\ (= \left(L \downarrow \omega_{i,j}, \sigma\right), = \left(R \downarrow \omega_{i,j}, W_{i,j}\right), = \left(prob \downarrow \omega_{i,j}, prob\right)), \varepsilon\ \right)\rangle.$$

And then the same as in $g_{4,4}$ – its creation flag $u = true$, is set, the rule receives the uniqueness index in the idone layer and the executor moves to the next sentence

$$g_{6,4} = \langle +(n,1,n), = (done \downarrow \text{W}_{\text{i,j}}, idone), = (u, true)\rangle.$$

And when the end is reached, rule s_7 is triggered

$$s_7 = \langle W_{i,j\ \tau 2}\rightarrow \perp, \varepsilon\rangle.$$

Reaching the end of a layer means the end of rule formation and transition to the formation of another. For this, the flags $\tau 2 = false$ and $\tau 1 = true$ are changed, which will close the rules $s_4 - s_7$ and open the rules $s_1 - s_3$ to search for other matches and count them. To reflect the operation of another rule in the layer, the rule's uniqueness number for the *idone* layer is increased

$$g_{7,1} = \langle == (u, true, y), <> (y, 0, \cdot \left(= (\tau 2, false), = (\tau 1, true), +(idone, 1, idone), = (t, 0)\right), \varepsilon\rangle.$$

If the work with the layer is completed and rules have been formed for all the words in it, the constructor moves to the next layer, starting again from the first sentence i=1 to search for a match. The calculation of the uniqueness of the rules in the layer also starts from the beginning of idone=1. If the final layer is reached (the last word in the longest sentence $\text{W}_{i,max}$ is processed), the work of the performer with the first layer will be completed $\tau 3 = true$, $\tau 1 = false, \tau 2 = false$

$$g_{7,2} = \langle == (u, false, y), <> (y, 0, \cdot\ (= (i,1), = (idone, 1), = (t,0), = (\tau 3, true), = (\tau 2, false), = (\tau 1, false)), \varepsilon\rangle.$$

To continue forming rules from tagged text, operating the consecutive pairs of words in each sentence. The transition from word to word does not occur along the sentence, but according to the number of words in them. In this way, the constructor considers a pair of consecutive words in a sentence

$$s_8 = \langle W_{i,j}W_{i,m\ \tau 3}\rightarrow W_{n,j}W_{n,m}, \varepsilon\rangle.$$

To consider an existing pair of words as similar, you need to check the following parameters: the words have not yet been processed; the attributes of the selected sequence of two words (part of speech, gender and number) match the numbered words in the next sentence, the previous string of words must also match, which is checked by jdone

$$g_{8,1} = \langle == (0, \text{done} \downarrow \text{W}_{\text{i,j}}, \text{x1}), * (\text{W}_{\text{i,j}}, \text{W}_{\text{n,j}}, \text{x2}), * (\text{W}_{\text{i,m}}, \text{W}_{\text{n,m}}, \text{x3}), -(\text{k}, \text{j}, 1), <$$
$$> (\text{j}, 1, == (\ \text{done} \downarrow \text{W}_{\text{i,k,}}\text{jdone}, \text{x4}\ ), \text{x4} = \text{true})\rangle.$$

If the attributes match, the pair is counted in the total number of similar sequences and the value of t and the value of n are increased to move to the next sentence in the layer

$$g_{8,2} = \langle \&(\text{y}, \text{x1}, \text{x2}, \text{x3}, \text{x4}), <> (y, 0, +(t, 1, t), \varepsilon\ ), +(n, 1, n)\rangle.$$

The following rule is triggered when the end of the sentence is reached, the sentence number n is incremented for further viewing of the words in the layer

$$s_9 = \langle W_{i,j}W_{i,m\ \tau 3}\rightarrow W_{n,j}\ \perp, \varepsilon\rangle,$$
$$g_9 = \langle +(n,1,n)\rangle.$$

The following rule is executed when it is impossible to move further along the sentences due to reaching the end of the layer. Under this condition, the probability of the appearance of the selected sequence in the text is calculated, rules $s_8 - s_{10}$ become unreachable, and rules $s_{11} - s_{14}$ become available for processing

$$s_{10} = \langle W_{i,j\ \tau 3}\rightarrow \perp, \varepsilon\rangle,$$
$$g_{10} = \langle -(n,n,1), : (prob,t,n), = (\tau 3, false), = (\tau 4, true), = (i,1), = (n,1)\rangle.$$

The next step is to revisit the current layer and create rules ω_k

$$s_{11} = \langle W_{i,j}W_{i,m\ \tau 4}\rightarrow W_{n,j}W_{n,m},\ \omega_k\rangle.$$

To form the appropriate rule, the check from the first rule is repeated and we additionally check whether the word is the first in the sentence (x4) for the correct formation of the initial rules

$$g_{11,1} = \langle == (0, \text{done} \downarrow \mathrm{W_{i,j}}, \mathrm{x1}), * (\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}), * (\mathrm{W_{i,m}}, \mathrm{W_{n,m}}, \mathrm{x3}), -(\mathrm{k,j},1), <> (\mathrm{j},1, == (\ \text{done} \downarrow \mathrm{W_{i,k}}, \mathrm{jdone}, \mathrm{x4}\ ), \mathrm{x4} = \mathrm{true})\rangle.$$

If everything matches and the word is not the first in the sentence, a new rule ω_k is built. $\alpha_{i,j}$ is written in the left part of the rule, $W_{i,m}\alpha_{i,m}$is written in its right part, where $\alpha_{i,j}$ – is the non-terminal of the newly formed rules, and the probability of its activation for the text $prob$ is written.

$$g_{11,2} = \langle \&(\mathrm{y}, \mathrm{x1}, \mathrm{x2}, \mathrm{x3}, \mathrm{x4}),$$
$$<> (y, 0, \cdot (= (L \downarrow \omega_{i,j}, \alpha_{i,j}), = (R \downarrow \omega_{i,j}, W_{i,m}\alpha_{i,m}), = (prob \downarrow \omega_{i,j}, prob)), \varepsilon)\rangle.$$

After creating a rule, its creation flag $u = true$, is set, the rule receives a unique index in the idone layer, and the constructor moves to the next sentence

$$g_{11,3} = \langle +(n,1,n), = (done \downarrow \mathrm{W_{i,j}}, idone), = (u, true)\rangle.$$

The next rule is similar to rule s_9, does not make calculations and is responsible for increasing the sentence number n to move further through the layer

$$s_{12} = \langle W_{i,j}W_{i,m\ \tau 4}\rightarrow W_{n,j}\ \perp, \varepsilon\rangle,$$
$$g_{12} = \langle +(n,1,n)\rangle.$$

If the last word in the sentence is in the layer, rule s_{13} is triggered

$$s_{13} = \langle W_{i,j}\ \perp_{\tau 4}\rightarrow W_{n,j},\ \omega_k\ \rangle.$$

To form a rule in this case, the following checks will be carried out: whether the word is included in another done rule, whether the corresponding attributes of the words match, whether the word is the first in the sentence and, if not, whether the previous chain matches

$$g_{13,1} = \langle == (0, \text{done} \downarrow \mathrm{W_{i,j}}, \mathrm{x1}), * (\mathrm{W_{i,j}}, \mathrm{W_{n,j}}, \mathrm{x2}), <> (\mathrm{j},1, == (\ \text{done} \downarrow \mathrm{W_{i,j-1}}, \mathrm{jdone}, \mathrm{x3}\ ), \mathrm{x3} = \mathrm{true})\rangle.$$

If the sentence is not the first, we form the corresponding rule ω_k. $\alpha_{i,j}$ is written in the left part of the rule, $W_{i,m}\alpha_{i,m}$is written in its right part, where $\alpha_{i,j}$ – is the non-terminal of the newly formed rules, and the probability of its activation for the text $prob$ is written.

$$g_{13,2} = \langle \&(\mathrm{y}, \mathrm{x1}, \mathrm{x2}, \mathrm{x3}),\ <> (y, 0, \cdot (= (L \downarrow \omega_{i,j}, \alpha_{i,j}), = (R \downarrow \omega_{i,j}, W_{i,j}), = (prob \downarrow \omega_{i,j}, prob)), \varepsilon)\rangle.$$

Next, just like in $g_{11,3}$ – its creation flag *u = true* is set, the rule receives the uniqueness index in the idone layer and the constructor moves to the next sentence

$$g_{13,3} = \langle +(n,1,n), = (done \downarrow \mathrm{W_{i,j}}, idone), = (u, true)\rangle.$$

When the end is reached, rule s_{14} is triggered

$$s_{14} = \langle W_{i,j\ \tau 4}\rightarrow \perp, \varepsilon\rangle.$$

Reaching the end of a layer means the end of rule formation and transition to the formation of another. For this, the flags $\tau 4 = false$ and $\tau 3 = true$ are changed, which will close the rules $s_{11} - s_{14}$ and open the rules $s_8 - s_{10}$ to search for other matches and count them. To reflect the operation of another rule in the layer, the rule's uniqueness number for the *idone* layer is increased

$$g_{14,1} = \langle = (\tau 4, false), = (\tau 3, true), +(idone, 1, idone), = (t, 0), \leq (r, j, l \downarrow \mathrm{W_{i,j}})\rangle.$$

The procedure of counting coincidences and calculating the probability of their occurrence for building rules on its basis continues until all words in the layer have been processed. To work with all chains, at each layer pass, the *jdone* uniqueness index is increased to check the calculation condition

$$g_{14,2} = \langle <> (\ am \downarrow \mathrm{W_{1,1}}, i, \circ (r, \cdot (= (i,1,i), == (u, true, 1) <> (1,0, +(jdone, 1, jdone))))\rangle.$$

If the work with the layer is completed and rules have been formed for all the words in it, the constructor moves to the next layer by increasing j and starting again from the first sentence *i = 1* to look for a match. The calculation of the uniqueness of rules in a layer also starts from the beginning of *idone = 1* and *jdone = 1*. On the condition that the final layer is reached (the last word in the longest sentence $\mathrm{W}_{i,max}$ is processed), the work of the constructor will be completed $\tau 5 = true$

$g_{14,3} = \langle <> (max \downarrow \mathrm{W}_{1,1}, j, \cdot (+(j,1,j), = (i,1), = (idone,1), = (jdone,1)\ ), \cdot (= (\tau 5, true), = (\tau 3, true)))\rangle$.

As a result of the work of the constructor-converter with $\Omega(\mathrm{C_{PAIK}}(TT))$, we get a set of rules that reflects the style of the author's language in the corresponding text $\Omega(\mathrm{C_T}(R))$.

Realization

The implementation of the structure is the formation of language constructions from the elements of its carrier through the execution of algorithms related to signature operations according to the rules of axionomics:

$$\mathrm{C_{P\,K}}_{\ R} \mapsto \bar{\Omega}(C_{P\,K}),$$

where $\bar{\Omega}(C_{P\,K}) \subset \Omega(\mathrm{C_{P\,K}})$.

For example, let's take sentences that have the form:

«Ми були дуже схожі.
Я любила читати книжки.
А ти захоплювався виставами.
Але..
Між нами було й багато різниці».

The tagged text for this example:

$W_{1,1} = {}_{pron,plur}$Ми $W_{1,2} =_{v,plur}$ були $W_{1,3} =_{adv,sing}$ дуже $W_{1,4} =_{adj,plur}$ схожі

$W_{2,1} = {}_{pron,sing}$Я $W_{2,2} =_{v,sing}$ любила $W_{2,3} =_{v,sing}$ читати $W_{2,4} =_{n,plur}$ книжки

$W_{3,1} = {}_{conj}$А $W_{3,2} =_{pron,sing}$ ти $W_{3,3} =_{v,sing}$ захоплювався $W_{3,4} =_{n,plur}$ виставами

$W_{4,1} = {}_{pron}$Але

$W_{5,1} =$

${}_{conj}$Між $W_{5,2} =_{pron,plur}$ нами $W_{5,3} =_{v,sing}$ було $W_{5,4} =_{conj}$ й $W_{5,5} =_{adj,sing}$ багато $W_{5,6} =_{adj,plur}$ різного.

The result of the designer's work will be presented in the form of relevant rules:

$\sigma \overset{0.2}{\rightarrow} W_{1,1}\alpha_{1,1};\ \alpha_{1,1} \overset{0.2}{\rightarrow} W_{1,2}\alpha_{1,2};\ \alpha_{1,2} \overset{0.2}{\rightarrow} W_{1,3}\alpha_{1,3};\ \alpha_{1,3} \overset{0.2}{\rightarrow} W_{1,4};$

$\sigma \overset{0.2}{\rightarrow} W_{2,1}\alpha_{1,1};\ \alpha_{2,1} \overset{0.2}{\rightarrow} W_{2,2}\alpha_{2,2};\ \alpha_{2,2} \overset{0.6}{\rightarrow} W_{2,3}\alpha_{2,3};\ \alpha_{2,3} \overset{0.4}{\rightarrow} W_{2,4};$

$\sigma \overset{0.4}{\rightarrow} W_{3,1}\alpha_{3,1};\ \alpha_{3,1} \overset{0.2}{\rightarrow} W_{3,2}\alpha_{3,2};\ \alpha_{3,2} \overset{0.6}{\rightarrow} W_{3,3}\alpha_{3,3};\ \alpha_{3,3} \overset{0.4}{\rightarrow} W_{3,4};$

$\sigma \overset{0.2}{\rightarrow} W_{4,1};$

$\sigma \overset{0.4}{\rightarrow} W_{5,1}\alpha_{5,1};\ \alpha_{5,1} \overset{0.2}{\rightarrow} W_{5,2}\alpha_{5,2};\ \alpha_{5,2} \overset{0.6}{\rightarrow} W_{5,3}\alpha_{5,3};\ \alpha_{5,3} \overset{0.2}{\rightarrow} W_{5,4}\alpha_{5,4};\ \alpha_{5,4} \overset{1}{\rightarrow} W_{5,5}\alpha_{5,5};\ \alpha_{5,5} \overset{1}{\rightarrow} W_{5,6}$.

Constructor-measurer of the similarity degree

In order to establish the similarity degree of the two texts according to the syntactic style of the author's language, a comparison of the text models is carried out with the help of a constructor-meter.

The purpose of construction is to establish the degree of similarity of texts by comparing stochastic constructors built according to their syntactic structure.

The initial conditions for constructing a model of two texts in the form of a set of substitution rules with the probability of its activation $\Omega(\mathrm{C_T}(R_1))$ and $\Omega(\mathrm{C_T}(R_2))$, which represent the text of certain technical works $\Omega(\mathrm{C_{PAIK}}(TT_1))$ and $\Omega(\mathrm{C_{PAIK}}(TT_2))$, which is the result of the execution of previous constructors.

Construction completion condition: $\tau 3 = true$, getting a number from 0 to 1 that reflects the similarity of two works after comparing all rules in two text models.

The designer has the following specialization:

$$C = \langle M, \Sigma, \Lambda\rangle\ _S \mapsto C_E = \langle M_E, \Sigma_E, \Lambda_E\rangle,$$

where M_E – is a medium that includes a set of rules describing the language of the author in a certain text R_i, Σ_E – are operations and relations on the elements M_E and CIS Λ_E.

We interpret the structure C_E using the algorithmic structure C_A:

$$\langle C_E, C_A\rangle\ _\mathrm{I} \mapsto \langle \mathrm{C}_E = \mathrm{M_{PI}}, \Sigma_{\mathrm{PI}}, \Lambda_{\mathrm{PI}}\rangle,$$

where $V_A = \left\{A_i^0|_{X_i}^{Y_i}\right\}$ – is the set of forming algorithms of the basic algorithmic structure, $\mathrm{X_i}$ and $\mathrm{Y_i}$ – are the set of definitions and values of the algorithm $A_i^0|_{X_i}^{Y_i}$, $M_A = \bigcup_{A_i^0 \in V_A}(X(A_i^0) \cup Y(A_i^0))$ – he carrier of the algorithmic structure, Σ_I – the set of operations linking algorithms, Λ_I – the axiomatics of the algorithmic structure, $\Omega(\mathrm{C_A})$ – a set of algorithms constructed in $\mathrm{C_A}$.

Next, the operation on attributes is presented.

The operation min(m, a, b) compares the numbers a and b, and stores the smallest in m;

The operation –(c, a, b) – is subtraction *c = a – b;*

The operation *(c, a, b) – is multiplication *c = a * b;*

M_{PI} – algorithmic structure for comparing rules, Σ_{PI} – operations of connecting algorithms, $\Lambda_{TI} \supset \Lambda_I \cup \Lambda_1 \cup \Lambda_2 \cup \Lambda_3$.

$\Lambda_3 = \{ (A_{15}|_{a,b}^{m} \downarrow "min"); (A_{16}|_{a,b}^{m} \downarrow " - "); (A_{17}|_{a,b}^{m} \downarrow " * "); (A_{18}|_{a,b}^{m} \downarrow "max") \}$.

The C_{PAI} structure includes the following algorithms:

- $A_0, A_1, A_2, A_3, A_6, A_7, A_8, A_9, A_{10}$ – similar algorithms of C_{PAI} and C_{TAI} structures;
- $A_{15}|_{a,b}^{m}$ – finding the minimum among the numbers a and b;
- $A_{16}|_{a,b}^{c}$ – subtraction *c = a – b;*
- $A_{17}|_{a,b}^{c}$ – multiplication *c = a * b;*
- $A_{18}|_{a,b}^{m}$ – finding the maximum among the numbers a and b.

Specifics C_T:

$$C_{E\ K} \longmapsto C_E(\Omega(\mathrm{CT}(R_1)), \Omega(\mathrm{CT}(R_2))) = \langle M_K, \Sigma_K, \Lambda_K \rangle,$$

where $\Lambda_K \supset \Lambda_I$, $\Lambda_K \supset \{M_K = T_T \cup N\}$ the terminals T include all the words in the rules of both constructors that compare ω and $\dot{\omega}$, the non-terminals N – include the auxiliary symbol τ.

In terms of constructive-synthesizing modeling, the set of rules comparing process for the formation of two texts (T_1 and T_2, respectively) and obtaining the final value of their similarity.

The first rule starts by comparing the rules of two constructors $\Omega(C_T(R_1))$ and $\Omega(C_T(R_2)$ describing two texts that are examined for their similarity, i =1, j=1.

If the same rules or rules exist, the degree of their statistical structural similarity will be determined as the product of the minimum difference in the probabilities of applying the corresponding rule

$$\rho(\vartheta_i, \dot{\vartheta}_j) = \prod_{m=1}^{l} min(prob_m - \dot{prob}_m),$$

where ϑ_i – i-th sentence in T_1 text and $\dot{\vartheta}_j$ – j-th sentence in T_2 text.

The degree of statistical structural similarity of T_1 and T_2 texts:

$$\rho(T_1, T_2) = \sum_{i=1}^{N} \rho(\vartheta_i, \dot{\vartheta}_j),$$

Initial conditions: *rule* = 1, i = m = j = n = 1, where i and m are numbers of chains (sentences) in the text, j and n – are numbers of rules in chains. $max \downarrow \omega_{i,1}$, where max = 0 is the product of the difference in probabilities. $max\_ch \downarrow \omega_{i,1}$, max_ch = 0 is the maximum length of the chain, res = 0 – is the total similarity of two texts, $k = n + 1, h = j + 1$, these are the next rules in the chain concerning j and n, respectively. And the flags for triggering s_1 and s_2 $\tau 1 = true, \tau 2 = false,$ as well as the flag for completing the comparison $\tau 3 = false$.

The first rule is used to compare the first rules in all strings of text

$$s_1 = \langle \sigma_{\tau 1} \rightarrow \vartheta_{i,1}; \sigma_{\tau 1} \rightarrow \vartheta_{m,1} \rangle.$$

For each rule, if their right parts match and the length of the chain is only one rule (that is, the sentence consists of only one word)

$$g_{1,1} = \langle * (R \downarrow \vartheta_{i,1}, R \downarrow \dot{\vartheta}_{m,1}, x1), == (l \downarrow W_{i,1} \downarrow \vartheta_{i,1}, 1, x2), == (l \downarrow W_{m,1} \downarrow \vartheta_{m,1}, 1, x3), \&(y, x1, x2, x3) \rangle.$$

If all conditions are met, the product of the difference in their probabilities is calculated, and the result is stored in the first element of the chain. And until the end of the second text is reached, the products are added up in *res*. If the chains from the first text end, the first rule is closed and the second is opened

$$g_{1,3} = \langle <> (y, 0, \cdot (* \left(max \downarrow \vartheta_{i,1}, min \left(-(r, prob \downarrow \vartheta_{i,h}, prob \downarrow \dot{\vartheta}_{m,k})\right)\right), <> (max \downarrow W_{1,1} \downarrow \vartheta_{m,1}, m, \varepsilon, \cdot \left(+(res, max \downarrow \vartheta_{i,1}, r), +(i, 1, i), = (m, 1)\right), <> (max \downarrow W_{1,1} \downarrow \vartheta_{i,1}, i, \varepsilon, \cdot (= (\tau 1, false), = (\tau 2, true)))), \varepsilon) \rangle.$$

The second rule sequentially traverses all strings longer than one rule, advancing along their length for both texts under investigation. All rules of the second text are reviewed (m varies from 1 to the end of the text). For each sentence, a sequential review of all rules is performed

$$s_2 = \langle \vartheta_{i,j\ \tau 2} \rightarrow \vartheta_{i,h}, \dot{\vartheta}_{m,n\ \tau 2} \rightarrow \dot{\vartheta}_{m,k} \rangle.$$

To start work and calculate similarities, the right parts of the first rules in both texts are compared and we perform operations on the attributes

$$g_{2,1} = \langle * (R \downarrow \vartheta_{i,j}, R \downarrow \dot{\vartheta}_{m,n}, x1), <> \left(l \downarrow W_{i,1} \downarrow \vartheta_{i,j}, j, = (x2, true)\right), = (x2, false)), <> (l \downarrow W_{m,1} \downarrow \dot{\vartheta}_{m,n}, n, = (x3, true), = (x3, false)), == (j, 1, x4) \rangle.$$

If all conditions are met, the first rule in the chain is processed: the length of the chain that matches *ch* is calculated, the product of the difference in the probabilities of the rules from both sim texts is found, and the maximum length of the matching chain and the result of calculating their coincidence are stored in the first element of the chain

$$g_{2,2} = \langle \&(y, x1, x2, x3, x4), <> (y, 0, \cdot (+(ch \downarrow \vartheta_{i,1}, 1, ch \downarrow \vartheta_{i,1}), * (rule, min(-(r, prob \downarrow \vartheta_{i,j}, prob \downarrow \dot{\vartheta}_{m,n})))), = (sim \downarrow \vartheta_{i,1}, rule), <> (ch \downarrow \vartheta_{i,1}, maxch \downarrow \vartheta_{i,1}, \cdot (= (maxch \downarrow \vartheta_{i,1}, ch \downarrow \vartheta_{i,1}), = (max \downarrow \vartheta_{i,1}, sim \downarrow \vartheta_{i,1})), +(j, 1, j), +(n, 1, n)), \cdot (= (j, 1), = (n, 1), +(m, 1, m)), \varepsilon) \rangle.$$

Then all subsequent chains and their rules are processed under the same conditions

$$g_{2,3} = \langle * (R \downarrow \vartheta_{i,h}, R \downarrow \dot{\vartheta}_{m,k}, x1), <> (l \downarrow W_{i,1} \downarrow \vartheta_{i,h}, h, = (x2, true)), = (x2, false)), <> (l \downarrow W_{m,1} \downarrow \dot{\vartheta}_{m,n}, k, = (x3, true), = (x3, false)), \&(y, x1, x2, x3) \rangle.$$

If the chain of coincidences is broken, the comparison of the rules of the 2nd text begins already for the next chain of rules of the first text. If the chain has ended, the transition to the next one is performed, and each of the rules in both texts is similarly checked for coincidence. If the rules in the text end, we close the possibility of executing the second rule $\tau 2 = false$ and end the calculations using the flag $\tau 3 = true$

$$g_{2,4} = \langle <> (y, 0, \cdot (+(ch \downarrow \vartheta_{i,1}, 1, ch \downarrow \vartheta_{i,1}), * (rule, min(-(r, prob \downarrow \vartheta_{i,h}, prob \downarrow \dot{\vartheta}_{m,k})))), = (sim \downarrow \vartheta_{i,1}, rule), +(h, 1, h), +(k, 1, k)), <> (max \downarrow W_{1,1} \downarrow \vartheta_{m,1}, m, \varepsilon, \cdot (+(res, max \downarrow \vartheta_{i,1}, r), +(i, 1, i), = (n, 1), = (m, 1), +(k, n, 1), = (j, 1), +(h, j, 1)), <> (max \downarrow W_{1,1} \downarrow \vartheta_{i,1}, i, \varepsilon, \cdot (= (\tau 2, false), = (\tau 3, true)))\rangle.$$

Note that $\rho(T_1, T_2) = \rho(T_2, T_1)$ $\rho(T_1, T_1) = 1$ is a complete match, $\rho(T_1, T_2) = 0$ – if there are no sentences of the same structure in texts T_1 and T_2.

Realization

The implementation of the structure is the language constructions formation from the elements of its carrier through the execution of algorithms associated with signature operations according to the rules of axionomics:

$$C_{P\,K\;\;R} \mapsto \bar{\Omega}(C_{P\,K}),$$

where $\bar{\Omega}(C_{P\,K}) \subset \Omega(C_{P\,K})$. As the constructor's work result is a number $\bar{\Omega}(C_{P\,K}) \in [0; 1]$, is obtained, which reflects the degree of similarity of the text.

Conclusions

In this paper, constructors are developed and presented that model a natural language text in the form of a stochastic grammar that displays the structures of sentences in it. This approach allows you to highlight the syntactic features of the construction of phrases by the author, which is a characteristic of his speech. Working with a sentence as a unit of text for analyzing its construction will allow you to more accurately capture the author's style in terms of the words use, their sequences and speech style characteristic. It allows you not to be tied to specific parts of speech, but reveals the general logic of constructing phrases, which can be more informative in terms of the author's style characteristics for any text.

The presented work is a theoretical basis for solving the problems of the text authorship establishing and identifying borrowings. Experimental studies have also been carried out, the results of which are partially presented in [3]. The statistical similarity of solutions to the problems of establishing authorship and identifying borrowings was experimentally revealed, which will be presented in the next article of the authors.

It is planned to use the created model in the future to determine the authorship of natural language texts of various directions: fiction and technical literature.

References

1. Kuropiatnyk, O., Shynkarenko V. Automation of template formation to identify the structure of natural language documents. CEUR Workshop Proceedings. 2021. Vol. 2870: 5th International Conference on Computational Linguistics and Intelligent Systems. Vol. I: Main Conference, COLINS 2021, 22–23 April 2021. P. 179–190.
2. V. I. Shynkarenko, V. M. Ilman. Constructive-Synthesizing Structures and Their Grammatical Interpretations. I. Generalized Formal Constructive-Synthesizing Structure / Cybernetics and Systems Analysis, 2014, Volume 50, Issue 5, pp 655-662.
3. V. I. Shynkarenko, I. M. Demidovich Natural Language Texts Authorship Establishing Based on the Sentences Structure, in: Proceedings of the 6th International Conference on Computational Linguistics and Intelligent Systems (COLINS 2022), Volume I: Main Conference, Gliwice, Poland, May 22-23, 2022, pp. 328-337.
4. I. Demidovich, V. Shynkarenko, O. Kuropiatnyk, O. Kirichenko, Processing Words Effectiveness Analysis in Solving the Natural Language Texts Authorship Determination Task, XVI International Scientific and Technical Conference (CSIT'2021). September 22-25, 2021, Lviv, Ukraine.
5. Плющ М. Я. Граматика української мови. Морфеміка. Словотвір. Морфологія. Підручник / М. Я. Плющ.
6. Viktor Shynkarenko and Olena Kuropiatnyk. Constructive Model of the Natural Language. Acta Cybernetica, 23(4):995-1015, 2018.
7. Foltýnek, Tomáš & Meuschke, Norman & Gipp, Bela. Academic Plagiarism Detection: A Systematic Literature Review. ACM Computing Surveys. 52. 2019. 1-42. 10.1145/3345317.

8. Mohammad Al-Smadi, Zain Jaradat, Mahmoud Al-Ayyoub, and Yaser Jararweh. 2017. Paraphrase identication and semantic text similarity analysis in arabic news tweets using lexical, syntactic, and semantic features. Inf. Process.Manag. 53, 3 (2017), 640–652. DOI:10.1016/j.ipm.2017.01.002.

9. Asli Eyecioglu and Bill Keller. 2015. Twitter paraphrase identication with simple overlap features and SVMs. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15). 64–69.

10. Helena Gómez-Adorno, Grigori Sidorov, David Pinto, and Ilia Markov. 2015. A graph based authorship identication approach — Notebook for PAN at CLEF 2015. In Proceedings of the Conference and Labs of the Evaluation Forum and Workshop (CLEF'15).

11. Tenney, Ian et al. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. ArXiv. 1-17. https://doi.org/10.48550/arXiv.1905.06316.

12. Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content Attention Model for Aspect Based Sentiment Analysis. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1023–1032. https://doi.org/10.1145/3178876.3186001.

13. Ravfogel, Shauli et al. 2020. Unsupervised Distillation of Syntactic Information from Contextualized Word Representations. BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP.

14. Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19). Association for Computing Machinery, New York, NY, USA, 985–988. https://doi.org/10.1145/3331184.3331303.

15. Wang, Wei, et al. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. arXiv. https://doi.org/10.48550/arXiv.1908.04577

16. Eva Hasler, Felix Stahlberg, Marcus Tomalin, Adri de Gispert, and Bill Byrne. 2017. A comparison of neural models for word ordering. arXiv. https://doi.org/10.48550/arXiv.1708.01809.

17. Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. arXiv. https://doi.org/10.48550/arXiv.1802.05365.

18. White, L., Togneri, R., Liu, W., Bennamoun, M. 2019. Sentence Representations and Beyond. In: Neural Representations of Natural Language. Studies in Computational Intelligence, vol 783. Springer, Singapore. 93–114. https://doi.org/10.1007/978-981-13-0062-2_5.

19. Liu, Z., Lin, Y., Sun, M. 2020. Sentence Representation. In: Representation Learning for Natural Language Processing. Springer, Singapore. 59–89. https://doi.org/10.1007/978-981-15-5573-2_4.

20. Cheng, Z., Yuan, C., Li, J., Yang, H. 2018. TreeNet: Learning Sentence Representations with Unconstrained Tree Structure. In IJCAI. 4005-4011.

21. Fereshteh Jafariakinabad and Kien A. Hua. 2022. A Self-Supervised Representation Learning of Sentence Structure for Authorship Attribution. ACM Trans. Knowl. Discov. Data 16, 4, Article 68 (August 2022), 16 pages. https://doi.org/10.1145/3491203.

22. Conneau Alexis, Kiela Douwe, Schwenk Holger, Barrault Loïc, and Bordes Antoine. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 670–680.

23. Walenski, M., Europa, E., Caplan, D., & Thompson, C. K. 2019. Neural networks for sentence comprehension and production: An ALE-based meta-analysis of neuroimaging studies. Human brain mapping, 40(8), 2275-2304. https://doi.org/10.1002/hbm.24523.

24. Liu, P., Chang, S., Huang, X., Tang, J., Cheung, J. C. K. 2019. Contextualized non-local neural networks for sequence learning. In Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33, No. 01. 6762-6769. https://doi.org/10.1609/aaai.v33i01.33016762.

Віктор Шинкаренко **Viktor Shynkarenko**	DrS, Professor of Computer and information technologies department, Ukrainian State University of Science and Technologies, Dnipro, Ukraine, e-mail: shinkarenko_vi@ua.fm https://orcid.org/0000-0001-8738-7225	Український державний університет науки та технологій
Інна Демидович **Inna Demidovich**	PhD student of Computer and information technologies department, Ukrainian State University of Science and Technologies, Dnipro, Ukraine, e-mail: 2019demidovichinn@gmail.com https://orcid.org/0000-0002-3644-184X	Український державний університет науки та технологій

https://doi.org/10.31891/csit-2023-3-11
UDC 004.9: 004.05

Iryna ZASORNOVA, Mykola FEDULA, Olga PAVLOVA, Tetiana KYSIL
Khmelnytskyi National University

AUTOMATED TESTING OF WEB PROJECT FUNCTIONALITY WITH USING OF ERROR PROPAGATION ANALYSIS

Automated testing is indispensable in the area of software engineering, particularly for web project functionality, as the complexity of software systems continues to surge. This paper delves into the pivotal role of automated testing and how the integration of error propagation analysis, grounded in chaos theory, can elevate its efficacy. The objective is to elucidate the significance of this methodology and its application in bolstering the reliability and performance of web projects. Automated testing automates the execution of predefined test cases, offering efficiency gains, reduced human error, and swift defect detection in software development. Various testing approaches, including unit testing, integration testing, and regression testing, cater to distinct facets of software functionality, ensuring seamless operation of all components. Web project functionality is integral to the user experience, encompassing navigation menus, forms, and search features. Testing this functionality is imperative to unearth inconsistencies or errors that could compromise user satisfaction and task completion.

This paper proposes a methodology for automated testing coupled with error propagation analysis, which involves scrutinizing how errors evolve through a system over time. Chaos theory, a branch of mathematics examining complex systems' behavior, is employed to understand how minor variations in initial conditions can precipitate substantial system behavior shifts.

Traditional error propagation analysis hinges on linear, deterministic models, but real-world systems often exhibit non-linear, chaotic characteristics, rendering such models inadequate. Chaos theory's non-linear dynamics model the intricate interactions between input variables and their effects on outputs, capturing the sensitivity of chaotic systems to initial conditions. This approach appreciates system complexity and intricate feedback loops, enhancing error analysis's robustness and accuracy. However, the application of chaos theory introduces complexity and computational demands, necessitating a balance between model intricacy and practicality. The proposed methodology unveil valuable insights into error propagation within web projects' functionality, pinpointing vulnerable components and areas ripe for improvement. The methodology's advantages include the ability to identify potential issues and vulnerabilities, ultimately enhancing web project reliability.

Keywords: web project functionality testing, automated testing, error propagation analysis, chaos theory, instability.

Ірина ЗАСОРНОВА, Микола ФЕДУЛА, Ольга ПАВЛОВА, Тетяна КИСІЛЬ
Хмельницький національний університет

АВТОМАТИЗОВАНЕ ТЕСТУВАННЯ ФУНКЦІОНАЛЬНОСТІ ВЕБ-ПРОЄКТУ З ВИКОРИСТАННЯМ АНАЛІЗУ ПОШИРЕННЯ ПОМИЛОК

Автоматизоване тестування є незамінним у сфері розробки програмного забезпечення, особливо для функціональності веб-проєктів, оскільки складність програмних систем продовжує зростати. У цій статті розглядається ключова роль автоматизованого тестування та те, як інтеграція аналізу поширення помилок, заснованого на теорії хаосу, може підвищити його ефективність. Мета полягає в тому, щоб з'ясувати значення цієї методології та її застосування для підвищення надійності та продуктивності веб-проєктів. Автоматизоване тестування автоматизує виконання попередньо визначених тестів, пропонуючи підвищення ефективності, зменшення людських помилок і швидке виявлення дефектів у розробці програмного забезпечення. Різні підходи до тестування, включаючи модульне тестування, інтеграційне тестування та регресійне тестування, задовольняють різні аспекти функціональності програмного забезпечення, забезпечуючи безперебійну роботу всіх компонентів. Функціональні можливості веб-проєкту є невід'ємною частиною взаємодії з користувачем, охоплюючи навігаційні меню, форми та функції пошуку. Тестування цієї функції є обов'язковим, щоб виявити невідповідності або помилки, які можуть поставити під загрозу задоволення користувачів і виконання завдань.

У статті пропонується методологія автоматизованого тестування в поєднанні з аналізом розповсюдження помилок, що передбачає ретельний аналіз того, як помилки розвиваються в системі з часом. Теорія хаосу - це розділ математики, що вивчає поведінку складних систем, використовується для розуміння того, як незначні варіації початкових умов можуть спричинити значні зміни поведінки системи.

Традиційний аналіз розповсюдження помилок базується на лінійних, детермінованих моделях, але реальні системи часто демонструють нелінійні, хаотичні характеристики, що робить такі моделі неадекватними. Нелінійна динаміка теорії хаосу моделює складну взаємодію між вхідними змінними та їхній вплив на результати, фіксуючи чутливість хаотичних систем до початкових умов. Цей підхід оцінює складність системи та заплутані цикли зворотного зв'язку, підвищуючи надійність і точність аналізу помилок. Однак застосування теорії хаосу створює складність і потребує обчислень, що вимагає балансу між складністю моделі та практичністю. Запропонована методологія розкриває цінну інформацію про розповсюдження помилок у функціональності веб-проєктів, визначаючи вразливі компоненти та області, які потребують вдосконалення. Переваги методології включають можливість виявлення потенційних проблем і вразливостей, що в кінцевому підсумку підвищує надійність веб-проєкту.

Ключові слова: тестування функціональності веб-проєкту, автоматизоване тестування, аналіз поширення помилок, теорія хаосу, нестабільність.

Introduction

Automated testing of web project functionality [1] with error propagation analysis is a crucial aspect of software engineering [2,3]. As technology continues to advance rapidly, software systems become increasingly complex, and ensuring their functionality becomes more challenging. Testing plays a vital role in identifying and

rectifying any issues or vulnerabilities that may arise during the development process [2,4,5]. In this paper, we will explore the significance of automated testing and how error propagation analysis incorporating chaos theory can be applied to enhance this method [6].

Automated testing refers to the use of tools and frameworks to execute predefined test cases automatically. It offers several benefits in software development, including increased efficiency, reduced human error, and faster detection of defects or bugs. There are various approaches to automated testing, such as unit testing, integration testing, and regression testing [1,7–9]. Each approach focuses on different aspects of software functionality and helps ensure that all components work together seamlessly.

Web project functionality refers to the ability of a website or web application to perform its intended tasks accurately and efficiently [9]. The user experience heavily relies on the proper functioning of web projects' features such as navigation menus, forms, search functionalities, etc. Testing web project functionality is essential for identifying any inconsistencies or errors that may hinder user satisfaction or impede successful completion of tasks [2,3].

The primary objective of web project functionality test is the evaluating whether the features and functions of a web application align with the project's requirements [1]. This testing phase is instrumental in identifying defects, anomalies, and inconsistencies within the web project, thereby ensuring that it meets user expectations and business goals.

The functionality of the web application is rigorously evaluated, including user interactions, data processing, and system operations. Each feature's behavior is meticulously assessed against predefined test cases. The web project can be tested across various browsers, devices, and operating systems to guarantee seamless functionality for a diverse user base. Assessing the responsiveness and scalability of the web project under varying loads and conditions is crucial to ensure an optimal user experience. Detecting vulnerabilities and ensuring the protection of sensitive user data is a pivotal component of functionality testing to safeguard against potential threats. Functionality testing is indispensable in ensuring the basic usability of a web project [10,11]. However, it may not provide a comprehensive view of the application's resilience to unforeseen issues and potential error propagation.

Error propagation analysis extends the evaluation of web projects beyond functionality testing. There are sophisticated techniques employed to understand how errors or issues in one part of the web project can ripple through other interconnected components. Often used in conjunction with functionality testing, error propagation analysis offers a deeper understanding of the application's reliability. The analysis begins by mapping out the intricate dependencies between different modules, components, and functions of the web project, shedding light on how data and errors can flow between these interconnected elements [2,3].

Various types of errors or faults are simulated at different points within the web application to gauge how they propagate through the system. These errors may include input validation issues, database connection failures, or network errors. Once errors are introduced, analysts assess their impact on the web project's functionality, performance, and security. They ascertain how errors influence the user experience and whether they lead to data corruption or system instability.

Error propagation analysis aids developers and testers in understanding the web project's ability to handle errors gracefully. It facilitates the identification of potential error recovery mechanisms and the reinforcement of error-handling processes. By pinpointing critical points of error propagation, web project teams can prioritize the resolution of high-risk areas, reducing the likelihood of system failures and data breaches [3].

Thus, web project functionality testing is an indispensable step in ensuring the usability of web applications. However, it may fall short in identifying vulnerabilities stemming from error propagation. Error propagation analysis serves as a vital complement to functionality testing, providing a holistic view of a web project's resilience to errors and failures. By combining these two approaches, web developers and quality assurance teams can create robust and dependable web applications that not only meet user expectations but also withstand the challenges of the digital landscape. This, in turn, enhances the overall user experience and security of the web project, making it a critical aspect of modern web development.

Therefore, automated testing of web project functionality is *an actual task* today.

The purpose of this study is to develop methodology for automated testing with error propagation analysis.

Methodology for Automated Testing with Error Propagation Analysis

Error propagation analysis involves studying how errors propagate through a system over time. In software engineering, it helps identify potential issues or vulnerabilities by analyzing how errors affect different components within a system or application stack[12]. By understanding how errors propagate throughout a system's architecture, developers gain valuable insights into areas that require attention during the testing phase [6,11].

Chaos theory is a mathematical concept that explores the behavior of complex systems that are highly sensitive to initial conditions [13,14]. It has found application in various fields, including physics, biology, and economics. When applied to web project functionality testing, chaos theory enables us to understand how slight variations in initial conditions can lead to significant differences in system behavior [15–17].

Error propagation analysis is a fundamental concept in various scientific disciplines, including physics,

engineering, and computer science. It involves assessing how uncertainties in input variables propagate through a system to affect the uncertainty in the output or result. While traditionally approached through deterministic and linear models, the application of chaos theory to error propagation analysis offers a fresh perspective that acknowledges the inherent complexity and non-linearity of real-world systems. In this paper, we will explore the integration of chaos theory principles into error propagation analysis, highlighting its benefits and implications. Traditionally, error propagation analysis relies on linear and deterministic models, such as the Gaussian error propagation formula, which assumes that errors in input variables are normally distributed and that the relationships between variables are linear. This approach has been successful in many scenarios, especially when dealing with simple systems and small uncertainties. However, real-world systems are often non-linear, chaotic, and subject to complex interactions, making deterministic and linear models inadequate for capturing their behavior accurately [6,18].

Chaos theory, initially developed to understand the behavior of dynamic systems that appear random but are governed by underlying deterministic processes, offers a powerful framework for error propagation analysis [19]. Chaos theory deals with non-linear dynamics, where small changes in initial conditions can lead to significant variations in outcomes, a phenomenon known as the butterfly effect. By embracing chaos theory, error propagation analysis can better account for the complexity and unpredictability inherent in many systems.

Chaos theory enables the modeling of non-linear interactions between input variables and their effects on the output. This is particularly valuable when dealing with systems where small errors in initial conditions or input variables can lead to dramatic and unexpected consequences [10,20]. The theory acknowledges the sensitivity of chaotic systems to initial conditions, highlighting that seemingly insignificant changes in inputs can propagate into substantial differences in outcomes. Error propagation analysis benefits from this perspective by accounting for the sensitivity of real-world systems. The methods of nonlinear dynamics emphasize the complexity of dynamic systems and the presence of complex feedback loops. When applied to error propagation analysis, this recognition allows for more accurate modeling of the interactions between variables, which may not adhere to linear assumptions.

By embracing chaos theory, error propagation analysis can provide a more robust understanding of the uncertainties associated with a system's outputs. This can be particularly crucial in critical applications such as aerospace, finance, and healthcare, where small errors can have significant consequences.

While integrating chaos theory into error propagation analysis offers significant benefits, it also presents challenges. Chaos theory often involves the use of complex mathematical models, which may require sophisticated computational tools and resources. Moreover, the implementation of such models may be more time-consuming and computationally intensive than traditional linear approaches. Therefore, it is essential to strike a balance between model complexity and practicality when applying chaos theory to error propagation analysis.

Thus, the integration of chaos theory into error propagation analysis represents a promising avenue for improving our understanding of complex, non-linear systems. By embracing the principles of chaos theory, error propagation analysis can better account for the inherent uncertainty, non-linearity, and sensitivity to initial conditions present in real-world systems. While challenges exist in implementing these models, the potential benefits, including more accurate predictions and enhanced robustness, make chaos theory a valuable tool for error propagation analysis in various scientific and engineering disciplines. As our understanding of chaos theory and its applications continues to evolve, it promises to provide deeper insights into the behavior of complex computer systems and how they propagate errors.

Our proposed methodology for automated testing using error propagation analysis builds upon the principles of chaos theory. By analyzing the sensitivity of web project functionality to initial conditions, we can identify potential points of failure or instability, as described in [21]. The steps involved in implementing this methodology include:

Experiments conducted using our methodology have yielded insightful findings regarding error propagation within web projects' functionality [22]. We observed certain trends where specific components exhibited higher vulnerability to errors than others, highlighting potential areas for improvement during development.

The proposed method for automated testing with error propagation analysis offers several advantages but also presents limitations worth considering. One limitation is determining an optimal balance between complexity and feasibility of the test scenarios. Additionally, the method heavily relies on accurate modeling of initial conditions and system behavior.

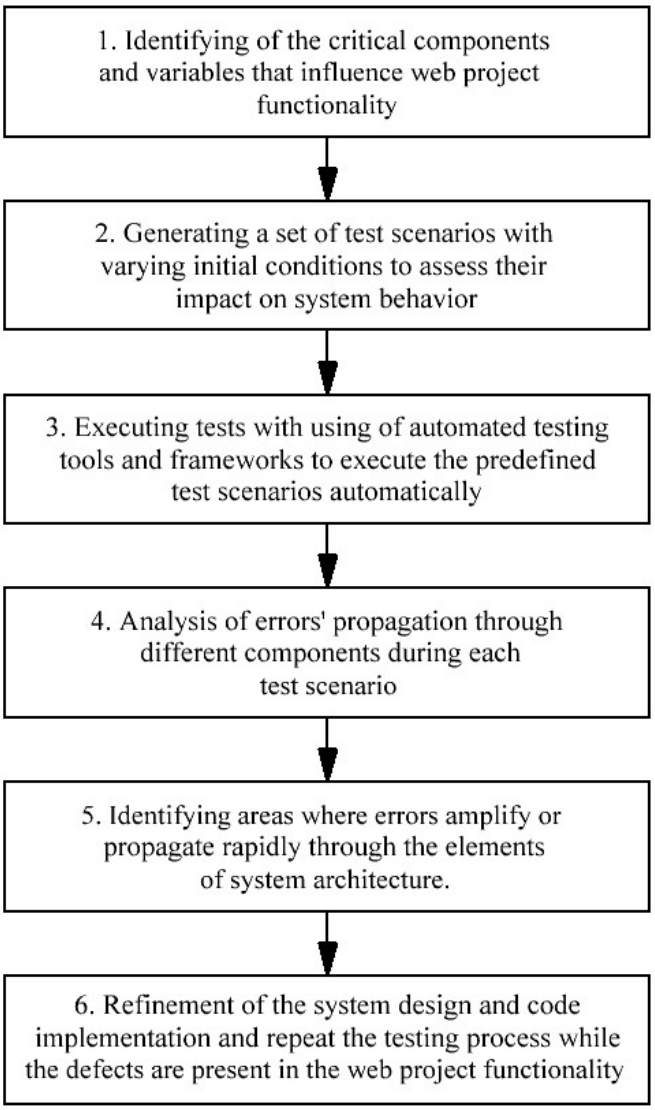

Fig. 1. The proposed methodology for automated testing using error propagation analysis

Conclusions

Automated testing with error propagation analysis is a valuable approach to ensure web project functionality. By incorporating chaos theory principles, this methodology allows developers to identify potential issues or vulnerabilities within a system's architecture and take appropriate measures for improvement [12]. The significance of automated testing cannot be overstated in improving user experience and overall software quality. Moving forward, further research can focus on refining the proposed methodology, addressing its limitations, and exploring additional applications in different domains. As software engineering continues to evolve rapidly, automated testing will remain an essential practice in ensuring the reliability and functionality of web projects.

In conclusion, automated testing combined with error propagation analysis is a powerful tool in software engineering that enables developers to identify weaknesses and enhance the functionality of web projects effectively. By leveraging chaos theory principles and considering initial conditions sensitivity, this methodology provides valuable insights into how errors propagate through complex systems. With further research and refinement, it holds great potential for improving software development practices across various domains.

References

1. Lenka RK, Satapathy U, Dey M. Comparative Analysis on Automated Testing of Web-based Application. Proceedings - IEEE 2018 International Conference on Advances in Computing, Communication Control and Networking, ICACCCN 2018 2018:408–13. https://doi.org/10.1109/ICACCCN.2018.8748374.

2. Bertolino A. Software testing research: Achievements, challenges, dreams. FoSE 2007: Future of Software Engineering 2007:85–103. https://doi.org/10.1109/FOSE.2007.25.

3. Li S, Lu M, Kong S. Research and Implementation of a Software Runtime Fault Propagation Path Acquisition Method. Proceedings of 2019 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, QR2MSE 2019 2019:695–702. https://doi.org/10.1109/QR2MSE46217.2019.9021165.
4. Lee J, Kang S, Lee D. Survey on software testing practices. IET Software 2012;6:275–82. https://doi.org/10.1049/IET-SEN.2011.0066.
5. Garousi V, Mäntylä M V. When and what to automate in software testing? A multi-vocal literature review. Inf Softw Technol 2016;76:92–117. https://doi.org/10.1016/J.INFSOF.2016.04.015.
6. Abdelmoez W, Nassar DM, Shereshevsky M, Gradetsky N, Gunnalan R, Ammar HH, et al. Error propagation in software architectures. Proceedings - International Software Metrics Symposium 2004:384–93. https://doi.org/10.1109/METRIC.2004.1357923.
7. Rafi DM, Moses KRK, Petersen K, Mäntylä M V. Benefits and limitations of automated software testing: Systematic literature review and practitioner survey. 2012 7th International Workshop on Automation of Software Test, AST 2012 - Proceedings 2012:36–42. https://doi.org/10.1109/IWAST.2012.6228988.
8. Vila E, Novakova G, Todorova D. Automation testing framework for web applications with selenium WebDriver: Opportunities and threats. ACM International Conference Proceeding Series 2017;Part F131200:144–50. https://doi.org/10.1145/3133264.3133300.
9. Debroy V, Brimble L, Yost M, Erry A. Automating Web Application Testing from the Ground Up: Experiences and Lessons Learned in an Industrial Setting. Proceedings - 2018 IEEE 11th International Conference on Software Testing, Verification and Validation, ICST 2018 2018:354–62. https://doi.org/10.1109/ICST.2018.00042.
10. Correya A, Bogdanov D, Joglar-Ongay L, Serra X. Essentia.js: A JavaScript Library for Music and Audio Analysis on the Web 2020.
11. Zasornova Iryna, Hovorushchenko Tetiana, Fedula Mykola, Buzyl Viktoriia. Intelligent Information Technology for Identification of Remaining Defects in Software Products. The 12th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, Dortmund: IEEE; 2023.
12. Jones N, Rosenthal C. Chaos Engineering. 1st ed. O'Reilly Media; 2020.
13. Peng Y, Sun K, He S, Peng D. Parameter Identification of Fractional-Order Discrete Chaotic Systems. Entropy 2019, Vol 21, Page 27 2019;21:27. https://doi.org/10.3390/E21010027.
14. Rosenthal C. Chaos Engineering: System Resiliency in Practice 2020:400.
15. Principles of Chaos Engineering. PrinciplesofchaosOrg n.d.
16. Fedula M, Hovorushchenko T, Nicheporuk A, Martynyuk V. Chaos-based signal detection with discrete-time processing of the Duffing attractor. Eastern-European Journal of Enterprise Technologies 2019;4:44–51. https://doi.org/10.15587/1729-4061.2019.175787.
17. James J, Clarke G, Mathew R, Mulkeen B, Papakostas N. On Reducing the Localisation Error of Modern Mobile Robotic Platforms. Procedia CIRP 2022;112:168–73. https://doi.org/10.1016/j.procir.2022.09.067.
18. Sánchez A. Modern control systems, 9th edition, Richard C. Dorf and Robert H. Bishop, Prentice-Hall, Upper Saddle River, NJ, 2001, 831 pages, ISBN 0-13-030660-6. International Journal of Robust and Nonlinear Control 2006;16:372–3. https://doi.org/10.1002/RNC.1054.
19. Fotiadis A, Vlachos I, Kugiumtzis D. Detecting Nonlinear Interactions in Complex Systems: Application in Financial Markets. Entropy 2023, Vol 25, Page 370 2023;25:370. https://doi.org/10.3390/E25020370.
20. Bai J, Perron P. Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica 1998;66:47. https://doi.org/10.2307/2998540.
21. Diz Felipe A, Ziegl A, Hayn D, Schreier G. Development of Algorithms for Automated Timed Up-and-Go Test Subtask and Step Frequency Analysis. Stud Health Technol Inform 2022;289:367–70. https://doi.org/10.3233/SHTI210935.
22. Psujek M, Radzik A, Kozieł G. Comparative analysis of solutions used in Automated Testing of Internet Applications. Journal of Computer Sciences Institute 2021;18:7–14. https://doi.org/10.35784/JCSI.2373.

Iryna Zasornova **Ірина Засорнова**	Candidate of Technical Sciences, Associate Professor of the Department of Computer Engineering and Information Systems, Khmelnytskyi National University https://orcid.org/0000-0001-6655-5023 e-mail: izasornova@gmail.com	Кандидат технічних наук, доцент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Mykola Fedula **Микола Федула**	Candidate of Technical Sciences, Associate Professor of the Department of Automation, Computer Integrated Technologies and Robotics, Khmelnytskyi National University https://orcid.org/0000-0002-3765-2016 e-mail: mailfm2000@gmail.com	Кандидат технічних наук, доцент кафедри автоматизації, комп'ютерно-інтегрованих технологій та робототехніки, Хмельницький національний університет
Olga Pavlova **Ольга Павлова**	PhD, Associate Professor of the Department of Computer Engineering and Information Systems, Khmelnytskyi National University https://orcid.org/0000-0001-7019-0354 e-mail: olya1607pavlova@gmail.com	Доктор філософії, доцент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет
Tetiana Kysil **Тетяна Кисіль**	Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Computer Engineering and Information Systems, Khmelnytskyi National University https://orcid.org/0000-0002-4094-3500 e-mail: kysil_tanya@ukr.net	Кандидат фізико-математичних наук, доцент кафедри комп'ютерної інженерії та інформаційних систем, Хмельницький національний університет

Full requirements for the design of the manuscript
Повні вимоги до оформлення рукопису
http://csitjournal.khmnu.edu.ua/

No editorial responsibility is required for the content of messages sub.
За зміст повідомлень редакція відповідальності не несе

To print 29.09.2023. Mind. Printing. Arch. 11,27. Obl.-vid. Arch. 10,32
Format 30x42 / 4, offset paper. Another risography.
Overlay 100, deputy. №

Підп. до друку 29.09.2023. Ум. друк. арк. 11,27. Обл.-вид. арк. 10.32
Формат 30х42/4, папір офсетний. Друк різографією.
Наклад 100, зам. №

Replication is made from the original layout, made edited
by the magazine "Computer Systems and Information Technology"

Тиражування здійснено з оригінал-макету, виготовленого
редакцією журналу “Комп’ютерні системи та інформаційні технології”

Editorial and publishing center of Khmelnytskyi national university
29016, Khmelnytskyi, street Institutska, 7/1, tel. (0382) 72-83-63

Редакційно-видавничий центр Хмельницького національного університету
29016, м. Хмельницький, вул. Інститутська, 7/1, тел. (0382) 72-83-63